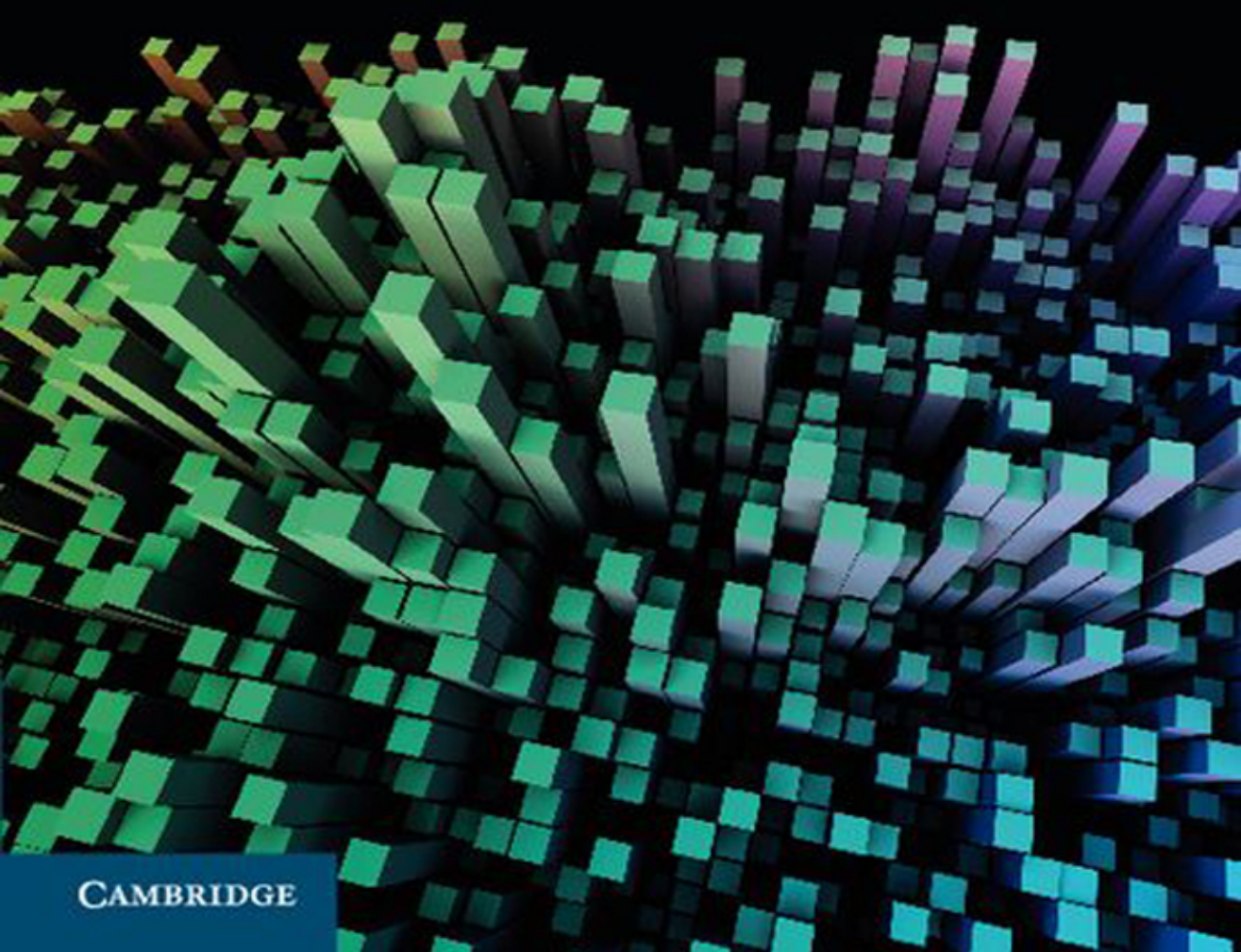


EDITED BY FA-LONG LUO

Digital Front-End in Wireless Communications and Broadcasting

Circuits and Signal Processing



CAMBRIDGE

CAMBRIDGE

more information – www.cambridge.org/9781107002135

This page intentionally left blank

Digital Front-End in Wireless Communications and Broadcasting

Circuits and Signal Processing

Covering everything from signal processing algorithms to integrated circuit design, this complete guide to digital front-end is invaluable for professional engineers and researchers in the fields of signal processing, wireless communication, and circuit design. Showing how theory is translated into practical technology, it covers all the relevant standards and gives readers the ideal design methodology to manage a rapidly increasing range of applications. Step-by-step information for designing practical systems is provided, with a systematic presentation of theory, principles, algorithms, standards, and implementation. Design trade-offs are also included, as are practical implementation examples from real-world systems. A broad range of topics is covered, including digital pre-distortion (DPD), digital up-conversion (DUC), digital down-conversion (DDC), and DC-offset calibration. Other important areas discussed are peak-to-average power ratio (PAPR) reduction, crest factor reduction (CFR), pulse-shaping, image rejection, digital mixing, delay/gain/imbalance compensation, error correction, noise-shaping, numerically controlled oscillator (NCO), and various diversity methods.

Fa-Long Luo, Ph.D., is Chief Scientist of two leading international companies on software defined radio and wireless multimedia with headquarters in Silicon Valley, California. He has 27 years of research and industrial experiences in multimedia, communication, and broadcasting with real-time implementation, applications, and standardizations with worldwide high recognition. He has authored two books, more than 100 technical papers, and 18 patents in these and closely related fields.

Digital Front-End in Wireless Communications and Broadcasting

Circuits and Signal Processing

Edited by

FA-LONG LUO

Element CXI, California



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org

Information on this title: www.cambridge.org/9781107002135

© Cambridge University Press 2011

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2011

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

Library of Congress Cataloging in Publication data

Digital front-end in wireless communications and broadcasting : circuits and signal
processing / edited by Fa-Long Luo.

p. cm.

ISBN 978-1-107-00213-5 (hardback)

1. Radio – Transmitters and transmission. 2. Digital communications. 3. Signal processing – Digital
techniques. 4. Radio circuits. I. Luo, Fa-Long.

TK6561.D54 2011

621.384'12–dc22

2011010608

ISBN 978-1-107-00213-5 Hardback

Cambridge University Press has no responsibility for the persistence or
accuracy of URLs for external or third-party internet websites referred to
in this publication, and does not guarantee that any content on such
websites is, or will remain, accurate or appropriate.

Contents

List of contributors
Preface

page xii
xvii

Part I Introduction to digital front-end	1
1 Introduction to wireless communications and digital broadcasting	
Hongxiang Li, Guanying Ru, Siqian Liu, Hui Liu, and Jenq-Neng Hwang	3
1.1 Evolution of mobile cellular networks	4
1.2 Evolution of broadcast network	9
1.3 Key technologies	15
1.4 Conclusions	23
References	24
2 Basic principles and protocols of digital wireless communications	
Dimitra I. Kaklamani and Panagiotis K. Gkonis	28
2.1 Introduction	28
2.2 Digital processing, transmission, and reception	29
2.3 The WCDMA/OFDMA physical layer protocols	32
2.4 General principles of MIMO systems	39
2.5 Space-time and space-frequency transmission over MIMO networks	44
2.6 Summary	48
References	48
3 Digital front-end and system overview in wireless communication	
Eduard Bertran and Maurizio Pagani	50
3.1 Introduction	50
3.2 New objectives and technologies in modern wireless communication systems	56
3.3 Other emerging standards beyond 3G	76
3.4 DFE introductory design considerations	79
3.5 Conclusions	91

	List of abbreviations (Chapter 3)	91
	References	95
4	System overview and front-end technologies in digital broadcasting	
	Francesc Tarres and Eduard Bertran	98
	4.1 Introduction	98
	4.2 Broadcast mobile multimedia services	102
	4.3 Common technologies for digital mobile broadcast services	103
	4.4 Most relevant standards for digital broadcasting to mobile devices	105
	4.5 Technology aspects of DFE in transmitters	111
	4.6 Conclusions	116
	List of abbreviations (Chapter 4)	116
	References	118
5	Digital front-end for multiple standards and multimode compatibility	
	Eduard Bertran	120
	5.1 Introduction	120
	5.2 Receiver architectures	122
	5.3 Analog-to-digital converters	126
	5.4 Preselect filters and LNA interface for multiband receivers	128
	5.5 Power amplifier architectures in multimode-multiband transmitters	130
	5.6 Conclusions	135
	List of abbreviations (Chapter 5)	136
	References	137
	Part II DPD and CFR	141
6	General principles and design overview of digital predistortion	
	R. Neil Braithwaite	143
	6.1 Introduction	143
	6.2 Power amplifier nonlinearity	143
	6.3 Digital predistortion overview	150
	6.4 Digital predistortion basics	160
	6.5 Advanced topics in DPD	177
	6.6 Concluding comments	187
	References	189
7	Power amplifier nonlinear modeling for digital predistortion	
	Gabriel Montoro and Pere L. Gilabert	192
	7.1 Introduction	192
	7.2 Principles of predistortion linearization	193

7.3	Power amplifier behavioral models for digital predistortion	196
7.4	Discussion	210
	References	211
8	Look-up table based digital predistortion schemes and implementation	
	Pere L. Gilabert and Gabriel Montoro	214
8.1	Introduction	214
8.2	Look-up table based schemes	215
8.3	FPGA implementation of an LUT-based DPD	219
8.4	Discussion	240
	References	241
9	Digital predistortion and its combination with crest factor reduction	
	Anding Zhu	244
9.1	Principles of operation	245
9.2	DPD models	246
9.3	Model extraction structures	255
9.4	Bandwidth and sampling rate requirements	257
9.5	System characterization procedures	259
9.6	Implementation considerations	264
9.7	Combination with crest factor reduction	270
9.8	Conclusions and future outlook	276
	References	277
10	Adaptive digital baseband predistortion: design and implementation	
	Hua Qian and G. Tong Zhou	280
10.1	Introduction	280
10.2	Power amplifier nonlinearity behavior and modeling	281
10.3	Digital predistortion modeling	286
10.4	Predistortion implementation challenges	298
10.5	Conclusions	305
	References	305
11	Crest factor reduction techniques	
	Nuno Borges Carvalho and Wohoon Jang	309
11.1	Introduction	309
11.2	Multi-carrier communication systems	311
11.3	CF reduction using clipping plus filtering	319
11.4	Conclusion	326
	References	326

Part III DUC, DDC, ADC, DAC, and NCO	331
12 Digital down-conversion and up-conversion: theory and implementation	
Yue Zhang, Taiwen Tang, Kok-Keong Loo, Ben Allen, and Dayou Li	333
12.1 Introduction	334
12.2 Multi-rate, multi-stage, and filter-banks in DDC and DUC	339
12.3 Mixing and combining for multi-channels and multi-carriers	348
12.4 Hardware implementation for digital down-conversion	352
12.5 Hardware implementation for digital up-conversion	370
12.6 Summary	377
References	378
13 A/D and D/A data conversion for wireless communications transceivers	
Jerzy Dąbrowski	380
13.1 A/D conversion in RF receiver	381
13.2 ADC for heterodyne and low-IF receiver	386
13.3 Digital down-conversion	389
13.4 Implementation of ADCs in communications receivers	392
13.5 D/A conversion in RF transmitter	397
13.6 Digital up-conversion	402
13.7 Implementation of DACs in communications transmitters	402
13.8 Summary	409
References	410
14 Advanced quadrature sigma-delta modulator designs for A/D interface	
Mikko Valkama, Jaakko Marttila, and Markus Allén	413
14.1 Sigma-delta basics	413
14.2 Considerations on further modulator concepts	427
14.3 Advanced modulators	431
14.4 Nonidealities in quadrature sigma-delta modulators	434
14.5 Examples on complex multiband transfer function design	439
14.6 Conclusion	445
References	446
15 Digital suppression of A/D interface nonlinearities	
Mikko Valkama, Markus Allén, and Jaakko Marttila	450
15.1 Introduction to ADC nonlinearities	452
15.2 Look-up table	456
15.3 Dithering	458
15.4 Model inversion	459
15.5 Adaptive interference cancellation	460

15.6	Clipping compensation using interpolation	465
15.7	Performance examples of online post-processing methods	467
15.8	Conclusions	470
	References	471

Part IV Digital calibration, imbalance compensation, and error corrections 473

16	Digital compensation and calibration of I/Q gain and phase imbalances	
	Lauri Anttila, Yaning Zou, and Mikko Valkama	475
16.1	Introduction	475
16.2	I/Q modulation and direct-conversion radio architecture	476
16.3	I/Q imbalance signal modeling	477
16.4	Compensation and calibration methods	482
16.5	Design and testing examples	494
16.6	Conclusions	497
	References	498
17	Joint digital predistortion of I/Q modulator and power amplifier impairments	
	Mikko Valkama and Lauri Anttila	502
17.1	Introduction	502
17.2	Effects of IQM impairments on PA intermodulation products and predistorter estimation	504
17.3	Joint PA and IQM predistortion	510
17.4	Predistorter parameter estimation	514
17.5	Simulation and measurement examples	519
17.6	Conclusions	527
	References	528
18	Diversity and error compensation in OFDM transceivers: principles and implementation	
	Tomohisa Wada	531
18.1	Introduction	531
18.2	OFDM transceivers: principles and design	532
18.3	Adaptive antenna array and diversity	547
18.4	Error detection and compensation	556
18.5	Hardware implementation and system integration	563
18.6	Conclusions	571
	References	571

19	Front-end architectures and impairment corrections in multimode and multi-antenna systems	
	Guillaume Villemaud, Jacques Verdier, Matthieu Gautier, Ioan Burciu, and Pierre-François Morlat	573
19.1	Introduction	573
19.2	State of the art of multi-* receivers	574
19.3	Mixed analog–digital front-end architecture for multiband simultaneous reception	580
19.4	Some methods for simplifying analog parts of multi-antenna front-ends	585
19.5	Front-ends combining multiband and multi-antenna capabilities	591
19.6	Conclusions	598
	References	599
	Part V Circuits and system integration in digital front-end	601
20	Integration and interface of digital front-end with analog front-end	
	Steven Loh	603
20.1	Introduction	603
20.2	Mixed signal transceiver architectures	604
20.3	Digital–analog system integration and interfaces	610
20.4	Conclusions and discussions	618
	References	619
21	Circuits and systems for digital front-ends to support multiple wireless standards	
	Imtinan Elahi and Khurram Muhammad	620
21.1	Introduction	620
21.2	Major Functions of the DFE	621
21.3	Determination of DFE specifications	622
21.4	Design of DFE to support WCDMA and GGE	626
21.5	Considerations to support LTE	646
21.6	Conclusions	648
	References	648
22	Programming models and implementation platforms for software defined radio configuration	
	Tanguy Risset, Riadh Ben Abdallah, Antoine Fraboulet, and Jérôme Martin	650
22.1	Introduction	650
22.2	Programming environment and tools for SDR	650
22.3	An existing radio virtual machine implementation	658
22.4	Conclusion	668
	References	668

23	Programmable discrete-time integrated receiver front-ends	
	Norm Filiol, Robert Klett, and Thomas Riley	671
23.1	Background	671
23.2	Theory	675
23.3	Zero intermediate frequency receivers	684
23.4	Low intermediate frequency receivers	688
23.5	Case study of a super-heterodyne AM/FM radio receiver	692
23.6	Summary and conclusions	703
23.7	Acknowledgements	704
	References	704
24	Multi-port front-end and transceivers for V-band multi-gigabit/s communication systems	
	Serioja Ovidiu Tatu, Emilia Moldovan, and Sofiene Affes	707
24.1	Introduction	707
24.2	Multi-port interferometer theory	709
24.3	Multi-port V-band practical implementation, modeling, and analysis	718
24.4	Proposed transceiver architectures	722
24.5	Advanced system simulations	727
24.6	Conclusion	731
	References	731
25	Cross-layer design and digital front-end for cognitive wireless link	
	Matthieu Gautier, Guillaume Villemaud, Cedric Lévy-Bencheton, Dominique Noguét, and Tanguy Risset	733
25.1	Introduction to flexible cognitive wireless link	733
25.2	A high-level optimization for SDR: multimode relaying	736
25.3	Implementation of the PHY sensing block: example of TV white space sensing	747
25.4	Conclusion	755
	References	756
	<i>Index</i>	758

Contributors

Ben Allen

University of Bedfordshire, UK

Markus Allén

Tampere University of Technology, Tampere, Finland

Sofiene Affes

INRS-EMT, Montreal, Quebec, Canada

Lauri Anttila

Tampere University of Technology, Tampere, Finland

Riadh Ben Abdallah

Université de Lyon, INRIA, INSA-Lyon, CITI, France

Eduard Bertran

Technical University of Catalonia -- UPC, Spain

R. Neil Braithwaite

Powerwave Technologies, Santa Ana, CA, USA

Ioan Burciu

Université de Lyon, INRIA, INSA-Lyon, CITI, France

Nuno Borges Carvalho

Universidade de Aveiro, Portugal

Jerzy Dąbrowski

Linköping University, Sweden

Imtinan Elahi

Masdar Institute, Abu Dhabi, UAE

Norm Filiol

Kaben Wireless Silicon Inc., Ottawa, Canada

Antoine Fraboulet

Université de Lyon, INRIA, INSA-Lyon, CITI, France

Matthieu Gautier

CEA, LETI, Minatec, Grenoble, France

Pere L. Gilabert

Technical University of Catalonia -- UPC, Spain

Panagiotis K. Gkonis

National Technical University of Athens, Greece

Jenq-Neng Hwang

University of Washington, Seattle, USA

Wohoon Jang

Universidade de Aveiro, Portugal

Dimitra I. Kaklamani

National Technical University of Athens, Greece

Robert Klett

Kaben Wireless Silicon Inc., Ottawa, Canada

Cedric Lévy-Bencheton

Université de Lyon, INRIA, INSA-Lyon, CITI, France

Dayou Li

University of Bedfordshire, UK

Hongxiang Li

North Dakota State University, USA

Hui Liu

University of Washington, Seattle, USA

Siqian Liu

North Dakota State University, USA

Steven Loh

Samsung Telecommunications America, Inc., Richardson, TX, USA

Kok-Keong Loo

Middlesex University, UK

Jerome Martin

CEA-Leti, Minatec, Grenoble, France

Jaakko Marttila

Tampere University of Technology, Tampere, Finland

Emilia Moldovan

INRS-EMT, Montreal, Quebec, Canada

Gabriel Montoro

Technical University of Catalonia -- UPC, Spain

Pierre-François Morlat

Silicom, Toulouse, France

Khurram Muhammad

Research In Motion, Irving, TX, USA

Dominique Noguét

CEA, LETI, Minatec, Grenoble, France

Maurizio Pagani

Ericsson, Vimodrone (MI), Italy

Hua Qian

Shanghai Institute of Microsystem and Information Technology, China

Thomas Riley

Kaben Wireless Silicon Inc., Ottawa, Canada

Tanguy Risset

Université de Lyon, INRIA, INSA-Lyon, CITI, France

Guanying Ru

North Dakota State University, USA

Taiwen Tang

University of Toronto, Canada

Francesc Tarres

Technical University of Catalonia -- UPC, Spain

Serioja Ovidiu Tatu

INRS-EMT, Montreal, Quebec, Canada

Mikko Valkama

Tampere University of Technology, Tampere, Finland

Jacques Verdier

Université de Lyon, CNRS, INSA-Lyon, CITI, France

Guillaume Villemaud

Université de Lyon, INRIA, INSA-Lyon, CITI, France

Tomohisa Wada

University of the Ryukyus, Okinawa, Japan

Yue Zhang

University of Bedfordshire, UK

G. Tong Zhou

Georgia Institute of Technology, Atlanta, GA, USA

Anding Zhu

University College Dublin, Ireland

Yaning Zou

Tampere University of Technology, Tampere, Finland

Preface

With the rapid development and worldwide deployment of broadband wireless communication and digital broadcasting infrastructures, the use of digital processing technology in the front-end and radio frequency unit is growing explosively. Digital processing technology for front-end in transmitters and receivers of wireless communication and digital broadcasting covers a broad range of topics including digital predistortion (DPD), digital up-conversion (DUC), digital down-conversion (DDC), DC-offset calibration, peak-to-average power ratio (PAPR) or crest factor reduction (CFR), pulse-shaping, delay/gain/imbalance compensation, noise-shaping, numerically controlled oscillator (NCO), and conversion between the analog and digital domains. These digital processing technologies offer a number of advantages in power efficiency, cost reduction, time-to-market, and flexibility for software defined radio (SDR) so as to support multiple standards and multimode applications. Unlike baseband processing, front-end is tightly connected to the radio frequency layer, therefore it imposes great limitations and difficulties on digital processing speed, memory, computational capability, power, size, data interfaces, and bandwidths. This suggests that digital processing and circuit implementation of front-end are very challenging tasks and require the huge efforts of the related industry, research, and regulatory authorities.

From an application and implementation design point of view, this book aims to be the first single volume to provide a comprehensive and highly coherent treatment on digital front-end and its system integration for multiple standards, multi-carrier and multimode in both broadband communications and digital broadcasting by covering basic principles, signal processing algorithms, silicon-convergence, design trade-off, and well-considered implementation examples.

This book is organized into twenty-five chapters in five parts.

Part I Introduction to digital front-end

The first part consists of five chapters aiming to present overviews of all the processing components of the digital front-end in both transmitters and receivers of broadband wireless communications and digital broadcasting with emphasis on orthogonal frequency-division multiplexing (OFDM) based systems. General principles and basics of digital wireless communications and digital broadcasting are also provided in the

first two chapters. [Chapter 1](#) first reviews the wireless evolution by looking at two major applications: cellular networks and TV broadcast networks. Then the discussion of [Chapter 1](#) focuses on some key techniques including multiple-input multiple-output (MIMO) technique, OFDM, and cognitive radio. The purpose of [Chapter 2](#) is mainly to provide an overview of the basic principles and protocols of digital transmission in wireless networks. The emphasis of this chapter is on transmission over MIMO networks with space-time and space-frequency transmission techniques being presented in alignment with the WCDMA/OFDMA physical layer protocols.

[Chapter 3](#) is devoted to the front-end and system overview of broadband wireless communications. Firstly digital front-end and its key processing units are reviewed. Then this chapter presents new objectives and technologies for transmission and reception of modern wireless communication systems with emphasis on the front-end part. In particular, this chapter introduces some recent mobile communications standards, suitable for SDR developments because of their novelty and capability of reconfiguration including the standards in development under the umbrella of the IMT-Advanced Program. Also, this chapter addresses some new standards on the basis of cognitive radio strategies. More importantly, this chapter focuses on various aspects of how to design and implement digital front-end, covering transceivers and related power amplifiers, chip circuits, and their hardware processing platforms.

[Chapter 4](#) presents a system overview and front-end technologies in digital broadcasting. Concepts such as multicast, return channel, video quality, and transcoding are presented in the context of broadcast mobile multimedia services, covering common modulation technologies, different alternatives for audio and video coding, most relevant standards, and proprietary systems for digital multimedia broadcasting. However, the emphasis of this chapter is on key technologies such as analog-to-digital conversion (ADC), digital-to-analog conversion (DAC), DDC, DUC, and power amplifier linearization for digital front-end processing in broadcast transmission and reception.

Digital front-end processing has begun to play an increasingly important role in wireless communication and broadcasting systems to support multiple standards and multimode compatibility, which is the focus of [Chapter 5](#). In this chapter, an overview of the key structures of digital front-end for multiple bands and multiple modes is first presented. Then, this chapter outlines some important aspects regarding the analog-to-digital conversion for multimode and multiband applications as well as the preselect filters and low-noise amplifier (LNA) interface in multimode-multiband transmitters. Power amplifier architectures in multiple standards are also addressed at the end of this chapter.

Part II DPD and CFR

Six chapters in [Part II](#) mainly deal with principles, theory, algorithms, circuit designs, and hardware implementations on various computing platforms for digital predistortion and crest factor reduction so as to compensate for power amplifier nonlinearities and to reduce peak-to-average power ratio. With a tutorial style, [Chapter 6](#) presents general principles and design overview of digital predistortion techniques. [Chapter 6](#) first reviews

various nonlinear behaviors and their modeling of a power amplifier (PA) as well as the impact of the nonlinearity on the output spectrum and entire performance of a PA taking memory effects into account. Moreover, this chapter is devoted to providing the details of the basic algorithms, learning-rules, and model selections in designing and implementing a DPD scheme, and also provides comprehensive discussions on some advanced topics in digital predistortion technology.

Modeling of power amplifiers with nonlinearity and/or memory effects is the starting point to develop effective and practical digital predistortion algorithms and hardware implementations, which is the focus of [Chapter 7](#). In this chapter, a comprehensive overview of the most widely used PA behavioral models is first presented on the basis of neural network modeling and Volterra series based modeling methods also taking dynamics of PA into account. [Chapter 7](#) then provides various adaptive algorithms to estimate the parameters (coefficients) of these models so as to make a better compromise among performance, complexity, and accuracy when using these models to obtain the desired DPD schemes.

[Chapter 8](#) mainly deals with look-up table (LUT) based digital predistortion schemes and implementations. The LUT-based DPD has been considered as the most efficient scheme because of its simplicity in computations and circuit implementations. This chapter covers the principles of LUT-based DPD schemes, the LUT organizations (the 1-D or 2-D table architecture), the optimum size of the LUT (trade-off between the data memory size and accuracy), and the indexing and spacing between entries within the LUT. As design references, some implementation examples using an FPGA platform for the LUT-based DPD schemes are also presented in this chapter.

Digital predistortion and its combination with crest factor reduction are addressed in [Chapter 9](#). This chapter first discusses various issues in designing and implementing a practical DPD scheme including model extraction structures, the bandwidth and sampling rate requirements, and system characterization procedures. Then, the chapter introduces some of the widely used crest factor reduction techniques which are proposed to achieve higher power efficiency by reducing the peaks of the transmit signal to a satisfactory level before digital predistortion and allowing a PA to be operated at higher average power.

[Chapter 10](#) is devoted to one of the most important aspects in compensating for the nonlinearity and memory effect of a PA: that is, design and implementation of adaptive digital baseband predistortion. Through implementation examples and their performance, this chapter covers learning architectures for adaptively modeling nonlinear power amplifiers, adaptive estimation algorithms of DPD coefficients, adaptive DPD architecture designs (filter selection, delay compensation, and signal scaling), and fixed-point implementation issues which take DSP, ASIC, FPGA, or other system-on-chip as computing platforms.

In [Chapter 11](#), crest factor reduction techniques are extensively discussed from design to implementation so as to attack the high PAPR problem encountered in multiband-multi-carrier communication and broadcasting systems. This chapter lists various crest factor reduction techniques and makes detailed comparisons in terms of throughput, error vector magnitude (EVM), bit error ratio (BER), and system complexity in order to help readers make the best compromise in designing and implementing a practical CFR scheme.

Part III DUC, DDC, ADC, DAC, and NCO

Organized into four chapters, **Part III** is devoted to technology and practice of digital up-conversion, digital down-conversion, analog-to-digital conversion, digital-to-analog conversion, and numerically controlled oscillator including re-sampling theory, filtering algorithms, aliasing cancellation, combination, mixing, quantization noise, bit-width effect analyses, circuit designs, and hardware implementations. These are the key processing parts that connect radio frequency signals to baseband signals in all broadband wireless communication and digital broadcasting systems.

Chapter 12 deals with up-conversion of the discrete baseband signal stream into a high-resolution radio signal at the transmitter, and down-conversion of a high-resolution radio signal back into a baseband signal at the receiver by covering the basic principles and functionality of DUC and DDC in relation to conversion between intermediate frequency and baseband with the emphasis on the implementation of the DDC and DUC for standard wireless communication systems. Furthermore, this chapter discusses the multi-rate, multi-stage and filter-banks design, I/Q (In-phase/Quadrature) modulation and demodulation, and NCO design in DDC and DUC.

Chapter 13 discusses A/D and D/A data conversion used in the transceivers of wireless communication systems. For ADC, this chapter covers the fundamental specifications like sensitivity, selectivity, dynamic range (DR), equivalent number of bits (ENOB) and linearity constraints. For DAC, this chapter focuses on the fundamental transmitter specifications which include the EVM and the adjacent-channel-power-rejection (ACPR). To show how to meet the above specifications for practical transceivers in RF application systems, this chapter provides design and implementation examples of the state-of-the-art DAC and ADC using CMOS processes.

Design and implementation of advanced quadrature sigma-delta modulator for A/D interface are addressed in **Chapter 14**. This chapter first outlines the basics of sigma-delta modulation and then extends the discussions to some further modulator concepts (number of bit, stage and band, and related noise shaping). Next, the selected advanced quadrature structures with multi-stage and multiband are provided. In presenting the illustrating examples and their experimental results, this chapter further deals with implementation and design issues related to nonideal factors such as nonlinearity, jitter, and I/Q imbalance.

Chapter 15 deals with ADC nonlinearities and their digital suppression for radio transceivers. Some significant sources of ADC nonlinearities are first analyzed including gain error, offset error, clipping, differential nonlinearity (DNL), and integral nonlinearity (INL). This chapter then discusses the impact of these nonlinearities on ADC performance and presents several digital processing based methods to suppress them. These digital suppression methods mainly employ look-up table, dithering and model inversion. Moreover, some effective adaptive interference cancellation and interpolation techniques are provided in this chapter giving related performance testing examples.

Part IV Digital calibration, imbalance compensation, and error corrections

In addition to the nonlinearities and nonideal factors discussed in [Part II](#) and [Part III](#), a number of other nonideal factors exist in transceivers of wireless communications and digital broadcasting from RF to baseband which include channel imbalance (gain, offset, and delay), I/Q mismatch, and synchronization error (RF, sampling rate, FFT window position, and symbol timing). More importantly, these nonideal factors impact on one another. Hence, more comprehensive digital processing techniques on the basis of joint-level, cross-layer, system-level, diversity (time, array, frequency), and error-tolerance principles are highly desirable in order to simultaneously perform digital calibration, imbalance compensation, and error correction in a practical and efficient way. This is the topic addressed in the four chapters of [Part IV](#).

[Chapter 16](#) is devoted to digital compensation and calibration of the I/Q gain and phase imbalances which are the main resource of the resulting mirror-frequency interference in direct-conversion type radio transmitters and receivers. After reviewing the I/Q modulation and direct-conversion type radio architectures, this chapter discusses behavioral modeling of the I/Q imbalance problem in radio transceivers, covering both frequency-independent and frequency-dependent I/Q imbalance cases. Moreover, various approaches for imbalance estimation and calibration are presented with design and testing examples by covering both digital predistortion type techniques on the transmitter side and digital post-correction methods on the receiver side. Also, complete link models including imbalanced transmitter, multipath radio channel, and imbalanced receiver are given in [Chapter 16](#).

In [Chapter 17](#), joint digital predistortion is presented to compensate for both I/Q modulator (IQM) and power amplifier impairments with covering principles, modeling, algorithms, designs, implementation and testing results. This chapter first illustrates that I/Q mismatch and LO leakage interact with PA nonlinearity such that extra intermodulation distortion products appear at the PA output and affect the estimation and performance of PA predistorters provided by the DPD methods listed in [Part II](#) of this book. A new predistorter structure is then presented so as to be able to jointly mitigate both IQM and PA impairments including frequency-dependent behavior (memory) of the impairments.

[Chapter 18](#) presents principles and implementation of diversity techniques and various error compensations in OFDM-based transceivers. From a system-level point of view, this chapter first presents the principles and designs of OFDM transceivers and then focuses on adaptive antenna array and several diversity techniques in time-domain, frequency-domain, spatial-domain, and the cross-layer level. Furthermore, this chapter deals with the detection and compensation of various error sources and practical factors. At the end, this chapter presents the hardware implementation in ASIC, DSP, and FPGA platforms with the related chip performance.

The emphasis of [Chapter 19](#) is on radio front-end architectures and related impairment corrections for multiband, multi-antenna, and multimode receivers so as to better support multiple standards and multimode applications. This chapter presents various mitigation

methods of RF impairments such as the phase noise due to local oscillator errors, frequency offset due to error between receiver's and emitter's local oscillators, and IQ mismatches (gain and phase). Different ways of multiplexing signals and sharing resources (hardware, frequency-spectrum, and structure) in multiple band/channel/modes are also described in this chapter.

Part V Circuits and system integration of digital front-end

The last part of this book deals with the system integration, interface, and convergence-to-silicon of digital front-end with analog front-end, baseband processing, and related cross-layer processing to support multiple standards and multimode applications in broadband wireless communication and digital broadcasting. A number of implemented systems are provided in this part to motivate readers that digital and programmable front-end processing can provide better performance, more flexibility, and less-power consumption than analog-based front-end processing and will replace more and more processing parts required from RF to baseband in advanced systems such as MIMO, ultra wideband (UWB), SDR, and cognitive radio.

[Chapter 20](#) describes the integration and interface between the digital front-end and analog front-end focusing on wireless terminals ASICs applications by showing how to minimize costs and size, and also how to optimize power efficiency in designing new devices. This chapter presents several transceiver mixed signal architectures with various analog-to-digital interfaces that have been popularized in today's wireless terminals' ASICs design. Furthermore, this chapter addresses the system aspects of integration and interfaces between digital front-end and analog front-end in detail and also discusses in detail the future directions of these designs.

In [Chapter 21](#), circuits and systems for digital front-end to support multiple wireless standards are addressed with emphasis on receiver front-end. In this chapter, three major functions of digital front-end are first outlined, namely, the sampling rate conversion, channel selection (or filtering), and compensation of various analog/RF impairments. Then, this chapter presents detailed circuit design considerations and system integration specifications of digital front-end to support three major wireless standards: WCDMA, GGE(GSM/GPRS/EDGE), and LTE.

Implementation (computing) platforms and the corresponding programming models are playing a critical role in both digital front-end processing and baseband processing, which is the focus of [Chapter 22](#). This chapter first reviews various platforms and programming models/languages and then presents an efficient programming machine called the radio virtual machine so as to achieve the following functions for SDR dynamic configurations in both baseband processing and digital front-end processing, namely: a programming language that permits an easy expression of physical layers and can be compiled into an executable form (bytecode), an abstraction based on the component model paradigm, mechanism to handle real-time constraints with easy access to hardware, and an arbitrary bit-width arithmetic.

Chapter 23 presents a programmable discrete-time integrated receiver front-end which can greatly reduce the power and other processing costs that are associated with analog-to-digital conversion used in integrated software defined radio. This chapter first discusses sampling theory and various programmable filtering circuits (filter types and transforms, their design and selection). Furthermore, this chapter reviews the programmable zero intermediate frequency and low intermediate frequency discrete-time receiver front-ends outlining the advantages and disadvantages of these receivers. Illustration implementation examples and testing results are also presented in this chapter.

The emphasis of Chapter 24 is on multi-port front-end and UWB transceivers for V-band multi-Gigabit/s communication systems. This chapter shows that multi-port circuits can successfully be used in quadrature down-converters, antenna arrays, and direct modulation of millimeter-wave signals by presenting that the important advantage in using multi-ports is the reduced LO power requested for down-conversion. This is particularly true in millimeter-wave applications where the received RF signal is considerably low, reducing both the cost of LO and the leakage between LO and the RF input.

Chapter 25 presents some algorithms and techniques in designing a flexible or cognitive radio link considering all cross-layer issues from RF, front-end, baseband, and the media-access-control layer. This chapter first introduces four processing steps needed in an SDR transceiver, namely, sensing, analyse, decision, and act; then it presents various advanced algorithms and related implementation examples in performing these four processing steps. How to better use digital front-end processing remains a hot topic in the emerging cognitive radio systems with large number of degrees of freedom and with multimode capabilities. This chapter serves as a starting point for future research and development in this topic.

For whom is this book written?

It is hoped that this book serves not only as a complete and invaluable reference for professional engineers, researchers, manufacturers, network operators, software developers, content providers, service providers, broadcasters, and regulatory bodies interested in broadband wireless communications and digital broadcasting system developments and applications, but also as a textbook for graduate students in circuits, signal processing, wireless communications, microwave technology, antenna and propagation, and system-on-chip implementation.

Fa-Long Luo, Ph.D.
Silicon Valley, California, USA

Part I

Introduction to digital front-end

1 Introduction to wireless communications and digital broadcasting

Hongxiang Li, Guanying Ru, Siqian Liu, Hui Liu, and Jenq-Neng Hwang

The purpose of communication engineering is to transmit information from its source to the destination over some distance away. A basic communication system mainly consists of three essential components: transmitter, channel (wired or wireless), and receiver. [Figure 1.1](#) shows a typical point-to-point one-way communication system. For a two-way system, a receiver and a transmitter are both required on each side.

The transmitter transforms the input signal to a transmission signal that is suitable for the characteristic of the channel. Since a channel is always varying with time, and the input signal to the system differs, the transmitter processes the input signal to produce a suitable signal for transmission. This generally includes the modulation and the coding. After being processed by the transmitter, the transmitted signal goes into the channel. The channel can be any medium or interface suitable for the transmission and it connects the transmitter and the receiver. The channel may be the laser beam, coaxial cable, or radio wave. During the transmission, there are various unwanted effects on the signals. Attenuation and power loss reduce the signal strength and make the detection difficult at the receiver. Besides the power loss and the attenuation, the channel may always introduce some undesired signals. These signals may be random and unpredictable signals that exist in nature, such as solar radiation, or the signals produced by other transmitters or machines. We call the former type of undesired signal “noise” and the latter type “interference.” If the interfering signals occupy different frequencies of the desired signal, using proper filters can remove these interferences. In comparison, the random noise that superimposes on the information-bearing signals is hard to completely eliminate by the filter and the remaining noise will inevitably corrupt the desired signal. The power ratio of received signal over noise decides the channel capacity which is one of the basic system performance parameters. After the receiver picks up the signal from the channel, it will do some filtering to compensate for the power loss, followed by demodulation and decoding to recover the original input signals.

Wireless transmission of information has experienced fast growth in the past few decades. A basic wireless system generally contains radio frequency (RF) front-end and baseband processing. From the application point of view, wireless systems can be categorized into two types: *unicasting* and *broadcasting* [1]. Unicasting is used for symmetric

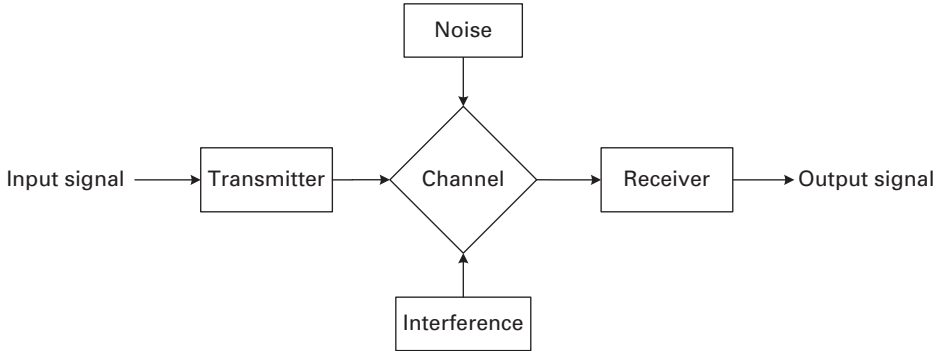


Figure 1.1 Elements of a communication system.

applications such as voice telephony and wireless data access in a bi-directional channel. Broadcasting is used for distribution of audio/video and other multimedia content over a reliable and unidirectional channel. As a major unicast application, the cellular network emerged from the first generation analog voice system, followed by its second-generation digital system providing basic data services. Now we are at the stage between the third generation and pre-fourth generation system. It offers both high-speed Internet data access and high-quality voice services. Owing to the different application characteristics, broadcasting evolves along a different trajectory which will also be discussed later in this chapter. Along with other networks, the ubiquitous wireless applications make multi-radio coexistence inevitable and the hybrid network converging both unicasting and broadcasting is becoming a notable trend [1].

Organized into four sections, this chapter provides a brief overview of the general principles of wireless broadband communications and broadcasting. We first review the wireless evolution by looking at two major applications: cellular network and TV broadcast network. Then, the discussion focuses on some key techniques including multiple-input multiple-output technique (MIMO), orthogonal frequency division multiplexing (OFDM), and cognitive radio. In the [final section](#), we will present a summary and further discussions.

1.1 Evolution of mobile cellular networks

The cellular telephone system is one the most successful applications of wireless communication. A cellular network consists of multiple cells. Each cell contains one base-station (BS) and several mobile stations. The base-station of each cell is connected and controlled by the mobile switching center (MSC) through a dedicated backhaul link. The MSC keeps track of users' information. The MSCs are connected by the gateway MSC (GMSC) to connect the cellular users with customers using other wireless providers. The GMSC is the termination of the public switched telephone network to provide wide range communication. The general system model for cellular systems is shown in [Figure 1.2](#).

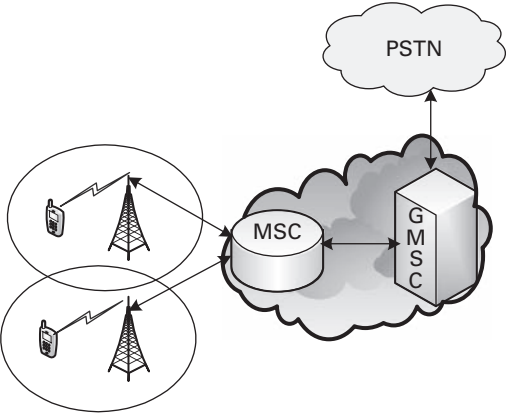


Figure 1.2 The general cellular system model.

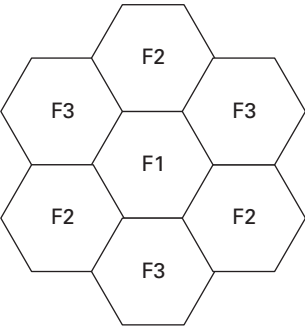


Figure 1.3 An example of frequency reuse in cellular networks.

In 1915, wireless voice transmission was established between New York and San Francisco, a distance of 2500 miles, which is a milestone in wireless communication. Wireless communication was a priority for the military until 1946 when AT&T was allowed to provide the world's first mobile telephone service. In 1946, 25 cities in the United States had built up a public mobile telephone service. These initial systems used one high-power BS to cover an entire metropolitan area without the frequency reuse which is called macrocell technology. Besides the deficiency of radio technologies, the system capacity was severely limited at that time [2]. To solve this problem, the cellular concept was proposed by AT&T Bell Lab researchers in the 1960s [3]. Cellular systems have the property that the power of a signal decreases with distance. Hence, the interference between two users that are spatially far from each other can be neglected. This means that two users far from each other can use the same frequency, which can greatly enhance the system capacity. An example of the frequency reuse is shown in [Figure 1.3](#), wherein F1, F2, and F3 represent different frequency bands. Each hexagon represents a cell, in which users share a certain frequency band as depicted in the figure.

1.1.1 First-generation cellular systems

The two landmark technologies of the first-generation cellular systems (1G) are the application of the cellular concept, and the relative handoff technology. The handoff technology is used when users travel between cells during a conversation in order to maintain the service. The first generation of cellular systems was designed in the late 1960s. However, it was not launched until the early 1980s because of regulatory delays. First-generation networks use Frequency Division Multiplexing Access (FDMA) to transmit analog voice signal through wireless channels. In 1979, NTT DoCoMo established the first commercially automated cellular network. Nordic Mobile Telephone (NMT) was the first to support international roaming. In 1982, the Advanced Mobile Phone System (AMPS) went operational, which was the first commercial analog system in the United States [4]. However, the 1G system only provides voice service over circuit switching.

1.1.2 Second-generation cellular systems

The second-generation (2G) mobile phone systems emerged in the early 1990s. It is based on digital communication, which is the key difference between 2G and 1G communication. The digitalization is realized by adding a digital-to-analog (D/A) converter before the RF transmitter, and an analog-to-digital (A/D) converter after the RF receiver. Compared to analog systems, digital systems have the following advantages:

- (1) Better security. Digital signals can be easily encrypted, and all digital cellular systems have their encryption part. In contrast, analog systems have no security which means an eavesdropper can easily intercept a user's identification number.
- (2) Higher communication quality. In digital systems, error detection and correction by encoding and decoding can be applied to achieve a higher communication quality. However, analog data form is susceptible to interference which results in a highly variable reception quality.
- (3) Higher frequency efficiency. Digital data can be compressed, so digital systems have higher spectrum efficiency.

Furthermore, 2G digital systems also have longer battery life, higher capacity, higher data rate, and cheaper digital equipment than analog systems. The cellular phones' great market potential has led to a proliferation of 2G standards. While 2G cellular systems initially supported only voice services, later on these systems also supported data services such as short messaging and Internet access. In the same way as 1G systems, 2G systems also use circuit-switching for data transmission. There are four main standards for 2G systems: Global System for Mobile (GSM) communications; Digital AMPS (D-AMPS); Code Division Multiple Access (CDMA) IS-95; and Personal Digital Cellular (PDC). However, these standards are all incompatible.

1.1.3 Third-generation cellular systems

With the development of the 2G technique and the wide application of the cellular phone, the demands for data services were growing. This trend stimulated the emergence of the third-generation (3G) technology. While 2G systems were designed to carry speech and low bit-rate data, 3G systems target at higher data rate services. The use of packet-switching for data transmission distinguishes the 3G technology from 2G technology. The 3G standards were defined as International Mobile Telecommunications-2000 (IMT-2000) by the International Telecommunications Union (ITU). The IMT-2000 standard aims to support high-speed voice and data services, and to support seamless global roaming and seamless delivery of services. According to the IMT-2000 specification, a 3G system must simultaneously support speech and data services, and provide at least 200 Kbps peak data rates. The three dominant 3G standards are Wideband CDMA (WCDMA), CDMA2000, and Time Division-Synchronous CDMA (TD-SCDMA). The first 3G network was deployed in Japan in 2001. The 3G networks can provide data rate ranging from 384 kbps to 2 Mbps; while 2G networks support data rate ranging from 9.6 kbps to 28.8 kbps. Broadband should be with instantaneous bandwidths greater than 1 MHz and support data rates greater than 1.5 Mbps [5]. Intel defines it as a continuum of co-existing, overlapping technologies that enable high-speed communications [6]. Hence, 3G belongs to broadband communications [7].

During the development of 3G systems, CDMA2000 1X and GPRS (also called 2.5G) were developed as extensions of 2G networks. 2.5G techniques have some features of 3G systems. For example, CDMA2000 1X can theoretically support 307 kbps as the maximum data speed. Beyond CDMA2000 1X, EDGE system theoretically meets the requirements of 3G systems [8]. In the mid 2000s, High-Speed Downlink Packet Access (HSDPA), as an evolution of 3G technology, began to be implemented. High-Speed Downlink Packet Access allows Universal Mobile Telecommunications System (UMTS) based networks to have a higher data rate and capacity. The downlink speed of current HSDPA can be 1.8, 3.6, 7.2, and 14.0 Mbps, while HSPA+ will provide a higher downlink speed of up to 84 Mbps, according to Release 9 of the 3rd Generation Partnership Project (3GPP) standards [9].

1.1.4 Future broadband wireless communications

In the future, different wireless communication standards are expected to be integrated into one unique communication system, and such a system is called the fourth generation (4G). Fourth generation will enable a comprehensive and secure all-IP based solution, such as IP phone, ultra-broadband Internet access, HDTV broadcast, and stream multimedia. The 4G system will provide full mobility and connectivity, which requires free roaming from standard to standard or from service to service. In July 2003, International Telecommunication Union (ITU) made a requirement for 4G system known as IMT-Advanced standard: First, the transmission data rate should be above 1 Gbps at the stationary condition. Second, the transmission

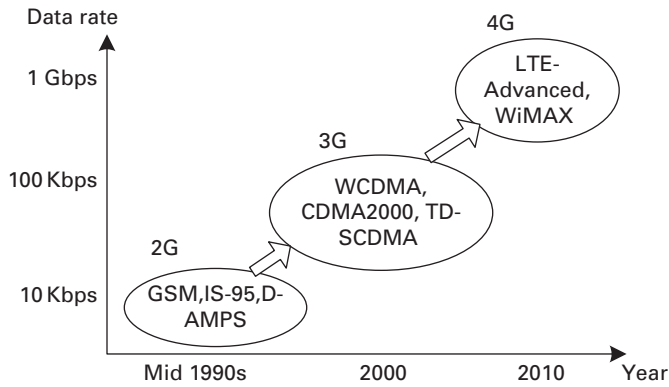


Figure 1.4 The evolution of communication systems.

data rate should be above 100 Mbps at moving speed [10]. Both 3GPP Long Term Evolution (LTE) and Worldwide Interoperability for Microwave Access (WiMAX) are candidates for 4G systems. The evolution of communication systems is shown in Figure 1.4.

To meet the high-level performance requirement, the first 3GPP LTE relies on advanced physical layer technologies, such as MIMO techniques, OFDM techniques, and cognitive radio [11]. Long Term Evolution cannot fully meet the IMT Advanced requirements. Hence, LTE-Advanced was proposed as a candidate for IMT Advanced standard. LTE advanced was formally submitted to ITU-T in fall 2009 by the 3GPP organization and is expected to be released in 2012. Parallel with LTE, IEEE 802.16m will be one of the 4G version and will be backward-compatible with the existing IEEE 802.16e. Many 4G air interfaces are now being examined. The 4G systems have to be cost-effective and quality of service (QoS) driven. Potential 4G technologies include:

- (1) Efficient modulation techniques, such as the OFDM technology which can easily combat the multipath effect in broadband systems;
- (2) Advanced antenna technologies, such as MIMO techniques that can combat interference and greatly enhance the system capacity;
- (3) Intelligent systems, such as cognitive radio that can adjust with the varying transmission conditions;
- (4) Advanced encoding and decoding techniques, such as turbo coding and low-density parity-check codes (LDPC);
- (5) Wireless access technologies, such as Orthogonal Frequency-Division Multiple Access (OFDMA) and Multiple Carrier Code-Division Multiple Access (MC-CDMA);
- (6) Multiple systems' cooperation and convergence, such as broadcast and cellular network hybrid.

1.2 Evolution of broadcast network

In this section, a brief overview of current existing broadcast systems is given. As the major broadcast application, TV networks provide video and audio services to a large number of users. Therefore, most of the discussions are based on existing TV technologies.

1.2.1 Analog broadcast system

The earliest wireless broadcast technology can be traced back to the year 1900 when Guglielmo Marconi sent his famous Morse code from England to Canada. At the early stage, broadcasting was mainly used for radio and wireless telegraph until the emergence of TV, which utilized the high-frequency radio wave to transmit TV signals.

In the early stages, all the information was encoded and transmitted as analog signals. Analog transmission is the traditional method for wireless broadcasting. It uses a continuous carrier signal whose amplitude, frequency, or phase varies in proportion to the analog message (voice, image). Frequency modulation (FM) and amplitude modulation (AM) are widely used in analog communications [12]. Different analog TV broadcast systems have different frequencies, frame rates, and resolutions. In 1950s, the International Telecommunication Union (ITU) standardized the monochrome combination for black-and-white TV by using the capital letters A to N. Later on, information of the hue and saturation in the picture were added as the signal parameters for color television. Nowadays, there are three standards for the coding and transmission of color TV signals, ANTSC (American National Television Systems Committee), PAL (Phase Alternation Line) and SECAM (Séquentiel Couleur Avec Mémoire). The ANTSC is mainly used in North America while the other two are applied in European and Asian countries.

Compared to digital systems that require complicated multiplexing and timing synchronization, analog transmission is a low-cost alternative for broadcasting. Nowadays, analog broadcasting is still very popular, especially for short-distance broadcasting. However, in long-distance and high-throughput transmissions, analog broadcasting has poor performance due to severe signal attenuation. Recently, many countries have completely ceased analog transmissions on certain media, such as television, for technological and economical reasons.

1.2.2 Digital broadcast system

In the 1990s, with the development of digital technology, digital communication attracted more and more attention. Comparing to analog signals, the transmission of information by digital signals can ensure better quality and has higher transmission rates. Also, digital signals can better resist the interference than analog signals. Some problems caused by channel noise in analog television have been solved, such as ghosting images, unstable images, and reduction of the image quality. In digital television systems, the

audio and video are synchronized before the transmission, so the received signals are almost complete. Otherwise, the received TV signals cannot be decoded and displayed. With the completeness of the signals, the digital TV program can still be watched when it experiences interference. Recently, the digital signal has gradually replaced the traditional analog signal in TV broadcasting. In some developed countries, including the United States and Germany, this replacement has been made. In the United States, since June 11, 2009, all on-the air TV signals have adopted the Advanced Television Systems Committee Standards (ATSC), and high-power analog signal transmissions have been ceased by Federal Communications Commission (FCC) [13].

Digital broadcasting uses digital data rather than analog waveform to carry broadcast information over television channels or assigned radio frequency bands. Unlike analog communication, digital broadcasting only transmits discrete messages. In the baseband transmission, information is represented by a sequence of pulses while a finite set of predetermined waveforms are used for presentation in passband transmission. The modulation as well as the corresponding demodulation and detection are carried out by modem. The data transmitted are in the form of digital symbols. An analog signal such as a voice signal is first sampled into a stream of bits using PCM (pulse-code modulation) or some other source-coding technique. Generally, the source coding and decoding are completed by encoder and decoder at the transmitting and receiving ends, respectively [14].

Besides its robustness to noise and interference, digital TV also has several other advantages over analog TV. One of the most significant advantages is its high spectrum efficiency. The bandwidth needed for digital TV broadcasting can be flexibly determined by the compression level and the transmitted image resolution. In this way, the service providers can broadcast more digital information or offer the high-definition television service (HDTV) within the same spectrum bandwidth. Furthermore, digital TV also supports some non-television services such as multimedia, multiplexing (playing more than one program simultaneously), and translating into other languages.

Today, most countries have adopted digital television as the main direction of future TV services. There are four different families of digital TV standards used in different regions, as is shown in [Figure 1.5](#) [15].

ATSC Terrestrial

In the early 1990s, the Advanced Television Systems Committee developed a set of standards known as ATSC used for digital television transmission over terrestrial networks. Now, it is the standard of digital TV service in North America. ATSC specifies the design of a system transmitting the digital video, audio, and data over 6-MHz channels and the information rate is 19.29 Mbps. The same bandwidth is used in the design of the ATSC as traditional analog NTSC TV channels. After MPEG-2 compression and multiplexing, the signals are output as streams that can be modulated in various ways based on the different transmission methods.

- (1) TV broadcasters use 8-level vestigial sideband modulation (8VSB) to deliver the video and audio information at a maximum rate of 19.39Mbps.

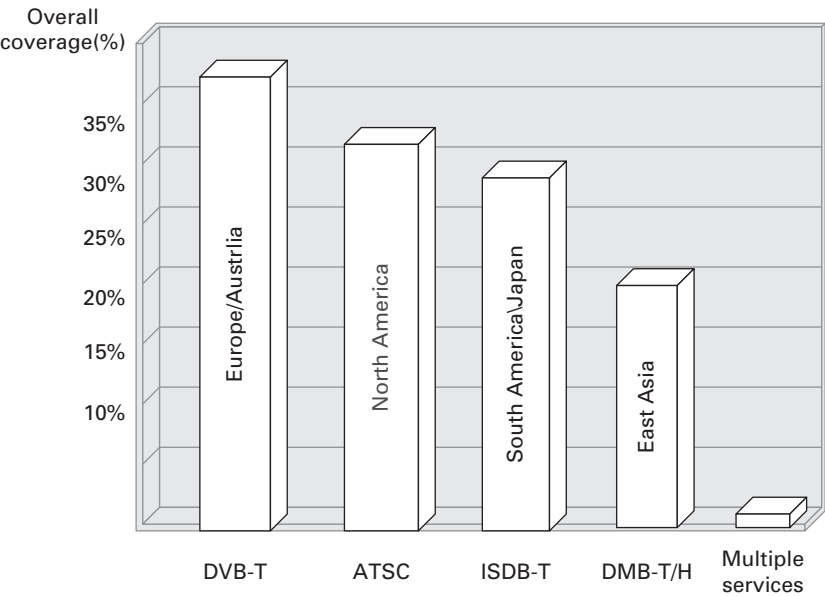


Figure 1.5 Different standards of DTT broadcasting systems.

(2) Cable TV stations operating in higher signal-to-noise ratio (SNR) environment can use the 16VSB defined by ATSC or the QAM to achieve a rate of 38.78 Mbps.

For HDTV, the ATSC defined a series of standards for wide screen 16:9 images and up to 1920×1080 pixels, which is over six times the resolution in the previous standards. Besides, it also supports many other image sizes. The type of formats can be broadly divided into two categories: High-definition television (HDTV) and standard definition TV (SDTV).

High-definition TV, transmitted through digital TV, can support several different formats. For example, 1280×720 pixels (known as *720p*) and 1920×1080 pixels (known as *1080p*). Each of these formats uses 16:9 ratio. However, traditional analog channels cannot transmit HDTV signals.

In contrast, SDTV can only use one format which depends on the technology used for transmission in different countries. In NTSC countries, the 640×480 format is used for 4:3 broadcasts and the 704×480 format is applied in 16:9 broadcasts. Meanwhile in PAL countries, the 720×576 is used in PAL for both. However, in order to save the bandwidth, broadcasting service providers may reduce resolutions [16].

DVB-T

In the 1990s, a set of standards for digital broadcasting of television, sound, and data services were adopted by the European Telecommunications Standards Institute. They have been used for cable TV, broadcast signal transmission, and satellite communication. Among all the standards, ETS 300 744 is assigned to Digital Video Broadcast-Terrestrial (DVB-T). The DVB-T standard specifies a system transmitting high-quality digital

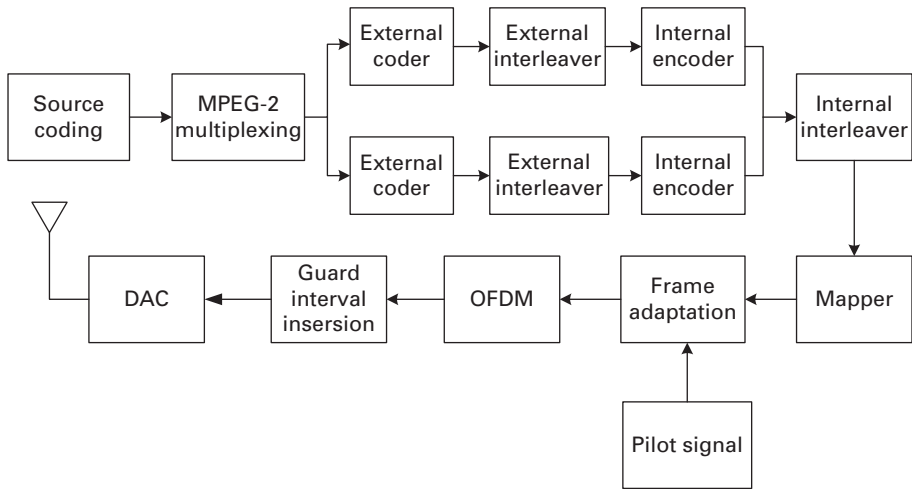


Figure 1.6 Scheme of a DVB-T transmission system.

video, audio, and data over the existing 7MHz or 8MHz channels. The digital information rate ranges from the minimum of 4.98 Mbps to the maximum of 31.67 Mbps.

In DVB-T, the data is transmitted as a series of blocks at some symbol rate during the transmission. This technique is called Coded Orthogonal Frequency-Division Multiplex (COFDM) transmission and a guard interval is used. The build-in of the guard intervals and equalization in the COFDM system greatly enhance the reception efficiency and the receiver is able to deal with the severe multipath attenuation. In some specific areas, DVB-T supports the single-frequency network (SFN) operation which allows several base-stations to simultaneously transmit the identical information on channels at the same frequency. In this case, the signals from different base-stations need to be accurately synchronized. The synchronization is based on the information attached in the bit stream and completed when the transmitter is referenced to GPS. Figure 1.6 gives the major components of a DVB-T system.

Although DVB-T is quite similar to the ATSC on the channel coding and transmission techniques, it uses a completely different modulation method. Due to the unique requirements of the base-stations and networks in Europe, SFN is used in most European countries since it can utilize the available channels more effectively. The COFDM system is selected as the most suitable. In SFN, a program is broadcasted by all base-stations on the same channel. The signals from different stations are precisely synchronized to the same reference signal. At the receiving end, a receiver may receive signals from more than one base-station with different delays, and the strength of the signal varies with time [17].

ISDB-T

Integrated Services Digital Broadcasting (ISDB) is the standard developed for Japanese digital TV and radio stations. It consists of ISDB-S, ISDB-T, ISDB-C for usages in satellite, terrestrial, and cable TV. All these three schemes are based on MPEG-2 coding

Table 1.1 Main characteristics of the three standards.

	Modulation scheme	Inner coding	Outer coding	Inner interleaver	Outer interleaver	Source coding
DVB-OFDM	8VSB, 16VSB	2/3 Trellis Code		12 Trellis Code	52 RS blocks	MPEG-2,
ATSC	COFDM, QPSK, 16/64 QAM	Multi-Rate Punctured	Reed-Solomon (RS) Code	frequency interleaving	12 RS blocks	Dolby AC-3
ISDB	BST-COFDM, DQPSK/QPSK, 16/64 QAM	Convolution Code		frequency interleaving, time interleaving		MPEG-2, MPEG-2 ACC audio

for video. The stream for transmission is as described in MPEG-2 which is capable of HDTV. Also, it supports the mobile broadcasting at the 2.6 GHz band. When using 64QAM, the system can deliver the information at the rate of 23 Mbps [18] [19].

Specifically, ISDB-T used for broadcasting has the following characteristics:

- (1) ISDB-T uses the frequency band between 470 MHz and 770 MHz, each channel is assigned 6 MHz bandwidth and there are 50 channels in total. They are called “physical channel.” On each single channel, there are 5617 carriers, each with interval of 0.99206 kHz.
- (2) In ISDB-T, each single channel is further divided into 13 segments and each segment can be flexibly allocated to the HDTV or SDTV program. For example, if one segment is used, the remaining 12 segments can be either used in one HDTV program or three SDTV programs.
- (3) ISDB-T also supports SFN which efficiently utilizes the spectrum resources and the on-channel repeater technique.

Compared to the ATSC and DVB-T, ISDB-T claims that it allows for the reception of HDTV on moving vehicles at 100 km/h while DVB-T only supports SDTV on fast-moving vehicles. For ATSC, neither HDTV nor SDTV can be received on moving vehicles. [Table 1.1](#) lists the main characteristics of the three digital TV broadcasting systems.

Although DTV is a great improvement compared to analog TV, some factors may still lead to poor performance. Due to the efficiency of the compression algorithms such as MPEG-2 and the limitation of bandwidth in DTV, the images may have some “mosquito noise” not existing in analog TV. Also, the influence brought by the multipath interference is much more significant at the receiver. In high SNR conditions, multipath can be more severe for DTV. These are problems that need to be further investigated in future research work [20].

1.2.3 Mobile TV

Mobile TV, as its name suggests, is the technology supporting users to watch TV on their cell phones or some other portable device from a service provider. The TV signals can be transmitted either through an existing cellular network or a broadcast network. Through

mobile TV, users can not only watch TV programs on portable devices such as cell phones or laptops, but also realize additional interactions conveniently in the cellular network.

The technologies used in mobile TV can be grouped into two categories that use either the cellular network (HSDPA) or broadcast network (DVB-H) to transmit the mobile TV signal. In a cellular network, the mobile TV traffic must be shared with voice and data over the network. Otherwise, a terrestrial broadcast network designed for the mobile TV is required. Nowadays, there are four commonly used standards for mobile TV: DVB-H, MediaFLO, CMMB, and MBMS. In 2010, the DVB-SH (Satellite to Handhelds) has adopted the ATSC M/H format; and the next generation of DVB-H, known as DVB-H2, is aiming to replace DVB-H as the new standard which has higher spectra efficiency and more flexible modulation.

The DVB-H standard provides a technical specification on the incorporation of the broadcast TV and the portable devices. Based on the DVB-T used for digital terrestrial TV, some new characteristics are added to meet the requirement of low-power handheld devices such as the mobile phone. In 2002, the industry agreed on the requirement for the DVB-H. It was required to provide a typical user environment, acceptable quality of broadcast and data services for portable devices, stable connection with the network during high-speed moving, and compatibility to the existing DVB-T so that the network and base-stations can be shared. For broadcasters and service providers, the DVB-H provides a way for the broadcasting service to collaborate with the cellular mobile market [21].

MediaFLO is developed by Qualcomm and is mainly used in the United States now. Instead of using an existing cellular network as in DVB-H, the data in MediaFLO is transmitted through a separate broadcast network consisting of high-power broadcast base-stations. The coverage area of each transmitter is up to 40 kilometers. During the transmission, the information is encrypted and sent on the frequency band 716–722 MHz using OFDM. However, since the transmission does not need to convey high resolution for larger screens, the speeds of MediaFLO streams are only 200–250 kb/s [22] [23].

China Multimedia Mobile Broadcasting (CMMB) is one of the main mobile TV standards in China (the other one is DTMB-H, which is developed from ATSC). The CMMB standard is developed by the State Administration of Radio, Film, and Television (SARFT). The broadcast signal is transmitted at the frequency band 470–862 MHz while each channel has bandwidth 2 or 8 MHz. Within the bandwidth, it can support 30 radio channels and 25 video channels. The CMMB is operated on the Satellite and Terrestrial Interactive Multiservice Infrastructure (StiMi), which is developed by the Chinese Academy of Broadcasting Science. As of June, 2010, the CMMB network has covered more than 317 cities in China [24].

Besides the above three standards, multimedia broadcast and multicast services (MBMS) is another broadcasting technique applied in mobile TV. It is offered via existing 3G cellular networks. Instead of point-to-point links between the service and the users, the MBMS deployed the multicast in its core network. In this way, the uplink channels can be used for interaction between users and service. One of the major benefits of the MBMS is its low cost for implementation since it is based on the existing 3G network. Many network equipment and operators in 3G network can easily support

MBMS. Furthermore, it also supports the broadcasting of the same information to multiple cellular subscribers simultaneously [25].

1.2.4 Future of wireless broadcast

Since broadcast communication has been developed over more than 100 years, by now with the application of the OFDM and appropriate coding schemes, it is very close to the Shannon capacity which is the maximum achievable transmission rate. In recent years, more and more research focus has shifted to the incorporation of the broadcast and unicast in a single network. A hybrid network which supports both cellular and broadcast communication has become the trend of the future network. In today's wireless network, with the fast growing number of users, the existing systems can hardly meet the requirement for future services. Even though the 3G network is much more powerful than the GPRS and GSM, there are still many limitations such as the transmission of large files or the support of the large data throughput. By adopting the hybrid network, it is possible to overcome those limitations. Also, at the same time, the hybrid IPTV combines the traditional broadcast TV with the IP network to deliver the program through the Internet. Besides TV services, the hybrid IPTV also supports video on demand (VOD) and catch-up TV. Many Internet applications, including gaming, online chatting, and videophone can be accessed via the IPTV network. It is believed that the Ethernet, cellular network, and broadcast network will be incorporated into one unified network in the near future.

1.3 Key technologies

To improve the system performance and provide better services, there are many advanced techniques in modern wireless communications and digital broadcasting. In this section, we discuss three key techniques: MIMO, OFDM, and cognitive radio.

1.3.1 MIMO

Multiple-input multiple-output (MIMO) technology is a breakthrough in wireless communication systems. This technology can improve the system capacity as well as performance. In a MIMO system, multiple antennas are used at the transmitter and receiver as illustrated in [Figure 1.7](#). In addition to the time domain and frequency domain, spatial resources are also exploited in a MIMO system. Due to the attractive advantages of MIMO techniques, it has already entered 3G standards; MIMO is also employed in IEEE 802.11n WLAN standard, IEEE 802.20 Mobile broadband wireless access system, and 3GPP LTE of W-CDMA.

The main advantage of the MIMO technique is that the signal at the transmit (TX) antennas and the receive (RX) antennas are “combined” such that the bit error rate (BER) or each MIMO user's data rate is improved [26]. Multiple-input multiple-output systems can improve the performance by exploiting array gain, interference reduction, spatial multiplexing gain, and diversity gain [27] [28]. Both the array gain and interference reduction can be achieved by

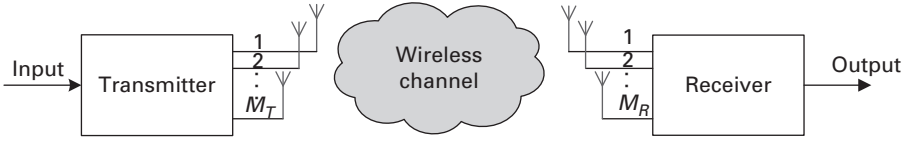


Figure 1.7 The general structure of MIMO system.

beamforming techniques. Specifically, beamforming can be deemed as a spatial filter which can filter out the undesired interference of certain directions. The concept of antenna arrays with beamforming is called “smart antennas,” which originally appeared in military communication, such as radar for target tracking. However, due to the required equipment and processing cost, the smart antenna technologies are currently mainly used in fixed stations, such as base-stations. We will briefly introduce the array gain and interference reduction, and investigate the spatial multiplexing gain and diversity gain in detail. In the following, we will consider a system with M_T transmit antennas and M_R receive antennas.

Array gain

The improvement of receiver SNR achieved by beamforming is called array gain. In a wireless communication environment, transmitted signals often propagate via a few dominant propagation paths. If these paths are known at the receiver side, the receiver can use spatial processing techniques to adjust the beam pattern such that it has high gain towards the dominant angles of reception; if these paths are known at the transmitter side, the transmit power can be concentrated within the effective angular regions. Moreover, beamforming can be used to reduce the delay spread caused by multipath propagation. Specifically, the transmitter or receiver beam pattern can be adjusted such that it can filter out the directions of dominant reflectors. In this way, the echoes with excessively large delays can be eliminated from the received signal. In practical systems, the directions of dominant propagation paths must be estimated by some techniques, such as the well-known multiple signal classification (MUSIC) algorithm [29] or the estimating signal parameter via rotational invariance techniques (ESPRIT) [30].

Interference reduction

Interference arises due to multiple users sharing time and frequency resources. When multiple antennas are applied, the spatial dimension can be exploited to reduce the interference between users. Again both transmitter and receiver beamforming techniques can be used. At the transmitter side, each user can adjust his beam pattern such that his signal is independent of other users and has a high directivity towards the desired directions. At the receiver side, each user can adjust his beam pattern such that directions of other cochannel interferers are suppressed and the desired directions are enhanced. In this way, each user can enhance his own signal-to-interference-plus-noise (SINR) ratio. Hence, the use of smart antennas for interference cancellation offers an opportunity to accommodate multiple users within the same frequency band, which is called space-division multiplexing (SDMA).

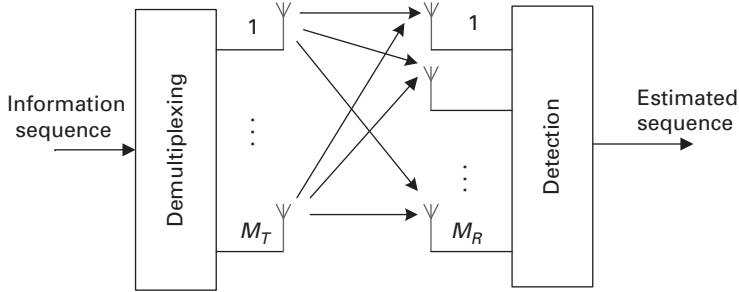


Figure 1.8 Basic structure of spatial multiplexing.

Spatial multiplexing gain

Spatial multiplexing is closely related to multi-user communications and aims to enhance the system capacity. The concept of spatial multiplexing was first proposed in [31]. From the information-theoretic point of view, the capacity of a MIMO system with M_T transmit and M_R receive antennas grows linearly with the minimum of M_T and M_R [32] [33]. In contrast, the single antenna systems' capacity can only be increased logarithmically with SNR. Multiplexing gain can be obtained by transmitting data sequences independently from the individual antennas. The basic structure to realize spatial multiplexing is shown in Figure 1.8.

The information sequence is split into M_T subsequences which we call demultiplexing, and then the subsequences are modulated and transmitted simultaneously over the M_T antennas under the same frequency band and time slot. At the receiver side, the received signals are separated by interference cancellation techniques. The optimum MIMO receiver design is an important issue. The existing detection algorithms include: zero-forcing (ZF) criterion, minimum-mean-square-error (MMSE) criterion, and maximum likelihood (ML) criterion, etc. For the transceiver structure, the Bell Laboratories Layered Space-Time (BLAST) detector provides an effective detection strategy that is similar to the successive interference cancellation method. Specifically, the BLAST detector first selects the subsequence with the largest SNR among M_T subsequences, and estimates this subsequence while treating other layers as interference. Then it subtracts this subsequence from the received signals, and repeat the above steps until the bits of all M_T subsequences are detected. All above detectors require M_R at least equal to M_T .

To guarantee the error performance for spatial multiplexing schemes, channel coding is usually required. For spatial multiplexing schemes, three different channel coding methods can be used: horizontal coding, vertical coding, and diagonal coding. Horizontal coding means channel encoding is done after the de-multiplexer and channel decoding is done individually for each information subsequence. For vertical coding, the channel encoding is done before the de-multiplexer which can offer an additional spatial diversity gain at the compensation of decoding complexity. The reason is that under vertical coding, the received information has to be decoded jointly at receivers. Diagonal coding, which is the combination of horizontal and vertical coding, can offer the same spatial

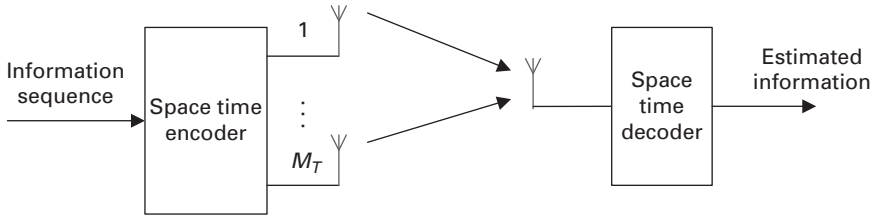


Figure 1.9 The structure of transmit diversity.

diversity advantage as vertical coding while having the same receiver complexity as horizontal coding [34].

Diversity gain

Different from the multiplexing techniques, diversity mainly aims to enhance the system error performance. The spatial diversity has two types: macroscopic diversity which is related to shadowing effects; microscopic diversity which is available in rich scattering environments with multipath fading [35]. Diversity can be realized both at receiver side and transmitter side.

The diversity reception techniques are carried out in systems with a single transmit antenna and multiple receive antennas. They perform a combining of the individual received signals, in order to provide a microscopic diversity gain. In a frequency flat fading environment, the optimum combining techniques include maximum ratio combining (MRC), equal gain combining (EGC), and selection diversity (SD). The MRC requires perfect channel knowledge at the receiver side, and the combining weight is proportional to its channel state; the EGC adds up the received signals with equal weight; while the SD only selects the received signal with the maximum instantaneous SNR.

The transmit diversity is realized by sending redundant signals representing the same information sequence over multiple ideally independent fading paths in time/frequency/space domain. To facilitate the coherent detection at the receiver, a preprocessing is required prior to transmission. The basic structure of transmit diversity techniques is shown in Figure 1.9.

The preprocessing is performed by the space-time encoder at the transmitter; the corresponding decoding process is carried out by the space-time decoder at the receiver. Among those space-time coding schemes, two famous transmit diversity techniques are Alamouti coding [36] and space-time trellis codes (STTCs) [37]. Alamouti performs an orthogonal space-time transmission; at the receiver side ML detection can be easily realized by simple linear processing. Space-time trellis codes can be deemed as trellis-coded modulation to transmit antennas. Optimum decoding of STTC is based on the Viterbi algorithm which has high complexity. Motivated by the simple receiver structure of Alamouti code, orthogonal space-time block codes (OSTBCs) were proposed [38], which can have more than two transmit antennas. Moreover, the STTCs and OSTBCs can be combined. All the above space-time codes offer at most the same data rate as an uncoded single-antenna system. In contrast, several high-rate space-time transmission schemes

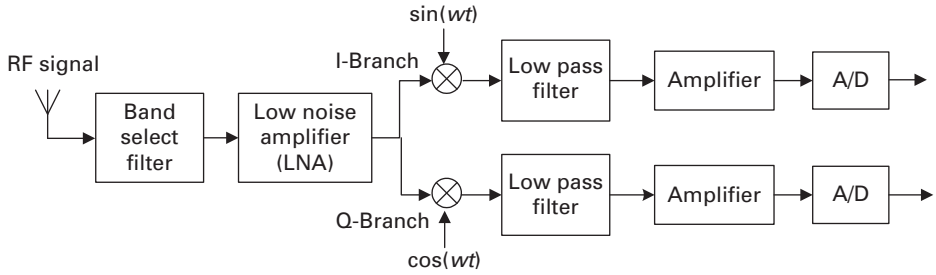


Figure 1.10 Full-parallel MIMO RF receiver.

with higher speed than an uncoded single antenna system have been proposed. Some of these schemes combine the space-time codes and the BLAST scheme, such as linear dispersion codes [39]. Some schemes generalize the channel codes to multiple antennas, such as repeat-accumulate (RA) codes [40]. Recently, distributed MIMO systems are becoming a hot topic, where multiple transmitting nodes and receiving nodes cooperate in terms of a joint transmission/reception strategy. In distributed MIMO systems, multiple nodes can share their transmit antennas by employing distributed space-time-coding schemes.

Basic work on MIMO RF receiver architectures is reported in [41] [42]. In conventional MIMO systems, transmit antennas carry uncorrelated signals, which indicates that the number of receiver RF front-ends should be the same as the number of transmit antennas to recover the uncorrelated signals. Hence, the direct approach of MIMO RF front-end implementation is to use one dedicated receiver per antenna. This structure is called a full-parallel MIMO RF receiver. Figure 1.10 is the receiver structure for each antenna [43]. Moreover, the distance between each antenna should be at least a half-wavelength to ensure the uncorrelated fading characteristics which will result in a large receiver size [44]. Obviously, the cost of implementing multiple antenna structure with multiple RF chains is too high for cost-sensitive applications. Furthermore, small-sized mobile terminals fit poorly for multi-antenna implementations [45]. Thus, the RF front-end design in MIMO system is an important task.

1.3.2 OFDM

Orthogonal Frequency-Division Multiplex is a multi-carrier transmission technique to achieve high data rate in a multipath-fading environment. The key idea is to divide the total available wireless channel into many subchannels. With the decrease of the bandwidth in each subchannel, the symbol duration per subchannel increases. Hence, the overall wideband frequency-selective fading can be treated as flat fading on each subchannel. At the receiver side, only a trivial frequency domain single-tap equalizer is needed to overcome the overall frequency selective fading. Comparing with the traditional FDM technique, OFDM has higher frequency efficiency by allowing the adjacent independent (orthogonal) subcarriers to overlap with each other without individual carrier guard band, as is shown in Figure 1.11.

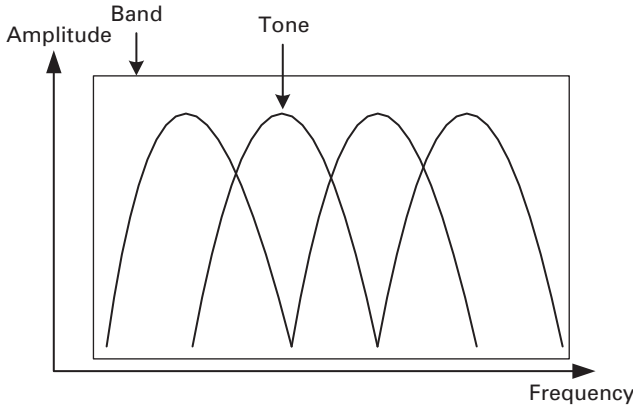


Figure 1.11 Subcarrier overlap in OFDM system.

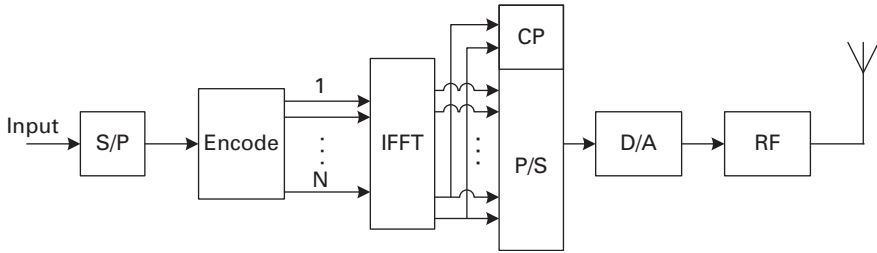


Figure 1.12 OFDM transmitter structure.

The concept of OFDM was first proposed in 1966 [46]. However, to realize subcarriers' orthogonality was a bottleneck for the development of the OFDM technique. Until 1971, when Weinstein and Ebert [47] introduced the inverse discrete Fourier transform (IDFT) and discrete Fourier transform (DFT) for the baseband modulation and demodulation, respectively. In a frequency selective channel, inter-symbol-interference (ISI) cancellation was still a problem. In 1980, Peled and Ruiz proposed cyclic prefix (CP) to maintain the subcarriers' orthogonality [48]. Today the modern OFDM concept and system model have been formed. The OFDM system transmitter and receiver structures are shown in Figure 1.12 and Figure 1.13, respectively.

One distinctive advantage of OFDM is that the orthogonality between subcarriers is easily maintained with low complexity FFT. In particular, only a trivial frequency domain single-tap equalizer is needed to combat the frequency selective fading effect, which requires costly equalizers in single carrier systems [49]. The size of the FFT in an OFDM design should be chosen carefully. With a larger size of FFT, the system can have better protection against the frequency selective fading; however, the OFDM system will be more sensitive to Doppler shift, and the total design cost will also increase. To further improve OFDM systems' performance, OFDM resource allocation is extensively

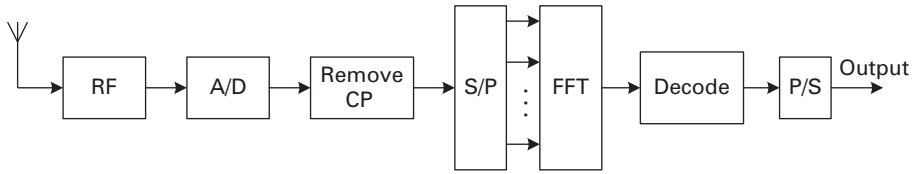


Figure 1.13 OFDM receiver structure.

investigated. By adapting each subchannel's transmission power and data rate, OFDM resource allocation can exploit the frequency selectivity in a wideband channel.

In OFDM systems, the channel state information (CSI) is required in coherent detection and resource allocation. Usually, the channel estimation is realized by inserting training symbols known at both the transmitter and receiver. It includes two types of channel estimation: pilot aided channel estimation (PACE) and decision directed channel estimation (DDCE). For PACE, the training symbols (also known as pilot) can be inserted according to certain patterns. In channel estimation, the CSI at the pilot symbols' positions is first obtained, and then by applying some interpolation methods, the CSI of the rest of the subchannels can be estimated. For the DDCE method, the training symbols are arranged at the first few OFDM blocks. Then the CSI corresponding to the first few blocks can be estimated and used to detect and demodulate the symbols at the following few data blocks. In this case, the training symbols should be inserted periodically to keep up with the channel variation. Comparing these two methods, DDCE has lower complexity, though it may result in error propagation in fast fading channels; the PACE can reliably estimate channels both in static channels and time-varying channels with comparatively higher complexity than DDCE [50].

Although frequency selective fading can be combated by choosing proper CP length and with proper CSI estimation, OFDM is very sensitive to the time selective fading which breaks the orthogonality between OFDM subcarriers and results in inter channel interference (ICI). The time-varying impairment is caused by the frequency offset or/and Doppler shift. To combat frequency offset, multiple frequency offset estimation and correction methods were proposed; to mitigate the ICI caused by the Doppler Spread, frequency equalization, time domain windowing, coding, etc. were investigated [51] [52] [53] [54].

Regarding the RF section, OFDM modulation designs face many key challenges, which include power consumption, linearity, image rejection, phase distortion, and phase noise [55]. The time domain OFDM signal is the superposition of signals of all the subcarriers. Hence, the distribution of time domain OFDM signal approximately belongs to Gaussian distribution of the complex baseband signal. When all of the subcarriers have the same phase, it will result in a peak in the OFDM signal power. The peak-to-average power ratio (PAPR) is defined as the ratio of the peak signal power to the average signal power. According to this definition, the PAPR of OFDM signal is $10 \log(N)$ dB, in which N represents the number of total subcarriers. To solve the high PAPR problem, the OFDM system requires not only high resolution of DA and AD converters but also a wide linear range high-power amplifier which is hard to implement [56]. To minimize the OFDM system performance degradation due to PAPR, various approaches have been proposed

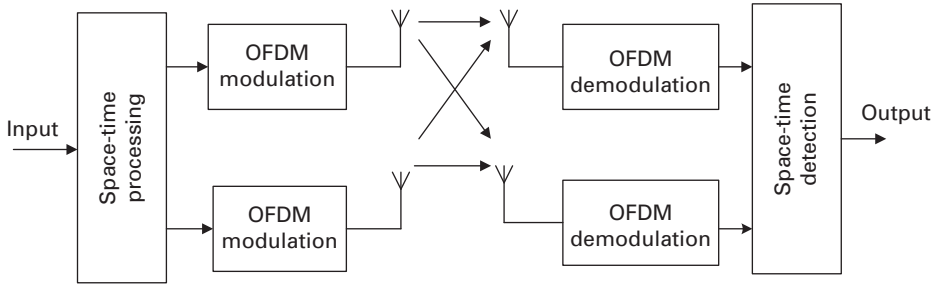


Figure 1.14 MIMO-OFDM system.

which include: signal clipping and filtering [57], [58], selected mapping [59], partial transmit sequence [60], and symbol-scrambling technique [61], as well as digital predistortion and crest factor reduction, which will be extensively discussed in the other related chapters of this book.

Orthogonal Frequency-Division Multiplex has been adopted in multiple wireless standards, such as: WLAN 802.11a/g/n [62], WMAN 802.16 d/e [63], UWB 802.15.3a [64], DAB [65], and DVB-T [66]. Currently, OFDM is used in 3GPP LTE techniques and IMT-advanced systems as a promising technique for 4G wireless communication. It is also being used for dedicated short-range communications, such as road side to vehicle communications.

Most of the aforementioned MIMO techniques are for flat fading channels. For broadband communication that suffers greatly from frequency selectivity, the combination of OFDM with MIMO will become a very promising choice. Because OFDM can easily combat frequency selectivity and MIMO can achieve much higher capacity or bit error performance than SISO system; hence, the MIMO-OFDM system has become a hot topic in the past few years [67] [68] [69] [70]. The general structure of the MIMO-OFDM system is shown in Figure 1.14.

1.3.3 Cognitive radio

In recent years, the communication demands have increased dramatically, and this increase challenges the traditional fixed spectrum assignment policies. According to the present frequency usage, some licensed frequency bands are seldom used, some bands are partially occupied, and only part of the frequency bands are heavily used [71]. This indicates that the fixed spectrum assignment policy greatly affects the frequency efficiency. Hence, cognitive radio (CR) which can use a primary network's frequency band in an opportunistic way or a sharing way was proposed to improve the spectrum efficiency. Specifically speaking, the CR should intelligently detect the primary network's spectrum holes as well as communication requirements and then adjust its own system parameters in real-time. The parameters adjusted in a CR system could be transmission power, frequency band, and the modulation scheme. Cognitive radio systems should meet two basic objectives: first is to support highly reliable communications anytime and anywhere; second is to enhance the spectrum efficiency [72].

Cognitive radio systems are expected to enhance the frequency efficiency through opportunistic spectrum access techniques. To share the spectrum with primary users without degrading their communication quality and to meet the different applications' diverse quality of service requirements, each CR user must have following functions [73]:

- (1) Spectrum sensing. Cognitive users should be able to detect spectrum holes and intelligently use the spectrum without causing harmful interference to other users.
- (2) Spectrum management. The detected spectrum should be allocated properly among cognitive users to best meet the communication requirement as well as the related internal and external policies.
- (3) Spectrum mobility. When a primary user has a communication requirement on a frequency band, the corresponding cognitive user needs to switch to another available spectrum while maintaining the seamless communication.
- (4) Spectrum sharing. The spectrum should be scheduled properly among coexisting users to avoid the collision.

One of the key challenges of cognitive radio is to accurately detect the existence of licensed users over a wide spectrum range which indicates that the challenge of cognitive radio lies in the RF front-end's wideband sensing capability. An RF front-end is composed of wideband antenna, power amplifier, and adaptive filter. Owing to the primary network's wide band property, a cognitive RF front-end should work effectively at any part of the wide frequency band. Furthermore, a real-time sensing ability is required to maintain the communication quality of the primary network. Also, because signals in a primary network have various power levels, bandwidth, and transmission environment, the RF front-end should have a wide dynamic range to detect primary network signals. However, this wide dynamic range requires A/D converters with GHz speed, which has been infeasible till now. One solution to this problem is to filter out strong signals first. Since strong signals can exist at anywhere in the wide spectrum range, tunable notch filters are needed [74]. Another possible solution is to use MIMO so that the filtering can be realized in the spatial domain which is easier comparing with the filtering in the frequency domain [75] [76].

The first cognitive radio standard is IEEE 802.22 that is designed to operate in the TV broadcast bands. This technology is used under the premise that the present operation and low-power licensed devices won't be affected. Recently, cognitive radio has already been used in femtocells for interference avoidance [77].

1.4 Conclusions

In this chapter, we provide a brief overview of wireless communication and digital broadcasting. We classify the wireless applications as unicasting and broadcasting and focus the discussion on cellular network and TV broadcast network. We first introduce the evolution of these two networks. Then the basic principles of MIMO, OFDM, and

cognitive radio are reviewed. In addition to these techniques, there are many other technology advancements in wireless communication and digital broadcasting that will be discussed in detail in this book.

In the past few decades, the success of wireless communication has substantially changed the way we live thanks to the technological innovations. However, there are still plenty of challenges ahead when wireless industry moves forward. For example, current wireless providers have started providing multiple services (voice, data, and TV) simultaneously to accommodate the various increasing customer demands [78] [79]. As a result, more and more radios are being integrated into a single wireless platform to allow maximum connectivity, making multi-radio coexistence inevitable. Till now, all existing solutions assume that multi-radios are operated independently over different frequency bands, which are in nature spectrum inefficient and make further radio integrations difficult. To solve such problems, cooperative hybrid transmission and network convergence become notable future evolution trends. As pointed out in [1], by exploiting the possible inter-network cooperation among different networks, the spectrum efficiency and the overall quality of service can be further improved.

References

- [1] H. Li B. Liu, and H. Liu, Transmission schemes for multicarrier broadcast and unicast hybrid systems, *IEEE Transaction on Wireless Communications*, **7** (2008), 4321–30.
- [2] A. Goldsmith, *Wireless Communications* (Cambridge University Press, 2005).
- [3] V. H. McDonald, *The Cellular Concept*, *Bell System Technical Journal* (1979).
- [4] http://en.wikipedia.org/wiki/History_of_mobile_phones.
- [5] IEEE 802.16 – 2004, *IEEE Standard for Local and Metropolitan Area Network – Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, June 24, 2004.
- [6] Intel, *Broadband Wireless: The New Era in Communications*, Technical White Paper, 2004.
- [7] S. J. Uma, and P. Ramjee, *OFDM Towards Fixed and Mobile Broadband Wireless Access* (Artech House, 2007).
- [8] www.matrixgulf.com/History.aspx.
- [9] http://en.wikipedia.org/wiki/High-Speed_Downlink_Packet_Access
- [10] W. C. Y. Lee, *Wireless and Cellular Communications* (McGraw-Hill, 2005).
- [11] Motorola Inc., *Long Term Evolution (LTE): A Technical Overview*, Motorola Technical White Paper, 2007.
- [12] R. L. Freeman, *Telecommunication System Engineering* (John Wiley and Sons, 2004)
- [13] FCC, *The Digital TV Transition: Will You Be Affected?*, www.dtv.gov/affected.html, 2009.
- [14] H. Sari, G. Karam and, I. Jeanclaude, Transmission techniques for digital terrestrial TV Broadcasting, *IEEE Communication Magazine*, **33** (1995), 100–9.
- [15] Algorith, *Traditional Broadcasting in a World with New Options*, White Paper, 2008.
- [16] ATSC, *ATSC Digital Television Standard – ATSC Specifications*, 1995.
- [17] Digital Video Broadcasting (DVB), *A Guideline for the Use of DVB Specifications and Standards*, 2008.
- [18] S. Nakahara, S. Moriyama, T. Kuroda *et al.*, Efficient use of frequencies in terrestrial ISDB system, *IEEE Transaction On Broadcasting*, **42** (1996), 173–8.

- [19] S. Moriyama, M. Takada S. Nakahara and H. Miyazawa, *Progress Report of ISDB-T System*, Broadcast Asia, 2000.
- [20] Y. Wu, Performance comparison of ATSC 8-VSB and DVB-T COFDM transmission systems for digital television terrestrial broadcasting, *IEEE Transactions on Consumer Electronics*, **45** (1999), 916–24.
- [21] European Broadcasting Union (EBU), *DVB-H – The Emerging Standard for Mobile Data Communication*, EBU Technical Review, 2000.
- [22] Qualcomm, *Live Datacasting* (Qualcomm Press, 2006).
- [23] M.R. Chari and F-y Ling, FLO physical layer: an overview, *IEEE Transaction on Broadcasting*, **53** (2007), 145–60.
- [24] Rohde and Schwarz, CMMB Information, 2010.
- [25] F. Hartung and U. Horn, Delivery of broadcast services in 3G networks, *IEEE Transaction on Broadcasting*, **53** (2007), 188–99.
- [26] D. Gesbert, M. Shafi, D. S. Shiu, P. J. Smith and A. Naguib, From theory to practice: an overview of MIMO space–time coded wireless systems, *IEEE Journal on Selected Areas in Communications*, **21** (2003), 281–302.
- [27] E. Biglieri, R. Calderbank and, A. Constantinides *et al.*, *MIMO Wireless Communications* (Cambridge University Press, 2007).
- [28] J. Mietzner, R. Schober, L. Lampe, W. H. Gerstacker, and P. A. Hoeher, Multiple-antenna techniques for wireless Communications – a comprehensive literature survey, *IEEE Communications Surveys & Tutorials*, **11** (2009), 87–105.
- [29] R. O. Schmidt, Multiple Emitter Location and Signal Parameter Estimation, *Proc. RADC Spectral Estimation Workshop*, 1979, 243–58.
- [30] R. H. Roy and T. Kailath, ESPRIT – estimation of signal parameters via rotational invariance techniques, *IEEE Transaction on Acoustic, Speech and Signal Processing*, **37** (1989), 984–95.
- [31] A. Paulraj and T. Kailath, *Increasing Capacity in Wireless Broadcast Systems Using Distributed Transmission/directional Reception*, US Patent 5345599. 1992-02-21.
- [32] G. J. Foschini, Layered space–time architecture for wireless communication in a fading environment when using multi-element antennas, *Bell System Technical Journal* (1996) 41–59.
- [33] I. E. Telatar, Capacity of multi-antenna Gaussian channels, *European Transaction on Telecommunication*, **10** (1999), 585–95.
- [34] G. J. Foschini, D. Chizhik, M. J. Gans, C. Papadias, and R. A. Valenzuela, Analysis and performance of some basic space-time Architectures, *IEEE Journal on Selected Areas in Communications*, **21** (2003) 303–20.
- [35] S. N. Diggavi, N. Al-Dhahir, A. Stamoulis, and A. R. Calderbank, Great expectations: The value of spatial diversity in wireless networks, *Proceedings of IEEE*, **92** (2004) 219–70.
- [36] S. M. Alamouti, A simple transmit diversity technique for wireless communications, *IEEE Journal on Selected Areas in Communications*, **16** (1998) 1451–8.
- [37] V. Tarokh, N. Seshadri, and A. R. Calderbank, Space-time codes for high data rate wireless communication: Performance criterion and code construction, *IEEE Transactions on Information Theory*, **44** (1998), 744–65.
- [38] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, Space-time block codes from orthogonal designs, *IEEE Transactions on Information Theory*, **45** (1999) 1456–67.
- [39] B. Hassibi and B. M. Hochwald, High-rate codes that are linear in space and time, *IEEE Transactions on Information Theory*, **48** (2002), 1804–24.

- [40] S. ten Brink and G. Kramer, Design of repeat-accumulate codes for iterative detection and decoding, *IEEE Transactions on Signal Processing*, **51** (2003) 2764–72.
- [41] R. Krenz and K. Wesolowski, Comparative study of space-diversity techniques for MLSE receivers in mobile radio, *IEEE Transactions on Vehicular Technology*, **46** (1997), 653–63.
- [42] T. Hashimoto, A list-type reduced-constraint generalization of the Viterbi algorithm, *IEEE Transactions on Information Theory*, **33** (1987), 866–76.
- [43] K. Thomas, *Smart Antenna: State of the Art*. (Hindawi Publishing Corporation, 2005).
- [44] E. Biglieri, J. Proakis, and S. Shamai, Fading channels: information-theoretic and communications aspects, *IEEE Transactions on Information Theory*, **44**, (1998) 2619–92.
- [45] A. Kalis, A. Kanatas, and C. Papadias, A novel approach to MIMO transmission using single RF front end, *IEEE Journal on Selected Areas in Communication*, **26** (2008), 972–80.
- [46] R. W. Chang, Synthesis of band-limited orthogonal signals for multichannel data transmission, *Bell System Technical Journal*, **45** (1966) 1775–97.
- [47] S. B. Weinstein and P. M. Ebert, Data transmission of frequency division multiplexing using the discrete frequency transform, *IEEE Transactions on Communications*, **19** (1971), 623–34.
- [48] A. Peled and A. Ruiz, Frequency domain data transmission using reduced computational complexity algorithm, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Denver 1980, 964–67.
- [49] J. Z. Wang, *Broadband Wireless Communications: 3G, 4G, and Wireless LAN* (Kluwer Academic, 2001).
- [50] T. Hwang, C. Yang, G. Wu, S. Li, and Y. Li, OFDM and its wireless applications: a survey, *IEEE Transactions on Vehicular Technology*, **58** (2009), 1673–94.
- [51] Y. Mostofi and D. C. Cox, ICI mitigation for pilot-aided OFDM mobile systems, *IEEE Transactions on Communication*, **4** (2005), 765–74.
- [52] A. Seyedi and G. J. Saulnier, General ICI self-cancellation scheme for OFDM systems, *IEEE Transactions on Vehicular Technology*, **54** (2005), 198–210.
- [53] J. Armstrong, Analysis of new and existing methods of reducing intercarrier interference due to carrier frequency offset in OFDM, *IEEE Transactions on Communication*, **47** (1999), 365–9.
- [54] Y. Fu, C. Tellambura, and W. A. Krzymien, Transmitter precoding for ICI reduction in closed-loop MIMO OFDM systems, *IEEE Transactions on Vehicular Technology*, **56** (2007), 115–25.
- [55] www.eetimes.com/electronics-news/4140979/The-OFDM-Challenge.
- [56] G. Fettweis, E. Zimmermann, V. Jungnickel, and E. A. Jorswieck, Challenges in future short range wireless systems, *IEEE Vehicular Technology Magazine*, **1** (2006), 24–31.
- [57] X. Li and L. J. Cimini, Effects of clipping and filtering on the performance of OFDM, *IEEE Communication Letters*, **2** (1998) 131–3.
- [58] J. Armstrong, Peak-to-average power reduction for OFDM by repeated clipping and frequency domain filtering, *Electronic Letters*, **38** (2002), 246–7.
- [59] R. W. Bäuml, R. F. H. Fischer, and J. B. Huber, Reducing the peak to-average power ratio of multicarrier modulation by selected mapping, *Electronic Letters*, **32** (1996), 2056–57.
- [60] S. H. Müller and J. B. Huber, OFDM with reduced peak-to-average power ratio by optimum combination of partial transmit sequences, *Electronic Letters*, **33** (1997), 368–9.
- [61] P. V. Eetvelt, G. Wade, and M. Thompson, Peak to average power reduction for OFDM schemes by selective scrambling, *Electronic Letters*, **32** (1996), 1963–4.

- [62] IEEE Part11: Wireless LAN Medium Access Control and Physical Layer specifications. High-speed Physical Layer in the 5 GHz Band, 1999.
- [63] IEEE Part16: Air Interface for Fixed Broadband Wireless Access Systems Amendment 2: Medium Access Control modifications and additional physical layer specifications for 2–11 GHz, 2003.
- [64] A. Batra, J. Balakrishnan, A. Dabak *et al.*, Multi-band OFDM physical layer proposal for IEEE 802.15 Task Group 3a, *MultiBand OFDM Alliance SIG*, 2004.
- [65] World DAB Forum. www.worlddab.org/.
- [66] ETSI, ETSI EN 300 744 v1.5.1: *Digital Video Broadcasting (DVB); Framing Structure, Channel Coding and Modulation for Digital Terrestrial Television*, European Broadcasting Union, 2004. Available: www.etsi.org
- [67] H. Bölcskei, D. Gesbert, and A. J. Paulraj, On the capacity of OFDM based spatial multiplexing systems, *IEEE Transactions On Communication*, **50** (2002) 225–34.
- [68] H. Yang, A road to future broadband wireless access: MIMO-OFDM based air interface, *IEEE Communication Magazine*, **43** (2005), 53–60.
- [69] M. Jiang and L. Hanzo, Multiuser MIMO-OFDM for next-generation wireless systems, *Proceedings of IEEE*, **95** (2007), 1430–69.
- [70] H. Li; G. Ru; Y. Kim, and H. Liu, OFDMA capacity analysis in MIMO channels, *IEEE Transactions on Information Theory*, **56** (2010), 4438–46,.
- [71] G. Staple and K. Werbach, The end of spectrum scarcity, *IEEE Spectrum*, **4** (2004) 48–52.
- [72] S. Haykin, Cognitive radio: brain-empowered wireless communications, *IEEE Journal on Selected Areas in Communications*, **23** (2005), 201–20.
- [73] I. F. Akyildiz,; W. Y. Lee, M. C. Vuran, and S. Mohanty, NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey, *Elsevier Computer Networks*, **50** (2006), 2127–59.
- [74] D. Cabric, S. M. Mishra, and R. W. Brodersen, Implementation issues in spectrum sensing for cognitive radios, *38th Asilomar Conference on Signals, Systems and Computers 2004*, 2004, 772–6.
- [75] D. Cabric and R. W. Brodersen, Physical layer design issues unique to cognitive radio systems, *IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, **2** (2005), 759–63.
- [76] I. F. Akyildiz, W. Y. Lee, M. C. Vuran, and S. Mohanty, NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey, *Elsevier Computer Networks*, **50** (2006), 2127–59.
- [77] Node-H Shows Cognitive Radio Femtocell Technology. www.node-H.com/MWC2010.html, 2010.
- [78] J-N. Hwang, *Multimedia Networking: from Theory to Practice* (Cambridge University Press, 2009).
- [79] J-N. Hwang, Wireless MediaNets: application driven next generation wireless IP networks (invited), *ACM Multimedia System Journal*, 2010, 85–114.

2 Basic principles and protocols of digital wireless communications

Dimitra I. Kaklamani and Panagiotis K. Gkonis

2.1 Introduction

Recent advantages in wireless communications have enabled reliable long-range coupling, high data rate services at mobile users and support of multimedia content. The vast majority of currently operating wireless networks is based on digital processing at both transmission ends, where incoming analog signals (e.g. voice) are converted into digital bit streams. Digital transmission offers increased reliability and flexibility when compared to analog transmission, as well as compatibility among different digital systems. The boost in wireless digital communications came along with the development of small-scale integrated circuits, that made feasible the deployment of cost-effective and time-efficient implementations of complex operations related with digital processing (e.g. sampling, quantization, matrix inversion, fast Fourier transform, etc.). However, the performance of a wireless network depends not only on the installed hardware at both transmission ends but on a number of issues related to system protocols and architecture, such as transceiver signal processing techniques and resource management strategies. The main challenging task in the design of a wireless link is the provision of acceptable quality of service at fixed or mobile users, despite the channel conditions. The latter term refers to channel variations either due to multipath propagation or due to pathloss. Moreover, in wireless networks different users share the same physical access medium; hence another challenging task is the implementation of effective physical layer protocols that allow the simultaneous transmission and reception from different users as well as the maximization of per user capacity.

In 1998, the Wideband Code Division Multiple Access (WCDMA) physical layer protocol was adopted for third-generation (3G) mobile networks, which offer higher data rates and multimedia communications when compared to second-generation mobile networks. However, as the demand for even higher data rates is constantly increasing, the research on spectrum efficient physical layer architectures continues to be important. In April of 2009, the Orthogonal Frequency Division Multiple Access (OFDMA) physical layer protocol was proposed for the Long Term Evolution (LTE) of the currently operating 3G networks. Fourth-generation (4G) wireless networks will be in position to provide among other features peak data rates up to 300 Mb/s and a radio-network delay less than 5 ms.

Another promising technique that increases the overall capacity of a wireless network without additional spectrum requirements is the use of antenna arrays at both transmission ends, also known as MIMO (multiple-input multiple-output). The MIMO techniques can provide array gain and diversity gain as well as spatial multiplexing gain. However, these benefits cannot be obtained at the same time due to contradictory demands. The integration of MIMO transmission techniques in WCDMA/OFDMA networks has attracted scientific interest over the last few years, where a considerable number of studies have focused on the deployment of MIMO technology versus the complexity of the transceiver.

The goal of this chapter is to provide an overview of the basic principles of digital transmission in wireless networks. Emphasis is given in transmission over MIMO networks where space-time and space-frequency transmission techniques are presented in alignment with the WCDMA/OFDMA physical layer protocols. In [Section 2.2](#), the basic principles of digital processing and transmission are described, while in [Section 2.3](#) a brief analysis of the WCDMA/OFDMA protocols is introduced. In [Section 2.4](#), an overview of MIMO technology is provided, while space-time and space-frequency techniques applicable in WCDMA and OFDMA networks are described in [Section 2.5](#). Finally, [Section 2.6](#) summarizes the main parts described in this chapter.

2.2 Digital processing, transmission, and reception

A typical diagram of the basic procedures in digital processing and transmission is shown in [Figure 2.1](#). The incoming signal from an analog source is sampled at discrete time intervals and the output samples are quantized. Denoting $f_{\max}$ the maximum frequency of the analog input signal, T_s the sampling interval, and $f_s = 1/T_s$ the sampling frequency, then according to Nyquist theorem the following condition must be satisfied in order to reconstruct the signal at the receiver:

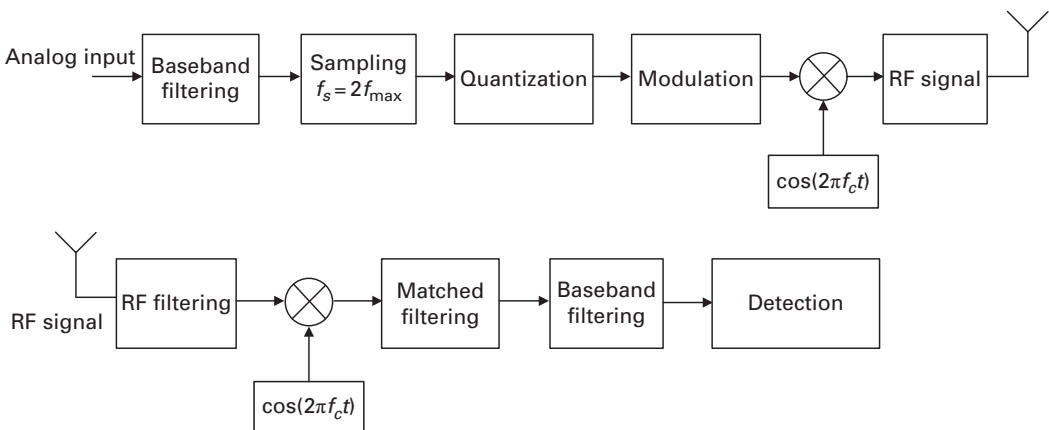


Figure 2.1 Block diagram of the basic transmit–receive procedures in digital communications.

$$f_s \geq 2f_{\max}. \quad (2.1)$$

In the quantization procedure, each sample is converted to the closest value from a set of discrete amplitude levels whose number is usually a power of two. Consider for example an analog signal ranging in space $[-A, A]$ where A is the maximum amplitude. Assuming 2^L discrete amplitudes where $L \in \mathbb{N} - \{0\}$ and denoting Q the absolute space between two quantization levels, then the following condition is satisfied:

$$(2^L - 1)Q = 2A \Rightarrow Q = \frac{2A}{2^L - 1}. \quad (2.2)$$

For example, if $A = 7$ and $L = 3$, then $Q = 2$. These amplitudes are then converted to their binary representations. For eight quantization levels, the corresponding amplitudes are $-7, -5, -3, -1, 1, 3, 5, 7$ and their binary representations $000, 001, 010, 011, 100, 101, 110, 111$. If for example the input value to the quantizer is 5.3 , then the output value will be 5 and its binary representation 110 . This procedure is depicted in Figure 2.2.

The next step after quantization is digital modulation. In BPSK modulation for example, the mapping of digital bit streams is $1 \rightarrow 1$ and $0 \rightarrow -1$ while in QPSK modulation

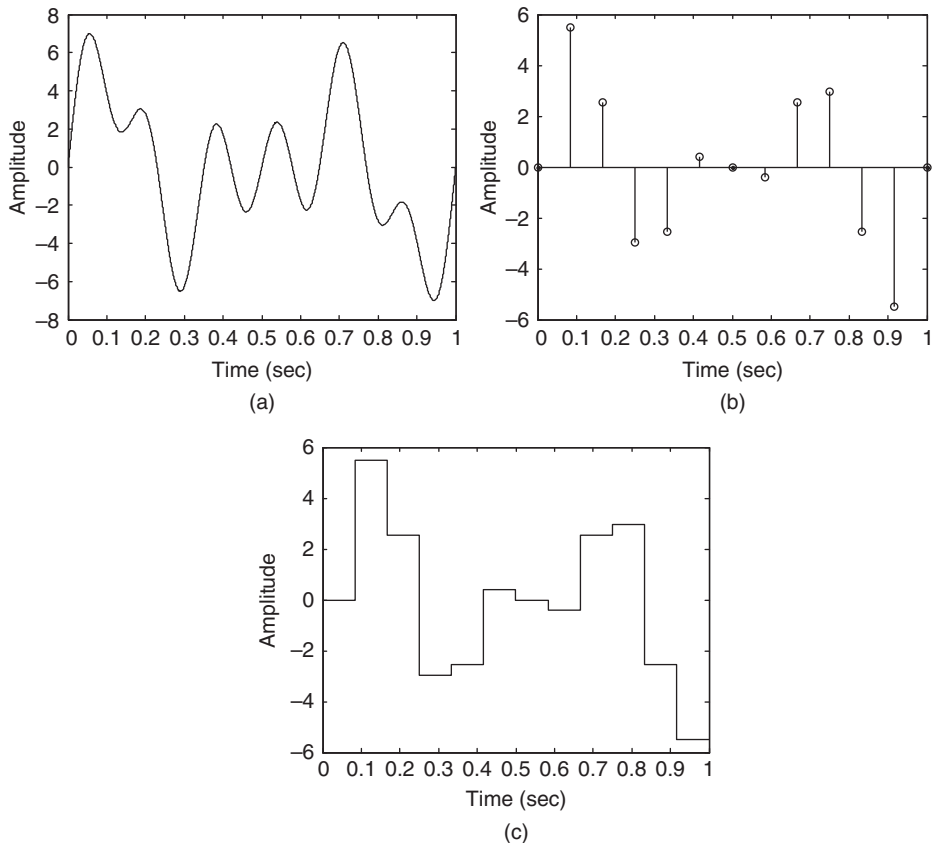


Figure 2.2 Sampling and quantization of an analog input signal.

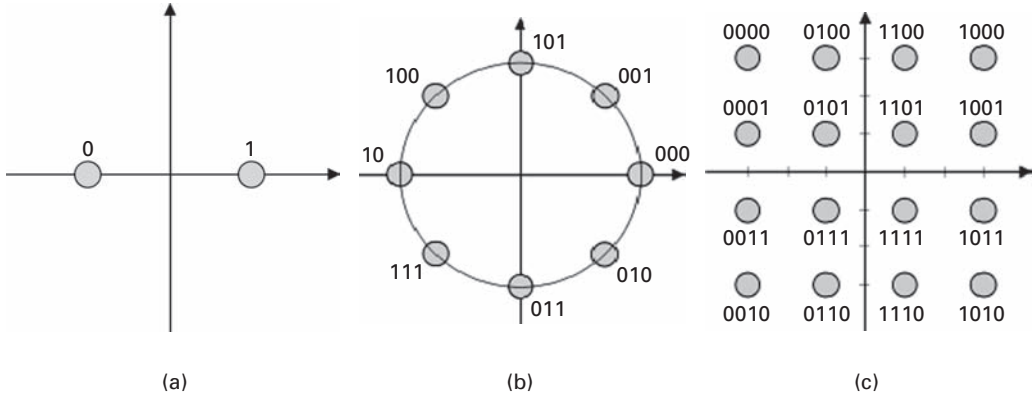


Figure 2.3 (a) BPSK, (b) 8-PSK, and (c) 16-QAM signal constellations.

the mapping is $00 \rightarrow 1$, $01 \rightarrow j$, $11 \rightarrow -1$, $10 \rightarrow -j$ where j is the imaginary unit. In general, a signal constellation is given by $Ae^{j\varphi}$, where A is the amplitude and φ the phase. In PSK modulation for example, A is fixed at a certain value and the phase varies. In BPSK $\varphi = \{0, 90^\circ\}$, in QPSK $\varphi = \{0, 90^\circ, 180^\circ, 270^\circ\}$ while in 8-PSK $\varphi = \{0, 45^\circ, 90^\circ, 135^\circ, 180^\circ, 225^\circ, 270^\circ\}$. The signal constellations for BPSK, QPSK and 16-QAM modulations are depicted in Figure 2.3. In general, high order modulations (such as QAM) provide increased data rates, at the cost however of increased signal power due to the decreased Euclidean distance of the constellations [1] [2].

So far, the analog input signal has been converted to a modulated digital bit stream. This bit stream is further converted to an RF signal through multiplication with a high-frequency cosine generated by a local oscillator and then transmitted to the wireless medium. The received signal can be written as:

$$y(t) = ax(t) \cos(2\pi f_c t) + n(t) \quad (2.3)$$

where $x(t)$ is the baseband transmit digital signal after modulation, f_c the carrier frequency, a the attenuation factor due to pathloss, shadowing, and fading, and $n(t)$ the additive white Gaussian noise. To reconstruct the signal, coherent demodulation is employed at the receiver:

$$\begin{aligned} z(t) &= 2y(t) \cos(2\pi f_c t) = 2ax(t) \cos^2(2\pi f_c t) + 2n(t) \cos(2\pi f_c t) \\ &= ax(t) + ax(t) \cos(4\pi f_c t) + 2n(t) \cos(2\pi f_c t). \end{aligned} \quad (2.4)$$

In BPSK modulation for example, the decision variable for the n th transmitted bit is derived by integrating $z(t)$ over a symbol period:

$$r = \frac{1}{T_b} \int_{nT_b}^{(n+1)T_b} z(t) dt \quad (2.5)$$

where T_b is the symbol period and r the decision variable. Assuming that $x(t) = b_n$ in $[nT_b, (n+1)T_b]$, then (2.5) can be equivalently written as:

$$r = \frac{1}{T_b} \int_{nT_b}^{(n+1)T_b} ab_n dt + \frac{1}{T_b} \int_{nT_b}^{(n+1)T_b} ab_n \cos(4\pi f_c t) dt + \frac{1}{T_b} \int_{nT_b}^{(n+1)T_b} 2n(t) \cos(4\pi f_c t) dt. \quad (2.6)$$

If we further assume for simplicity that $n = 0$, then (2.6) becomes:

$$r = ab_0 + ab_0 \frac{\sin(4\pi f_c T_b)}{4\pi f_c T_b} + \frac{2}{T_b} \int_0^{T_b} n(t) \cos(4\pi f_c t) dt. \quad (2.7)$$

A practical assumption is that $f_c T_b \gg 1$, as in wireless digital communications the carrier frequency is much larger compared to the bandwidth of the baseband signal. Therefore, the denominator of the middle term of (2.7) is very large; hence this term can be neglected. Finally, (2.7) can be approximated by:

$$r \approx ab_0 + \frac{2}{T_b} \int_0^{T_b} n(t) \cos(4\pi f_c t) dt. \quad (2.8)$$

If $r > 0$ then the output is 1, otherwise the output of the decision is -1 .

Up to this stage, the digital bit stream has been recovered at the receiver. To reconstruct the initial analog signal, the inverse operation of quantization is employed, where for every L bits an amplitude value is assigned. The output signal to the receiver will be a distorted version of the initial transmit signal, mainly due to three factors: quantization error, channel attenuation factor, and thermal noise level. Modern wireless digital systems are designed so as to minimize this distortion. A typical metric to express the severity of error in digital communications is bit error rate (BER), expressed as the ratio of badly recovered bits to the total number of bits transmitted. Currently operating digital networks are designed with a tolerable BER from 10^{-3} to 10^{-6} , depending on the requested service. Usually, bit error probability (BEP) curves are provided versus the effective signal-to-noise (SNR) ratio, expressed as the ratio of the desired signal power to the thermal noise level at the receiver. Based on these curves it is possible to define the minimum transmission power level for acceptable quality of service for different channel conditions.

2.3 The WCDMA/OFDMA physical layer protocols

In this section the WCDMA and OFDMA physical layer protocols are described, which have been adopted for a majority of wireless digital applications. Unlike FDMA or TDMA where different frequency bands or time slots respectively are assigned to different users, in CDMA all users can transmit simultaneously occupying the whole transmission bandwidth (Figure 2.4). This is achieved with the use of different codes per user that multiply the baseband transmission signal. Consider for example a BPSK modulated

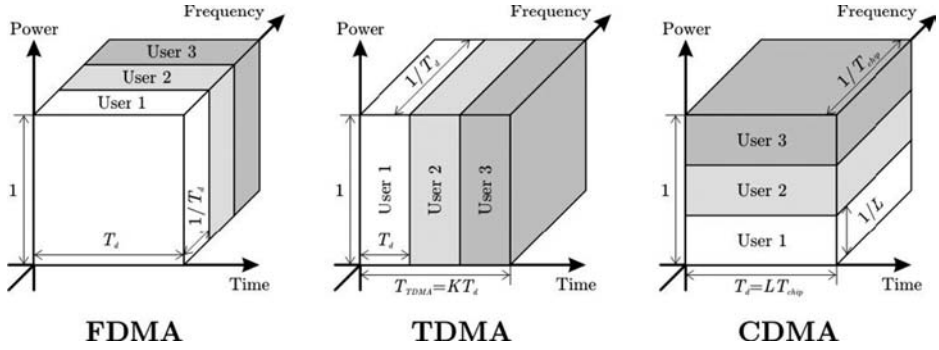


Figure 2.4 FDMA, TDMA, and WCDMA physical layer protocols in time, frequency, and power domain.

signal with values ± 1 of rate R_b and symbol duration equal to $T_b = 1/R_b$. If this signal is multiplied with a high rate coding sequence with symbol duration T_c and rate $R_c = 1/T_c$, then the output signal will also be of rate R_c . Therefore, the initial bandwidth is now spread to a larger bandwidth and the ratio of R_c to R_b that indicates the amount of spreading (usually an integer number), is called spreading factor (SF):

$$SF = \frac{R_c}{R_b}. \quad (2.9)$$

This procedure is illustrated in Figure 2.5(a) for $SF = 8$. In this case, eight bits with duration T_c multiply a bit symbol with duration T_b . These bits are also referred as chips and T_c as chip duration. The power spectral densities of the baseband BPSK modulated signal and the high rate coding sequence are given by (2.10) and (2.11) respectively:

$$S_b(f) = T_b \frac{\sin^2(\pi T_b f)}{(\pi T_b f)^2}, \quad (2.10)$$

$$S_c(f) = T_c \frac{\sin^2(\pi T_c f)}{(\pi T_c f)^2}. \quad (2.11)$$

These spectral densities are plotted on a common diagram in Figure 2.5(b). As it can be observed, although they have a similar form the bandwidth of the coding sequence is significantly larger compared to the bandwidth of the baseband signal. The spreading factor can be alternatively expressed as the ratio of the peak of the baseband signal power spectral density to the peak of the coding sequence power spectral density:

$$SF = \frac{S_b(0)}{S_c(0)}. \quad (2.12)$$

The transmitted signal can be expressed as:

$$s_k(t) = \sqrt{P_k} b_k(t) c_k(t) \quad (2.13)$$

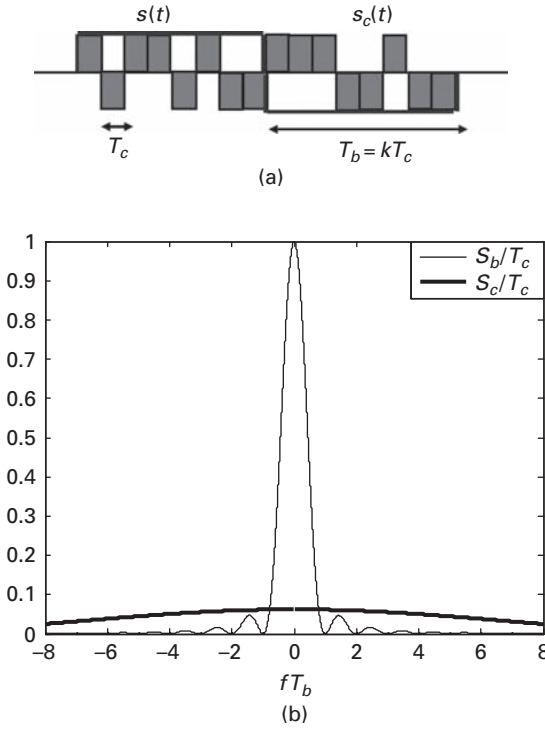


Figure 2.5 (a) WCDMA modulation and (b) power spectral densities for the baseband and the spread-spectrum signal.

where $c_k(t)$ is the transmission coding sequence for the k th user ($1 \leq k \leq K$). Considering for simplicity only the uplink (i.e. the link from the mobile terminals to the base-station), then the received signal will be a superposition of K transmitted signals:

$$r(t) = \sum_{k=1}^K \sqrt{P_k} b_k(t - \tau_k) c_k(t - \tau_k) + n(t) \quad (2.14)$$

where τ_k is the time delay for the k th user's transmission signal. Assuming perfect time synchronization at the receiver for the desired signal, then the received signal is correlated with the desired coding sequence. An ideal set of spreading sequences would have zero cross-correlation and auto correlation function equal to one for zero time offset and zero for non-zero time offset [3]; hence capacity would be limited only by the available number of coding sequences. However, in the design of realistic coding sequences it is desirable to minimize the amount of cross-correlation.

The received signal after multiplication with the desired coding sequence can be expressed as:

$$c_j(t)r(t) = \sqrt{P_j} b_j(t) c_j^2(t) + \sum_{k=1, k \neq j}^K \sqrt{P_k} b_k(t) c_j(t) c_k(t) + c_j(t) n(t). \quad (2.15)$$

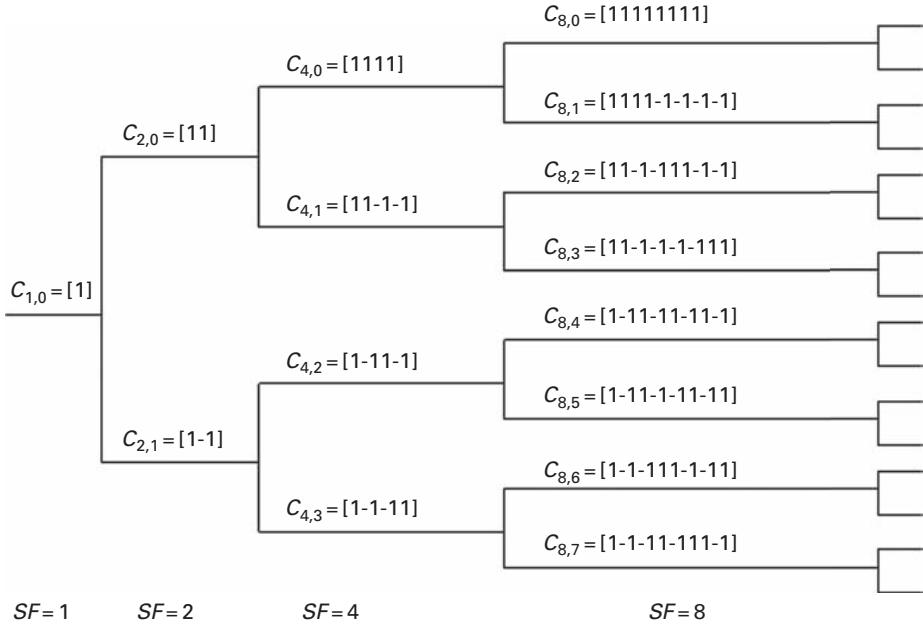


Figure 2.6 OVFS coding tree.

The second term of (2.15) denotes multiple access interference (MAI). Since $c_j^2(t) = 1$ and the cross-correlation of $c_k(t)$ and $c_j(t)$ is low, the sequence $b_k(t)$ can be acquired as described in Section 2.2. However, for increasing number of users, the MAI term becomes non-negligible, thus limiting capacity.

In the currently deployed CDMA networks, the spreading codes are based on the orthogonal variable spreading factor (OVFS) technique, as illustrated in Figure 2.6. For example, the code $c_{4,1}$ has $SF = 4$, while the code $c_{8,6}$ has $SF = 8$. Therefore, for a fixed total bandwidth W equal to 3.84 MHz as standardized in 3G wideband CDMA (WCDMA) networks, in the first case data rates up to 960 kbps can be supported while in the second case data rates up to 480 kbps. However, to maintain orthogonality, if a specific code is selected then no other code stemming from the branch of the previous code can be selected. For example, if $c_{4,3}$ is selected, $c_{8,6}$ and $c_{8,7}$ will have to be excluded as they are non-orthogonal with $c_{4,3}$. The main advantages of OVFS codes are their variable spreading factor which provides bandwidth on demand, as well as the fact that they maintain orthogonality in non-multipath environments. However, in multipath propagation the high values of their autocorrelation function (Figure 2.7(a)) can have severe effects on capacity.

In the standardized 3G networks, short or long Gold codes are used, that are scrambled with the OVFS codes [4]. Usually one Gold code is used per sector, to distinguish users among different sectors. The concatenated orthogonal/scrambling codes have improved autocorrelation characteristics, as shown in Figure 2.7(b).

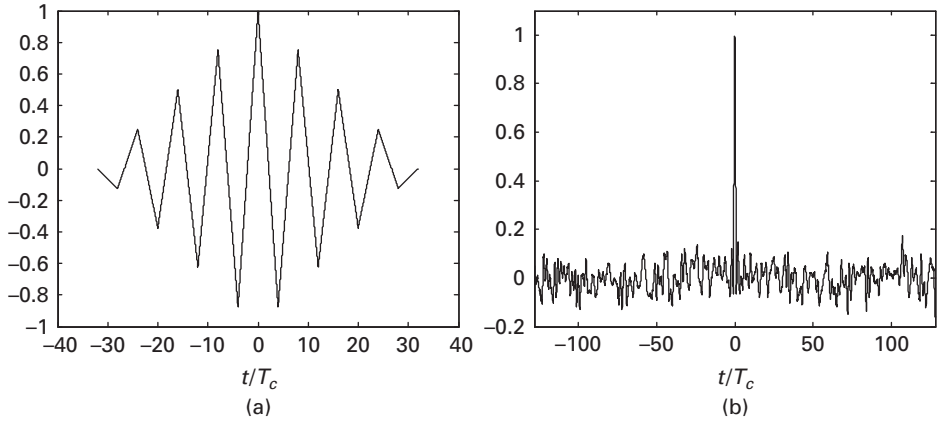


Figure 2.7 (a) Autocorrelation of an OVSF code and (b) autocorrelation of a concatenated orthogonal/spreading code.

The OFDM physical layer protocol is based on multi-carrier transmission, where the incoming data stream is divided into a number of parallel data streams that are transmitted over different subcarriers. In the multi-user OFDM access, also known as OFDMA, each user is assigned with a different set of subcarriers. Each subcarrier has bandwidth B/N_{sc} , where B is the total bandwidth and N_{sc} the number of subcarriers, while the symbol duration T_s is increased by a factor of N_{sc} . Therefore, up to N_{sc} higher data rates can be supported for a given delay spread [5].

In Figure 2.8, a block diagram is shown for multi-carrier transmission where the function $g(t)$ denotes a pulse-shaping filter. The filtered signals are modulated on the different subcarriers and summed up before transmission. The complex baseband signal is given by:

$$s(t) = \sum_k e^{j2\pi f_k t} \sum_l s_{kl} g(t - lT_s) \quad (2.16)$$

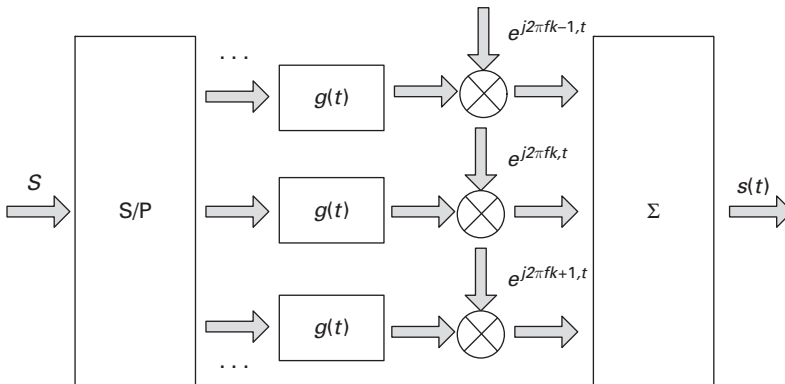


Figure 2.8 Block diagram for multi-carrier transmission.

where s_{kl} are the complex modulation symbols drawn from a signal constellation (e.g. PSK, QAM), k is the frequency index and l the time index. Denoting $g_k(t) = e^{j2\pi f_k t} g(t)$, then the transmitted signal can be alternately given by:

$$s(t) = \sum_l \sum_k s_{kl} g_k(t - lT_s). \quad (2.17)$$

If we further denote:

$$g_{kl}(t) = g_k(t - lT_s) = e^{j2\pi f_k(t-lT_s)} g(t - lT_s) \quad (2.18)$$

then we get the expression:

$$s(t) = \sum_{kl} s_{kl} g_{kl}(t). \quad (2.19)$$

In OFDM transmission, the goal is to select a set of functions g_{kl} , that are orthogonal both in time and frequency domain. Hence, the following condition must be satisfied:

$$\langle g_{kl}, g_{k'l'} \rangle = \delta_{kk'} \delta_{ll'} \quad (2.20)$$

where $\langle x, y \rangle$ denotes the inner product of signals x and y . Two pulses are orthogonal if they do not overlap either in time or in frequency domain. An approach would be to choose time-limited pulses that are orthogonal in frequency domain. In this approach, $g(t)$ is a pulse that is strictly limited to the time interval $[-T_s/2, T_s/2]$ of duration T_s in such a way that $|g(t)|^2$ has a raised-cosine shape with roll-off factor α . The frequency shifted pulse can be described as $g_k(t) = e^{j2\pi f_k t} g(t)$ and in frequency domain $G_k(f) = G(f - f_k)$. It follows that:

$$\langle G_k, G_{k'} \rangle = \langle g_k, g_{k'} \rangle = \delta_{kk'} \quad (2.21)$$

Furthermore, if we define $g_{kl}(t) = g_k(t - lT_s)$, then using the fact that these pulses are strictly separated in time for different l , it can easily be verified that the condition (2.20) is satisfied. The spectrum of these pulses is shown in Figure 2.9(a), for $\alpha = 0$. Although there is a spectral overlap of the subcarriers, they can be separated due to their orthogonality. In this case $f_k = k/T_s$, and the frequency shifted pulses are given by:

$$g_k(t) = \sqrt{\frac{1}{T_s}} e^{j2\pi \frac{k}{T_s} t} \Pi\left(\frac{t}{T_s}\right) \quad (2.22)$$

where the pulse $\Pi(t/T_s)$ equals one for $|t| < T_s/2$ and zero for $|t| > T_s/2$. For a specific carrier frequency and ideal transmission, the symbol s_{k0} can be recovered by:

$$s_{k0} = \sqrt{\frac{1}{T_s}} \int_0^{T_s} e^{-j2\pi \frac{k}{T_s} t} r(t) dt \quad (2.23)$$

where $r(t) = s(t)$, disregarding noise and channel fading. In Figure 2.9(b) the power spectral density of an OFDM signal is shown for $N_{sc} = 101$ subcarriers, which looks familiar to a rectangular spectrum.

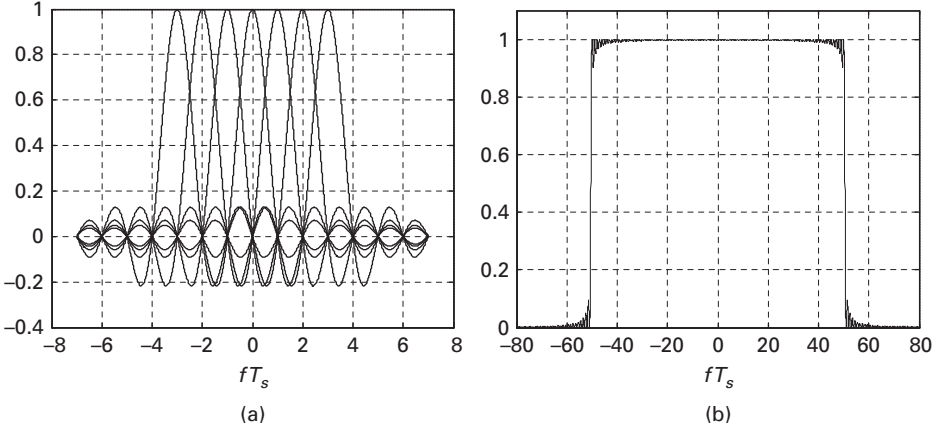


Figure 2.9 (a) Spectrum of OFDM subcarriers and (b) OFDM signal power spectral density for $N_{sc} = 101$ subcarriers.

If we consider one time interval, then the transmitted signal is given by:

$$s_k(t) = \sqrt{\frac{1}{T_s}} \sum_{n=-N_{sc}/2}^{n=N_{sc}/2} s_k e^{j2\pi \frac{k}{T_s} t} \Pi\left(\frac{t}{T_s}\right) \quad (2.24)$$

which means that for each time interval T_s the OFDM transmitted signal is a Fourier synthesis for that period. In practical digital implementations, the Fourier synthesis is implemented with the fast Fourier transform (FFT). In this approach, the digitally modulated symbols s_{kl} are divided into blocks of N_{sc} symbols and then discretely Fourier transformed by the use of the inverse FFT (IFFT). Usually the length N_{FFT} of the FFT is chosen to be larger than N_{sc} , in order to avoid edge effects at half the sampling frequency.

The analysis of OFDM signals so far, assumes perfect synchronization between the transmitter and the receiver, which is not always the case in realistic mobile networks. In a frequency selective multipath environment, the transmit pulses of the OFDM signal and its delayed versions are no longer orthogonal which can lead to severe intersymbol interference (ISI). This can be avoided, by introducing a guard interval, also known as cyclic prefix. With respect to Figure 2.10(a), for a specific OFDM symbol a part with duration G is copied and placed at the beginning of the symbol. By doing so, the symbol duration is now increased to $T = T_s + G$. In this case, the frequency shifted pulses are given by:

$$g'_k(t) = \sqrt{\frac{1}{T}} e^{j2\pi \frac{k}{T_s} t} \Pi\left(\frac{t + \Delta}{T}\right). \quad (2.25)$$

Note that the complex exponential remains exactly the same, while the energy of the OFDM pulse is still normalized to one. The transmit pulses are now given by $g'_{kl}(t) = g'_k(t - lT)$, which are not pair wise orthogonal. However, at the receiver side

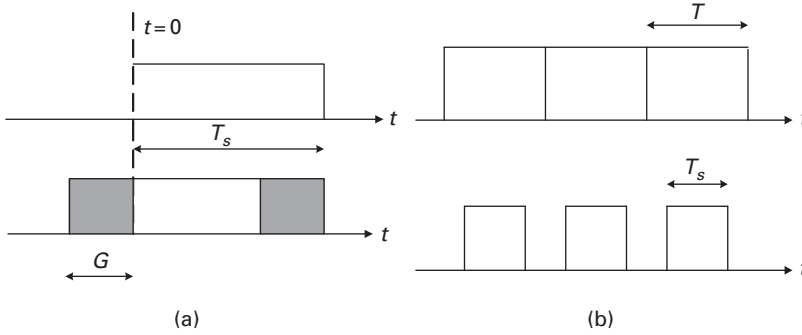


Figure 2.10 (a) Guard interval for OFDM transmitted symbols and (b) Fourier analysis window.

the set of orthogonal detector pulses is given by $g_{kl}(t) = g_k(t - lT)$ which means that Fourier analysis is still performed on a window of length T_s . As shown in Figure 2.10(b), there is now a gap of length G among two adjacent analysis windows that allows a synchronization mismatch of duration up to G . The output of the detector will be given by:

$$\langle g_{kl}, g'_{k'l'} \rangle = \sqrt{\frac{T_s}{T}} \delta_{kk'} \delta_{ll'}. \quad (2.26)$$

Note from (2.26) that if $k = k'$ and $l = l'$, then the value of the inner product is $\sqrt{T_s/T}$ which is less than one. This represents a loss of energy by transmitting a part of the symbol that is not used for detection. If now the transmitted pulse is delayed by $\tau < \Delta$, then (2.26) becomes:

$$\langle g_{kl}, g'_{k'l',\tau} \rangle = \sqrt{\frac{T_s}{T}} e^{-j2\pi f_k \tau} \delta_{kk'} \delta_{ll'} \quad (2.27)$$

which means that the orthogonality between transmit and receive pulses is still preserved as long as the delay offset remains smaller than Δ . The output of (2.27) depends only on a frequency-depended phase factor.

2.4 General principles of MIMO systems

A multi-antenna configuration (MIMO) is comprised of M_t antennas at the transmitter and M_r antennas at the receiver. The use of multiple antennas at both transmission ends improves either the diversity order of the system or increases the overall network throughput through spatial multiplexing ([6], [7]). A typical MIMO configuration is shown in Figure 2.11. The channel from the transmitter to the receiver can be represented as an $M_r \times M_t$ matrix:

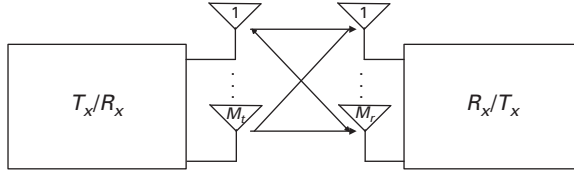


Figure 2.11 Typical MIMO configuration with M_t transmit antennas and M_r receive antennas.

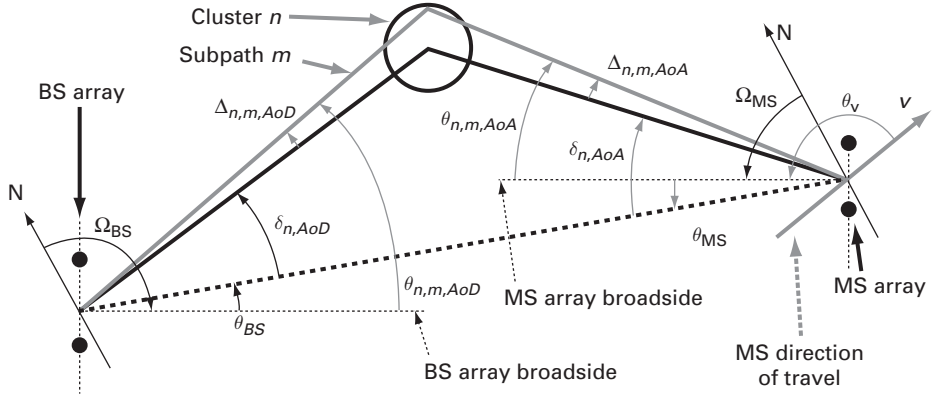


Figure 2.12 3GPP channel model representation (from [8]).

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1M_t} \\ h_{21} & h_{22} & \dots & h_{2M_t} \\ \dots & \dots & \dots & \dots \\ h_{M_r1} & h_{M_r2} & \dots & h_{M_rM_t} \end{bmatrix} \quad (2.28)$$

where the h_{ij} entry represents the channel from the j th transmit antenna to the i th receive antenna. Each element h_{ij} can be modeled as a zero mean Gaussian random variable, where the real and imaginary parts are normally distributed with zero mean and variance equal to $1/2$. However, this representation is not applicable in realistic wireless networks, as it presumes independent signals among different transmit and receive antennas. In a multipath environment, the 3GPP model specifies the channel for a pair of transmit-receive antennas with the use of subpaths. The channel representation is given by [8] and shown in Figure 2.12:

$$h = \sqrt{\frac{P_l \sigma_{SF}}{M}} \sum_{m=1}^M \left(\frac{\sqrt{G_{BS}(\theta_{l,m,AoD})} \exp(j[k_w d_q \sin(\theta_{l,m,AoD}) + \Phi_{l,m}]) \times \sqrt{G_{MS}(\theta_{l,m,AoA})} \exp(j(k_w d_u \sin(\theta_{l,m,AoA}))) \right) \quad (2.29)$$

where j is the imaginary unit, P_l is the power of the l th path, σ_{SF} is the lognormal shadow fading, M is the number of sub-paths per path, $\theta_{l,m,AoD}$ and $\theta_{l,m,AoA}$ are the angles of departure (AoD) and arrival (AoA) respectively for the m th subpath of the l th path, $G_{BS}(\theta_{l,m,AoD})$ is the BS antenna gain for each array element, and $G_{MS}(\theta_{l,m,AoA})$ is the MS antenna gain for each array element for the AoD and AoA respectively. Moreover, $\Phi_{l,m}$ is

the phase of the m th subpath of the l th path, uniformly distributed in $[0, 2\pi]$. Finally, k_w is the wave number $2\pi/\lambda$ where λ is the carrier wavelength in meters, d_q is the distance in meters of BS antenna element q from the reference ($q=1$) antenna, and d_u and is the distance in meters of MS antenna element u from the reference ($u=1$) antenna.

The $M_r \times 1$ signal at the receiver can be expressed as:

$$\mathbf{y} = \sqrt{P}\mathbf{H}\mathbf{w} + \mathbf{n} \quad (2.30)$$

where $\mathbf{w}$ can be either a vector matrix in diversity combining mode or a square $d \times d$ matrix in spatial multiplexing mode, where $d = \min(M_r, M_t)$ is the dimension of the channel. Moreover, P is the transmission power and $\mathbf{n}$ represents the $M_r \times 1$ thermal noise vector, where each element is modeled as a Gaussian process. If the transmitter is unaware of the channel conditions, then two common approaches to transmit weight selection are the uniform and random power allocation. Denoting $\mathbf{w}_{sp}$ the transmit matrix in spatial multiplexing mode and $\mathbf{w}_{div}$ the transmit weight vector in diversity combining mode, then in uniform power allocation $\mathbf{w}_{sp}$ and $\mathbf{w}_{div}$ will be given by:

$$\mathbf{w}_{sp} = \left(\frac{1}{\sqrt{d}} \right) \mathbf{I}_d \mathbf{w}_{div} = \left(\frac{1}{\sqrt{M_t}} \right) \mathbf{1}(M_t) \quad (2.31)$$

where $\mathbf{I}_d$ is the $d \times d$ identity matrix and $\mathbf{1}(M_t)$ is an $M_t \times 1$ vector matrix of ones. Note that in both cases $\|\mathbf{w}_{sp}\|_F = \|\mathbf{w}_{div}\|_F = 1$, where $\|\mathbf{x}\|_F$ is the Frobenius norm of $\mathbf{x}$.

If the transmitter is aware of the channel conditions through channel feedback, then the optimum transmit solution in diversity-combining mode in a single-user environment is to set the transmit weight vector equal to the eigenvector corresponding to the maximum eigenvalue of the channel matrix, hence:

$$\mathbf{w}_{div} = X_{max}(\mathbf{H}^H \mathbf{H}) \quad (2.32)$$

where $X_{max}(\mathbf{A})$ is the eigenvector corresponding to the maximum eigenvalue of matrix $\mathbf{A}$. In this case the average power of the received signal can be expressed as:

$$E\{|y|^2\} = PE\{\mathbf{w}^H \mathbf{H}^H \mathbf{H} \mathbf{w}\} + E\{n^* n\} = P\lambda_{max} + \sigma^2 \quad (2.33)$$

where σ^2 is the noise variance and λ_{max} the maximum eigenvalue of matrix $\mathbf{H}^H \mathbf{H}$. Hence, the SNR is given by:

$$SNR = \frac{P\lambda_{max}}{\sigma^2}. \quad (2.34)$$

The expression in (2.34) is the maximum SNR value that can be obtained with full channel knowledge at the transmitter. In spatial multiplexing mode, $\mathbf{w}_{sp}$ is set equal to the right singular matrix of the singular value decomposition of the channel matrix. In this case, the channel matrix can be expressed as:

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H \quad (2.35)$$

where $\mathbf{U}$ and $\mathbf{V}$ are the left and right unitary matrices ($\mathbf{U}^H \mathbf{U} = \mathbf{I}$, $\mathbf{V}^H \mathbf{V} = \mathbf{I}$) of dimensions $M_r \times M_r$ and $M_t \times M_t$ respectively and $\mathbf{\Sigma}$ is a diagonal $M_r \times M_t$ matrix with diagonal elements being the eigenvalues of matrix $\mathbf{H}$ [9]. The received signal is given by:

$$\mathbf{y} = \sqrt{P} \mathbf{U} \mathbf{\Sigma} \mathbf{V}^H \mathbf{V} + \mathbf{n} = \sqrt{P} \mathbf{\Sigma} \mathbf{U} + \mathbf{n}. \quad (2.36)$$

If $\mathbf{y}$ is multiplied at the receiver with the matrix $\mathbf{U}^H$, then:

$$\mathbf{U}^H \mathbf{y} = \sqrt{P} \mathbf{U}^H \mathbf{U} \mathbf{\Sigma} \mathbf{V}^H \mathbf{V} + \mathbf{U}^H \mathbf{n} = \sqrt{P} \mathbf{\Sigma} + \mathbf{U}^H \mathbf{n}. \quad (2.37)$$

Since $\mathbf{\Sigma}$ is a diagonal matrix, it follows from (2.37) that with full channel knowledge at the transmitter the MIMO channel can be decoupled to $d = \min(M_t, M_r)$ parallel channels with zero interchannel interference.

However, even if the transmitter is unaware of the spatial modes of the channel matrix, transmit diversity can still be provided with the use of space time block codes (STBC). The scientific interest in the design of STBCs was mainly boosted after the pioneering work of Alamouti in 1998 [10], where a simple orthogonal code was proposed for two antennas at the transmitter. We assume that the symbol s_1 drawn from a complex constellation is transmitted from the first antenna and the symbol s_2 from the second transmit antenna during a symbol period. In the next symbol period it is assumed that the symbol $-s_2^*$ is transmitted from the first antenna and the symbol s_1^* from the second antenna, where x^* denotes the conjugate of x . Moreover, let h_1 denote the channel from the first transmit antenna to the receiver and h_2 the corresponding channel from the second transmit antenna. Assuming that h_1 and h_2 remain constant during two consecutive symbol periods, the received signal over two symbol periods can be expressed as:

$$r_1 = h_1 s_1 + h_2 s_2 + n_1 \quad (2.38)$$

$$r_2 = h_1 (-s_2^*) + h_2 s_1^* + n_2 \quad (2.39)$$

where the variables n_1 and n_2 represent the noise level in each of the two links. Denoting the receive signal vector as $\mathbf{r} = [r_1 \ r_2^*]^T$, the transmit signal vector as $\mathbf{s} = [s_1 \ s_2]^T$ and the thermal noise vector as $\mathbf{n} = [n_1 \ n_2^*]^T$, then (2.38) and (2.39) can be equivalently written as:

$$\mathbf{r} = \mathbf{H} \mathbf{s} + \mathbf{n} \quad (2.40)$$

where the channel matrix $\mathbf{H}$ is given by:

$$\mathbf{H} = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix}. \quad (2.41)$$

The matrix $\mathbf{H}$ is orthogonal, as $\mathbf{H}^H \mathbf{H} = (|h_1|^2 + |h_2|^2) \mathbf{I}_2$, where $\mathbf{I}_2$ is the 2×2 identity matrix. Consider the modified signal vector given by:

$$\bar{\mathbf{r}} = \mathbf{H}^H \mathbf{r} = (|h_1|^2 + |h_2|^2) \mathbf{s} + \mathbf{H}^H \mathbf{n}. \quad (2.42)$$

In this case, the decoding decision can be represented as:

$$\bar{\mathbf{s}} = \operatorname{argmin}_{\bar{\mathbf{s}} \in S} \left\| \bar{\mathbf{r}} - \left(|h_1|^2 + |h_2|^2 \right) \bar{\mathbf{s}} \right\|^2 \quad (2.43)$$

where S is the set of all possible combinations of symbols $\{s_1, s_2\}$.

The Alamouti scheme achieves full rate, as two symbols are transmitted over two symbol periods. In general, the rate of the space-time block code is defined as the ratio between the number of symbols that the encoder takes as input (denoted as k) and the number of space-time coded symbols transmitted from each antenna (denoted as p), hence:

$$R = \frac{k}{p}. \quad (2.44)$$

The spectral efficiency (η) of the space-time block code is given by:

$$\eta = \frac{km}{p} \text{ bits/s/Hz} \quad (2.45)$$

where 2^m is the number of points of the complex signal constellation. For general complex constellations such as M-QAM or M-PSK, up to now there are no STBCs with full transmission rate and simple linear processing at the receiver that can achieve the maximum order of diversity for more than two transmit antennas. However, for non-full rates, such codes have been constructed. In (2.46) and (2.47) for example, two such codes are depicted for three and four transmit antennas respectively that achieve rate equal to 1/2.

$$\mathbf{C}_3 = \begin{bmatrix} s_1 & -s_2 & -s_3 & -s_4 & s_1^* & -s_2^* & -s_3^* & -s_4^* \\ s_2 & s_1 & s_4 & -s_3 & s_2^* & s_1^* & s_4^* & -s_3^* \\ s_3 & -s_4 & -s_1 & s_2 & s_3^* & -s_4^* & s_1^* & s_2^* \end{bmatrix} \quad (2.46)$$

$$\mathbf{C}_4 = \begin{bmatrix} s_1 & -s_2 & -s_3 & -s_4 & s_1^* & -s_2^* & -s_3^* & -s_4^* \\ s_2 & s_1 & s_4 & -s_3 & s_2^* & s_1^* & s_4^* & -s_3^* \\ s_3 & -s_4 & -s_1 & s_2 & s_3^* & -s_4^* & s_1^* & s_2^* \\ s_4 & s_3 & -s_2 & s_1 & s_4^* & s_3^* & -s_2^* & s_1^* \end{bmatrix} \quad (2.47)$$

In Figure 2.13(a), results are provided for four transmission schemes: uncoded QPSK, the Alamouti 2×1 space-time code as well as $\mathbf{C}_3$ and $\mathbf{C}_4$ codes for three and four transmit antennas respectively. In the vertical axis is the BER while in the horizontal axis the E_b/N_o ranging from 0 to 20dB. The spectral efficiency in all four cases is 2 bits/s/Hz for 16-QAM modulation in $\mathbf{C}_3$ and $\mathbf{C}_4$. As it can be observed, for BER equal to 10^{-3} , $\mathbf{C}_4$ code provides about 1 dB gain compared to $\mathbf{C}_3$ code and 3 dB gain compared to the Alamouti scheme. In Figure 2.13(b), $\mathbf{C}_4$ code is compared to the Alamouti transmission scheme for one and two receive antennas. Adding receive antennas can significantly improve transmission performance in terms of BER, rather than adding transmit antennas. As it can be seen, the Alamouti scheme with two receive antennas provides almost 3.5 dB gain compared to $\mathbf{C}_4$ code for target BER equal to 10^{-3} , although both configurations have the same number of antennas.

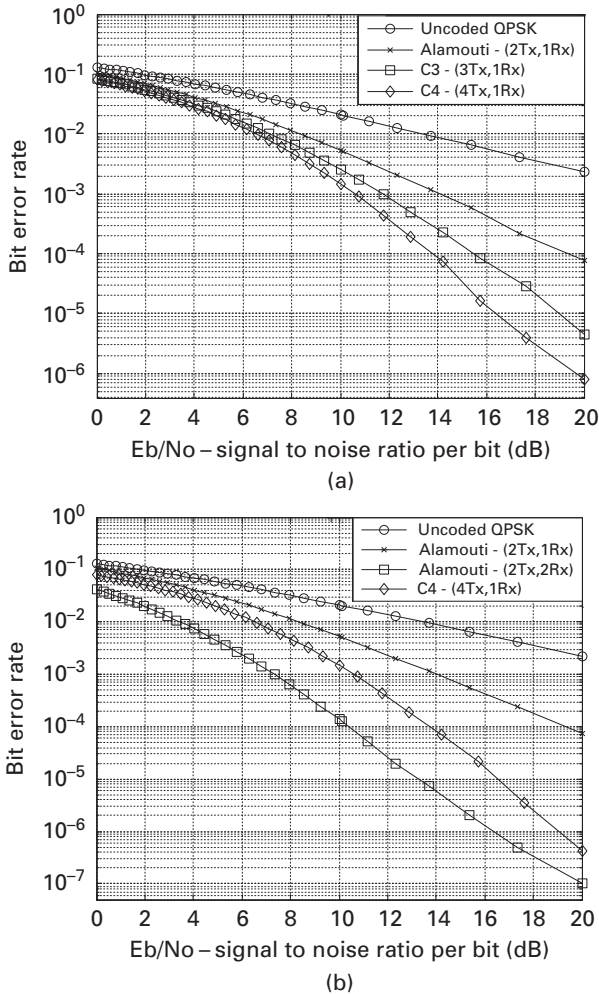


Figure 2.13 BER curves for various STBC orientations.

2.5 Space-time and space-frequency transmission over MIMO networks

In this section the basic physical layer structures are described for integrating MIMO transmission technology in WCDMA or OFDMA networks. In the design of a MIMO transceiver a number of issues must be taken into account, such as transmit and receive complexity, signaling burden, channel measurements at the receiver, and feedback to the transmitter as well as achievable data rates. In [11], several proposals are described for deploying MIMO transmission in the currently operating third generation networks. A very popular transmission technique is per antenna rate control (PARC), as shown in Figure 2.14. A high data rate stream is demultiplexed into a number of T substreams, where T is the number of transmit antennas. Each of these data streams is then encoded

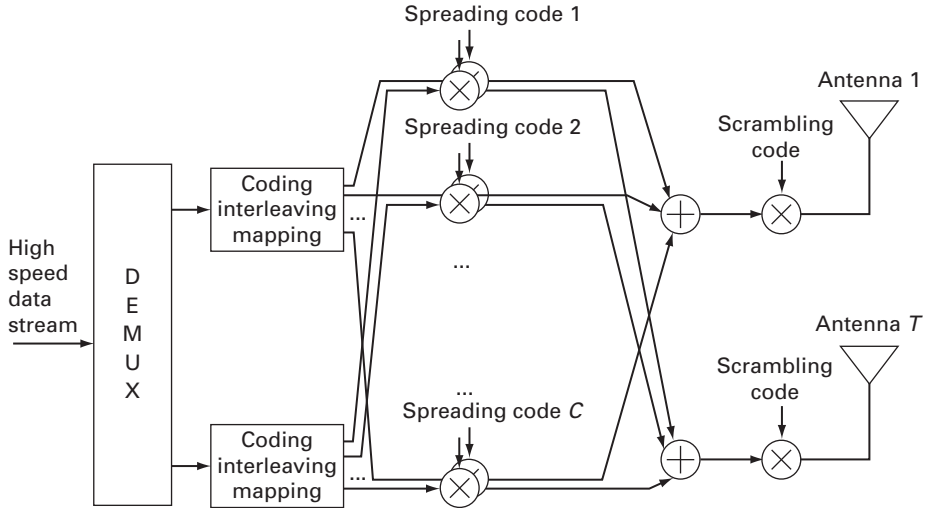


Figure 2.14 Per Antenna Rate Control (PARC) transmission over WCDMA networks (from [11]).

and modulated using either QPSK or 16-QAM modulation and further demultiplexed into a number of C substreams that are spread using distinct OVFS channelization codes. The individual substreams are then summed and scrambled using a distinct scrambling code per transmit antenna. At the receiver side, the effective signal-to-interference-plus-noise ratio (SINR) is measured separately from every transmit antenna across all receive antennas and the subset that provides the highest data rate is chosen. For two transmit antennas for example, the possible power allocation combinations are $\{P, 0\}$, $\{0, P\}$, $\{P/2, P/2\}$ where P is the total transmit power. In the first two cases only one transmit antenna is operating, while in the third case the total transmission power is equally divided among the two transmit antennas. Note that this architecture describes a closed-loop MIMO transmission technique, as the receiver must be in position to send the effective SINR per transmit antenna to the transmitter.

In Figure 2.15, a schematic diagram of the rate control multi paths (RC-MPD) transmission technique is shown. The initial data stream is divided into two data streams of equal rate that are separately spread and scrambled. Another copy of the same symbols is transmitted after one chip delay period and after encoding the symbols by the space time transmit diversity (STTD) code:

$$\begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \Leftrightarrow \begin{pmatrix} -s_2^* \\ s_1^* \end{pmatrix} \quad (2.48)$$

This encoding technique is equal to the Alamouti transmission scheme. The channel state information (CSI) is fed from the receiver to the transmitter which decides on the modulation and coding scheme for each stream. In a multipath environment, the delayed versions of the signal with time offset greater than T_c will be suppressed, due to the low values of the cross-correlation functions of the spreading/scrambling sequences. Hence if

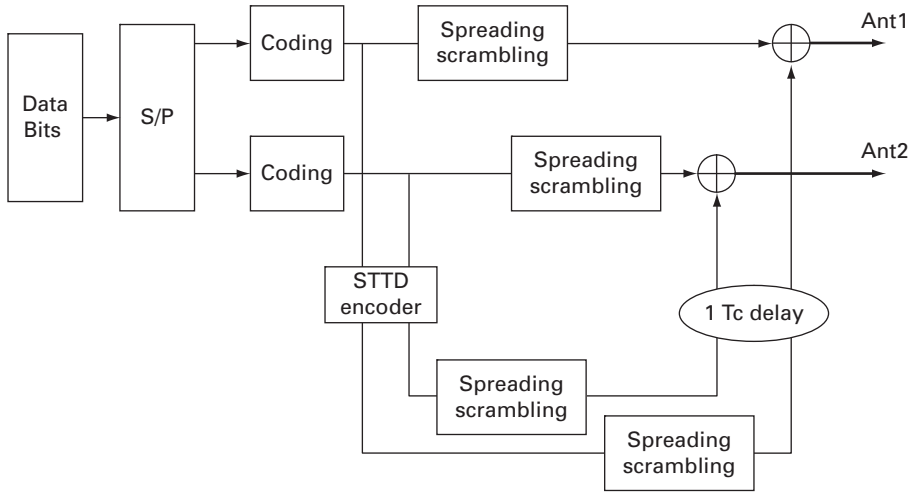


Figure 2.15 Rate Control Multipaths (RC-MPD) transmission architecture (from [11]).

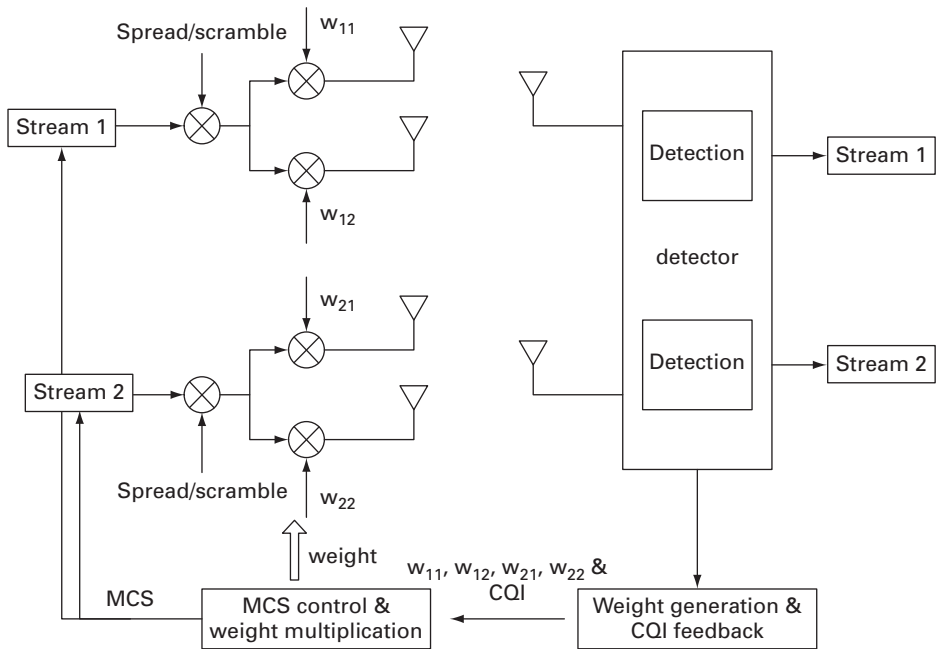


Figure 2.16 Double transmit antenna array (D-TxAA) transmission architecture (from [11]).

a RAKE receiver is employed ([3]) this architecture can provide both space and frequency diversity.

In Figure 2.16, the double transmit antenna array (D-TxAA) transmission is illustrated, which is based on the eigenvalue analysis of the MIMO channel matrix. For four transmit and two receive antennas as shown in Figure 2.16, the transmit antennas are divided into

two groups with two antennas per group. The data stream per user is split into two separate streams which are then individually spread and scrambled. Denoting:

$$\mathbf{H}_1 = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} \mathbf{H}_2 = \begin{bmatrix} h_{13} & h_{14} \\ h_{23} & h_{24} \end{bmatrix} \quad (2.49)$$

then the transmit weight vectors are calculated by the eigenvalue analysis of the covariance matrices $\mathbf{H}_1^H \mathbf{H}_1$ and $\mathbf{H}_2^H \mathbf{H}_2$. Note from Figure 2.16 that the weight calculation takes place at the receiver side and then these weights are sent to the transmitter through a feedback channel.

The integration of MIMO technology in OFDM networks is a very promising technique to improve capacity and provide high data rates. Transmit information can be encoded either in time or frequency domain, thus providing space-time or space-frequency diversity which enables more reliable and robust transmission. In Figure 2.17, a simplified block diagram of a MIMO-OFDM system is shown, for M_t antennas at the transmitter and M_r antennas at the receiver.

The incoming bit stream S is mapped into a number of complex data symbols (e.g. with QAM modulation) which are then transmitted through M_t antennas in T transmission blocks. Each block consists of N_{sc} subcarriers. Depending on the allocation of symbols per transmit antenna, transmission block and subcarrier, different space-time or space-frequency codes can be defined. In space-time coding for example, for a specific transmit antenna and subcarrier, different symbols are transmitted in consecutive transmission blocks. For the special case of two transmit antennas, the Alamouti scheme will be given by:

$$\mathbf{C}_2(k, n, n+1) = \begin{bmatrix} s_1 & -s_2^* \\ s_2 & -s_1^* \end{bmatrix} \quad (2.50)$$

where k is the subcarrier index ($1 \leq k \leq N_{sc}$) and $[n, n+1]$ the transmission period. In this case we have $T=2$. To further improve the diversity order of the system, the code in (2.50) can be repeated in other subcarriers as well, at the cost however of reduced data rate.

In space-frequency coding, (2.50) will be modified as:

$$\mathbf{C}_2(k, n) = \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \mathbf{C}_2(l, n) = \begin{bmatrix} -s_2^* \\ s_1^* \end{bmatrix} \quad (2.51)$$

where k and l are two different subcarrier indexes. Note that in this case, the transmitted symbols are sent during one transmission period, thus reducing the complexity of the receiver in terms of storage complexity.

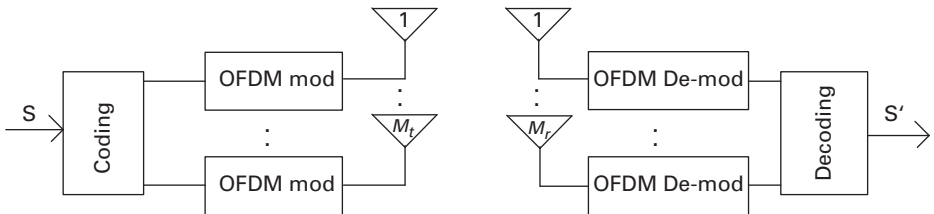


Figure 2.17 MIMO-OFDM simplified block diagram.

In space-time-frequency encoding, the transmit data symbols are spread across different transmit antennas and time periods as well as subcarriers ([12]), thus:

$$\mathbf{C}_2(k, n) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mathbf{C}_2(l, n+1) = \begin{bmatrix} x_3 \\ x_4 \end{bmatrix} \quad (2.51)$$

where x_1, x_2, x_3, x_4 different symbols drawn from a complex constellation.

2.6 Summary

In this chapter, the basic principles of signal processing procedures and physical layer protocol structures were described for broadband wireless digital communications. The MIMO transmission technology was introduced as well, which is a promising technique for increasing per user throughput without additional spectrum requirements. As the support of high data rates and reliable multimedia communications is of primary importance in next generation networks, the integration of MIMO transmission techniques in the currently operating WCDMA or OFDMA wireless networks becomes a challenging task. Space-time and space-frequency transmission architectures were described as well, that provide space-time or space-frequency diversity gain. Future challenges in the design and implementation of next generation broadband wireless networks include (among others) reduced complexity transceiver architectures, integration of MIMO technology, and efficient feedback techniques to acquire channel state information at the transmitter.

References

- [1] A. Goldsmith, *Wireless Communications*, Cambridge University Press, 2005.
- [2] M. K. Simon and M. S. Alouini, *Digital Communication over Fading Channels*, 2nd edition, Wiley, 2001.
- [3] H. Holma and A. Toskala, eds., *WCDMA for UMTS. Radio Access for Third Generation Mobile Communications*, Wiley, 2004.
- [4] 3GPP TS 25.213, *Spreading and Modulation (FDD)*, 2005.
- [5] Y. Li and G.L. Stüber, *Orthogonal Frequency Division Multiplexing for Wireless Communications*, Springer, 2007.
- [6] D. Gesbert, M. Shafi, D. Shiu, P. Smith, and A. Naguib, "From theory to practice: An overview of MIMO space-time coded wireless systems," *IEEE Journal on Selected Areas in Communications* vol. **21**, no. 3, pp. 281–302, 2003.
- [7] A. Paulraj, D. Gore, R. Nabar, and H. Bölcskei, "An overview of MIMO communications-A key to gigabit wireless," *Proceedings of IEEE*, vol. **92**, no. 2, 198–218, 2004.
- [8] 3GPP TR 25.996 v6.1.0¹, *Spatial Channel Model for Multiple Input Multiple Output (MIMO) Simulations*, September 2003.

¹ "© 2003. 3GPP™ TSs and TRs are the property of ARIB, ATIS, CCSA, ETSI, TTA, and TTC who jointly own the copyright in them. They are subject to further modifications and are therefore provided to you 'as is' for information purposes only. Further use is strictly prohibited."

- [9] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space Time Wireless Communications*, Cambridge University Press, 2003.
- [10] S. Alamouti, "A simple transmitter diversity technique for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. **16**, no. 8, 1451–1458, 1998.
- [11] 3GPP TR 25.876 v1.8.0², *Multiple Input Multiple Output in UTRA*, 2005.
- [12] W. Zhang, X. G. Xia, and K. B. Letaief, "Space-time/Frequency coding for MIMO-OFDM in next generation broadband wireless systems," *IEEE Wireless Communications*, vol. **14**, issue 3, 32–43, 2007.

² "© 2005. 3GPP™ TSs and TRs are the property of ARIB, ATIS, CCSA, ETSI, TTA, and TTC who jointly own the copyright in them. They are subject to further modifications and are therefore provided to you 'as is' for information purposes only. Further use is strictly prohibited."

3 Digital front-end and system overview in wireless communication

Eduard Bertran and Maurizio Pagani

3.1 Introduction

With the software defined radio (SDR) concept a new communication technology was created to improve interoperability among different wireless networks, systems, and devices and some of the typical limitations in flexibility of classical radio implementations were removed, aiming at the same time at reducing the cost of technology components. For example, in a classical radio implementation, some hardware components like filters are required to cancel unwanted emissions, but good analog filters are costly and inflexible. Instead, an SDR can be considered a wireless system whose main functionality and operational parameters, like output power, frequency band, air interface including modulation techniques, security and performance characteristics, are implemented by means of software programs running on programmable devices that can dynamically reconfigure the system. In such a way there is the possibility to create multimode, multiband, and multi-functional network equipment that can be reconfigured, enhanced, and upgraded during their operation through software updates and hardware reconfiguration, instead of physical changes in the hardware. As detailed in the sections below, standards for wireless communications include adaptive modulation and coding, as well as scalable bandwidths and transmit power control. These facts require, both in user terminals and in base-stations, a degree of flexibility and reconfigurability, software upgrading, and multimode-multiband compliance. Besides, some access techniques, especially CDMA, are very sensitive to precise frame synchronization. Software defined radio is a suitable tool to overcome issues like these.

This chapter is organized into five sections. In the rest of this section, we will give a brief introduction to digital front-end and its key processing units. [Section 3.2](#) is devoted to presenting new objectives and technologies for transmission and reception of modern wireless communication systems with emphasis on the front-end part. In particular, this section introduces some recent mobile communications standards, suitable for SDR developments because of their novelty and capability of reconfiguration. Also, an overview of the fundamentals of the main access techniques is presented, and the key terminology and most relevant parameters are also provided. There are a considerable number of abbreviations used in modern wireless communication systems, and this is

unavoidable. In [Section 3.3](#), emerging broadband wireless standards are reviewed. This includes the standards in development under the umbrella of the IMT-Advanced Program, as well as some new standards including some that are based on cognitive radio strategies. [Section 3.4](#) focuses on some considerations on how to design and implement DFE, covering transceiver and related power amplifier, chip circuit and its implementation platform as well as some technological comments. Finally, [Section 3.5](#) summarizes and concludes.

When the analog-to-digital converter (ADC) is directly connected to the antenna (or more realistically, closer to the antenna) the resulting structure is known as the “ideal receiver” or in some publications as the software radio (SR or SWR) receiver [32]. In a realistic approach ([Figure 3.1](#)), and regarding the quantization levels of the ADC (minimum sensitivity), a LNA (low-noise amplifier) is included as part of the SR front-end. Also an analog anti-aliasing filter should be included [20]. In a similar way, the “all-digital transmitter” has been proposed as a kind of system which might even avoid the use of the digital-to-analog converter (DAC) by using direct digital conversion techniques, where the baseband signal is generated by means of a digital modulator ([Figure 3.2](#)) and upconverted with a digital upmixer to the required radio frequency (i.e., by using software interpolation techniques, or digital logic in the case of ASICs – application-specific integrated circuits). The output of the digital up-converter is used to drive some kind of highly efficient switched-mode power amplifier (i.e., class E or class F). Digital predistorters (DPD) are included to deal with the linearity-efficiency trade-off in the power amplifier (PA). Alternatively, the inclusion of some switched PA (i.e., a class D) inside the delta-sigma modulator loop is another approach to the all-digital transmitter implementation.

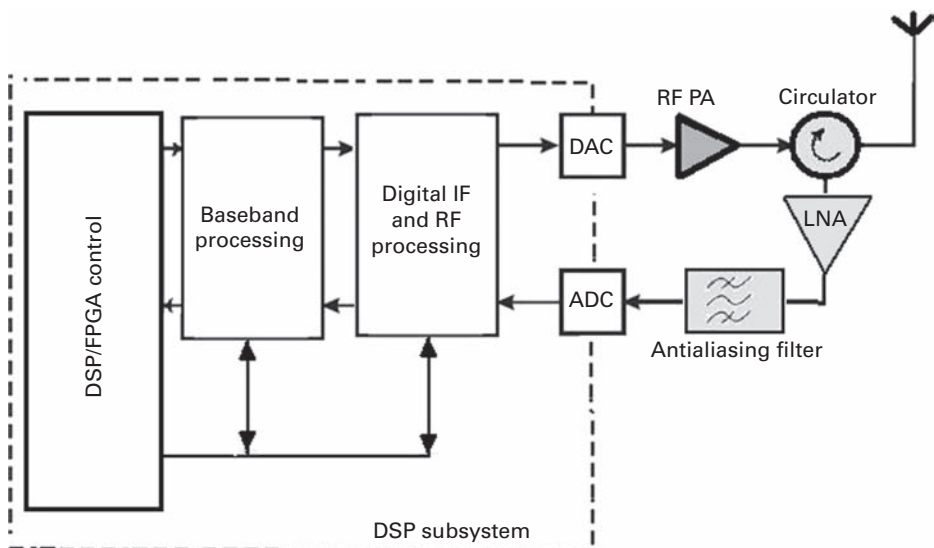


Figure 3.1 Ideal SDR architecture (full-duplex system).

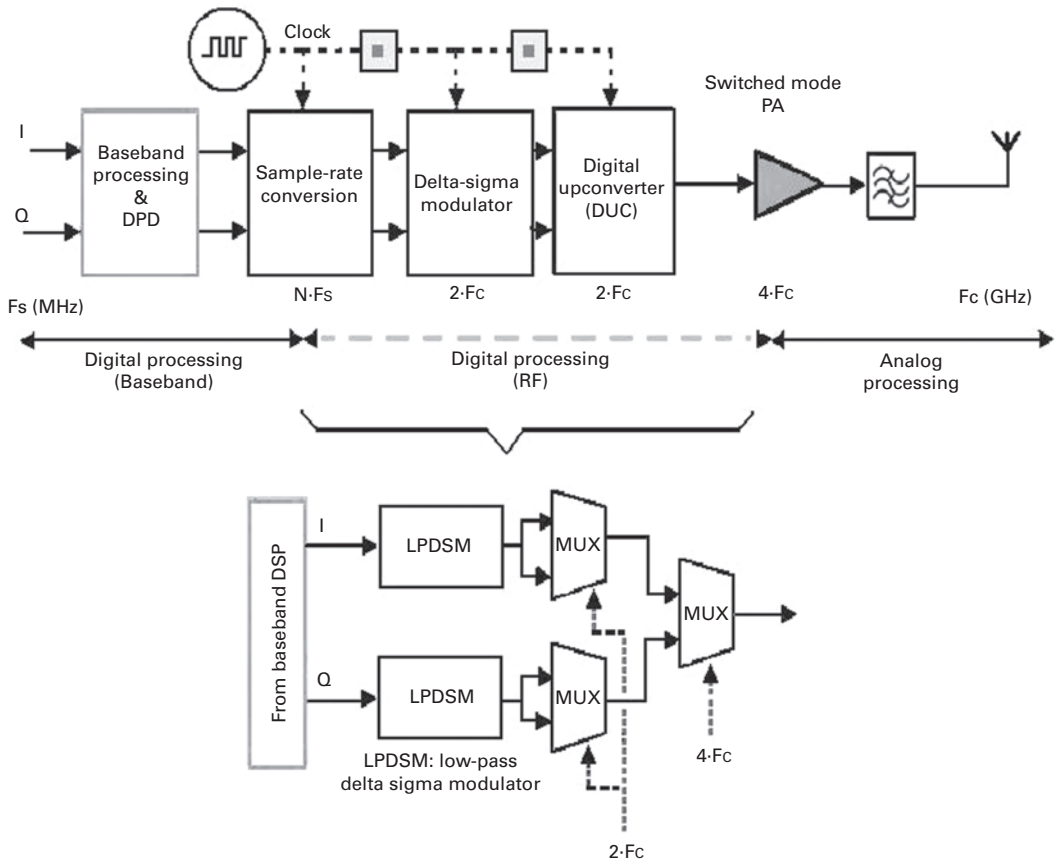


Figure 3.2 All-digital transmitter approach.

When not completely digital, the ideal scheme of an SDR transmitter contains a digital signal processor generating a stream of bits that is sent to a digital-to-analog converter connected to a radio antenna. However, this scheme is not completely realizable due to actual technology limits. The main problem is the difficulty of conversion between the digital and the analog domains, at a high enough rate and accuracy, at the same time as reducing costs. Indeed, there has been a strong advance in technology over the years and some SDR designs have been presented which work well for signal processing at low and medium bandwidths.

Digital front-end (DFE) that is allocated at or near the antenna is one of the most critical parts of a software radio transceiver [45]. In the receiver the DFE changes from analog signal processing (ASP) to digital signal processing (DSP), whilst in the transmitter it moves from DSP to ASP. Analog signal processing technologies are associated with the operation at radio-frequencies and microwaves, and, in the transmitter, they are also related with power processing and amplification. These are the current constraints for DSP application (or for application at a reasonable cost): high frequency and transmit

power. Nevertheless, DSP has a number of benefits in many aspects such as the increasing flexibility, reconfigurability, compatibility among standards, functionality upgrading, maintenance and adaptation to market requirements (download for local services, economy of scale, global roaming). These are actually the merits of the software radio technologies.

Unfortunately, even when merging both ASP and DSP in the same system, a true SDR realization is still facing numerous technical challenges from both hardware and software sides and many of them are still addressed by undergoing research. In addition, in order to be competitive and future-proof, commercial SDR systems have to be efficient in terms of power and spectrum, to be able to handle higher data rates, consume less energy, be more environmentally friendly and, last but not least, produced with low manufacturing costs. The industrial development of software defined radio systems with the possibility to be used across multi-standard applications, today represents one of the critical elements in enabling the convergence of different communication systems.

Current radios do not use the aforementioned ideal or quasi-ideal architectures, which may be considered, nowadays, as asymptotic objectives. In fact, actual structures combine analog and digital building blocks, in a combination of ASICs (usually used in the RF subsystems) and analog power devices, with pure digital components (i.e., DSP, digital signal processor, or FPGA, field-programmable gate array). These hybrid SR solutions where analog hardware coexists with software defined functions are referred to as *software defined radio* as mentioned at the beginning of the chapter. Another generic concept that appears in the literature [45] is the *digital radio* (DR), which simply means, in a wide sense, that the information is digitized at some point of the transceiver, and it does not necessarily mean that the radio is SDR based (i.e., if the digital processing is embedded in ASIC devices). Sometimes “defined” is identified with “fixed”, hence understanding SDR as a kind of radio where definite functionalities are moved from classical hardware to the software, proposing other names for describing the reconfigurable, flexible architectures, such as Adaptable Radio or Adaptive Intelligent Software Radio. Anyway, no standard definition for SDR exists. So, in the remaining of this chapter, we will use the term SDR in a wide sense, admitting all its variations.

The DFE design complexity depends on many factors and parameters; some of them are usually defined in the communication standards while others are simply dependent on the desired transceiver quality (i.e., manufacturer prestige). Besides, some DFE are required to be adaptable for each standard of operation, sometimes with slight adjustments, and occasionally needing a larger change (i.e., for multi-standards support). Despite most of the literature emphasizing the possibilities of the DFE design alternatives regarding the receiver section (kind of ADC, SFDR (spurious free dynamic range), receiver architecture (zero IF, low IF, heterodyne), filtering, adjustable tuning, noise management, jitter effects, undersampling techniques, multimode reception, synchronization, etc.), the transmit section also has different implementation alternatives [20][23], such as the architecture (direct conversion, multiple conversion, analog or digital IF up-converters), the way to face the PA linearity-efficiency trade-off, the use of diplexers/duplexers or circulators, ISI (inter-symbol interference) equalization and adjacent channel interference filters, power control blocks, multiband-multimode operation. Moreover,

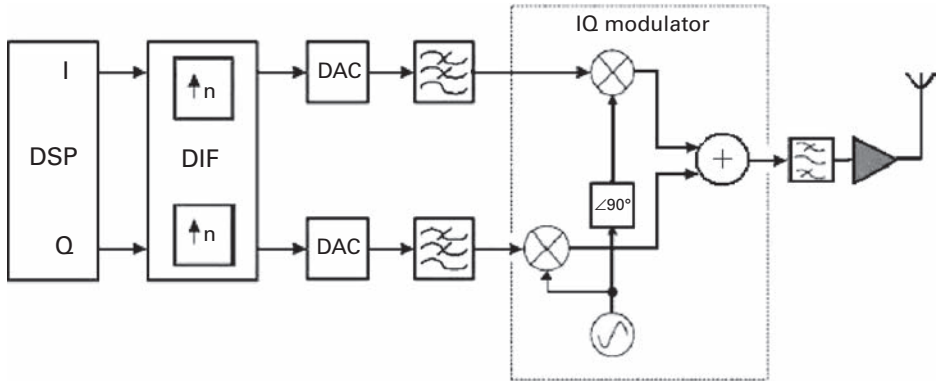


Figure 3.3 Direct up-conversion transmitter using analog up-converter and digital interpolation filters.

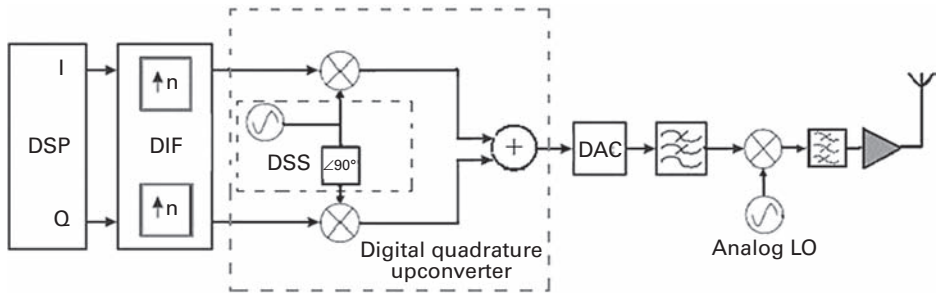


Figure 3.4 Digital IF output.

it is important to distinguish between DFE for fixed (base-stations) or mobile equipment (handsets); simply envisaging the PA requirements may clarify this last statement.

The direct up-conversion transmitter (Figure 3.3) reduces the number of mixers and oscillators. It is a desirable transmitter for its simplicity, flexibility, reduced power consumption and relative low cost of implementation. The use of an analog I&Q modulator (only at high frequencies; if lower it may become a digital IF structure) would produce a single sideband up-conversion assuming ideal operation. However, when faced with mal-adjustments in the I&Q modulator (gain or phase) the signal constellation is distorted and unsuppressed sidebands appear. Besides, the local oscillator (LO) in multiband transmitters must have variable frequency. This fact, along with the need to keep the impedances matched for the sake of a maximum power transfer, prevents the suppression of the LO feedthrough to the PA, which would produce interference problems in the antenna. Anyway, the direct conversion transmitter is of the preferred architectures in user terminal designs, as well as in most base-station ones, due to the cost and power consumption advantages.

The abovementioned leakage from the LO to the antenna may also appear in the IF transmitter (Figure 3.4), however the I&Q modulator here is digital, and also driven by an digital synthesizer. The multiple conversion transmitter of Figure 3.5 doesn't have the advantages of digital I&Q demodulation, though it improves (at higher cost and power

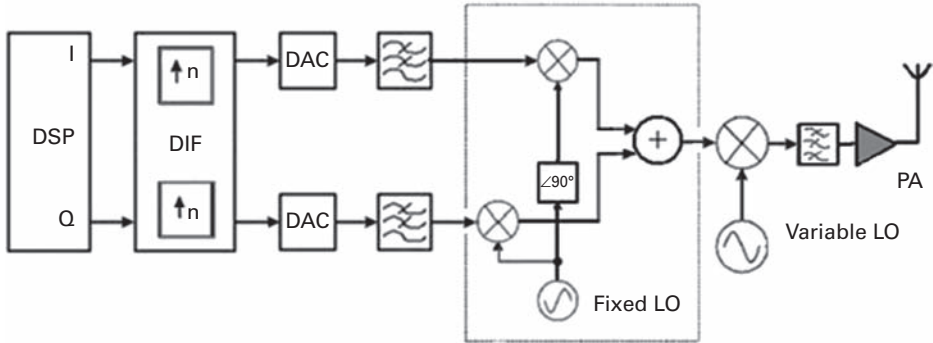


Figure 3.5 Multiple conversion transmitter.

consumption) the direct up-conversion transmitter by keeping constant the frequency of the LO used in the I&Q modulator.

The digital interpolation filters, used in the previous three transmitter schemes, are an interface that allows a lower sample rate operation at baseband (i.e., in the DSP or FPGA devices). Moreover, they increase the separation of the alias spectra from the baseband spectrum of the discrete time signal, so reducing complexity and cost of the posterior analog reconstruction filter, after the DAC converter. As it is known from the fundamentals of discrete-time signal processing, the use of sampled signals produces alias of the baseband spectrum that are repeated at frequency multiples of f_s , where f_s is the sampling frequency (update rate) of the signal at the input of the interpolation filters. After interpolation, the sampling frequency becomes f_{DAC} (update rate in the DAC), thus the alias is moved to this new higher frequency. If the filter is a digital low-pass, it may suppress the alias spectrum placed between the second and the third Nyquist zones, so the transition-band (filter slope) specification of the further analog reconstruction filter may be relaxed because the first unfiltered image has been pushed higher in frequency. It is important to assess whether the $\sin(x)/x$ distortion (or aperture distortion) in the DAC, which produces nulls at multiples of f_{DAC} , is relevant. Fortunately, it may be digitally compensated by using simple digital equalization techniques.

Interpolating with a rate of two results in more time-efficient structures, because they require less computational effort. Usual structures to implement the low-pass filter are Halfband filters (easy to implement due to certain symmetry properties, and showing optimum efficiency with filter coefficients scaled in inverted powers of 2), the cascaded integrator comb (CIC) interpolation filters (simplicity of implementation), and polyphase filters, where a large FIR (finite impulse response) filter is reduced to a set of smaller filters, so allowing lower sample rates. These filter architectures are computationally efficient implementations of the filter. Halfband and CIC filters do not use multipliers.

The interpolation process may be made by cascading basic interpolation filters. A single low pass interpolation filter is composed of two subsystems: first, the interpolator places zeros between successive samples of the signal (*zero packing*, in a number according to the desired interpolation rate), then the low-pass filter eliminates images

of the spectrum produced by the inserted zero-valued samples (for details, see related chapters of this book on resampling). Notice that trying to replicate the values of previous samples, instead of using the zero packing process, implies a zero-order hold function, and the consequent $\sin(x)/x$ could distort the signal. An alternative, only possible if the bandwidth of the DAC is large enough, is to implement the interpolation filter as a high-pass or band-pass. In these cases, the fundamental of the spectrum is suppressed, the filter output is the first alias signal, placed around f_{ups} , where f_{ups} is the resulting frequency after interpolation. This operation acts like a first IF translation, in a digital mixing process. When using high-pass or bandpass filters it is recommended to have a f_{DAC} at least twice f_{ups} in order to alleviate the $\sin(x)/x$ distortion in the DAC.

3.2 New objectives and technologies in modern wireless communication systems

3.2.1 New and emerging wireless communication standards

Market research studies made around 2005 [29] showed a prevision of revenue decrease in cellular communications based on circuit switching (i.e., GSM), a maintenance of the benefits of voice communications based on packet switching (i.e., GPRS), and a huge increase supported by the technologies enhancing data communications (i.e., CDMA-2000, UMTS). On the other hand, in the same period, wireless communications technologies were diverse in access techniques, modulation formats, adaptability levels, carrier frequencies, and bandwidths. Other studies [13] have shown a tendency to increase subscriptions using mobile broadband, with the mobile data expected to overtake voice traffic after 2010.

Cellular communications, Internet, WLAN, and WPAN are generally disjoint applications operating at different data rates, with complicated (if possible) interoperability, practically impossible mobility among technologies, and difficult operation among network operators; the last implying the payment of several bills to different operators or service providers. This scenario was the starting point for the visionaries of the *fourth generation* (4G) of mobile communications: the convergence of fixed (wired or wireless) networks with mobile and nomadic ones with end-to-end IP solutions (fully IP-based integrated system), at speeds between 100 Mbps and 1 Gbps. Apart from voice, the key application is streamed multimedia, to be served to users on an “anytime, anywhere” basis, according to a specified quality of service (QoS). Ideally the QoS should be the unique user-sensitive parameter, other technical parameters being transparent to the user. Besides, 4G systems should dynamically share and utilize network resources to meet the minimal user’s requirements.

Different standardization bodies (Figure 3.6) have contributed to define standards for mobile communication in the path towards 4G. The Institute of Electrical and Electronics Engineers (IEEE) has proposed the IEEE 802.16 series (a series of Wireless Broadband standards), and especially the IEEE 802.16e-2005 (commercially known as Mobile WiMAX, acronym for Worldwide Interoperability for Microwave Access) already

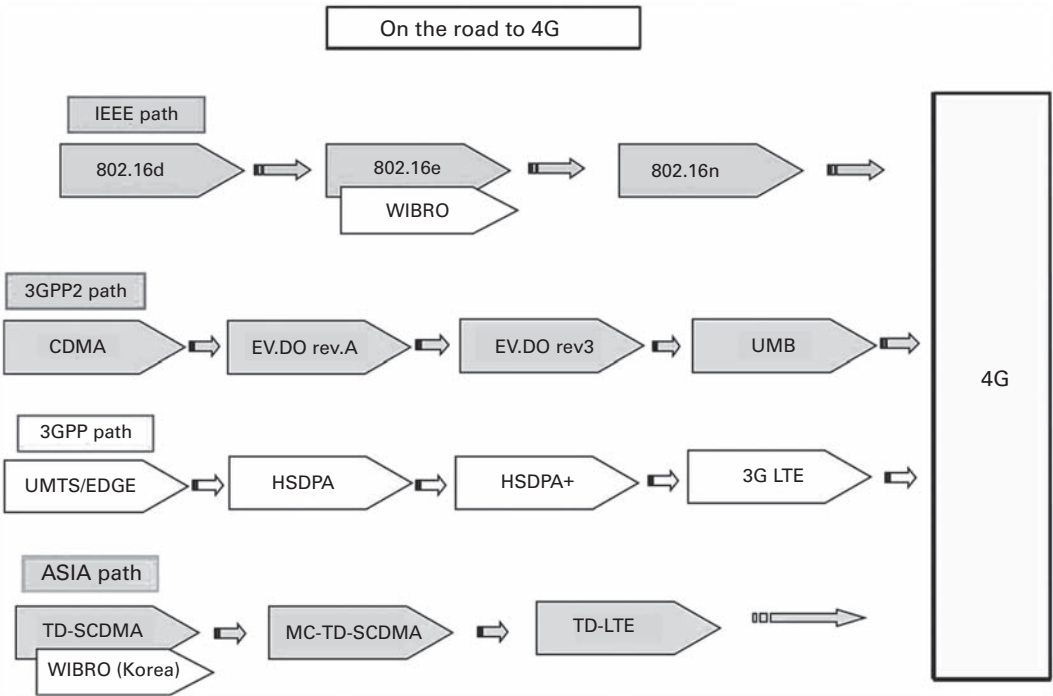


Figure 3.6 Different standards on the path to 4G of mobile telephony.

used for mobile telephony in some few countries. Currently, some standard amendments such as the IEEE 802.16m (advanced air interface with peak data rates of 100 Mbps for mobile or 1 Gbps for fixed or nomadic terminals, with spectral efficiency up to 10 b/s/Hz) are under development. IEEE 802.22 is another standard for wireless communications suitable for rural and remote areas employing the underutilized frequencies among the TV broadcast bands. On the other hand, standards of the family IEEE 802.11, although designed for wireless LAN/MAN, have to share equipment in mobile terminals (MT) with cellular applications and with WPAN subsystems (i.e., Bluetooth, IEEE 802.15.1, or UWB-Zigbee, IEEE 802.15.4).

Other standardization bodies are the 3rd Generation Partnership Project (3GPP) and the 3GPP2, both promoting the deployment of 3G (3rd Generation) under the guidelines of the International Mobile Telecommunications-2000 (IMT-2000), basically a family of standards for mobile telecommunications fulfilling directives of the International Telecommunication Union (ITU). IMT-2000 has been presented as the “umbrella specification” for all 3G systems, as is the case of UMTS or CDMA-2000.

The 3GPP was pioneered by the European Telecommunications Standards Institute (ETSI), with partners in Europe, America and Asia. The objective was to produce technical specifications and reports for a 3G mobile system based on evolved GSM core networks as well as the supporting radio access technologies (RAT), like the General Packet Radio Service (GPRS) and the Enhanced Data rates for GSM Evolution (EDGE),

as well as the Universal Terrestrial Radio Access (UTRA) that is the RAT for Universal Mobile Telecommunications System (UMTS) and uses both Frequency Division Duplex (FDD) and Time Division Duplex (TDD) modes. The most recent enhancement has been Long Term Evolution (LTE), to meet the IMT-Advanced program requirements.

The 3GPP2 has organizational partners from America and Asia, and their specifications are not based on GSM technology. It has published the specifications for CDMA2000, which represents a family of standards developed to maximize the use of pre-existing deployments of CDMA2000-1x and *HRPD* (High Rate Packet Data, known as 1xEV-DO, a high-speed CDMA-based wireless technology developed by Qualcomm), and includes:

- CDMA2000 1xEV-DO, EV-DO Rev. A, and EV-DO Rev. B (EV-DO: Evolution Data Optimized or Evolution Data (EV) only). CDMA2000 1xEV-DO Rev.A provides theoretical peak data rates of 3.1 Mbps (forward link) and 1.8 Mbps (reverse link), although under normal conditions in commercial networks users experience lower rates. CDMA2000 1xEV-DO Rev B provides improved support for multimedia content and VoIP (voice over IP) performance. It introduces dynamically scalable bandwidths. Typical deployments include two or three carriers: with three carriers the peak rate is of 9.3 Mbps, reaching speeds up to 5.4 Mbps in the downlink. The required bandwidth is 1.25 MHz per carrier.
- Ultra Mobile Broadband (UMB) is the 3GPP2 standard originally planned for the next generation of mobile phones. Designed for peak data rates up to 288 Mbps on the forward link, and up to 75 Mbps in the reverse link, it uses multiple-input multiple-output (MIMO) antennas, and Spatial Division Multiple Access (SDMA). Besides, it combines the best aspects of CDMA (Code Division Multiple Access), TDMA (Time Division Multiple Access) and OFDMA (Orthogonal Frequency Division Multiple Access). Initially it was seen as the competitor of the 3GPP standard LTE. Nevertheless the announcement of Qualcomm, the UMB's lead sponsor to abandon the development of UMB standard in November 2008, left LTE and Mobile WiMAX as the unique technologies for Evolved 3G (E3G).

Apart from the paths from IEEE, 3GPP and 3GPP2 to the 4G of cellular telephony, there is also the Asian path. This is mainly based on the Chinese standard *TD-SCDMA* ("S": synchronous) proposed in 2006, with the collaboration of Datang and Siemens AG. Initially designed to have a Chinese proprietary standard, it was later recognized by ITU and adopted by 3GPP (Rel-4) as "UTRA TDD 1.28 Mcps Option". It uses TDMA in addition to the CDMA to accommodate asymmetric traffic by dynamically adjusting the number of timeslots used for both downlink and uplink frames, splitting resources on the basis of real-time demand (reduction of the number of users in each timeslot, a smart solution in populated areas). The TD-SCDMA evolution towards TD-LTE (or LTE TDD, Long Term Evolution-Time Division Duplex) was initially pushed by China Mobile; TD-LTE is compatible with the specifications of 3GPP Release 8 LTE, and as TD-SCDMA, it uses TDD unpaired spectrum channels (1.6 MHz) that combine uplink and downlink.

In addition, the standard *WiBro*, compatible with Mobile WiMAX, was developed in South Korea in 2004. The commercial launch was made in 2006.

3.2.2 Overview of UMTS

Universal Mobile Telecommunication System (UMTS) is the European adaptation of IMT-2000, and it offers flexible bandwidth up to 2 Mbps (only for indoor or urban outdoor and for low-speed moving users). The terrestrial radio interface of UMTS [14] [1], UTRA (previously mentioned in this chapter) combines the FDD and TDD components to support the different symmetrical and asymmetrical service needs in a spectrum-efficient way. In *FDD* different frequencies are used for up-link and down-link, thus allowing two-way communication between radio stations or vice versa. Instead, *TDD* uses the same frequency to transmit signals in both the downstream and upstream directions, switching the communication in a very short period of time. FDD permits higher user mobility, while TDD is more spectrum-efficient.

Universal Mobile Telecommunication System uses CDMA. In CDMA each MT has assigned a unique digital code which identifies it, thus being a technology that allows simultaneous communication on the same frequency band among mobiles and the base-stations. Universal Mobile Telecommunication System uses spreading factors (SF, a ratio between bit-time and chip-time) ranging from SF=4 to SF=512 in both FDD and TDD modes. The spreading codes used in UMTS CDMA are called orthogonal variable spreading factor (OVSF) codes.

At a glance (Figure 3.7), CDMA is based on the *Direct Sequence Spread Spectrum* (DS-SS) technique [42][35], where an already modulated signal is again re-modulated by

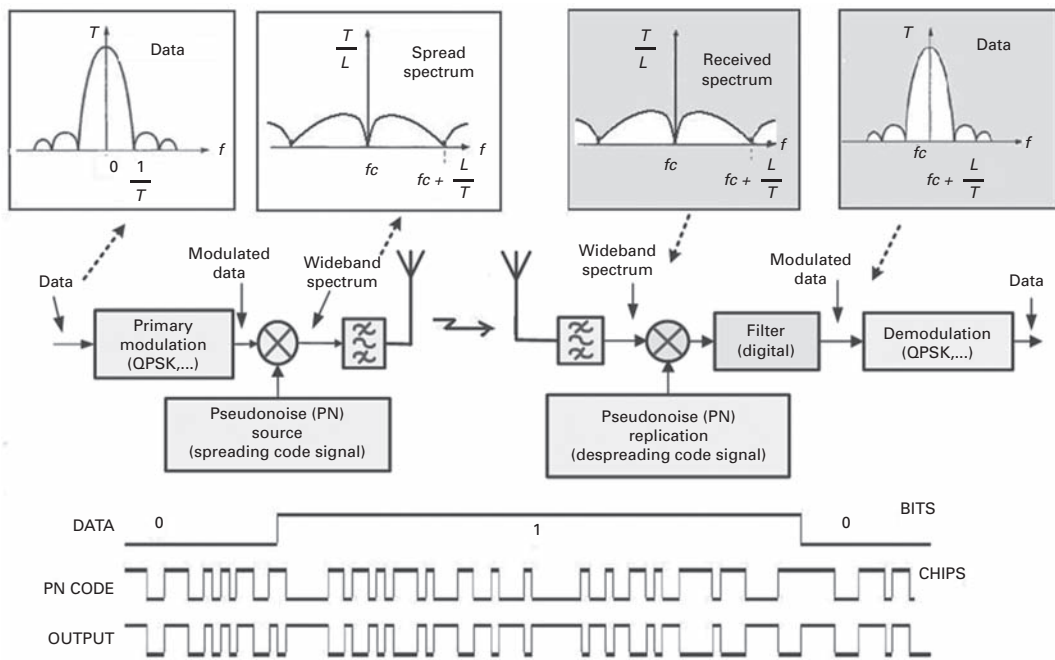


Figure 3.7 Spread spectrum (direct sequence). L/T = chip rate.

a *pseudo-random* (PR) stream of digits, decreasing the spectrum amplitude while enlarging its bandwidth (and so keeping the signal energy, which is somehow related with the area). Multi-user operation is possible by using a set of orthogonal PR sequences, with amplitudes ± 1 and a basic period called “chip,” shorter than the bits period. Each PR sequence is assigned to a different user. Only the receiver capable of reproducing (from a good frame synchronization) the correct transmit PR sequence may recover the right message (i.e., to recover the original modulated signal polarity, as it was before the multiplication for the PR sequence in the transmitter); the rest of the users perceive this transmission like a floor noise (PR noise). When faced with multipath distortion, the synchronization in the receiver is made by using the RAKE receiver, one of the most critical elements for correct CDMA reception requiring (if implemented in SDR) a large computational burden (so it is often developed as an ASIC). Basically, the *RAKE* receiver determines the time delays of the signal replicas due to multipath, and identifies the position at which are allocated the peaks of the energy. For fine tracking before data demodulation, *correlator-based tracking code loops* are also used, as is the *delay-locked loop* (DLL).

The CDMA-FDD mode is usually known as WCDMA (Wireless CDMA), and supports data rates up to 384 kbps for high mobility, whilst CDMA-TDD, also known as *TD-CDMA*, allows small cells with symmetrical or asymmetrical traffic up to 2 Mbps. Wireless WCDMA is the most used mode for public mobile network providers, while TD-CDMA mode, which supports low user mobility and needs lower power control regularity (control commands cadence), is well suited for local and indoor communications (UTRA TDD). Apart from TD-CDMA, the other variation of TDD based CDMA is TD-SCDMA.

Wireless CDMA fits into two 5 MHz paired channels (typical chip rate of 3.84 Mcps), while in a TD-CDMA the channel spacing is also 5 MHz, but an unpaired channel is enough. Data stream is distributed to one or several *dedicated physical channels* (DPCHs), with a basic frame of 10 ms organized in 15 time slots. With an SF of 256, this implies the sequence of 3.84 Mcps (2560 chips per slot).

Although in FDD mode the minimum required frequency band is 2×5 MHz, network operators need, at least, three or four paired channels of 5 MHz (2×15 or 2×20 MHz) to build a high-speed, high-capacity network. The UMTS Forum proposed to national spectrum regulation bodies to reserve a minimum of 2×40 MHz for UMTS deployments. In Europe, WCDMA uses a spectrum of 2×60 MHz in paired bands (1920–1980/2110–2170 MHz); TD-CDMA uses $20 + 15 = 35$ MHz, in unpaired bands within the frequency ranges 1900–1920 and 2010–2025 MHz.

Universal Mobile Telecommunication System operators in North America have to share spectrum with existing 2G services in the 1900 MHz band, or to use other bands (i.e., the 2100/1700 MHz band). Other countries (Canada, some South American countries, India, . . .) already use the 700–800 MHz spectrum, which is better to provide higher coverage at reduced transmit power. With the world tendency to use digital TV technologies for broadcasting, the classical 200–800 MHz spectrum is under the so-called refarming process, being 3G and Evolved 3G (e3G) serious candidates in countries where this band is not yet being used for mobile wireless communications. Some

multimedia broadcasting systems, such as DVB-H or ISDB-T, have already reserved frequencies within this UHF band.

The UMTS vocoder uses adaptive multirate (MR-ACELP), thus resulting in bit data rates between 12.2 and 4.75 kbps. The maximum number of voice channels on the UMTS bandwidth depends on the SF. Modulation (before spreading) is always QPSK, being also allowed 8-PSK in TD-SCDMA. The detection is coherent both in uplink and downlink channels. The UMTS uses convolutional coding with a rate of 1/2 or 1/3, or Turbo Codes with rate 1/3 for high rate data. In the transmitter, the impulse shaping filter is a root raised cosine (RRC), with a roll-off factor of 0.22.

Regarding the MTs, the maximum peak power is independent of the FDD or TDD mode, and it depends on the Power Class. For Power Class 1 it is +33 dBm (2W); for Class 2 +27 dBm, Class 3 +24 dBm, and Class 4 +21 dBm [15]. The adjacent channel leakage ratio (ACLR) affects other network operators in the neighbor frequency bands because of the spectral regrowth, which supposes a problem in the design of the PA because the trade-off between linearity (necessary for ACLR and EVM (error vector magnitude) compliance) and power efficiency (the *DPD* and the *crest factor reduction techniques* presented in other chapters of this book are good tools to resolve this trade-off). In the base-station transmitter the ACLR has a conformance limit between 44.2 dB and 49.2 dB (these figures are sensitive to confusion with the adjacent channel selectivity (ACS), defined at the receiver level, which must be better than 33 dB).

In the previous paragraph the term “base-station” has been used which perhaps needs clarification. In UMTS the term “Node B” is used instead of BTS (base transceiver station) used in other networks, such as GSM. Node B is a rack containing all the necessary hardware and software to connect the mobile handsets with the radio network controller (RNC). The RNCs are the governing elements in the UMTS Radio Access Network (UTRAN), and they connect various Node Bs to the core network (CN) through Iu interface. Although a Node B typically serves a cell site, network cells may be associated to different Node Bs thanks to the utilization of WCDMA technology that allows frequency reuse. The UTRAN access network is composed by one or several Radio Network Systems (RNS), which are responsible for the management of the radio resources. A RNS is composed of one RNC and one or various Node Bs. According to the vendors, a Node B transmitter (WCDMA) may range between 20 and 40 W, although some equipment with higher power is available on the market. A smaller number of vendors offer Node Bs for TDD technology, with lower transmit power.

The radio link quality requires *power control* to minimize *interferences* and to reduce the *near-far* effect. Power control ensures all devices transmitting at the same time are received with a similar power level. In WCDMA the power control frequency (1500 times per second) is 800 times higher than in GSM networks, and twice as high than in IS-95 uplink. In the closed loop power control mode the user Equipment (UE) transmitter has to be capable of adjusting the output power with step sizes of 1, 2, and 3 dB. In open loop mode, the power control tolerance is ± 9 dB (normal conditions) or ± 12 dB (extreme conditions). Isolation between the transmitter and the receiver sections in the base-station is specified at 80 dB, being relaxed to 55 dB in the UE.

The handover is soft (two active paths during transitions); the UE always keeps at least one radio link to the UTRAN.

Sampling values used in SDR developments are multiples of the typical chip rate of 3.84 Mcps. Regarding the DAC and analog filter within the transmit chain, the DAC input sample rate is higher than the rate used for the previous baseband digital signal processing blocks. This *oversampling* process separates images of the baseband spectrum, thus relaxing the analog filter requirements, as explained above. Oversampling at a factor of 16 results in 61.44 Msps, a value used in the technical literature (Figure 3.8). Other values used are the ones imposed by some application-specific standard processors (ASSP, i.e., TMS320TCI110) or application notes (as is the case of Xilinx Virtex-5 FPGA, [54]) which may oversample at rates of two, four, six, or eight samples per chip, resulting in sample rates of 7.68, 15.36, 23.04, or 30.72 Msps. Digital-to-analog converter devices capable of operating at minimum speeds of $2 \times 61.44 = 122.88$ Msps are usual (it is easy to find commercial DAC and ADC converters with higher values).

In the digital up-converter, the complex baseband signal at the chip rate (3.84 Mcps) is converted to a real passband signal capable of driving a final analog up-converter. In Figure 3.8, both in-phase and quadrature components of the baseband signal are filtered and interpolated by equal interpolating pulse-shaping filters; $P(z)$ is the square-root, raised-cosine (RRC) pulse-shaping filter (an FIR filter used to channelize and to minimize the ISI); $C(z)$ is another FIR filter to compensate for the in-band roll-off of the CIC interpolation filter. The CIC filters are time-efficient solutions to interpolate because they do not use multipliers, instead only adders, subtractors, and registers are necessary. Hence, the CIC filters' output have a higher sampling rate (3.84 Msps are transformed to 61.44 Msps) which are later digitally modulated with a complex carrier, DDS (direct digital synthesizer) generated, and added to produce the real up-converter signal [54]. The fundamentals of this process as well as CIC and RRC filters are explained in other chapters of this book on resampling and DUC.

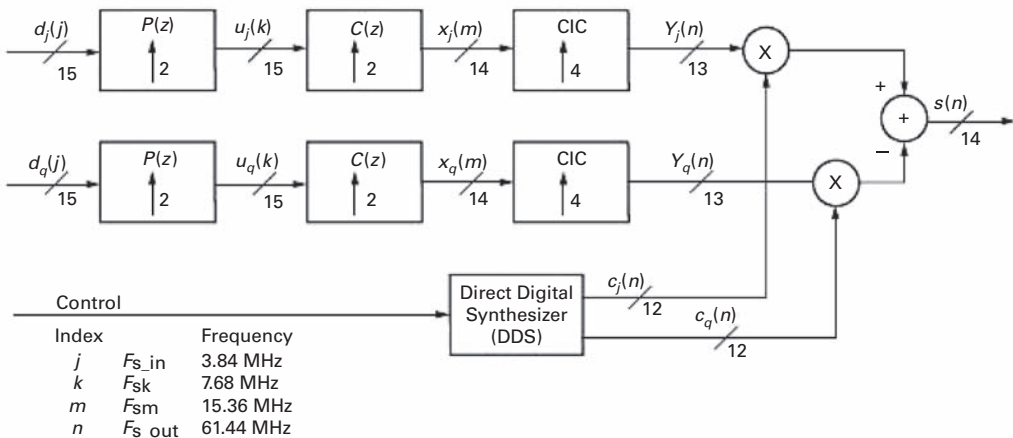


Figure 3.8 UMTS single-channel digital up-converter (DUC). Reproduced with permission from [54].

Noise figures are sensitive to the sampling rate, but this aspect is more relevant in the receiver than in the transmitter section. In the transmitter DACs, the minimum resolution is 10 bit for practical realization [37]; 12 or 14-bit converters being frequently used.

Finally, we should mention that UMTS technology allows improvements that increase its capacity. An example is the High Speed Downlink Packet Access (HSDPA), detailed in 3GPP Release 5, capable of increasing the data speed up to 14 Mbps and to shorten the latency time. High Speed Downlink Packet Access (commercially deployed in 2005), together with High Speed Uplink Packet Access (HSUPA, 2007) constitutes *HSPA*, a 3GPP packet-based technology that uses QPSK/16QAM modulation and turbo coding. It improves the performance of WCDMA systems and makes UMTS more competitive than CDMA2000 1xEV-DV. In Releases 5 and 6, HSPA supports peak rates up to 14 Mbps in DL (downlink) and 5.7 Mbps in UL (uplink); *HSPA Evolution* (or HSPA+) brings a maximum 28 Mbps in DL and 11 Mbps in UL.

3.2.3 Some physical (PHY) layer characteristics of LTE and Mobile WiMAX

With the withdrawal of UMB, the technologies for Evolved 3G (E3G, or 3G+) are LTE and Mobile WiMAX. Sometimes these technologies are commercially presented as 4G, and some operators are eager to announce the deployment of 4G networks by “using LTE technology.” Currently there are no official definitions for 4G; simply “it is the next stage of broadband mobile communications that will supersede 3G.” The ITU is working on a definition under the program *IMT-Advanced*, prior to 4G and according to Resolution UIT-R 57. Requirements related to technical performance for IMT-Advanced radio interface are in Report ITU-R M.2134.

Some general targets of IMT-Advanced are already defined: ubiquitous (functionality worldwide), compatibility of services within IMT and fixed networks, fully IP based (VoIP capacity), high-quality mobile services, gigabit speeds (up to 350 Km/h, with a minimum requirement of 100 Mbps for high mobility), etc. Anyway, the classification of LTE and WiMAX within 4G is inappropriate. They may be classified as Evolved 3G (E3G, or 3G+), but not as 4G [39].

The LTE Advanced (Long-Term-Evolution Advanced) technology is a candidate to be included among the IMT-Advanced standards. It should be compatible and share frequency bands with the previous releases of LTE equipment.

In the same way, Mobile WiMAX (IEEE802.16e-2005) [50][51] is evolving towards the IEEE 802.16m standard [22]. Mobile WiMAX operates at licensed bands between 2 and 6 GHz, can offer peak data rates of 63 Mbps downlink and 28 Mbps uplink over 10 MHz channels (using MIMO antennas), and supports user speeds up to 120 Km/h. IEEE 802.16m is designed to provide mobile broadband wireless access in the next decade, also with the objective to fulfill the IMT-Advanced criteria. It also aims at higher user speeds (up to 350 Km/h) according to IMT-Advanced objectives.

3.2.3.1 Long Term Evolution (LTE)

The LTE standard has been defined by 3GPP. Initiated in 2004, the LTE project focused on enhancing the UTRA and optimizing 3GPP's radio access architecture. Release 8 was

published in December 2008 and it has been the basis for the first LTE developments. According to Release 8, LTE supports six scalable bandwidths, from 1.4 MHz to 20 MHz (1.4, 3, 5, 10, 15, and 20 MHz) and supports both FDD and TDD duplex modes in the same platform. Although optimized for pedestrian speeds, it supports up to 120 km/h (speeds higher than 300 km/h are announced as possible, depending on the used frequency band). Data rate is 100 Mbps in the DL, and 50 Mbps in the UL in a 20 MHz bandwidth (in practice, between 70 and 90 Mbps). Using spatial multiplexing (MIMO) the user data-rate becomes 326.4 Mbps for 4×4 antennas and 172.8 Mbps for 2×2 antennas [38][41]. LTE supports at least 200 active users in every 5 MHz cell.

The enhanced UTRA (eUTRA) is the air interface for LTE, unrelated and incompatible with WCDMA, whose first trials started in 2008. In contrast with UMTS UTRA that uses fixed 5 MHz channels, eUTRA uses different, scalable, channel bandwidths between 1.25 and 20 MHz. Access multiplexing technique is OFDMA for the DL and Single Carrier-Frequency Division Multiple Access (SC-FDMA) for the UL. The air interface supports QPSK, 16QAM, and 64QAM data modulation schemes, and adaptive modulation and coding is employed in both DL and UL. It permits convolutional and turbo coding (coding rate of 1/3). Latency (roundtrip delay) is low, about 5 ms (10 ms in some sources) for small IP packets.

Long Term Evolution implements OFDMA in the downlink. Orthogonal Frequency Division Multiple Access (OFDMA) is a *multi-user* access based on OFDM (Orthogonal Frequency Division Multiplexing) technologies. Notice that OFDMA is not exactly OFDM; it is a multi-user version of the OFDM, designed to facilitate multiple users access to the same channel by assigning subcarriers to individual active users (Figure 3.9). So all active users can transmit and receive at the same time within a single channel. Instead, OFDM allows only one user on the channel at any given time; OFDM is the technique to combat multipath problems, while OFDMA is for multiple access.

Neither is OFDM a modulation format. Rather, it's a multi-carrier transmission technique where data are transmitted on multiple adjacent subcarriers, each subcarrier being previously modulated with a linear modulation scheme such as QAM or QPSK. The basic principle of the OFDM technique [35][46] is to split a high-rate data-modulated stream (of rate R) into N parallel low-rate data streams (thus resulting in lower rates of R/N) each being a narrowband signal modulated by a different subcarrier. As a result, it is more robust to multipath and ISI. Each of these subcarriers experiences

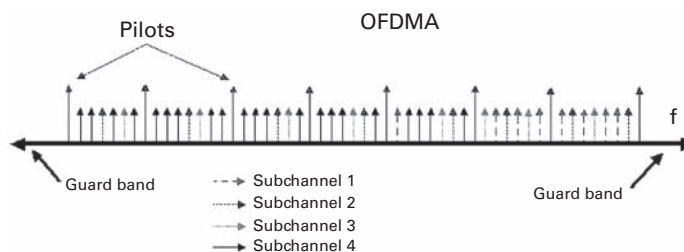


Figure 3.9 OFDMA assignment of subcarriers to different users.

flat fading as they have a bandwidth smaller than the mobile channel coherence bandwidth. This fact obviates the need for complex equalizers. Figure 3.10 presents the underlying idea, and Figure 3.11 shows the general structure of an OFDM system.

In OFDM the serial data stream, sent at a rate of R bps, is first transformed into parallel, resulting in N streams of rate R/N . The OFDM multiplex process supposes the multiplication of each parallelized data for a sinusoid of the corresponding subcarrier frequency (Figure 3.11). Analytically, this is made by means of the inverse discrete Fourier transform (IDFT) while the demodulation is a DFT operation (again the samples are multiplied by a set of complex exponentials whose frequencies are placed on the subcarriers). This operation, facilitated by the disposal of efficient FFT algorithms along with very large scale integration (VLSI) technologies, permits *ASIC* developments avoiding the impractical use of banks of oscillators and coherent demodulators at each subcarrier frequency.

A useful OFDMA symbol period is $T_u = 66.7 \mu\text{s}$. The subcarrier spacing is the inverse of the symbol period, $\Delta f = 1/T_u = 15 \text{ kHz}$. Each 15 kHz subcarrier is capable of transmitting 15 ksps, resulting in a raw symbol rate of 18 Msps in a 20 MHz system bandwidth with 1200 subcarriers. Using 64QAM – the most efficient of the LTE modulation formats – in which one symbol represents six bits, the theoretical raw capacity is 108 Mbps.

Long Term Evolution time intervals are expressed in technical documentation [3] as multiples of a basic time unit $T_s = (1/30720000)\text{s}$. Time scheduling (Figure 3.12) uses the subframe as the computational unit, which consists of two equally sized slots of 0.5 ms ($T_{\text{slot}} = 15360 \cdot T_s$, $T_{\text{subframe}} = 30720 \cdot T_s$). In *frame structure Type 1*, used in both FDD and H-FDD (Half Duplex-Frequency Division Duplexing) modes, the radio frame has a length of 10 ms (it consists of 10 subframes or, equivalently, 20 slots) [36]. In *frame structure Type 2*, used in TDD modes, the frame length is also 10 ms, but among the slots are included some special fields.

The composite signal obtained after the IFFT block is extended by repeating the initial part of the signal (called the *cyclic prefix*, CP). The CP is basically a guard time to combat channel multipath (ISI), during which reflected signals will reach the receiver. This

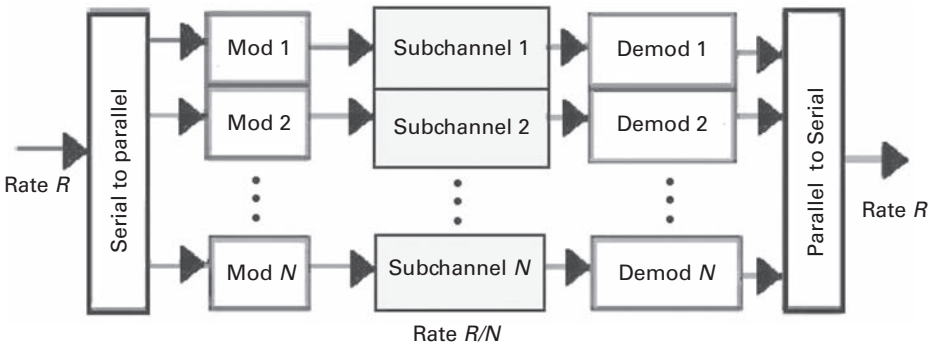


Figure 3.10 Channelization into N parallel channels.

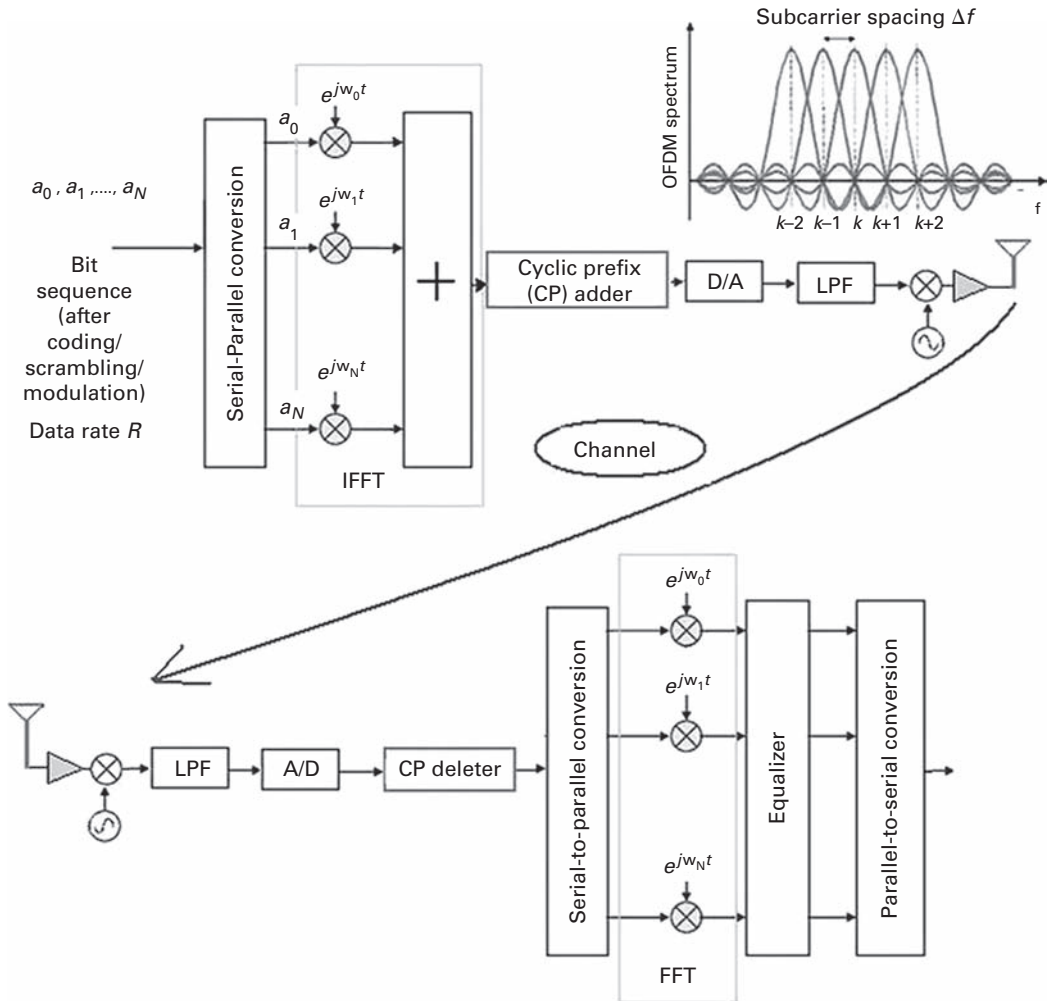


Figure 3.11 General structure of an OFDM system.

extended signal represents an OFDM symbol. For readers keen for analytical details, just to mention that the CP allows the circular convolution between the symbol and the channel impulse response to result in a linear convolution.

The CP may be short in normal mode ($4.69 \mu\text{s}$, or $5.2 \mu\text{s}$ for symbol 0 of the subframe) or long ($16.67 \mu\text{s}$) in extended mode. The useful part of the OFDM symbol occupies a time of $66.7 \mu\text{s}$ ($T_u = 2048 \cdot T_s$), so, to fulfill the 0.5 ms of the time-slot, seven symbols are necessary in normal mode $[(5.2 + 66.7) \mu\text{s} + 6 \cdot (4.69 + 66.7) \mu\text{s} = 0.5 \text{ ms}]$, or six in the extended one. The short CP permits the system to cope with path delays variations up to about 1.4 km , and it is the one used in urban cells. Extended CP is used in special cases (multi-cell broadcast, rural areas, low data rate applications).

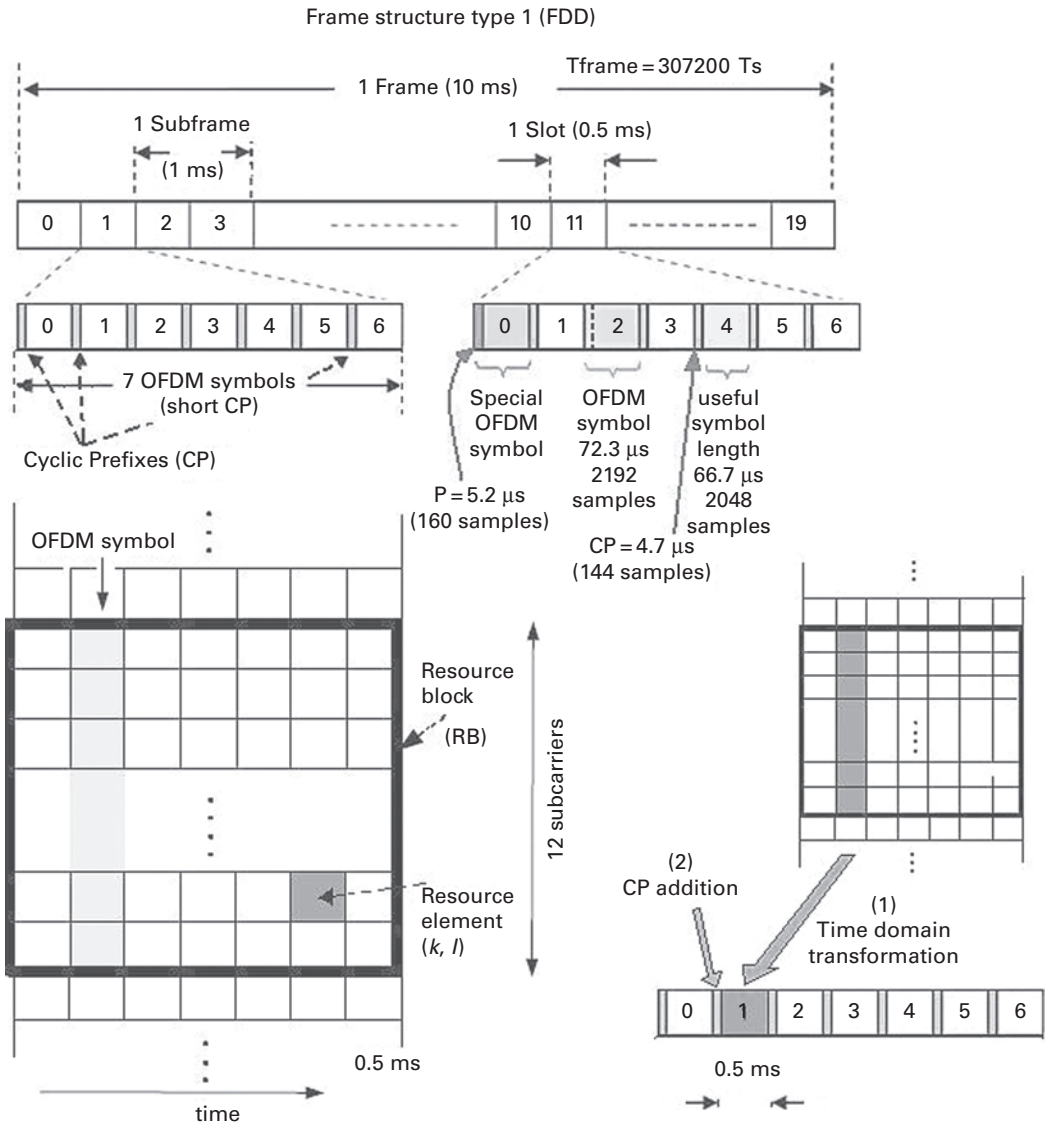


Figure 3.12 LTE frame structure Type I. (The OFDM symbol is the parallel transmission, over the subcarriers, of a source symbol. The physical wave is the result of the addition of the subcarriers after IFFT block in Figure 11.)

An LTE resource block (RB) is a set of consecutive carriers, and each RB consists in a number of *resource elements* (RE) which are uniquely defined by index pairs (k, l), k and l being the indices for the frequency and time-domain, respectively. After modulation, coding and scrambling, the symbols are mapped to RE [25][21], which may be also filled from the mapping of data from the control channels. An RB corresponds to 180 kHz in the frequency domain, and one slot in time domain. The physical layer of LTE is designed

Table 3.1 Some key parameters in LTE (each RB contains 12 subcarriers).

Spectrum allocation	1.4 MHz	3 MHz	5 MHz	10 MHz	15 MHz	20 MHz
Frame duration				10 ms		
Sub-frame duration				1 ms		
Slot duration				0.5 ms		
Useful symbol duration				66.7 μ s		
Subcarrier spacing				15 kHz		
FFT size	128	256	521	1024	1536	2048
Number of RBs per subframe	6	15	25	50	75	100
Number of subcarriers	72	180	300	600	900	1200
DL and UL modulation			QPSK, 16-QAM, and 64-QAM			
Max. practical data rate per carrier	4.5 MHz	11 MHz	19 MHz	37 MHz	56 MHz	75 MHz
Normal CP length		Symbol 0: 5.21 μ s, Symbols 1 to 6: 4.69 μ s				
Extended CP length		Symbols 0 to 5: 16.7 μ s.				

to support bandwidths in increments of 180 kHz starting from a minimum bandwidth of 1.08 MHz. The minimum number of RB is six, so the minimum frequency is 1.08 MHz (bandwidth configuration). Consequently, only 77% of the bandwidth is used in a 1.4 MHz transmission bandwidth, the remainder is the guard band. In a 20 MHz transmission bandwidth, the bandwidth configuration occupies 18 MHz (90%). Table 3.1 shows some key LTE parameters.

The reason to abandon CDMA technology, key in UMTS, and favor OFDMA and SC-FDMA in LTE is because of the robustness of these technologies when faced with multipath. Besides, power control in OFDMA systems is less critical than in WCDMA systems since in OFDMA the UEs are separated in time and frequency whereas in CDMA they share the same physical channel, being only differentiated by the spreading code, which requires much tighter control on received power. However, OFDM has also some drawbacks. One, also present in CDMA technologies used in UMTS, is its high peak-to-average power ratio (PAPR, a ratio between the maximum power peak of the modulated signal and its mean value), considered below in the section devoted to design considerations, which makes the efficient operation of the PAs difficult.

The PAPR problem of an efficient biasing of the PAs may be relaxed in the DL path, because in the base-station (known as the *evolved Node B*, eNB, for similarity with UMTS terminology) the transmit equipment is fed from the electrical grid. So the lifetime of the batteries is somehow not critical (the electrical bill and the environmental cost, because of the energy spoil in the PA, as well as the increase of the working time of the air conditioners is another, not minor, issue). But in the UL, the PAs of the MT transmitters have to be battery operated, so it is critical to extend their lifetime. Apart from power control and other strategies, the UL uses a relatively new improvement of OFDMA, well suited for hand-held devices, introduced earlier in this chapter and called Single Carrier Frequency Division Multiple Access (SC-FDMA) that reduces the PAPR with regard to OFDMA (and also with regard to 3G systems using CDMA), in a typical factor of 2 to 6 dB.

The term “single carrier” in the definition of SC-FDMA is often confusing. Actually, in SC-FDMA, the transmitted signal is not just a single carrier; it is also multi-carrier. But unlike OFDM transmitters where each data symbol is carried by individual subcarriers, the SC-FDMA transmitter first spreads data symbols over a group of subcarriers to be transmitted simultaneously. This group can be seen as a single frequency band carrying data sequentially. Besides, in the “single carrier” modulation (SC-FDMA), the resulting PAPR values somehow remember the ones obtained with a basic single carrier modulation. The SC-FDMA system structure is practically the same as that of the OFDM one, with an additional DFT computation in the transmitter and an additional IDFT in the receiver (Figure 3.13).

In OFDM each subcarrier is modulated with a data symbol: the additional DFT in SC-FDMA spreads the input data symbols over all the subcarriers (so the SC-FDMA is occasionally referred to as DFT-spread OFDM). Regarding Figure 3.13, SC-FDMA first feeds the M outputs of the mappers into an M -point DFT processor, thus creating subcarriers in the form of complex numbers; then these complex numbers are processed by an N -point ($N > M$) IDFT to generate a time signal.

Long Term Evolution permits spatial processing techniques based on the use of multiple antennas at the receiver, at the transmitter, or both, together with spatial signal processing and coding. The spatial processing techniques specified in LTE are: spatial diversity, multiple-input, multiple-output (MIMO), and adaptive antenna system (AAS). As shown in Figure 3.14 and Table 3.2, MIMO uses multiple antennas at both the transmitter and receiver to improve data throughput and spectral efficiency. The transmitter sends multiple streams, taking advantage of the multipath to produce a constructive interference at the receiver antennas. In a 10 MHz bandwidth LTE system, a 2×2 MIMO allows 173 Mbps in the DL, and 58 Mbps in the UL, whereas a 4×4 MIMO increases speed up to 326 Mbps in DL and 86 Mbps in UL. The importance of power control grows when using MIMO.

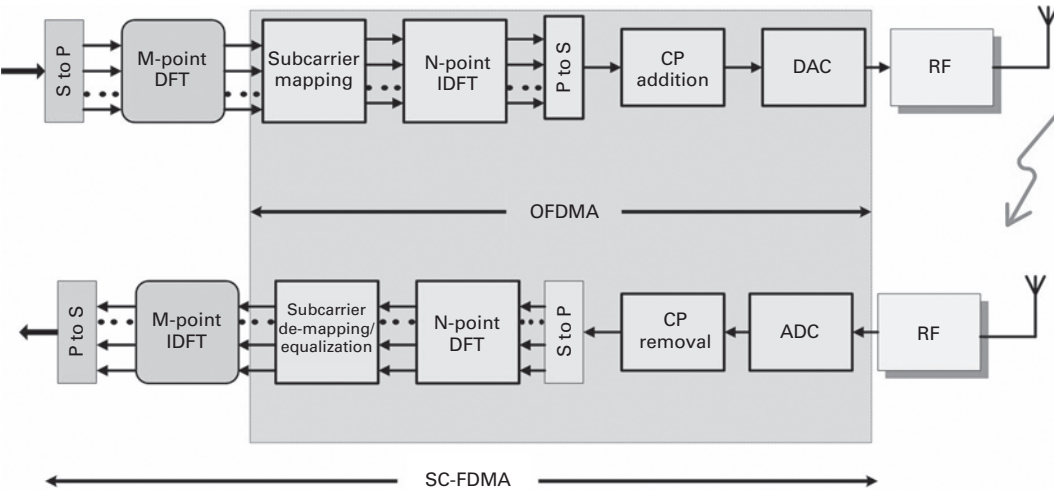


Figure 3.13 OFDMA and SC-FDMA structures. Notice the additional M-point DFT and IDFT in SC-FDMA.

Table 3.2 Peak data rates for UE categories (derived from 3GPP TS 36.306).

UE category	Peak DL data-rate (Mbps)	Peak UL data-rate (Mbps)	DL antenna diversity	Support for 64-QAM in uplink
Category 1	10.296	5.16	1 × 2	No
Category 2	51.024	25.456	2 × 2	No
Category 3	102.048	51.024	2 × 2	No
Category 4	150.752	51.024	2 × 2	No
Category 5	299.552	75.376	4 × 2	Yes

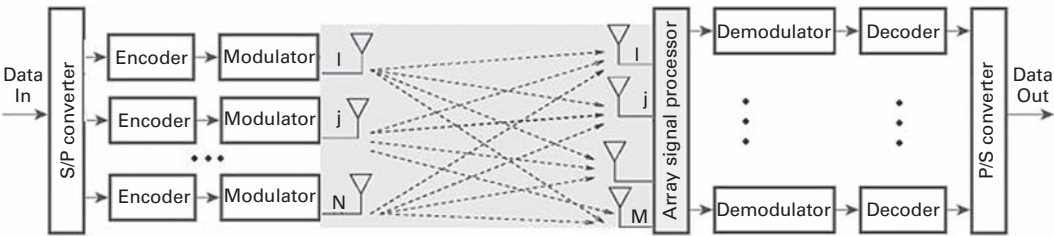


Figure 3.14 General scheme of a $N \times M$ MIMO system.

Three duplex schemes are utilized in LTE systems: FDD, TDD, and H-FDD. The TDD frame structure allows backward compatibility to TD-SCDMA systems. In H-FDD, the DL and UL channels are separated, but the mobile stations (MS) do not simultaneously transmit or receive, while the base-station (BS) continuously transmits and receives in a full-duplex mode. In the time-periods that the MS connects the BS by using the UL channel, the BS may use the DL channel for transmission to another user. The MS antenna is controlled by a switch.

The IMT-2000 frequency bands were defined by the ETSI and 3GPP. Most of the frequency bands are recognized by all the ITU regions, although it should be noted that definition of a band does not imply its availability for deployment since not all the defined bands are available in all countries for IMT-2000. The bands cover from frequencies of 600 MHz (band number 12) to 2690 MHz (band 7). Furthermore, the current transition from analog TV to digital TV in most countries has freed frequency bands in parts of the VHF and UHF spectra. This spectrum, called digital dividend (located between 200 MHz and 1GHz) offers an excellent balance between transmission capacity or power consumption and distance coverage, including a better penetration in buildings.

Several frequency bands support LTE (currently, 13 for FDD and 8 for TDD). The initial deployments were expected in the 700 MHz for the USA, 2.1 GHz for Japan, and 2.6 GHz in Europe. Although the first LTE network in USA (MetroPCS Communications deployed in Las Vegas and in the Dallas metropolitan areas) works at the AWS (Advanced Wireless Services) band (1710–1755 and 2110–2155 MHz), other American operators (i.e., Verizon), based their LTE development plans on the 700 MHz band.

The 2.3/2.5 GHz TDD band is available for deployment in China and India. In Europe the 2.5 GHz band is also under consideration and, moreover, the re-use of the old 900 MHz or to take advantage of the “re-farming” of some UHF frequencies freed from classical analog TV bands. Even 2.1 GHz is considered in some European countries. Also, in Japan, new bands of 1.5 GHz (DoCoMo, among others) and 1.7 GHz (eMobile) are allocated for LTE. In September 2010, mobile operators from Russia and some nearby countries, running CDMA networks in the 450 MHz band, asked the ITU for permission to deploy LTE networks in the band.

There are four Power Classes defined for the LTE UE. At the current time, only Power Class 3 (100 MHz DL and 50 MHz UL) is defined, and specified as 23 dBm (± 2 dB) for all bands. The power may be reduced depending on the modulation depth (signal peaks) and the transmitted bandwidth [3GPP07a]. The RF specifications for LTE are covered in two 3GPP technical specification documents: 36.101 for the UE and 36.104 for the base-station (evolved Node B, eNB).

The LTE emission limits are defined in document ITU-R SM 329. Over 1 GHz, and based on the Category A limit for spurious emission (America and Japan), the allowed emission limit is -13 dBm in a measurement bandwidth of 1 MHz. For category B (Europe), the limit is -30 dBm/MHz. Additional unwanted emission limits have been imposed by US FCC, between -13 and -15 dBm according to the bandwidth, so the spectral mask requirement for the different bandwidths, respectively, can be calculated from the combination of the general LTE emission limits and the additional LTE emission ones. A limit for ACLR is around 45 dB, and the maximum BS power is 43 dBm for 5-MHz carrier offset and 46 dBm for 10 MHz. To follow the power requirements and the ACLR limits is slightly cumbersome because of the number of documents. However, a gate may be found in [2], where the main regulatory references for LTE emission limits are summarized.

Regarding the in-band distortion, measured by means of the EVM, this is 7–8% for 64 QAM modulation, relaxed to 12.5% and 17.5% for 16 QAM and QPSK, respectively. In UMTS the limit is 17.5%.

In a 20 MHz bandwidth, to fulfill the 0.5 ms of the LTE time-slot (Table 3.1 and Fig 3.12), 7×2048 samples are necessary (useful symbol length in a 20 MHz bandwidth), plus 160 samples in the first CP, and 6×144 samples in the remaining short CP. This results in 15 360 samples in 0.5 ms, 30.72 Msps. Table 3.3 summarizes the sampling frequencies for different transmission bandwidths.

Table 3.3 Sampling frequencies for different bandwidths (data partially extracted from [56]).

Transmission BW (MHz)	1.25	2.5	5	10	15	20
Sub-frame duration (ms)				0.5		
Subcarrier spacing (kHz)				15		
FFT size	128	256	512	1024	1536	2048
Sampling frequency (MHz)	192	3.84	7.68	15.36	23.04	30.72
	($1/2 \times 3.84$)		(2×3.84)	(4×3.84)	(6×3.84)	(8×3.84)

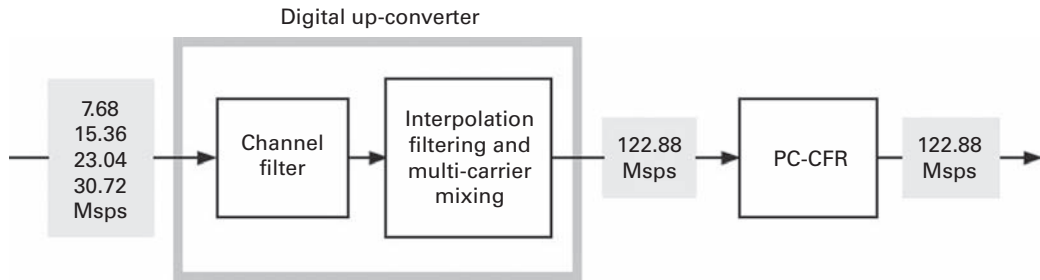


Figure 3.15 Digital front-end architecture for transmit downlink (reproduced with permission from [44]).

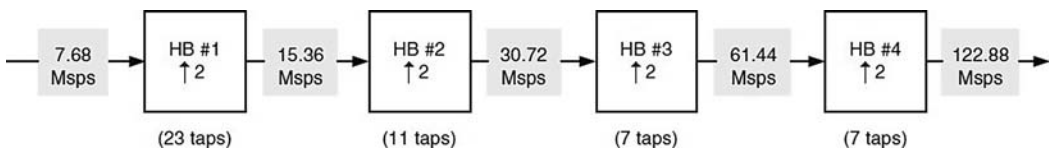


Figure 3.16 Interpolation filter structure for 1x5 MHz configuration (reproduced with permission from [44]).

In an LTE transmitter (Figure 3.15), after baseband processing, the input sample rates at the digital up-converter (DUC) are 7.68, 15.36, 23.04, 30.72 Msps for channel bandwidths of 5, 10, 15, 20 MHz, respectively. After the DUC (composed by channel and interpolation filters) the output sample rate becomes 122.88 Msps [44]. The interpolation filter (Figure 3.16) can be made with a chain of *halfband* (HB) interpolation filters: duplicating the frequency in each HB from 7.68 Msps, the whole set of different input sampling rates is obtained, thus resulting in a general structure for all LTE bandwidths.

Although LTE symbol length is 256 times larger than WCDMA (3.84 Msps symbol rate), the use of this set of sample rates makes them worthy for both UMTS and LTE developments.

3.2.3.2 Mobile WiMAX

As introduced previously in this chapter, WiMAX stands for Worldwide Interoperability for Microwave Access, and it is not a wireless standard. WiMAX is the short name for WiMAX Forum, an industry-led non-profit organization formed to certify and promote the compatibility and interoperability of broadband wireless products based upon the IEEE 802.16/ETSI HiperMAN standard. So WiMAX promotes a family of interoperable technologies, officially called WirelessMAN in IEEE. The Forum is mainly composed by operators, component, and equipment vendors. IEEE is not a member of the Forum: IEEE makes the standards while WiMAX commercially exploits them.

In 1998 IEEE formed the group 802.16, which was in charge of developing a standard for the global deployment of broadband Wireless Metropolitan Area Networks (WirelessMAN). Over the past decade, the initial 802.16 standard evolved, and the following are some significant amendments:

- IEEE Std 802.16-2001: Line-of-sight fixed operation in 10 to 66 GHz
- IEEE Std 802.16a-2003: Air interface support for 2 to 11 GHz
- IEEE Std 802.16d-2004: High-speed data rates for fixed wireless and nomadic access
- IEEE Std 802.16e-2005: Improved air interface and capability for limited mobility
- IEEE Std 802.16-2009: Air Interface for Fixed and Mobile Broadband Wireless Access System
- IEEE Std 802.16j-2009: Multihop relay

Currently IEEE Std 802.16m is under development, which is a candidate for consideration in the IMT-Advanced evaluation process.

An important step in the evolution of IEEE 802.16 happened with the release of the IEEE 802.16-2004 standard. It added multiple radio interfaces, including one based on OFDM-256 and another based on OFDMA. Like the original version (2001) of the standard, operation is fixed, meaning that subscriber terminals are immobile. Fixed WiMAX currently has both FDD and TDD certification profiles. Originally it was only seen as solution for “last mile” broadband connections, as well as a backhaul complement for WiFi (IEEE 802.11) networks. Later, some public and private entities developed WiMAX in unlicensed bands for local connectivity or for very specific applications. But, it was the completion of the IEEE 802.16e-2005, the mobile broadband amendment of the standard that strongly increased the commercial interest, and operators have already deployed some networks in different countries. Mobile WiMAX networks are not backward-compatible with IEEE 802.16-2004 networks. In [53] a survey of WiMAX deployments around the world is presented. According to this document, in the Russian WiMAX deployment (at 2.5–2.7 GHz bands) the DL speed is 10 Mbps. Similar values (ranging from 8 to 16 Mbps) are reported in Taiwan, Ireland, Japan and the Netherlands. In USA some operators speed at 4 Mbps. In 2010 frequencies potentially available for Mobile WiMAX have been allocated to 140 countries, and Mobile WiMAX is adopted by 85 operators in 49 countries. The band of 2.3 GHz is used in 43 countries, the 2.5 GHz band in 27 and the 3.5 GHz in 107 countries.

IEEE 802.16e-2005 employs OFDMA techniques, as well as adaptive modulation and coding. It emphasizes TDD operation, which saves spectrum bandwidth but requires fine network synchronization and produces technical disadvantage in the size of the cells (a typical radius is 2–5 Km; it requires about 1.7 times more cell sites than HSPA). Mobile WiMAX and LTE are quite close in a lot of technical details. Both improve coverage by utilizing advanced antenna diversity schemes (MIMO); peak rates in Mobile WiMAX are approximately 63 Mbps in DL and 28 Mbps UL with 2×2 MIMO in a 10 MHz channel [49]. A primary difference is that LTE uses SC-FDMA in the UL whereas Mobile WiMAX uses OFDMA, a fact that affects the lifetime of the UE batteries. Besides, the lower OFDMA subcarrier spacing (10.94 kHz) in WiMAX, compared with the 15 kHz in LTE, decreases WiMAX susceptibility to multipath delay spread, but at the price of also decreasing robustness when faced with Doppler effects and phase noise in oscillators. So Mobile WiMAX is less suitable for high-speed subscriber terminals. At a glance, Table 3.4 gives a technical comparative between LTE and mobile WiMAX.

Table 3.4 Comparative between LTE and Mobile WiMAX (WiMAX 8.75 MHz channel bandwidth option is WiBRO). FDD is also proposed for Mobile WiMAX in some documents.

PARAMETER	Mobile WIMAX	LTE
Spectrum type	Licensed	Licensed
Duplex method	TDD	FDD, TDD
DL access	OFDMA	OFDMA
UL access	OFDMA	SC-FDMA
Channel Bandwidth (MHz)	3.5, 5, 7, 8.75, 10, (20)	1.4, 3, 5, 10, 15, 20
DL and UL Modulation	QPSK, 16-QAM, 64-QAM	QPSK, 16-QAM, 64-QAM
Subcarrier spacing (kHz)	10.94	15
Subcarriers in 10 MHz BW	720	600
Subcarrier total used BW	$702 \times 0.01094 = 7.88$ MHz	$600 \times 0.015 = 9$ MHz
User plane latency	<50 ms	<10 ms

Table 3.5 Some parameters for Mobile WiMAX.

Channel bandwidth (MHz)	1.25	5	10	20	3.5	7	8.75
Sampling factor	28/25	28/25	28/25	28/25	8/7	8/7	8/7
Sampling frequency (Fs, MHz)	1.4	5.6	11.2	22.4	4	8	10
FFT size	128	512	1024	2048	512	1024	1024
Subcarrier spacing (Δf , kHz)	10.93	10.93	10.93	10.93	7.81	7.81	9.77
Useful symbol time ($T_b = 1/\Delta f$, μs)	91.4	91.4	91.4	91.4	128	128	102.4
Guard time ($T_g = T_b/8$)	11.4	11.4	11.4	11.4	16	16	12.8
OFDMA symbol time ($T_s = T_b + T_g$)	102.8	102.8	102.8	102.8	144	144	115.2

Table 3.5 shows some detailed OFDMA parameters for Mobile WiMAX. Mobile WiMAX frame structure is 5 ms length for all options, while the CP is 1/8 of the useful time. For 1.25, 5, 10, and 20 MHz bandwidth, the number of OFDMA symbols within a frame is 48. The sampling factor (Table 3.5) relates sampling frequency to channel bandwidth, and it is computed as the product of subcarrier spacing and FFT size. That is, for a 10 MHz bandwidth, this last product is $10.93 \text{ kHz} \cdot 1024 = 11.2 \text{ MHz}$. Therefore, the sampling factor is $11.2/10 = 28/25$.

A typical distance between base-stations is up to 3 Km, and common base-station transmit power is around 35 dBm [28], the maximum being 43 dBm. Regarding user terminals, and using typical terminal Power Classes, the maximum output power of Mobile WiMAX terminals is 23 dBm. However, according to [52], the power limits among the Power Classes (Table 3.6) depend on the modulation format. Transmit spectral mask for WiMAX is defined in IEEE 802.16–2004 standard. Figure 3.17 illustrates the corner frequencies for the 10 MHz option. Notice the sharp transition between 4.75 MHz and 5.45 MHz (of at least 25 dB). In practice, the attenuation of the stop band after 10 MHz is designed higher to prevent interference.

Figure 3.18 presents a possible realization for the DFE of the LTE transmitter [19], as well as some practical sampling frequencies. In the figure, the minimum considered

Table 3.6 Power Classes.

Power Class	Transmit power (dBm) 16 QAM	Transmit power (dBm) QPSK
Power Class 1	18 – 21	20 – 23
Power Class 2	21 – 25	23 – 27
Power Class 3	25 – 30	27 – 30
Power Class 4	> 30	> 30

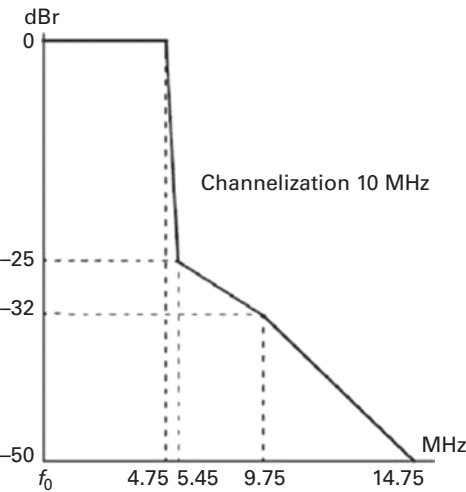


Figure 3.17 ACLR for a channelization of 10 MHz.

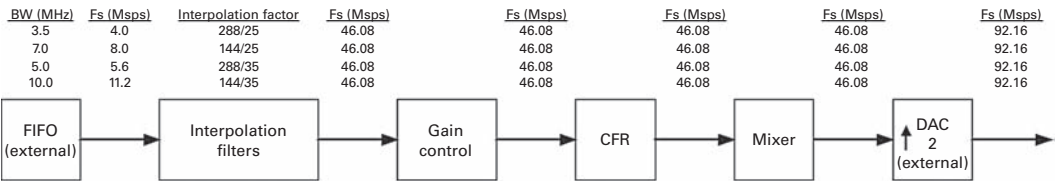


Figure 3.18 WIMAX transmitter DFE (reproduced with permission from [19]).

bandwidth is 3.5 MHz. With the sampling factor of 8/7, it becomes a sample rate of 4 Mbps. The same sampling factor is used for 8 MHz, while the other bandwidths in the figure, 5 and 10 MHz, use the sampling factor of 28/25 (Table 3.5).

The channel filter (CFR) in the DUC chain is designed to deal with the stringent out-of-band power spectral mask shown in Figure 3.17.

The DUC Interpolation filter (Figure 3.19) may be composed by a chain of polyphase and halfband filters, with a polyphase resampler.

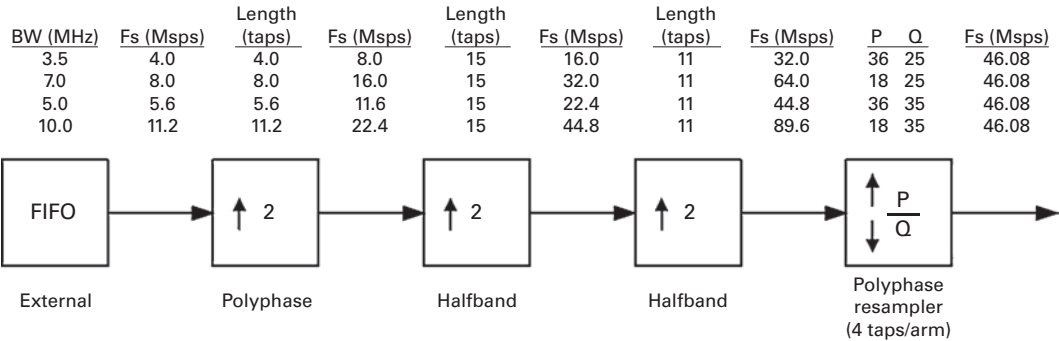


Figure 3.19 DUC interpolation filter chain (reproduced with permission from [19]).

3.3 Other emerging standards beyond 3G

Different IEEE standardization task groups are working on the development of new standards. This is the case of the *IEEE Std. 802.16m* [22] [5], whose preliminary definitions started on December 2006. The IEEE 802.16m is a new amendment of the IEEE 802.16 standard, proposing an advanced air interface to meet the requirements of the International Telecommunication Union (Radiocommunication/International Mobile Telecommunications, ITU-R/IMT) and it has been submitted for inclusion in the IMT-Advanced program as a compliant standard. As in the earlier IEEE802.16e, it also uses OFDMA for multiple access in both the DL and UL. It supports FDD, H-FDD, and TDD duplex schemes, and it is also based on MIMO structures. This new standard is defined as compliant with IEEE 802.16e-2005, and the modulation formats are the same. Operating bandwidth is 5 to 20 MHz (up to 100 MHz through band aggregation), and optimal performance is obtained at user speeds lower than 10 km/h, with little degradation up to 120 km/h and connectivity up to 350 km/h. At peak data rates, and with MIMO operation, the efficiency is 15 b/s/Hz in the DL and 6.75 in the UL. Their frames are 5 ms length (as in Mobile WiMAX), organized in superframes of 20 ms. Each frame has five OFDMA symbols, each one lasting 0.617 ms.

With similarities to IEEE 802.16e-2005, the IEEE 802.20 standard also uses OFDMA and it may provide flexible channelization up to 20 MHz, providing peak data rates over 80 Mbps. It also uses MIMO technologies. The air interface is optimized for the transport of IP based services and supports various vehicular mobility classes up to 250 Km/h. However, the current commercial viability of this standard is, at least, a risk financially, considering the strong interest of vendors and network operators on LTE and Mobile WiMAX technologies.

On the other hand, 3GPP is working on the *LTE-Advanced* standard (LTE Release 10 and beyond), and 3GPP partners made a formal submission to the ITU-R in the fall of 2009 proposing LTE-Advanced to be evaluated as a candidate for IMT-Advanced program.

The LTE-Advanced should be compatible and share frequency bands with the first LTE release, and it is optimized for heterogeneous networks including low power nodes (i.e., femtocells). It will support data rates up to 1 Gbps for low mobility users, and 100 Mbps for the high mobility case. System bandwidth is scalable, exceeding 20 MHz, and potentially up to 100 MHz. Some proposals for LTE-Advanced, regarding the objective of interworking with other radio access systems, also include the use of some Cognitive Radio technologies.

Cognitive Radio (CR), presented by Joseph Mitola III 1998 [34], is a kind of wireless communications strategy in which either a network or a terminal may change its parameters in order to communicate more efficiently or to make more flexible some communication resources among licensed or unlicensed bands. This implies an intelligent detection of the used communication channels, and either to share them or to move users into vacant channels. This optimizes the use of available RF spectrum while minimizing interference to other users. The alteration of radio parameters is based on the active monitoring of several factors in both external and internal radio environments. Cognitive Radio utilizes SDR and other technologies to automatically adjust its behavior. Once the decision is taken, the radio has to be reconfigured accordingly, by adjusting parameters such as the carrier frequency, the bandwidth, the modulation format or the transmit power, apart from other changes in the MAC layer.

The first proposal of J. Mitola III is known as Full Cognitive Radio, and within this approach “the radio autonomously observes the radio environment, infers context, assesses alternatives, generates plans, supervises multimedia services, and learns from its mistakes.” In Full Cognitive Radio it is assumed possible to modify every observable parameter in the nodes or in the network. Another approach searches for holes of available radio frequency bands in a crowded spectrum, and it is named Spectrum Sensing Cognitive Radio, Opportunistic Spectrum Utilization or Interwave.

A CR network consists of two sub-networks: the *primary network* composed by the primary users (PU) and the base-station, and the *secondary network*, where secondary users (SU), licensed or not, try to utilize the spectrum in an opportunistic manner. The secondary network might or might not include a base-station.

The cognitive capability requires adaptability. The steps involved in the adaptive operation (spectrum management) are spectrum sensing, spectrum analysis, spectrum decision, and spectrum handoff [33] [40] [6].

The role of *spectrum management* consists in the coordination of the network access in order to prevent multiple users from colliding in overlapping portions of the spectrum. This management may be either regulator dependent or independent. In regulator dependent management, the strategies may be: spectrum pooling, when two or more parties decide how to pool the resources given by the operator; *spectrum leasing*, which consists in a lease of spectrum by the regulator; spectrum sharing, when a portion of the spectrum is given to a subset of parties; or negotiated spectrum use. The dynamic spectrum access refers to the access to licensed channels based on policies. Regulator independent management is based on opportunistic spectrum strategies.

In the *spectrum sensing* task the user should monitor the available spectrum bands, capture their information, and then detect spectrum holes. This task may be carried out by

using some kind of information transmitted from the base-station (as in the case of FCCH (Frequency Correction Channel), BCCH (Broadcast Control Channel), and SCH (Synchronization Channel) in the GERAN Network (GSM/EDGE Radio Access Network), or may need the application of some signal processing techniques from the observed environmental signals [8]. Examples of different signal processing techniques to detect spectrum holes are matched filter detection (when a SU has a priori knowledge of PU signal), energy detection (radiometers, periodograms, ...) or feature detection techniques (spectral correlation: cyclostationarity may discriminate the noise from the modulated signal energy). In addition, spectrum sensing may be either *cooperative* or not.

Spectrum decision is based on the spectrum availability and determined on internal (and possibly external) policies. It usually consists of two steps: spectrum band characterization and selection of the most appropriate band.

When multiple users try to access the spectrum, spectrum sharing is the task responsible for network access coordination to prevent users colliding. Roughly, three techniques are possible in spectrum sharing: underlay (or interference-tolerant approach), overlay, and interweave.

The underlay approach is usually associated with UWB (ultra-wideband) devices, and it allows concurrent PU and SU transmission. The SU access the network by spreading their signals over a wide frequency band. Underlay protects PUs by enforcing a spectral mask on the secondary signals so that the interference generated by the SU is lower than the acceptable noise floor for the PU (Interference Temperature is a concept to measure the power generated by undesired emitters). In *overlay* approach concurrent PU and SU transmission is also permitted. Within the “known interference model,” two types of channel models have been proposed: mutual interference channel and mutual interference channel with relay (SU can use part of their power for secondary transmissions and the remainder of the power to relay the primary transmissions: it is a symbiotic approach). In interference overlay, SU must know the codes used by the PU to mitigate interference (dirty paper coding, DPC). *Interweave* is the technique just based on the opportunistic use of frequency holes.

Spectrum handoff happens when the specific portion of the spectrum in use for SU is required by a PU. The SU communication has to be moved to another vacant portion of the spectrum.

Apart from the abovementioned proposals considering the use of CR in both LTE-Advanced and IEEE 802.16m, other groups are also working to standardize this technology. This is the case of the IEEE 1900 group, composed of several subgroups: 1900.1 (standardize definitions and terminology related to CR), 1900.2 (testing and verifying the operation of CRs), 1900.3 (conformance evaluation of software modules), 1900.4 (optimized radio resource usage), 1900.5 (architectures for managing CR for dynamic spectrum access), and 1900.6 (spectrum sensing interfaces and data structures).

Also, in the IEEE 802.22 standard (previously introduced in this chapter), some CR techniques are under consideration to provide high bandwidth access in rural environments by using white spaces in the TV spectrum.

Other standards, not exactly formulated as CR, show portions of their protocols close to CR concepts. This is the case of the standard IEEE 802.11h (coexistence radar and

satellite) where, from an estimation of channel characteristics, the change in operation frequency is decided (dynamic frequency selection) and the transmit power is adjusted. Also IEEE 802.16j (relay radios), although not exactly cognitive, offers some CR characteristics. Anyway, CR seems to be the new paradigm for mobile communications developed in this new decade, and apart from standardization bodies, some research groups worldwide are active in promoting this technology [11].

3.4 DFE introductory design considerations

The continuous progress in electron device technologies is providing improved performance in terms of greater bandwidth, computational speed, reduced dimensions, energy consumption, and manufacturing cost, and it has also opened up in the field of wireless communication and broadcasting systems a wider range of applications for high-complexity integrated circuits. At the same time, the present worldwide diversification and fragmentation of markets is causing more cost in product development, as similar designs have to be performed, validated, and supplied many times, and in interoperability, as several similar products for different standards have to be made to work together. The higher cost of such developments and the multiple evolving standards are therefore increasing the risk in the traditional hardware-based design approach. Hence, there is a strong demand to achieve a higher flexibility in products than before that leads system designers to a new strategy and technical approach. The concept of implementing multiple standards in software instead of hardware can be developed with SDR techniques. A typical case is the requirement for reduced dimensions and costs of the MTs that have to allow interoperability among different standards of cellular communications, WLAN applications (i.e., 802.11), and WPAN (i.e., Bluetooth or Zigbee) and even to allow GPS location or mobile broadcasting reception. Besides, an additional challenge for MTs' design is their worldwide commercialization, so they have to cope with the needs of different and intercultural consumers at the same time.

A programmable RF transceiver, provided it is sufficiently frequency-agile, can simply be configured to operate in any frequency band, facilitating fast, reliable, and economical deployment of wireless communication and broadcasting networks. Later on, the same transceiver implementation can be used in different product ranges and geographical locations, without the need to develop new RF designs. So if a new frequency band appears, or OEMs wish to move into a different global market, the RF transceiver can be simply reconfigured accordingly. In this way, design reuse should be maximized, shortening the production cycle and allowing the rapid movement into new and emerging markets. However, if the redesign complexity is higher than the simple adaptation to a new frequency band, as may be the case in increased complexity because of the incorporation of new standards, a significant reuse of already developed blocks become more difficult. Hence it is not possible to search for a single solution taking a huge advantage from previous industrial development. In this way, SDR maximizes the reuse of the existing hardware.

The architecture of the MTs is critical [47] because of the need to achieve reduced costs in a high volume market, usually affected by cyclic business (e.g. development cycles have to fit fairs or Christmas sales). Besides, the inclusion of results from fast innovation cycles (upgradeable features) produces a compromise between the technological complexity (long development cycles) and the small time-to-market windows, along with the user requirements of small size, weight, and volume, as well as low power consumption (battery powered). Different structures are currently used for wireless communication transceivers, the most usual ones being shown in Figure 3.20.

Because of their low size and cost, the direct conversion and the IF-transceiver are recognized as the most suitable structures for UE developments. For base-stations, the structures are more diverse. A primary aspect to be considered in transceiver design is the use, off the shelf, of some chipset providing OEMs with tools for an efficient design solution. Radio Frequency Integrated Circuit (RFIC) may also be a flexible radio chip [10], with the RF part integrated together with digital parts of the radio in a single SoC (system on chip).

Some analog subsystems are also necessary if the transmit frequency is too high for cost-competitive DAC operation (Figure 3.20). The state of the art for inexpensive DAC converters is 10–14 bit devices, with SNR higher than 60–80 dB and sampling rates ranging from 100 to 400 Msps. For high bandwidth FPGAs some expensive ADC and DAC cards are specified for Gsps operation; DACs up to 800 Msps, 16 bits, may cost a few tens of dollars in large piece quantities. Even some DACs are really a SoC, including oversampling and interpolation filters, clocks and PLL. An example, among others, is in [7] where a 14-bit DAC is presented whose digital interface can support input data rates as high as 160 Msps.

The subsystems typically implemented in analog electronics designs where the DAC cost is prohibitive are analog mixers and oscillators, as well as analog tunable filters (although the tunable capability may be done by digital devices), detectors and, often, I&Q modulators. Current high-frequency voltage controlled oscillators (VCOs) are designed in analog technology. A common solution in the multiband case is the use of the VCO inside a crystal controlled PPL loop, where additional to the analog tuning voltage from the PLL the VCO may include a configurable resonator bank, controlled by a CMOS digitally switched capacitor matrix [43]. Regarding Figure 3.20, some subsystems may be bought as RFICs. Examples are the set composed by the driver and the PA (for UE transmitting at medium/small powers, and often including the temperature sensor), sets of mixers and quadrature modulators (including PLL), and programmable attenuators for power control.

If successful, SDR designs must have a significant utility in several communication and broadcasting services, with a capability to serve a large variety of changing radio protocols in real time. To achieve that, typical transceiver functions traditionally implemented in hardware, could instead be implemented by means of software programs on embedded computing devices. This concept is not new, but only recently, with the rapidly evolving capability of digital electronics, has it been made practical with many implemented processes previously only theoretically possible. A challenge in project planning is to decide what parts of the product under development have to be developed either in

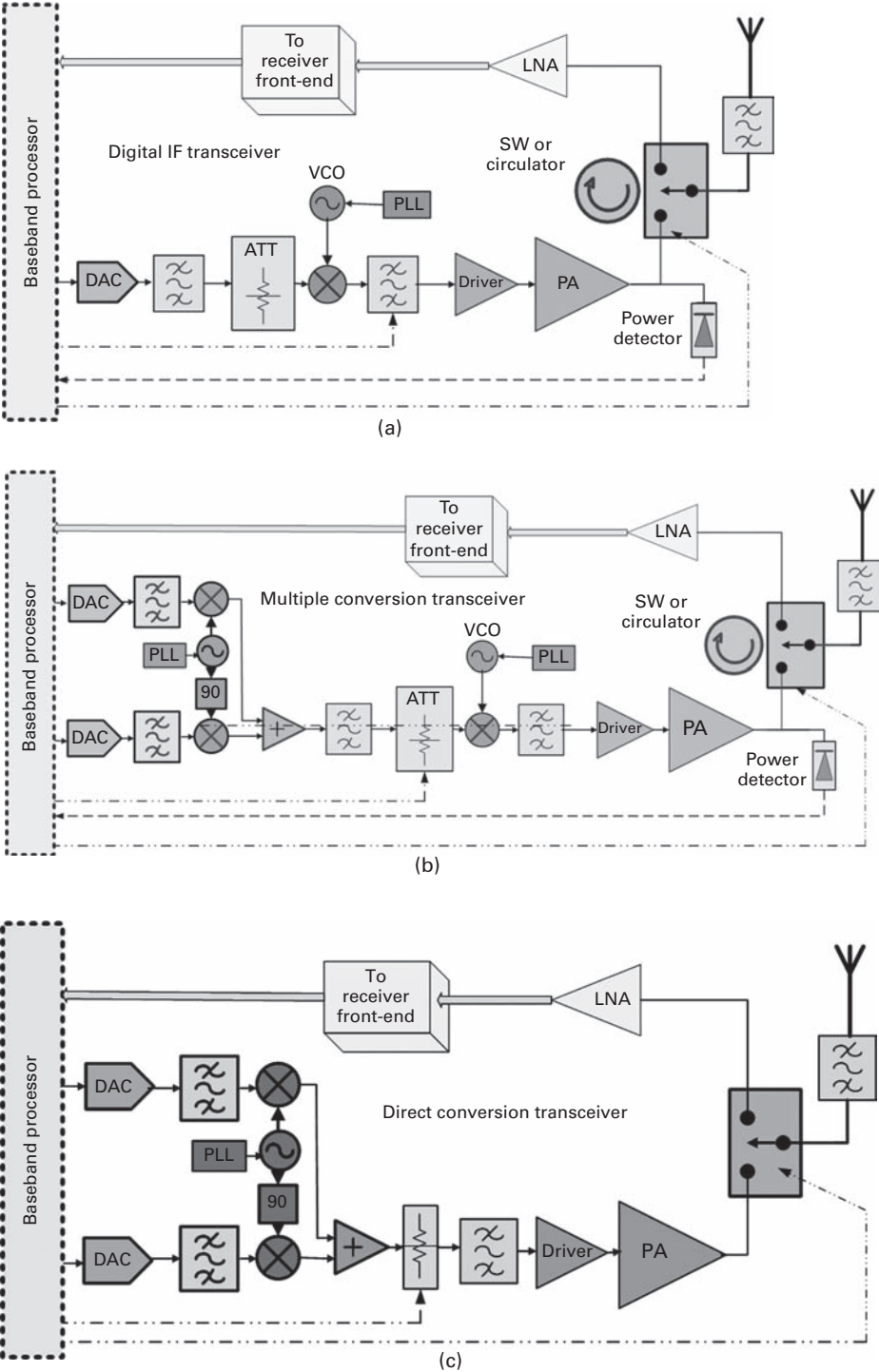


Figure 3.20 Block diagrams of some architectures for a wireless communication transceiver.

digital or in analog technologies. According to [47] digital implies long concept engineering and long verification periods leading to comparatively slow tape-out cycles. On the other hand, analog requires less pre-silicon verification, but a post-silicon centering and optimization is needed with comparatively fast tape-out cycles. A trade-off appears between the implementation of some subsystems in analog (or mixed analog–digital) ASIC, or to use SDR techniques. It is not easy to solve, moreover attending the constant evolution in cost and performance of ADC and DAC, mainly in sampling rate, DSP devices and programmable FPGAs, together with fast memory (flash) devices. Other subsystems, such as wideband tunable analog RF transceiver front-ends are nowadays implemented in RFIC devices. Anyway, important constraints for ideal SDR are the need for analog subsystems such as LNA and anti-aliasing filters, as well as the PAs, power duplexers or circulators, analog switches and antenna filters.

The software vision can be based on few simple foundations:

- to realize as much as possible component functionality by software
- to implement signal processing and protocol specifications for many standards and bandwidths
- to switch between different systems
- to implement new system specifications easily in the memory device of the terminal.

Wireless systems are continuously advancing, with new high-speed standards driving higher data bandwidths and spectral efficiency. The demand for ever higher bandwidths and spectral efficiency presents the need for more complex signal processing at higher speed and lower power consumption. To meet the tough requirements of modern wireless systems and to address typical needs of software defined radio communication technology, a versatile programmable hardware platform is required, typically containing a general purpose processor used as applications processors, complemented by highly efficient programmable DSP or FPGAs, RF modules, antenna, and a rich set of I/O devices.

In such a context, while the implementation with DSP techniques of SDR algorithms provides both high flexibility, and accurate and reliable operations with fault-tolerant SW/HW design, the main wireless system performances, like power consumption, signal-to-noise ratio, interference immunity, nonlinear distortion, bit error rate and dynamic range, are still dependent mainly on the characteristics of the electronic circuits (i.e. low-noise amplifiers, voltage controlled oscillators, and frequency synthesizers, PAs, ADC, and DAC) in transceiver front-ends. Further on, in order to be efficient and effective, an extremely important role is played by the design of a software-enabled, RF/DSP co-designed PA which is suitable for linear and highly efficient transmitters with capability of multi-standard and multiband applications.

Challenges for operation at higher data rate are twofold. On the one hand, new standards are including new modulation, coding and access schemes, flexible spectrum usage and smaller cells, optimum FEC, MIMO operation, and advanced antenna systems. On the other hand, in SDR user terminal designs, it is desired to put, at reduced costs, the ADC and DAC converters as close to the antenna as possible. A smart selection of the required ADC and DAC, regarding the necessary sample rates for every standard

and searching for a common basic rate for all the standards to be included in the same equipment, surely is a first step that enables better SDR designs. However the PA remains the most critical component of the transceiver and it has to be considered together with the whole system architecture, in order to ensure optimal system performance in terms of power efficiency, linearity and cost.

The development of low cost flexible hardware solutions for SDR surely remains a challenging task, but in the near future, software defined radios are expected to become a dominant technology in radio communications and along with software defined antennas they are to be considered as the enablers of the CR.

3.4.1 Transceiver issues associated with software defined radio

Some transmitter subsystems, such as source coders, modulators, channel coders, scramblers, interleavers, mappers, or FFT blocks, are well suited to be implemented in digital devices, as the design parameters are already detailed in the respective standards. Other components of the DUC chain, such as the interpolation filters (i.e., halfband, polyphase, or CIC filters), pulse-shaping filters (i.e., RRC) or channelization filters (to meet the usually stringent ACLR specification) are also implemented in digital. Furthermore, additional subsystems, such as gain controllers (for power control), digital signal synthesizers (DSS), I&Q, DPDs for PA linearity improvements, or crest factor reducers are also implemented in the digital section. However, the main problem in reaching the ideal scheme for a SDR transmitter is the difficulty of conversion between the digital and the analog domains at a high enough rate and a high enough accuracy at the same time. Even the analog RF components of SDR have special requirements if multiple radio systems are to be managed by common hardware.

In [previous sections](#) possible different sampling rates have been presented for the different standards. In UMTS and LTE, the use of sample rates that are multiples of the duration of the UMTS chip rate, 3.84 Mcps, should facilitate common implementations for both standards. Digital-to-analog converters capable of operating over a minimum sample rate of 130 Msps seem adequate for UMTS, LTE, and Mobile WiMAX, a value that could be reduced to 70 MHz if the oversampling ratio is relaxed. Also, some kind of pulse shaping and interpolation filters have been also introduced in [previous sections](#). Half-band and CIC filters are time-efficient solutions as interpolation filters because they do not use multipliers, so the computational burden is reduced. Polyphase filters are also efficient because they split an FIR filter into sub-sets of lower complexity FIR filters.

In the SDR transceiver, a circulator could be used to separate transmit and receive path signals. However, maximum isolations are typically of the order of 20 dB and circulators are very frequency sensitive devices, so the broadband impedance matching between the circulator, the antenna, and the PA is bandwidth limited in practice. Other alternatives are filter-based solutions (i.e., a conventional diplexer in FDD modes, or a duplexer in TDD), but the challenge of designing programmable analog bandpass filters remains to be fully resolved. In place of a restrictive analog filter for the duplexer in multiband systems different solutions can be used:

- *Transmitter/Receiver switch*: with this solution transmitting and receiving at exactly the same time instant is prevented by switching the connection to the antenna at distinct time slots to achieve full TDD operation. The switches can be made very broadband and have the advantage of avoiding filtering, placing no restriction on the frequency split. Low-cost switches may be implemented using PIN-diodes, although transmit–receive isolation or switching transient time may become an issue. Additionally, PIN diodes are nonlinear, have low capacity for handling power and require bias control lines. A new alternative is the use of MEMS (micro-electromechanical-system) switches, with higher isolation and operation frequencies, even though the switching time is of the order of μs .

These considerations also apply in multimode PAs, when the power transistors are used for multiband transmission, where it is necessary to match the device output to the antenna impedance with different matching networks according to the actual frequency band. In this case also power varicaps or varactors (i.e., to take advantage of the parasitic capacity of a power MOS device) may be considered.

- *Cancellation techniques*: such techniques, that remove the transmitter signal from the receive path, are still an area of research and are complicated by external reflections from the antenna back into the system.

Apart from the aforementioned use of circulators to separate transmit and receive path signals, in base-stations circulators are also placed between the PA and the antenna, in order to prevent reflecting waves (VSWR) capable of destroying the expensive PAs.

3.4.2 Power amplifier issues

Apart from its advantages concerning multipath, OFDM techniques used in Mobile WiMAX and LTE also have some drawbacks. One, not directly related to the PA, is due to the insertion of CP to face channel multipath and ISI, which produces an efficiency loss. Hence, it is desirable to have a relative reduction of the CP by transmitting long symbols, which requires a tight spacing of the OFDM subcarriers. But, the closer the subcarriers are, the more probable are the orthogonality failures (a carrier must be as independent as possible from another carrier) and frequency errors. Hence, a drawback in OFDM is the higher sensitivity to frequency offsets and phase noise. Another drawback related to the PA is the high PAPR value produced by both OFDM and CDMA access techniques. Related to PAPR (based on *power* measurements) the crest factor parameter (CF, based on signal *amplitude* measurements) is defined as:

$$CF = \sqrt{PAPR}$$

being a general approximation in an OFDM system with N subcarriers:

$$PAPR \leq 10 \cdot \log_{10} N \text{ (dB)}.$$

Hence, the increase of the number of OFDM subcarriers has a sensitive effect on signal PAPR that may reach theoretical values up to 15 dB (12 dB in systems with 256 subcarriers).

Regarding UMTS, in a band slice of 5 MHz, a number of simultaneous users have to be allocated. This number depends on the processing gain (or spreading factor) and the minimum SIR (signal-to-interference ratio) to decode the user data. While the chip-time is fixed, the bit-time has flexible rates, so different SF values are possible (from 4 to 512, as has been introduced in [previous sections](#)). In the base-station, when different categories of services are provided in the same area, or when the number of users is too high to be allocated in one band, each service category or group of users is assigned to a different carrier, thus resulting in the multi-carrier WCDMA. Typical figures for the PAPR are around 8 dB for a single carrier, and around 12–15 dB for several carriers.¹

The high PAPR figures are not only critical for PAs but also for ADC and DAC dynamic margins. A high PAPR requires the use of converters with a high dynamic range, thus preserving some quantification levels as statistically less used (just for signal peaks) and so increasing the quantification noise in the ADC converters from the theoretical SNR values.

In constant envelope modulations used in former wireless mobile communications (i.e., GMSK used in GSM), it was relatively easy to bias the PA at a fixed back-off level allowing all the dynamic margin of the signal to be seized while keeping controlled the power consumption. Even the use of class-C amplifiers or some kind of constant-amplitude power efficient switched PA, such as class E or F, was possible. However, new access techniques (access techniques are most critical for PAPR than the previous data modulation formats [9]) have produced an increase in PAPR values, so it is not possible to bias the PA in an energy-efficient back-off level. At reduced back-off (a power efficient solution) the transmit signal peaks will be distorted (or clipped). Conversely, if the PA bias point allows correct transmission of the signal peaks, then the power efficiency significantly shrinks.

High-performance PAs are expensive and typically account for over 35% of the cost of a MT, reaching up to 70% in some designs for mobile handsets, at power added efficiencies (PAE) of the order of 50% and less [12]. Typical efficiencies with 3G air interfaces on class-AB designed LDMOS PAs are in the range of 7 to 20%, and can reach around 40% if a *Doherty* structure is used.

Apart from power efficiency, the intrinsic PA nonlinear behavior generates IMD (intermodulation distortion) products, which produce out-of-band distortion. This distortion, when faced with multilevel modulated signals, produces spectral regrowth on adjacent channels, thus shrinking the mandatory ACLR. Besides, the nonlinear behaviour produces in-band distortion, that affects the transmit signal constellation quality, measured as the EVM. It is convenient to have a linear PA for reducing out-of-band

¹ These PAPR values, as well as the maximum values for OFDM, correspond to extreme situations with reduced probability and considering all the frame contents: preambles, headers, and data payload. Actually, PAPR values depend on the CCDF (complementary cumulative distribution function), which denotes the PAPR probability to surpass a given threshold. For probabilities between 10^{-3} and 10^{-4} , the PAPR typically ranges 8–10 dB.

emission and EVM, and to keep the desired data throughput. On the other hand, power efficiency affects the lifetime of batteries, the cost of the energy in base-stations, and the reduction of heat exchangers' requirements (including the size of the heat sinks for power transistors and the power consumption of the air conditioners in base-stations). Using advanced digital linearization algorithms, the power efficiency can be increased at or beyond 15%, and the adjacent channel interference rejection may be improved by 15–35 dBc.

Power amplifier linearizing structures are being employed to overcome the compromise between linearity and power efficiency [24]. There are many types that can be applied at different levels: device, circuit, or system. At device level, the strategies are based on the cancellation of nonlinearities from transconductance compensation, harmonic terminations, active bias, or thermal compensation. At circuit level the strategies are based on either classical feedback or analog predistortion. Classical circuital feedback has to equilibrate loop gain with system bandwidth in order to keep the stability. So it works in a design compromise between the operative bandwidth and the in-band loss of gain. The analog predistortion technique is simple and unconditionally stable, and it is well suited for simple weak nonlinearities. Each design is tailored for a specific power transistor. Diodes or low power transistors are used to generate IMD terms which, once phase and amplitude compensated, are combined with the PA input signal to cancel some nonlinear products at the output. This structure is very sensitive to these gain and phase adjustments.

At system level, there are two kind of linearizers: those aiming at avoiding distortion and those aiming at reducing distortion [55]. The first include the *LINC* linearizer, suitable for highly nonlinear (switched) amplifiers. Its implementation needs a DSP device to generate the LINC signal components, and requires two equilibrated branches, each one with a balanced RF PA. The pitfalls are the possible mismatches between gain and phase in both branches. CALLUM structure appears as a solution to compensate for gain and phase maladjustments in LINC amplifiers, but presents risks of instability (due to phase shifts in feedback loops). Currently, digital technologies allow the use of adaptive schemes to compensate LINC imbalances.

The EE&R (Envelope Elimination and Restoration), or Khan amplifier, suitable for analog transmitters, has drawbacks in both the time-constant and the sensitivity of the envelope detector, apart from the need for a sufficient bandwidth and efficiency in the amplitude amplifier used to drive the transistor drain (up to 20 MHz in some new standards). An alternative of the EE&R, suitable for SDR systems where the modulated signal is digitally generated, is the *Polar Transmitter* (Figure 3.21). In Polar Transmitters [17][27] the baseband amplitude and phase components of the complex signal are generated in separated ways and independently treated until the last stage, where the modulated envelope and phase signals are combined and amplified by some efficient switched-mode PA: the envelope adjusting the drain voltage (so this structure is also called drain-modulation) and the phase ($\Phi(t)$, constant amplitude) signal driving the PA input. This structure requires the bandwidth of the envelope to be low enough to be real-time processed by power converters, or the use of hybrid solutions which are the only possible candidates nowadays for new standards appliance.

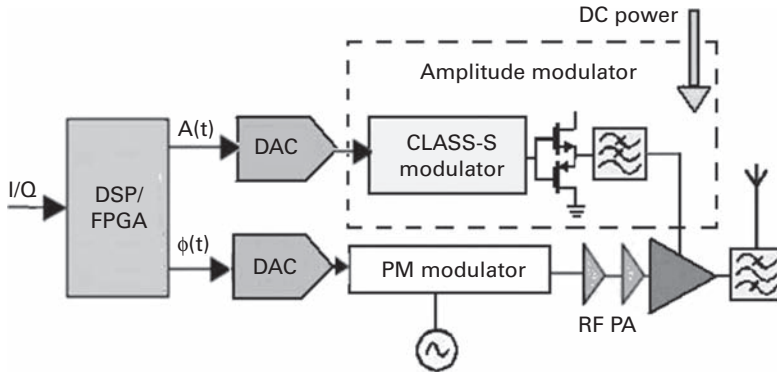


Figure 3.21 System model of a polar transmitter.

Also trying to avoid distortion are the system level *predistorters*, whose basic idea is to reproduce the inverse PA nonlinear behavior, in a kind of nonlinear equalization, so having linear amplification at the output of the PA. A DPD implemented at baseband or at IF frequencies, may be static (or memoryless, just compensating for static AM-AM and AM-PM curves in the amplifier) or also include memory effects. Digital predistorter performance is sensitive to this inclusion of the memory effects, especially in broadband amplifiers, as will be explained in several other chapters of this book. Digital predistorter, together with LINC and Polar Transmitters, are linearizing schemes requiring some degree of digital signal processing, and so, suitable for SDR if their implementation doesn't suppose an excessive computational burden on DSP or FPGA devices leading to an excessive increase in the sampling period.

Among linearizers aiming at reducing distortion there are two main strategies: Feedback and Feed-Forward. Feedback linearizers have some variants: RF, envelope, Cartesian, or polar. Cartesian feedback avoids some adjustment problems of other feedback structures, but it has instability risks. Hence it is just a solution for narrow-band systems and consequently not appropriate for modern wireless communication systems. The other strategy is Feed-Forward; a stable and efficient technique for wide-bandwidth systems, but its open-loop nature makes it too sensitive to delay mismatches and component tolerances, which produce imbalances among the different branches. In practice, the adjustments are difficult problems for mass production and, besides, it involves the cost of an auxiliary amplifier.

Most of practical linearizers but the non-adaptive DPD are more expensive than the strategies based on crest factor reduction, which are solutions that also reduce the problem of the PA biasing, and can be totally software implemented. However, linearizers have the advantage of not reducing the data-throughput. Crest factor reduction techniques [18] work well for reducing the back-off in the PA, however some techniques produce signal distortion (i.e., clipping), thus increasing the BER and relying on a further error correction procedure (from channel coders) that, even though keeping the data-rate, will reduce the data-throughput. Data-throughput diminution may also appear in both the

techniques that need some additional information to be sent to the receiver and the ones based on enlarged codewords.

An additional challenge in PA design is the inclusion in new standards of the transient mask specification, in addition to the usual spectrum mask. Nowadays, to cope with both masks requires a design procedure including some trial-and-error methodologies.

3.4.3 Technology aspects

Over the last few years, wireless communications transceivers have gone through an intense development process. The advent of inexpensive silicon microelectronics technology to implement RF circuitry at the speed of the Moore's law has driven the transceiver design to continuously increase its level of integration, thanks to the great potential of continuous improved processing power and cost advantages. The demonstration a few years ago that RF circuits such as LNAs, mixers, and VCOs could actually be designed for manufacturability in a relatively inexpensive CMOS or SiGe process has led to a tremendous level of integration in a single silicon IC and the system integration has progressed from chip on board to system in a package (SiP), to SoC.

In fact, programmable silicon is becoming increasingly flexible whilst remaining cost competitive. The ultimate goal is a single bill of materials for a reconfigurable product that the manufacturer configures to operate in the desired standard and frequency band. The use of just a single type of RF transceiver and baseband ICs for a whole range of products results in economies of scale by capturing larger volumes and provides also significant cost benefits associated with simplifying manufacturers' inventory.

SiGe (silicon germanium) bipolar transistor technology which offers high transit frequencies and excellent $1/f$ -noise behavior is an attractive solution for the receiver front-end, while SiGe BiCMOS technology is a good candidate technology for more advanced IC solutions, as it also allows the integration of digital functionality. Therefore, SiGe oscillators or even complete synthesizers with excellent phase noise performance can be developed with appropriate architecture definition, through innovative design and cost-efficient technology selection. Better still, the continuous reduction in CMOS feature size, from 0.5 microns a few years ago to 0.35, 0.25, 0.18, 0.13, and now 0.09 and 0.065 microns, has meant smaller wireless ICs that consume less power than their predecessors and are designed into ever-smaller handheld devices with extended battery lives.

Current advances in CMOS technology, that is currently capable of reaching up to 50 GHz in 0.18 μm CMOS, have recovered the suitability of this technology for high-frequency RFICs design. According to [30], realizing some digital signal processing functionalities on the RFIC in an advanced CMOS technology has advantages over pure analog RFICs in terms of production stability, power consumption, and cost for mass market applications. In this way, some prototypes of flexible RF transceiver in 40-nm, low-power digital-CMOS technology have been already presented [10]. Recent advances in silicon developments and component architectures have also enabled the implementation of tunable blocks which allow the possibility of a better optimization of the front-end designs.

Integrated analog circuitry is not as scalable as digital circuitry is. Hence, the ability to save cost and semiconductor design area is more difficult and requires designers to be updated in semiconductor technologies and components. More advanced commercial systems with high demands in operating frequency, output power, noise figure, and linearity need more powerful, multi-function monolithic microwave integrated circuit solutions (MMIC). In order to meet their performance requirements, radio designers must use GaAs, InP, GaN, SiGe and other heterostructure compound-semiconductor components which allow them to meet the stringent commercial goals for cost, size and time to market. Modern circuit designs must find new ways to overcome these demands, as opposed to implementations using more expensive semiconductor materials like GaAs or discrete realizations using “off-the-shelf” components. The circuit design process must begin with both the systems architecture and the technology choice to achieve efficient, cost-effective results.

Advances in new wide bandgap materials for PA design, such as SiC (silicon carbide) and GaN (gallium nitride), together with *sapphire* substrates, have allowed the first commercialization of SiC MESFET and GaN HEMT transistors, capable of operating at high breakdown voltages. Gallium nitride HEMTs offer efficiency (up to 60%), bandwidth, and power advantages compared to Si LDMOS FETs, even though their higher power density supposes a difficulty in removing the heat from the power transistor. Silicon carbide has higher thermal conductivity than GaN or Si and hence SiC devices can operate at higher temperatures and power densities, although its lower electron mobility reduces the operation bandwidth, and currently only a few samples are enough for 3G application. So its main use at microwave frequencies is as a substrate for GaN (GaN on SiC HEMTs).

3.4.4 Digital processing devices: DSP and FPGA

Digital devices enabling digital processing are scarcely diverse. Mainly, they are digital signal processors (DSPs) and field-programmable gate arrays (FPGAs). Appearing in the early 1980s, DSP devices were developed to improve the processing speed of general purpose microprocessors (μ P). Instead of the Von Newman structure used in μ Ps, DSPs are based on the Harvard architecture, so sacrificing some capacity of memory addressing in benefit of the computational speed. The inclusion of the MAC (multiply and accumulate) instruction was a key architectural innovation boosting its use for real-time signal processing. From 1980, DSPs have significantly changed the operation capacity (in MIPS, million instructions per second), as well as the usability of the development tools (C-compilers, assembly optimizers, debugger interfaces, or even interfaces with system level simulation programs). Texas Instruments estimates the DSP needs 200 MIPS to support development of the high-bandwidth 3G standards [31].

Field programmable gate arrays are basically chips containing arrays of configurable logic blocks, which can be programmed to perform different digital functions. They were first introduced in the mid 80s, being originally a low-cost, low-size alternative for combinational, and later sequential, logic circuits. They strongly reduced the size of the circuitry for address decoding and memory access in μ C developments. Later, FPGAs evolved including some mathematical functions (as dedicated multipliers, multiply

accumulators, or multiply-adder blocks) and improving memory access capacity. Nowadays, FPGAs are starting to displace ASICs in some traditional application areas; FPGA-derived ASICs make possible single-chip solutions in challenging applications because of the size or reliability issues. Compared with FPGAs, ASICs are neither flexible nor upgradeable; however, when its use is possible, they have advantages in power consumption, cost and bandwidth.

The Xilinx Virtex families include look-up tables (LUT), control logic, and arithmetic support for high-speed multipliers, configurable storage elements, and I/O capability. The Virtex-6 family can manage up to 566 784 logic cells and 6 370 kb of RAM memory. Virtex 6 devices have between 8 to 72 gigabit transceiver circuits, each transceiver being a combined transmitter and receiver capable of up to 6.6 Gbps or 11.18 Gbps, depending on the kind of transceiver. The main competitors of the Virtex families are the Stratix families from Altera.

In DSPs the execution of the algorithms runs in a serial way. Conversely, in FPGA the hardware supports pre-determined signal processing functions, using separate hardware circuits for different operations. This allows some parallelization or pipelining of the operations. Digital signal processors are more suitable than FPGAs in complex applications while FPGAs are prevalent at high data rates and low-complexity algorithms. Even their I/O flexibility may be an advantage in some applications. Field-programmable gate arrays devices are suitable for simple algorithms requiring a high computational rate for signal processing (i.e. FIR filtering or digital predistortion based on the use of LUT tables containing the predistortion values). However, FPGAs are more expensive. An FPGA design may cost an estimated factor of two to ten times more than a DSP to tackle a given problem. Regarding power consumption, FPGA have historically shown disadvantages, nevertheless this fact is shrinking in low-power FPGAs.

Digital signal processors may operate at fixed or floating point. Floating-point processors run algorithms slower than fixed-point processors, so they have to be seriously assessed as an option for high-performance equipment. On the other hand, floating point is better for operations requiring high dynamic range and precision. Typical functions implemented in floating point could be FFTs, matrix multiplication, or matrix inversion. In particular matrix inversion is very susceptible to arithmetic precision (because of the risk of dividing by zero), causing 16 and even 32 bit fixed-point operations to suffer in performance or simply not work well [16]. Another advantage of floating point DSPs is for IIR (infinite impulse response) filtering, especially in frequency selective filters with high Q factors: in this situation the filter coefficient truncations or rounding because of fixed point arithmetic may produce stability concerns. Roughly speaking, fixed point is preferable to increase DSP speed, while floating point is the best option for high dynamic ranges or for complicated arithmetic.

Finally, just to mention that both DSP and FPGA based designs are supported for real-time operating systems (RTOS) that manage real-time multi-tasks, which is necessary to handle different and simultaneous operations in the equipment. This implies all kind of functionalities, covering from the user manipulation of the MT to the automatic management of various communication protocols.

3.5 Conclusions

This chapter has presented a system overview in wireless communications attending the design of their DFE. Hence, the new and emergent wireless communications standards have been introduced, including UMTS, Mobile WiMAX, LTE, and some IMT-Advanced standards, as well as some others including some kind of cognitive radio capabilities. The fundamentals of the main access techniques for new standards, CDMA and OFDMA, have been revisited, and most of the key parameters regarding the SDR designer point of view have been presented for the main wireless communications standards. Actual engineering practice in wireless communications requires designers to be familiar with a lot of abbreviations, a fact that might be troublesome to some readers having a first contact with the wireless communications field. However, we have preferred a realistic approach and do not hide the use of the principal abbreviations.

Regarding technology aspects, some design considerations and issues have been presented regarding the DFE design. In this way, critical issues from analog subsystems, such as switches, filters or circulators, have been discussed. Moreover, the PA and associated design trade-off between linearity and power efficiency has been also discussed, and an overview of the linearizing structures has been presented. At the silicon level, some comments on technology aspects related to advances on silicon microelectronics have been made, and, finally, the main digital processing devices, DSP and FPGA, have been presented and commented on.

List of abbreviations (Chapter 3)

3G:	Third Generation (cellular telephony)
3GPP:	3rd Generation Partnership Project
4G:	Fourth Generation (cellular telephony)
AAS:	Adaptive Antenna System
ACLR:	Adjacent Channel Leakage Ratio
ACS:	Adjacent Channel Selectivity
ADC:	Analog-to-Digital Converter
ASIC:	Application-Specific Integrated Circuit
ASP:	Analog Signal Processing
ASSP:	Application-Specific Standard Processor
BCCH:	Broadcast Control Channel
BER:	Bit Error Rate
BiCMOS:	Bipolar-CMOS
BS:	Base-Station
BTS:	Base Transceiver Station
CCK:	Complementary Code Keying
CDMA:	Code Division Multiple Access

CF:	Crest Factor
CFR:	Channel Filter
CIC:	Cascaded Integrator Comb (filter)
CMOS:	Complementary Metal-Oxide Semiconductor
CN:	Core Network
CP:	Cyclic Prefix
CR:	Cognitive Radio
DAB:	Digital Audio Broadcasting
DAC:	Digital-to-Analog Converter
DCS:	Digital Cellular Service
DDC:	Digital Down-Converter
DDS:	Direct Digital Synthesizer
DFE:	Digital Front End
DFT:	Discrete Fourier Transform
DIF:	Digital Interpolation Filter
DL:	Downlink
DMB-T:	Digital Multimedia Broadcasting – Terrestrial
DPC:	Dirty Paper Coding
DPCH:	Dedicated Physical Channel
DPD:	Digital Predistorter
DSP:	Digital Signal Processor
DSS:	Digital Signal Synthesizer
DUC:	Digital Up-Converter
DVB-H:	Digital Video Broadcasting Handheld
E3G:	Evolved 3G
EE&R:	Envelope Elimination and Restoration
eNB:	evolved Node B
ENOB:	Effective Number of Bits
ETSI:	European Telecommunications Standards Institute
eUTRA:	Enhanced UTRA
EV-DO:	Evolution Data Optimized
EVM:	Error Vector Magnitude
FCCH:	Frequency Correction Channel
FDD:	Frequency Division Duplexing
FDMA:	Frequency Division Multiple Access
FEC:	Forward Error Correction
FIR:	Finite Impulse Response (filter)
FM:	Frequency Modulation
FPGA:	Field Programmable Gate Array
GaN:	Gallium Nitride (Semiconductor)
GERAN:	GSM/EDGE Radio Access Network
GMSK:	Gaussian Minimum Shift Keying (modulation)
GPRS:	General Packet Radio Service

GPS:	Global Positioning System
GSM:	Groupe Special Mobile
HB:	Halfband (filter)
HEMT:	High Electron Mobility Transistor
H-FDD:	Half Duplex-Frequency Division Duplexing
HRPD:	High Rate Packet Data
HSDPA:	High Speed Downlink Packet Access
HSUPA:	High Speed Uplink Packet Access
I&Q:	In-phase and Quadrature components
IC:	Integrated Circuit
ICI:	Interchannel Interference
IDFT:	Inverse Discrete Fourier Transform
IEEE:	Institute of Electrical and Electronics Engineers
IF:	Intermediate Frequency
IMD:	Intermodulation Distortion
IMR:	Image Rejection Mixer
IMT:	International Mobile Telecommunications
ISDB-T:	Integrated Services Digital Broadcasting – Terrestrial
ISI:	Inter-Symbol Interference
ITU:	International Telecommunications Union
LAN:	Local Area Network
LDMOS:	Lateral Diffused MOS (transistor)
LNA:	Low-Noise Amplifier
LO:	Local Oscillator
LTE:	Long Term Evolution
LUT:	Look Up Table
MAN:	Metropolitan Area Network
MEMS:	Micro-Electro-Mechanical System
MESFET:	Metal-Semiconductor-Field-Effect-Transistor
MIMO:	Multiple-Input Multiple-Output
MIPS:	Million Instructions per Second
MMIC:	Monolithic Microwave Integrated Circuit
MOS:	Metal Oxide Semiconductor
MOSFET:	Metal Oxide Semiconductor Field Effect Transistor
MS:	Mobile Station
MT:	Mobile Terminal
OEM:	Original Equipment Manufacturer
OFDM:	Orthogonal Frequency Division Multiplexing
OFDMA:	Orthogonal Frequency Division Multiple Access
OSR:	Oversampling Ratio
OVSF:	Orthogonal Variable Spreading Factor
PA:	Power Amplifier
PAPR:	Peak-to-Average Power Ratio

PGA:	Programmable Gain Amplifiers
PHY:	Physical (layer)
PLL:	Phase-Locked Loop
PR:	Pseudo-Random (sequence)
PU:	Primary User
QAM:	Quadrature Amplitude Modulation
QoS:	Quality of Service
QPSK:	Quadrature Phase Shift Keying
RAT:	Radio Access Technology
RB:	Resource Block
RE:	Resource Element
RFIC:	Radio Frequency Integrated Circuit
RNC:	Radio Network Controller
RNS:	Radio Network System
RRC:	Root Raised Cosine (filter)
RTOS:	Real-Time Operating System
SAW:	Surface acoustic wave (filter)
SC-FDMA:	Single Carrier Frequency Division Multiple Access
SCH:	Synchronization Channel
SDMA:	Spatial Division Multiple Access
SDR:	Software Defined Radio
SER:	Symbol Error Rate
SF:	Spreading Factor
SFDR:	Spurious Free Dynamic Range
SFN:	Single Frequency Network
SiGe:	Silicon-Germanium (Semiconductor)
SiP:	System in a Package
SNR:	Signal-to-Noise Ratio
SoC:	System on Chip
SR:	Software Radio
SU:	Secondary User
TDD:	Time Division Duplexing
TD-LTE (or LTE TDD):	Long Term Evolution-Time Division Duplex
TDMA:	Time Division Multiple Access
UE:	User Equipment
UL:	Uplink
UMB:	Ultra Mobile Broadband
UMTS:	Universal Mobile Telecommunications System
UTRA:	Universal Terrestrial Radio Access
UTRAN:	UMTS Radio Access Network
UWB:	Ultra-Wideband
VCO:	Voltage Controlled Oscillator
VLSI:	Very Large Scale Integration

VoIP:	Voice over IP
VSWR:	Voltage Standing Wave Ratio
WCDMA:	Wireless CDMA
WiMAX:	Worldwide Interoperability for Microwave Access (Industry association)
WirelessMAN (WMAN):	Wireless Metropolitan Area Networks
WLAN:	Wireless Local Area Network
WPAN:	Wireless Personal Area Networks)
$\Delta\Sigma$:	Delta-Sigma converter

References

- [1] 3GPP Technical Specification Group, "TS 25 – Series Specifications. Radio Access Network: UTRA." July 2010
- [2] 3GPP TR 36.804 v0.7.1 (2007–10), 3rd Generation Partnership Project, Technical Specification Group Radio Access Network, "Evolved Universal Terrestrial Radio Access (E-UTRA); Base-station (BS) Radio Transmission and Reception (Release 8)." 2007.
- [3] 3GPP TS 36.201 V8.1.0 (2007–11), 3rd Generation Partnership Project; Technical Specification Group Radio Access Network, "Evolved Universal Terrestrial Radio Access (E-UTRA); LTE Physical Layer – General Description (Release 8)." 2007.
- [4] *3GPP Long Term Evolution: System Overview, Product Development, and Test Challenges*. Agilent Application Note, September 2009.
- [5] Ahmadi, S. "An Overview of Next-Generation Mobile WiMAX Technology," *IEEE Communications Magazine*, June 2009.
- [6] Akyildiz, I.F., Lee, W.-Y., Vuran, M.C., and Mohanty, S. "A Survey on Spectrum Management in Cognitive Radio Networks," *IEEE Communications Magazine*, April 2008.
- [7] Analog Devices, "14-Bit, 160 MSPS TxDAC+ with $2\times$ Interpolation Filter." Data Sheet. 2008. www.analog.com
- [8] Arslan, H. *Cognitive Radio, Software Defined Radio, and Adaptive Wireless Systems*, Springer, 2007.
- [9] Bertran, E., Porta, O., Montoro, G., and Delgado-Penin, J.A. "On PAPR for Combined Modulation and Access Techniques in Configurable Radio." IEEE International Microwave Week Symposium: RF Front-ends for Software Defined and Cognitive Radio Solution, IMWS 2010, Aveiro. February 2010.
- [10] Craninckx, J. and Wambacq, P., "Reconfigurable Single-chip Radios," *EDN*, May 27, 2010.
- [11] www.elec.york.ac.uk/research/comms/wireless/cognitive.html
- [12] O'Droma, M. S., Portilla, J., Bertran, E. *et al.* "Linearisation Issues in Microwave Amplifiers." European Microwave Week, 2004.
- [13] *Long Term Evolution (LTE): An Introduction*. Ericsson White Paper, October 2007
- [14] "Proposed Radio Technologies: The ETSI UMTS Terrestrial Radio Access (UTRA) ITU-R RTT Candidate Submission," ETSI SMG2. January 29, 1998.
- [15] ETSI TS 125 461 V7.2.0 (2005–12). *Technical Specification. Universal Mobile Telecommunications System (UMTS)*. 2005.
- [16] Friedmann, A. *Enabling LTE Development with TI's New Multicore SoC Architecture*, White Paper, Texas Instruments, 2010.

- [17] Groe, J. "Polar Transmitters for Wireless Communications," *IEEE Communications Magazine*, vol. **45**, no. 9, September 2007
- [18] Han, S.H. and , Lee J.H. "An Overview of Peak-to-Average Power Ratio Reduction Techniques for Multicarrier Transmission," *IEEE Wireless Communications*, vol. **12**, no. 2, 2005.
- [19] Hemphill, E., Tarn, H., Hawke, D., and Seoane, J. "Xilinx Application Note: Virtex-4 and Virtex-5 FPGAs "WiMAX Digital Front End Reference Design". XAPP966C (v1.2) June 16, 2008.
- [20] Hentschel, T., Henker, M., and Fettweis, G. "The Digital Front-End of Software Radio Terminals", *IEEE Personal Communications*, August 1999.
- [21] Holma, H. and Toskala, A. *LTE for UMTS. OFDMA and SC-FDMA Based Radio Access*, John Wiley & Sons, 2009.
- [22] IEEE 802.16m-08/003r7, *IEEE 802.16m System Description Document*, Jan. 2009; <http://ieee802.org/16/tgm/index.html>
- [23] Kenington, P. B. *RF and Baseband Techniques for Software Defined Radio*, Artech House, 2005
- [24] Kenington, P. B. *High-Linearity RF Amplifier Design*, Artech House, 2000.
- [25] Khan, F. *LTE for 4G Mobile Broadband. Air Interface Technologies and Performance*. Cambridge University Press, 2009
- [26] Kim, J. and and Konstantinou, K. "Digital Predistortion of Wideband Signals Based on Power Amplifier Model with Memory," *IEE Electronics Letters*, vol. **37**, no 23, pp. 1417–1418, 8 November 2001
- [27] Kimball, D. F., Jeong, J., Hsia C. *et al.* "High-Efficiency Envelope-Tracking W-CDMA Base-Station Amplifier Using GaN HFETs," *IEEE Transactions on Microwave Theory and Techniques*, vol. **54**, Nov. 2006.
- [28] Chen, K.-C. and de Marca, J. R. B. (eds.), *Mobile WiMAX*, John Wiley & Sons, 2008
- [29] Lawton, G. "What Lies Ahead for Cellular Technology?", *Computer*, June 2005
- [30] Maurer, L. "Highly Flexible Digital Front-End Enhanced CMOS-Based RT Transceivers," *IEEE International Microwave Workshop Series on "RF Front-ends for Software Defined and Cognitive Radio Solutions"*, Aveiro, February 2010..
- [31] McMahan, M. L. Texas Instruments Application Report. "Evolving Cellular Handset Architectures but a Continuing, Insatiable Desire for DSP MIPS". SPRA650, March 2000
- [32] Mitola, J. "The Software Radio Architecture", *IEEE Communications Magazine*, May 1995
- [33] Mitola, J. "Cognitive Radio for Flexible Mobile Multimedia Communications," *IEEE Mobile Multimedia Conference*, 1999
- [34] Mitola, J. and Maguire, G. Q., Jr. "Cognitive Radio: Making Software Radios More Personal," *IEEE Personal Communications*, vol. **6**, no. 4, Aug. 1999.
- [35] Prasad, R. and Ruggieri, M. *Technology Trends in Wireless Communications*. Artech House, 2003.
- [36] Rayal, F. "An Overview of the LTE Physical Layer," *CommsDesign*, May 2010, www.commsdesign.com/article/printableArticle.jhtml?articleID=225000018
- [37] Razak, Z. and Arslan, T. "Nyquist-rate Analog-to-Digital Converter Specification for Zero-IF UMTS Receiver," *IEEE International Symposium on Circuits and Systems*, 2008. ISCAS 2008.
- [38] Rumney, M. "3GPP LTE: Introducing Single-Carrier FDMA," *Agilent Measurement Journal*. January 2008. <http://cp.literature.agilent.com/litweb/pdf/5989-7898EN.pdf>

- [39] "EDGE, HSPA and LTE. The Mobile Broadband Advantage." Rysavy Research. September 2007.
- [40] SDR Forum, "Cognitive Radio Definitions," Document SDRF-06-R-0011-V1.0.0, Nov. 2007
- [41] Sesia, S., Toufik, I., and Baker M. *LTE – The UMTS Long Term Evolution; From Theory to Practice*, John Wiley & Sons, 2009.
- [42] Sklar, B., *Digital Communications. Fundamentals and Applications*, Second Edition, Prentice Hall, 2001
- [43] Strohmenger, K., Kakerow, R., Henkel, F., and Köther, D. "Re-configurable Multi-mode Radio Architectures for Enhanced 3G Terminals," *10th Meeting of Wireless World Research Forum, WWRF*, New York, November 2003
- [44] Tam, H., Hemphill, E., and Hawke, D. Xilinx Application Note: "Virtex-5 FPGA 3GPP LTE Digital Front End Reference Design," XAPP1123 (v1.0), October 29, 2008.
- [45] Tuttlebee, W. *Software Defined Radio. Enabling Technologies*, John Wiley & Sons, 2002
- [46] Van Nee, R. and Prasad, R. *OFDM for Wireless Multimedia Communications*, Artech House, 2000.
- [47] Wenzel, D. "System Architectures in Multi Mode Mobile Terminals," *European Conference on Wireless Technologies*, October 2007.
- [48] Wilson, R. "Market for SiC and GaN to boom," *ElectronicsWeekly.com*, 08 February 2010.
- [49] WiMAX Forum, "Mobile WiMAX–Part I: A Technical Overview and Performance Evaluation," August, 2006
- [50] WiMAX Forum, "Mobile System Profile, Release 1.0, Approved Specification," Revision 1.7.1, Nov. 7, 2008; www.wimaxforum.org/technology/documents
- [51] *Part 16: Air Interface for Broadband Wireless Access Systems, P802.16Rev2/D9*, (Revision of IEEE Std 802.16-2004 and consolidates material from IEEE Std 802.16e-2005, IEEE Std 802.16-2004/Cor1-2005, IEEE Std 802.16f-2005, and IEEE Std 802.16g-2007), Jan. 2009.
- [52] WiMAX Forum, "Mobile System Profile, Release 1.5. Common Part," WMF-T23–001-R015v01, August 2009.
- [53] WiMAX Forum, "WiMAX: Success Stories from Around the World," 2010
- [54] Xilinx Application Note: "Spartan-3 FPGA Series, XAPP569 (v1.0.1), Digital Up and Down Converters for the CDMA2000 and UMTS Base-stations," August 10, 2006.
- [55] Zozaya A. J. and Bertran, E. "Passivity Theory Applied to the Design of Power-Amplifier Linearizers". *IEEE Transactions on Vehicular Technology*, vol. **53**, no. 4, July 2004.
- [56] Zyren, J., *Overview of the 3GPP Long Term Evolution Physical Layer*, White Paper, Freescale. July 2007

4 System overview and front-end technologies in digital broadcasting

Francesc Tarres and Eduard Bertran

4.1 Introduction

The first International standard for Terrestrial Broadcasting of Digital Television was published by the Advanced Television System Committee (ATSC) in 1995. The standard is known as ATSC and was adopted by the Federal Communications Commission (FCC) in 1996. The main purpose of this standard was the transmission of High Definition Television (HDTV) for home consumption, that is, to deliver the experience of viewing full motion pictures in fixed scenarios with large screens [15]. The modulation scheme selected in ATSC was 8-level Vestigial Sideband (8-VSB) that contrasts with the Coded Orthogonal Frequency Division Multiplexing (COFDM) scheme selected by the European Digital Video Broadcasting – Terrestrial (DVB-T) and Japanese Integrated Services Digital Broadcasting – Terrestrial (ISDB-T) standards. Since any terrestrial TV system has to overcome many channel impairments and interferences (ghosts, noise bursts, fading, etc.) to reach the home viewer, the selection of the RF modulation format is crucial.

The selection of 8-VSB modulation in Digital Television ATSC system was motivated by different reasons. First, 8-VSB can cover larger distances with fewer repeaters than COFDM, which supposes a considerable cost reduction for low population North America rural areas. Direct Broadcast Satellite (DBS) Television is not popular in North America and therefore rural areas had to be covered by terrestrial television.

The second reason is that the system was specifically designed for fixed reception and therefore robustness against Doppler shift and multipath were not considered. The reasoning under this assumption was that High Definition TV only had commercial interest when viewed on large screens and therefore transmission to mobile devices was not considered. However, in practice, the Digital Television ATSC system has been used not only for transmitting HD but also Standard Definition (SD) signals, and obviously those signals could fit into some mobile displays in public transport services or built-in rear seat car entertainment systems.

It is also important to keep in mind that, for the same level of complexity, 8-VSB is spectrally more efficient than COFDM, that is, the number of bits per second that can

be achieved for a given bandwidth is greater in 8-VSB than in COFDM.¹ This improved spectral efficiency is easy to understand if we take into account that robustness in COFDM is achieved through guard bands, pilots, and time guard slots that do not contain audiovisual information. Spectral efficiency was a very important argument at the beginning of digital television since the coding of audio and video signals was at a preliminary stage, with compression ratios quite below the numbers achievable today. Coding of High Definition video and audio, with acceptable quality, could require in the late 1990s about 17–18 Mbps while today, the same quality, may be achieved using around 10 Mbps or even less. Moreover, TV channels in North America have a bandwidth allocation of 6 MHz while in Europe the bandwidth is of 8 MHz, providing extra space for using a more robust modulation. As a summary, an 8-VSB ATSC modulation can accommodate 19.39 Mbps in a 6 MHz RF channel, which gave enough space for a high definition video signal plus its associated multi-channel audio and metadata (electronic program guides, teletext, service information).

As opposed to ATSC, the European DVB-T and the Japanese ISDB-T selected a modulation scheme based on a highly configurable COFDM. The main purpose of that selection was to obtain a more robust scheme that could cope with hard receiving conditions such as multipath, fast fading, interferences and Doppler shift due to motion of the receiver; DVB-T and ISDB-T were specifically designed to be robust and reliable even when the receiver was in a vehicle. In fact, DVB-T and ISDB-T has been demonstrated in public transportation services and several manufacturers produce receiver set-top-boxes for the automobile industry.

The DVB project is an industry-led initiative to standardize digital broadcasting worldwide, taking advantage of the spectral efficiency of the digital signals compared to analog television. During the 1990s, the DVB working groups produced different system specifications including satellite: DVB-S [EN 300421], cable: DVB-C [EN 300429], terrestrial: DVB-T [EN 300744], among others.

The standard DVB-T was introduced in 1997 and has been adopted by many countries around the world and is today the most widely spread digital television standard.

Digital Video Broadcasting – Terrestrial services were designed to operate within the existing UHF analog transmissions, thus requiring sufficient protection against high levels of co-channel interference (CCI) and adjacent-channel interference (ACI) emanating from existing PAL/SECAM services. Maximum spectral efficiency was an important requirement since the system had to be used within the UHF bands in scenarios sensitive to multipath. These requirements boosted the use of Single Frequency Network (SFN) operation and OFDM transmission with concatenated error correcting, named COFDM.

The DVB-T system is flexible, offering to select the type of modulation among three options (QPSK, 16QAM, 64QAM), the forward error correction (FEC) among

¹ COFDM in DVB-T has many selectable operating modes. Some modes have better performance in spectral efficiency than 8-VSB but its complexity (for the transmitter and the receiver) is also greater (it can use modes with 8K carriers, 64 QAM and puncturing codes).

five different rates, the guard intervals have four options, submultiples of the symbol duration, and it can operate in 6, 7 or 8 MHz channel bandwidths.

According to the number of OFDM carriers, two main *modes* of operation have been defined: “2K mode” and “8K mode.” The “2K mode” is suitable for single transmitter operation or for small SFN networks with limited transmitter distances. The “8K mode” can be used both for single transmitter operation and for small and large SFN networks, and it supports the worst multipath situations [8]. Transmitters for DVB-T typically range from 30 W to 10 kW (100 W to 1 kW for medium power transmitters), a reduction with regard to the transmitters for analog TV, with typical ranges of 50 W to 30 kW (500 W to 2 kW for medium power). Notice that, using directive antennas, the *Equivalent Isotropic Radiated Power* (EIRP) may be significantly higher than the power delivered by the transmitter.

Digital Video Broadcasting – Terrestrial is very flexible in configuring the contents of the carrier. It is possible to transmit a single HD program or several SD (four to six, depending on the country) in a single carrier. Although some equipment has been produced for the reception of digital television in personal mobile devices (smart phones, handheld) the constraints on power consumption and the poor performance of small reception antennas result in very poor performance and have not been commercially accepted. The DVB consortium produced the standard DVB-H (*Handheld*) as the technical solution to the personal and handheld mobile receivers.

Recently [9] the standard DVB-T2 (Digital Video Broadcasting – Second Generation Terrestrial) has been published that is an extension of DVB-T capable of increasing efficiencies of 30–50 percent in its use of spectrum. The standard DVB-T2 uses a new organization of services, allocating system inputs in one or more logical data streams, based on individual physical layer pipes (PLP). The maximum achievable throughput (6) rate, after deletion of null packets when applicable, is announced as over 50 Mbps (in 8 MHz channels).

The ISDB-T system was commercially introduced in 2003. Japan had an analog High Definition television standard (MUSE – Multiple sub-nyquist Sampling Encoding system) that justified the delay in introducing the digital television.

The ISDB-T system uses MPEG-2 or H.264 video coding and MPEG-2 advanced audio coding (AAC) for HDTV and H.264 for mobile reception. The transmission bit rate ranges from 3.65 Mbps to 23.23 Mbps (the last using a 64 QAM modulation). The transmission bandwidth is 5.6 MHz (wideband ISDB-T), and it operates at the UHF band by using OFDM access. For audio and data transmission, without TV, there are also narrowband channels (432 kHz).

There are three modes of operation, all using 13 OFDM *segments* but with different number of subcarriers. Mode 1 (2K) is suitable for mobile SDTV, while mode 3 (8K) is for fixed HDTV or SDTV.

The hierarchical transmission allows a maximum of three levels. This is based on a previous hierarchical separation (time interleaved) of up to the three segments (sources), with separated paths where each segment is delay-compensated, interleaved, coded, and modulated. Finally, these paths are combined into packets of 13 segments, prior to the OFDM generation. One of the segments contains the so-called 1seg (one segment)

Table 4.1 Comparison of four major multimedia broadcast standards for terrestrial communications.

Standard	Fixed	Vehicular mobile	Personal mobile	Region Countries
ATSC	x	—	—	North America Central America
DVB-T	x	x	—	Europe Australia Africa East Asia
ISDB-T	x	x	x	Japan South America
DMTB	x	x	x	China

signal that is designed to be delivered to personal mobile communications devices. The other 12 segments may be configured in one HD program, 3 SD programs (4 segments for each program) or one Medium Definition program plus one SD program (8 segments + 4 segments). This structure of the signal in temporal segments which are later modulated in COFDM allows ISDB-T to work in different scenarios such as fixed, vehicular mobile, and personal mobile. Connecting the receiver only in the segments of the selected program, the segmented structure can be used to optimize the power consumption. The ISDB-T standard has been adopted in Japan and many countries in South America (with slight modifications in some countries).

Finally, another major standard for digital broadcast of multimedia is the Chinese GB206000–2006, also known as the Digital Multimedia Terrestrial Broadcast (DMTB) or Digital Multimedia Broadcast for Terrestrial and Handheld (DMB-T/H). This standard was introduced in 2006 and demonstrated in 2008 with the Olympic Games in China. It is the newest international multimedia broadcast standard and includes many technological improvements such as newer modulation schemes (Time-Domain Synchronous OFDM, TDS-OFDM) and information encoding systems (Low Density Parity Check Code, LDPC). The system has been adopted in China, Macao, and Hong-Kong and can be used in HD fixed communications, HD vehicular, and low-definition handheld devices.

Table 4.1 summarizes the capabilities of these four major multimedia broadcast standards for terrestrial communications. It is obvious that the year of appearance of the standard determines its capabilities for use in fixed scenarios, vehicular mobile (that is mobile scenarios without heavy restrictions in power consumption and antennas), and personal mobile (handheld, smart phones, PDA, etc.). Due to those limitations, new versions or extensions of the standards or even proprietary solutions have appeared to facilitate the broadcasting of multimedia contents to any type of mobile user.

The rest of this chapter is organized into the following sections. In the [next section](#), concepts such as multicast, return channel, video quality, transcoding, etc. are presented in the context of broadcast mobile multimedia services. [Section 4.3](#) is devoted to present common modulation technologies in digital broadcasting systems. The different alternatives for audio and video coding in commercial systems are also explained in this

section. [Section 4.4](#) outlines most relevant standards and proprietary systems for digital multimedia broadcasting. In [Section 4.5](#), our emphasis is on key technologies such as ADC, DAC, and power amplifiers for digital front-end in broadcast transmission and reception. The [final section](#) of this chapter gives a summary and discusses perspectives of these systems.

4.2 Broadcast mobile multimedia services

There are different approaches for broadcast mobile multimedia services, which have to simultaneously consider bit-rate (necessary for giving enough image quality), power consumption in handhelds (because they are battery operated), small antennas, indoor reception, and, finally, the possibility of reaching the maximum possible number of users. Data rate is especially important because of the wide spread in mobile devices. The required video quality depends on the screen size. Smart phones, PDA, or car rear-seat entertainment systems have a wide variability in screen sizes. The broadcast mobile multimedia system has to be flexible enough to accommodate all these requirements.

The high power consumption in the receiver, together with the size of the antennas have been the main constraints in distributing DVB-T to portable devices. And regarding the number of users, 3G cellular systems are just suitable for distribution in the unicast or the point-to-point modes, not supporting broadcast. Hence, new standards for mobile multimedia (digital broadcasting) have been developed. Besides, it is the cost of standard patents (*IPR licences*) that has also been a booster for standards developments in different countries, with such developments sometimes showing quite similar technical characteristics.

Unicast means to send the messages to single users within the network. Conversely, in both multicast and broadcast techniques the messages are sent to several users, provided with permissions and device capabilities for decoding the stream. Broadcast is the transmission of the same data to all the users having MS with enough sensitivity to tune the transmission. Unicast is practically restricted to users using packet-switched streaming, and normally just for particular contents, not in real-time. On the other hand, unicast has the advantage of network simplicity, while broadcast needs intermediate gap-fillers and routers supporting this technology. Broadcast may be based on two different modes of operation, depending on the way the users get the transmitted information. If it is necessary to send an explicit request, then the mode is called “pull”. Instead, in the “push” mode, all the users continuously monitor the broadcasted information [12].

Apart from the aforementioned constraints, another factor of merit among the different standards is user mobility, not only at pedestrian speeds but also permitting the reception of users traveling in cars or high-speed trains. Radius of coverage of the transmitters for a given power consumption is also a very important design parameter that has to be combined with the selection of parameters in the OFDM modulation if SFN are to be used.

Another important aspect is the commercial model of the system. Broadcast Mobile Multimedia Services imply that the model for the content provider does not suffer

significant changes. The content broadcaster only has to *transcode* high definition signals to lower resolution and send the same contents to the fixed and portable devices scenarios. The portable device usually can interact with the provider through mobile data networks (3G, UMTS). On the other hand, the content distribution of video through streaming in data networks implies a change in the model where the content provider interacts with the user through the network provider, adding a new actor into the scene.

The minimum requirements for digital broadcasting are the content servers, the transmitters, the receivers, and the users. However, services on demand need a return (uplink) channel from the users (subscribers) to the server, which is usually implemented through the cellular network (i.e., GERAN or UMTS). The lack of an explicit return channel for the mobile multimedia standard introduces additional challenges. As opposed to wireless communications presented in the [previous chapter](#), it is not possible for the server to know the quality of the downlink, and hence, it cannot make a quality-based scheduling. Furthermore, the transmit power has to be fixed with enough link strength to reach the desired QoS even for the subscribers with the worse propagation scenarios.

There are a few exceptions among the mobile broadcasting standards that do not use a 3G based return channel, thus having a return channel explicitly defined in the same standard, and most of them are only suitable for fixed or portable equipment, but not for mobile handhelds. Among these exceptions is the standard Digital Video Broadcasting – Return Channel via Satellite (DVB-RCS), presented in the following sections and whose return channel is not operative at the users' level. It is used to feedback information from the earth station gateway that is responsible for broadcasting multimedia contents, in open loop, to the users. Because DVB-RCS has the ability to use DVB-S2 bursts in the return link back to the satellite, it can achieve variable and adaptive modulation and coding.

4.3 Common technologies for digital mobile broadcast services

As introduced in [Chapter 3](#) of this book, OFDM techniques have the advantage of being robust when faced with multipath effects. So, communication applications based on OFDM are very suitable for urban environments, where the multipath due to buildings is relevant. Furthermore, these environments are also the most commercially appreciated, because of the density of potential clients. The previous chapters of this book have presented the fundamentals of OFDM multiplex, and some drawbacks have been mentioned such as the sensitivity to *synchronization* errors (although in WCDMA techniques the effects of these errors are worse) and the increased peak-to-average power ratio (PAPR) values, a challenge also shared with CDMA. Synchronization errors produce both frequency offsets and time shifts. The first leads to the loss of *orthogonality* among OFDM subcarriers, so introducing intercarrier interference (ICI).

A design compromise in OFDM-based mobile systems is among the capability to combat multipath, the allowance of high mobility in users, and the suitability to develop

SFN. User's mobility causes *Doppler shifts*, which appear as carrier frequency shifts and symbol length variation. In this situation, what is preferable is the use in a given bandwidth of a reduced number of subcarriers (larger intercarrier spacing). Conversely, multipath effects are best reduced when using a large number of subcarriers. Besides, an increased number of subcarriers makes easier the development of SFN, where the reception from a secondary distant transmitter may be perceived as an enlarged delay. It also takes advantage of large guard intervals among OFDM symbols, a fact improved in OFDM systems with a high number of subcarriers.

Some recent standards such as DMTB use improved versions of OFDM. In this case TDS-OFDM can cope with severe signal fade-ins and fade-outs; TDS-OFDM improves the estimation of the channel.

In OFDM, the receiver processes the guard interval bits in the frequency domain in order to estimate the channel. When the receiver is moving the time required to decode the bits in the frequency domain may be too large. In TDS-OFDM the guard interval is filled with pseudo-noise bits that can be processed in the time domain, which is significantly faster. In OFDM it takes about 100 milliseconds for a receiver to lock in on a signal while in TDS-OFDM it takes only 5 ms.

A few systems do not use OFDM technology for digital mobile broadcast. The most relevant example is MBMS (Multimedia Broadcast/Multicast Service) that uses the UMTS cellular network and, consequently, is based on WCDMA technologies.

Transmission of Digital Television has been possible due to the development of very efficient methods for audio and video signal compression. A SD video sequence has an uncompressed data rate of about 125 Mbps. Uncompressed HD video may require, in the highest quality formats, about 3 Gbps. The first international standard for video compression in HD and SD television applications was MPEG-2 Video and was proposed as ISO/IEC 13818–2. The MPEG-2 standard is based on applying many compression tools to the video sequence. These tools include differential prediction, discrete cosine transform coding, Huffman coding, motion compensation, interlaced modes, etc. The MPEG-2 standard specifies the syntax and the semantics of the coded bit stream but does not impose the methods for estimating the different compression parameters. This flexibility has permitted many developers to compete for obtaining the best codec, that is, the codec providing the best image quality for the low data rate. Almost all the broadcasting television standards (ATSC, DVB-T, ISDB-T) adopted MPEG-2 as the video codec of HD and SD signals. Moreover, MPEG-2 has also been used as standard for other systems such as DVD-Video. Data rate of MPEG-2 can have a good compromise between data rate and quality, giving a certain flexibility to accommodate any stream in the payload.

The MPEG committee also proposed an audio codec that was backward compatible with the previous audio compression standard proposed in MPEG-1. The new standard could cope with multi-channel audio with acceptable quality and compression rates. However, ATSC selected the Dolby Digital AC-3, a proprietary solution instead of *MPEG-2 Audio* while DVB-T selected the MPEG Audio proposal. The DVD-Video system accepted both audio coding methods in the standard. Later, the MPEG committee reviewed the audio coding standard and proposed the Advanced Audio Coding (AAC)

that is considered today as an excellent trade-off between data rate and quality. The AAC has been revisited and improved many times by the MPEG-4 committee. The AAC is the audio coding standard used in ISDB-T. The High Efficiency AAC codec (HE-AAC) is an extension of the Low Complexity AAC (AAC-LC) that uses additional compression tools for achieving higher degrees of compression. The HE-AAC has a lower quality than AAC but data rate is reduced by a factor of two. It is used in coding audio for handheld digital television.

The MPEG-2 committee is also responsible for the introduction of the multiplexing standard for video, audio, and data sources. This standard is usually referred as MPEG-2 Transport Stream (MPEG-2 TS) and it is used to create the data stream, composed by several video, audio, and data programs that will be sent to the channel coder and modulator. The MPEG-2 TS standard is being used in both ATSC and DVB.

The video coding standard H.264 was proposed in 2003 as the best state-of-the-art technology in video compression, obtaining considerable improvements compared to MPEG-2 video. The standard H.264 was developed by the Joint Video Team (JVT), a group of experts formed with members from the ISO and the ITU. The name H.264 is the official name giving by the ITU, while ISO uses the name MPEG-4/AVC (MPEG-4 Advanced Video Coding). Some countries, like France, where digital television was introduced recently, have selected H.264 as the video codec standard. The flexibility of the DVB-T has made possible such implementations without any impact on the receivers front-end. In fact, the H.264 stream is transported, together with the audio and possibly other programs, through MPEG-2 TS. The only difference in the receiver is that the decoder has to accept the H.264 coded stream. The H.264 High Profile is used today in ISDB-T HD broadcasting. The H.264 Baseline Profile is being used for most of the video services to handheld devices.

Other improvements of the system, such as the introduction of 3D TV, are being accomplished with a similar strategy. For the specific case of 3D TV the standard H.264 has included multiview modes (H.264/MVC – MultiView Coding) that permit the information associated to multiple cameras to be transmitted. The compressor analyzes the data and exploits not only time and space redundancy but also similarities between the images captured by different cameras.

The DMTB standard does not specify any video or audio codecs, providing some flexibility to the content provider to select between any possible codecs (proprietary or standards). This flexibility is aimed towards the reduction of royalties that have to be paid for using some codecs.

4.4 Most relevant standards for digital broadcasting to mobile devices

4.4.1 DVB-RCT

The DVB-RCT (DVB-Return Channel Terrestrial) standard was published by the ETSI in 2001 [7], and offers a wireless interaction channel for Interactive DVB-T, even in the congested UHF/VHF bands. According to the DVB Organization, the reason for the RCT

is to allow the TV commerce, which requires a return path from user to the digital TV service provider. Besides, it supports within the same cell the simultaneous use of different types of modulations from 4QAM to 64 QAM (Dynamically Assignable Adaptive Modulation).

Also based in OFDM access, DVB-RCT can serve large cells, with radius up to 65 km, providing a typical bit-rate capacity of several kbps for each TV subscriber. It employs either 1024 (1k) or 2048 (2k) subcarriers [11].

4.4.2 DVB-H

The DVB-T standard, first published in 1997, was originally designed for fixed or vehicular receivers, but not targeted for mobile handheld receivers. To cover the market of mobile broadcasting, the Digital Video Broadcasting – Handheld was formally adopted as an ETSI standard in November 2004 [4]. The DVB-H standard is an extension to DVB-T with very similar physical layers and allowing some backward compatibility. It can operate in 6, 7, and 8 MHz channel bandwidths.

The DVB-H can share the same multiplex with DVB-T, so it is possible to have DVB-H services coexisting in the same DVB-T multiplex. In particular, the transmission chain allows DVB-H modulated signals to be transmitted over DVB-T, by multiplexing MPEG-2 TV services with DVB-H encapsulations. The multiplexed signal is transmitted to mobile terminals through the same DVB-T modulator. The return channel is configured by means of the base-stations of the cellular network.

A data rate up to 14 Mbps using one 8 MHz UHF channel is announced as possible with the 64QAM mode for high-speed users [10]. This challenge, along with the operation in wide SFN areas and the handheld autonomy, has to be solved by also preserving the basic compliance with the DVB-T standard. To do that, two main technical features have been introduced in DVB-H: the introduction of a new “4K mode” and the inclusion of a time-sliced transmission in order to increase power savings.

The operators can choose among 2K, 4K, and 8K modes. The 4K mode is a solution, *not compatible* with DVB-T networks, that combines the advantages of the 2K mode (high-speed users, but low SFN capability) with the advantages of the 8K mode to deploy larger area SFN, although at lower speeds because of the lower carrier spacing. For instance [5], in DVB-H, the guard intervals range from 224 to 28 μ s in the “8K mode,” from 112 to 14 μ s in the 4K mode, and from 56 to 7 μ s in the 2K mode. Obviously, the 8K mode allows larger separations between SFN transmitters minimizing interferences (Table 4.2). The “4K mode” practically doubles user’s mobility with regard to the 8K mode, and also doubles the size of the SFN regarding 2K mode.

The time-slicing mechanism allows increased power savings. This mechanism consists of the transmission of the data corresponding to a certain service (video or audio) continuously multiplexed with other services. Hence, for the desired service, the time intervals between active receptions are spread, allowing the radio components to be turned off during these intervals and, consequently, saving power in the receiver. According to [11], the quality of the video display in handheld terminals is compatible with relatively low transmission bit-rates (the average bit-rate is 500 kbps; one 8-MHz channel can

deliver 30–50 video streaming services to the small screen terminals). Conversely, in the former DVB-T standard, all the services are transmitted simultaneously.

In addition, the capability for mobile users to receive the same content when moving to another radio cell (handovers) is increased because of the time slicing. While waiting for the next burst of the selected service, the user terminal can assess the radio environment and manage the handover.

The DVB-H standard can operate in different frequency bands: VHF-III (170–230 MHz), UHF-IV/V (470–862 MHz) and L-band (1.452–1.492 GHz).

At the receiver level, the DVB-H demodulator is basically a DVB-T demodulator enlarged to have the additional possibility of receiving signals in the 4K mode (supplementary to the 8K and 2K modes already used for DVB-T), and of arranging the data bursts according to the time slicing structure.

4.4.3 DVB-SH

Another DVB standard is DVB-SH, which stands for Digital Video Broadcasting – Satellite to Handheld. It uses the S-band (2.17–2.2 GHz), next to the UMTS band. The DVB-SH standard may use channel bandwidths of 5, 6, 7 or 8 MHz, and FFT sizes of 2, 4 or, 8 k. The preferred choices seem to be 5 MHz and 2K, in order to preserve the UMTS channelization bands (adjacent in the spectrum), and to allow for high-mobility users.

Defenders of the DVB-SH standard declare that it could reduce infrastructure deployment cost with regard to the terrestrial DVB-H. Actually, DVB-SH is a hybrid satellite-terrestrial system that uses satellite to provide coverage in wide areas, reinforced by terrestrial gap fillers (or even cellular base-stations in some scenarios) to provide coverage in areas where direct reception is not possible from the space segment.

The DVB-SH standard combines the satellite component (SC) and the complementary ground component (CGC): in a cooperative mode, the SC ensures geographical global coverage while the CGC provides cellular-type coverage [16]. Apart from the OFDM access, the DVB-SH also uses a Time Division Multiplex (TDM), thus leading to two reference architectures: SH-A that uses OFDM in both the satellite and the terrestrial link, and SH-B that uses TDM in the satellite link and OFDM for the terrestrial.

Also based on satellite systems, the standard DVB-RCS includes a *return channel*, although it is not accessible at the user's level. It is used to feedback information from a gateway placed in a ground station, which is responsible for receiving signals from the satellite segment and for broadcasting them to the users. When possible, the user's feedback (return channel) has to be made through the *gateways*. It relies on the ability to use DVB-S2 (ETSI EN 302 307) bursts in the return link to the satellite.

The DVB-S2 satellite broadcasting system has two versions, with different degrees of compatibility with the previous DVB-S standard. The non-compatible version allows about 30 percent more data capacity. Thanks to the return channel, the DVB-S2 system has improved flexibility, thus allowing the downlink channel to be dynamically scaled according to user's requirements.

A typical value for the transmit power is 45 dBW (in the Ka band), a value that may range between 40 dBW and 55 dBW.

4.4.4 T-DMB

Terrestrial Digital Multimedia Broadcasting (T-DMB or DMB-T) is a standard based on *Eureka 147 Digital Audio Broadcasting* (DAB), and standardized by the ETSI (TS 102 427 and TS 102 428). The first launch of T-DMB services was in Korea, in 2005. The T-DMB standard enhances DAB by also applying MPEG-4 technology. Audio, video, and data services are transmitted using DAB infrastructure. It aims at cheaper receivers and infrastructure than DVB-H, a fact favored by the lower field strength and operational bandwidth. Although the acronyms are similar, T-DMB should not be confused with the Chinese DMTB Digital Broadcasting Standard that has been explained in the related previous sections.

The channel bandwidth is organized in blocks of 1.536 MHz (three blocks in 6 MHz) and data rates are between 1 and 1.5 Mbps, depending on the kind of convolutional coding. Also based in OFDM access (4K mode), the modulation is DQPSK. The images are reproduced in the mobile displays at 30 fps (frames per second). It may be operated in the VHF band III from 174 to 240 MHz or in the L-band from 1452 to 1490 MHz.

Transmission mode I (single frequency network) works at VHF frequencies, while mode II (multiple frequency networks, MFN) operates at the L-band. Mode III is appropriate for satellite or cable distribution, as well as for complementary terrestrial transmission at different frequencies since it can be operated up to 3 GHz. Mode IV is used in the 1.5 GHz band and allows larger transmitter spacing in SFNs. However it is the least resistant to degradation at higher vehicle speeds [14]. The differences between the modes are in the number of subcarriers, the symbol period, the length of the transmission frame and the guard time. Mode III has the maximum subcarrier separation (8 KHz), while the maximum guard time is done in mode I (246 μ s). The ranges are 96 km, 24 km, and 48 km for modes I, II, and III, respectively [12].

The T-DMB transmit power depends on the frequency band. Typically, for the VHF band III different power transmitters range from 50 W to 4 kW, and for L-band from 100 W to 2 kW.

4.4.5 MediaFLO

MediaFLO (sometimes shorted as FLO) stands for Media Forward Link Only, and it is a Qualcomm proprietary transmission system to broadcast data to portable devices. It has been recognized by ETSI. The return channel relies on cellular networks. It is planned to operate at different frequency bands, with channel bandwidths of 5, 6, 7, or 8 MHz. In order to reduce power consumption in the receivers, MediaFLO uses time division multiplexing, using a time-slicing mechanism similar to the one used in DVB-H. Also similar to DVB-H, MediaFLO access is based on OFDM transmission with either QPSK or 16QAM modulation of the carriers. Another similarity with DVB-H is the frequency range in the UHF band. The video source coding is H.264 and is in a reduced resolution QVGA (320 $\times$ 240).

An expected goal for viewing-time with 850 mAh battery-powered devices using 360 Kbps of data rate is 4 hours [12]. The number of subcarriers is the same (4096) for the

different channel bandwidths (5, 6, 7, or 8 MHz), so showing different values of the subcarrier spacing. The data-rate ranges 8 Mbps to 11.2 Mbps in a 6 MHz channel, with services per channel ranging from 17–19 in a 5 MHz bandwidth to 28–32 channels in 8 MHz.

In the US the *FCC* assigned licenses for 698–746 MHz in 6 MHz blocks for a variety of broadcasting, mobile, and fixed services, with a maximum transmit power of 50 kW Effective Radiated Power (ERP). According to [17], a transmit power of 1kW, in the 788 MHz band, could cover an area of 153 square km. The same transmit power, in the 2130 MHz band, covers 47 square km. With 50 KW, 716 MHz, the coverage is 1937 square km.

4.4.6 MBMS

Multimedia Broadcast/Multicast Service (MBMS) is a 3GPP standard for point-to-multipoint data distribution service within 3G cellular networks. The features are split into the MBMS *bearer* service, which addresses MBMS transmission features below the IP layer [12]. The MBMS bearer service includes broadcast and multicast modes. The multicast mode takes advantage of the whole 3G cellular operation, allowing both uplink and downlink directions. Hence, mobile terminals can inform the network about their state and, consequently, increase the *efficiency* of the radio resources [1].

The MBMS may be supported by UMTS Terrestrial Radio Access Network (UTRAN) and by GSM/EDGE Radio Access Network (GERAN).

When supported by UMTS, the MBMS Bearer Service provides different QoS. The MBMS uses the Multimedia Traffic Channel (MTCH) that enables point-to-multipoint distribution [1]. This channel is mapped to the Forward Access Channel (FACH), which is finally mapped to the Secondary-Common Control Physical Channel (S-CCPCH). The Broadcast Multicast Service Center (BM-SC) provides a set of functions for MBMS user services. In coherence with UMTS, radio frames last 10 ms, with a chip-rate of 3.84 Mbps and QPSK modulation.

MBMS Download Delivery Service has different options [12]. One is *Scheduled Distribution without Time-Limits*, whose drawback may be the long periods in accessing the files (an impatient user would abort the download). *Time-Constrained Distribution* puts a bound on the radio resources, the efficiency of the system being determined by maximum bit-rate, related to the quality of the media stream. In *Carousel Services* the download time is not bounded, and the user can get the files during the time window when they are being transmitted (*static mode*). In the *dynamic mode*, individual files may change dynamically, as in the case of Video-on-Demand services.

In MBMS the video resolution is QCIF (176×144 pixels) with a bit-rate of 128 kbps, a quality similar to the worst case in DVB-H. Data ranges may be up to 256 kbps per MBMS Bearer Service in UMTS and between 800 kbps and 1.7 Mbps per cell/band. When supported by GERAN, MBMS offers between 32 kbps and 128 kbps.

In Table 4.2 some mobile multimedia standards are compared.

Table 4.2 Comparison of some mobile digital multimedia standards.

	T-DMB	DVB-H	MediaFLO	ISDB-T	MBMS
Platform	EUREKA147	DVB-T	–	–	UMTS, GERAN
Modulation	$\pi/4$ DQPSK	QPSK, 16QAM, 64QAM	QPSK, 16QAM	DQPSK, QPSK, 16QAM, 64 QAM	QPSK, GMSK-8PSK
Transmission	OFDM (4K)	OFDM (2K,4K,8K)	OFDM (4 k)	OFDM (2K,4K,8K)	WCDMA (UMTS) 3.84 Mcps
Error control	RS + Convolutional code	RS + Convolutional code	RS(16,12 or 14 or 16) + Turbo codes	RS + Convolutional code	3G (turbo in UMTS)
Video compression	H.264	H.264	H.264	H.264	H.264
Audio compression	MPEG4 BSAC MPEG1 Layer2 (MUSICAM)	HE-AAC	HE-AAC	HE-AAC	HEA-AC
Bandwidth	1.536 MHz	5, 6, 7, 8 MHz	5, 6, 7, 8 MHz	6,7,8 MHz	5 MHz (UMTS)
Frequency	174–240 MHz (VHF)	470–862 MHz (UHF-IV/V)	716–722MHz (UHF) 698–746MHz (UHF)	Lower UHF channels (Japan)	1920–1980MHz 2110–2170 MHz (UMTS)
Power saving	Bandwidth shrinking	Time-slicing	Time division multiplexing	–	3G Power Control (return channel)
Display	30 fps	15–30 fps	15–30 fps	15–30 fps SDTV or HDTV	15–30 fps
Coverage (km)	96 (mode I) 24 (mode II) 48 (mode III)	17 (2K) – 67 (8K)	25 (typ). 1kW (788MHz): 158 1kW(2130MHz): 47	30–70	3G 500m – 3 Km
SFN: Max. transmitters separation from guard-time (Km)	18.6 (mode II)–74 (mode I)	2–67 Km	20 – 33 Km	2–75 Km	No OFDM

4.5 Technology aspects of DFE in transmitters

4.5.1 Introduction

Most of the design considerations already presented in [Chapter 3](#) of this book, with regard to *digital front-end* (DFE) for wireless communications transmitters, are applied to DFE for Digital Multimedia transmitters, including commercial and general design criteria.

The digital up-converter (DUC), and the kind of interpolation filters (halfband, poly-phase, CIC, . . .), which will be explained in more detail in other related chapters of the book, have similar considerations to wireless communications transmitters. Besides, some components of the DUC chain, such as the pulse-shaping or the channelization filters (which have to cope with the stringent *ACLR* specifications) are also implemented in digital domain, as well as gain controllers, signal synthesizers, crest factor reducers or some types of power amplifier linearizers.

Also the use of low-pass or bandpass filters after the zero-packing in the interpolation process depends on the same criteria: low-pass filters eliminate images of the spectrum produced by the inserted zero-valued samples, and the output of the interpolation process produces a first alias spectrum pushed up in frequency. Hence, the posterior analog filter specifications are relaxed. Using bandpass filters, a first IF translation may be done taking advantage of the first alias spectrum (the one placed at the end of the second Nyquist zone). In this case the image frequency (at the beginning of the third Nyquist zone) may be removed by using image rejection (IMR) structures based on Hartley-Llike schemes.

Furthermore, the applicability and the performance of *DSP* and *FPGA* devices are independent of the kind of transmitters between Wireless Communications and Digital Broadcasting ones. A system on chip (SoC) is still the best option when design flexibility is not an issue, because of the better cost (for medium-large series) and the lower power consumption.

4.5.2 ADC and DAC converters

The digital-to-analog (DAC) and analog-to-digital (ADC) converters will be revisited in the [next chapter](#) of the book, because of the special implications due to the multimode operation. However, it is critical to choose the right devices, so an introduction will be advanced in this section.

The minimum resolution is 10 bits, although 12–14 bits is the figure most recommended. The sampling rates for digital multimedia transmitters may be in the range 20–100 Msps, depending on the oversampling ratio. Anyway, the maximum channel bandwidth among the digital multimedia standards is 8 MHz, and it is easy to find converters up to 300 Msps at acceptable costs, so designers would permit higher interpolations in the DUC chain, thus alleviating the design of the analog filters.

Regarding ADC converters, they are necessary in power transmitters for functions such as to digitally monitor the power stages (i.e., transmit power, reflected power, temperature, or bias deviations) or to implement closed-loop predistorters.

The SNR in an ADC converter is

$$SNR = 6.02n + K + PG - \alpha$$

where PG is the processing gain, defined as $PG = 10 \cdot \log(f_s / 2f_m)$, where f_s is the sampling frequency and f_m the maximum frequency of the signal to be sampled. The parameter α corresponds to the PAPR (peak-to-average power ratio); and it is an indicator of the right exploitation of the ADC dynamic range. If the PAPR is high, some ADC bits may be spoiled, being statistically less used. The parameter K depends on the statistical distributions of the input signal amplitudes, becoming 1.76 for a pure sinusoid. And f_s/f_m is the oversampling ratio (OSR).

If the ADC clock produces jitter when sampling the signal, and the clock frequency (f_{clk}) is very high, then the SNR equation has to be corrected [Shi90] by a factor of

$$\Delta SNR = -20 \log(\pi \cdot f_{clk} \cdot \Delta t_s)$$

where Δt_s is the root mean square (rms) clock jitter.

The suitable ADC technologies are Flash, Pipelined and Folding, the last being for very fast ADC the cost of which, regarding the required bandwidth in digital multimedia is, at least, questionable. Analog-to-digital converters common in other fields are not possible in broadband transmitters, and neither is the cheap SAR (successive approximation register). The applicability of delta-sigma ADCs is still under research. These kind of converters, where oversampling is mandatory, are well exploited for audio and other low-frequency applications, however products for high bandwidth applications, although reported in technical papers, are not yet in mass industrial production. The commercialized delta-sigma converters typically operate below 10 Msps.

4.5.3 Some characteristics of the power amplifiers

Implementation differences among the different standards for digital multimedia are in the baseband processing stages (up to the modulator). There are no differences in the power stages apart from the ones derived from the transmit power levels.

Because the access technique is usually OFDM (or WCDMA in MBMS) the *PAPR* is expected to be high, thus leading to the aforementioned trade-off between the power amplifier (PA) linearity and efficiency. The solutions to this trade-off include the use of PA linearizing structures or to apply crest factor reduction techniques, the latter with some sacrifice in data throughput or in transmit power.

Also the abovementioned aspects are similar in *transistor technologies*. Lateral diffused metal oxide semiconductors (LDMOS) technology is broadly used; normally in *class AB* power amplifiers. Gallium Nitride (GaN) is a new technology, already on the shelf, which permits transistors to deliver powers of tens of watts, at high frequencies and increased breakdown voltages. Doherty structures also allow the power efficiency of the class AB transistors to be increased (up to 48 percent).

Currently, there are disposable balanced LDMOS power amplifiers (class AB) providing over 100W in the UHF band (470 MHz to 860 MHz), with power efficiencies up to 28 percent. Using LDMOS class AB power amplifiers, a 500 W amplifier may

need 2.2 KVA from the electrical grid, and a 7 kW one, up to 30 KVA, in both the UHF and VHF bands used for digital mobile broadcasting.

A special mention is due to Silicon Carbide (SiC), the application of which has been reduced to a substrate for GaN devices in the [previous chapter](#) on Wireless Communications transmitters, due to its low electron mobility that reduces the application bandwidth. Some digital multimedia standards allow operation at VHF and lower UHF frequencies, so SiC transistors may be used here. Currently, power transistor devices, capable of delivering power over 1kW, at frequencies lower than 450 MHz are disposable (at prices higher than one thousand dollars per transistor). They are a more robust alternative to the use of LDMOS transistors.

While in the base-stations for wireless communications the maximum transmit power per transceiver was as much as of the order of tens of watts, in broadcasting transmitters this power can range up to 10–30 kW (or beyond, although in practice maximum powers of 1–2 kW are most common). This significant increase in the transmit power leads to PA structures based on line-up power amplifiers at the circuit level, in the same box, and later also to line-up these boxes of power amplifiers at the system level. This implies the use of *splitters* after the driver amplifiers, as well as the use of posterior power combiners once the amplification has been made by two or more parallel power modules. These power combiners have to be as efficient as possible because of the managed power levels. Also hybrid combiners or circulators along with dummy loads are inserted in the power chain (normally close to the antenna) to reduce the risk of stationary waves that should damage the expensive power devices. Besides, the heat dissipation is an issue, leading to the use of liquid cooling systems inside air-conditioned refrigerated containers.

Typically, powers up to 300 W might be provided by a single power module (including internally various amplifiers in the printed circuit boards). Hence, a transmit power of 3 kW could require around 10 power modules. These power modules should be somehow “hot swappable”, i.e., to be removed without turning off the transmitter. This fact is convenient for maintenance purposes.

The output signal must be monitored, usually in a digital presentation, indicating operational aspects such as the status of transmit and the reflected powers. In addition, nonlinear distortion should be automatically corrected, measuring feedback signals from the transmitter output and updating the correction tables.

Transmitters and micro-transmitters for SFN networks take advantage of GPS signals to synchronize the transmitter’s clocks.

Regarding possible digital front-ends in transmitters of digital broadcasting, these can follow the same architectures as the ones presented for wireless communication transceivers ([Figure 3.20, Chapter 3](#)): the direct conversion, the IF-transmitter and the multiple conversion transmitter.

Among these structures, the PA may be implemented in different forms. In the [next chapter](#), some different power combinations will be presented, including the diplexer filters, depending on the number or power transistors. For multi-standard transmitters delivering powers of the order of tens of watts, or even a few hundreds, the traditional structure consisting of a number of parallelized PAs may be simplified to a single,

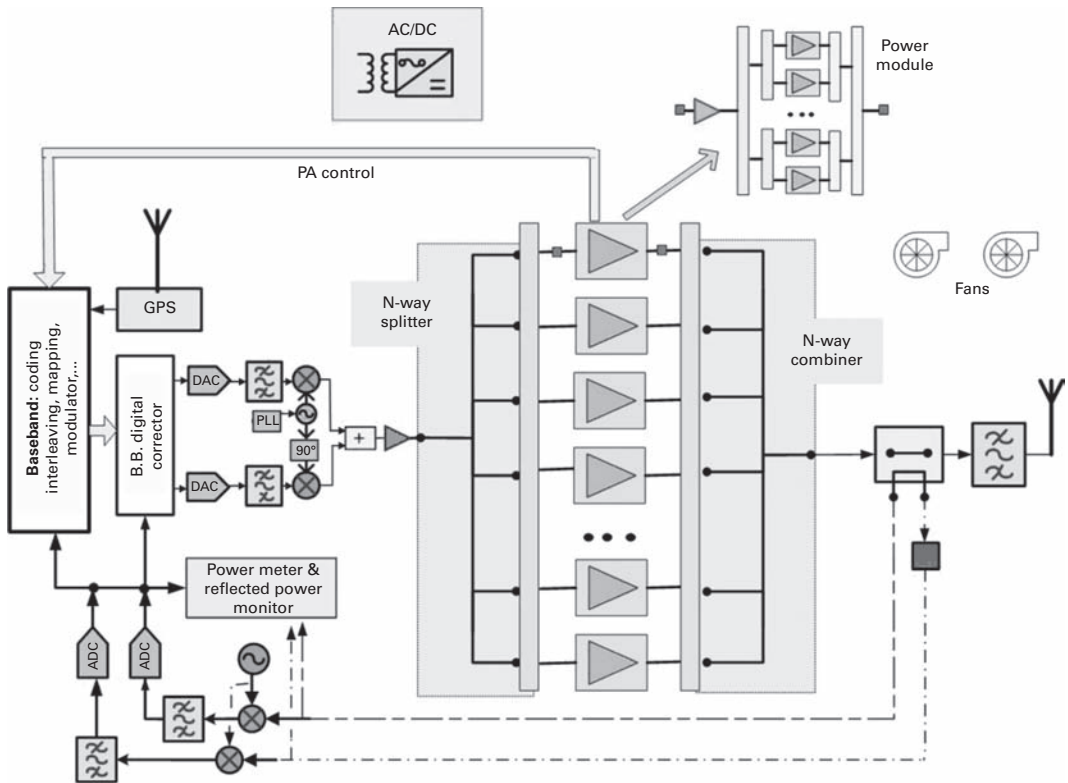
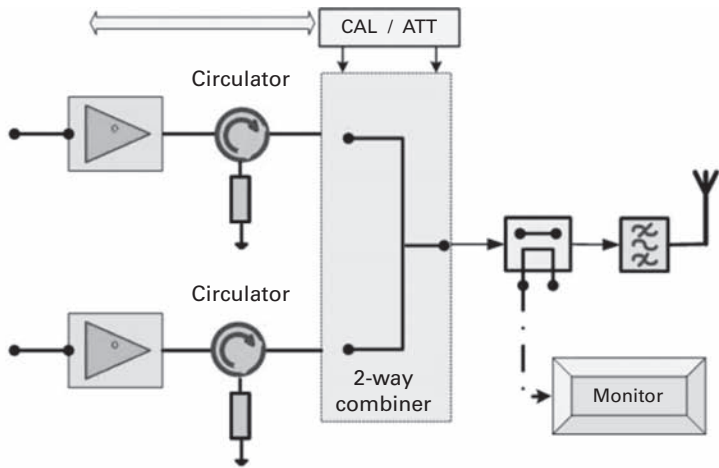


Figure 4.1 Schematic of a high power amplifier.

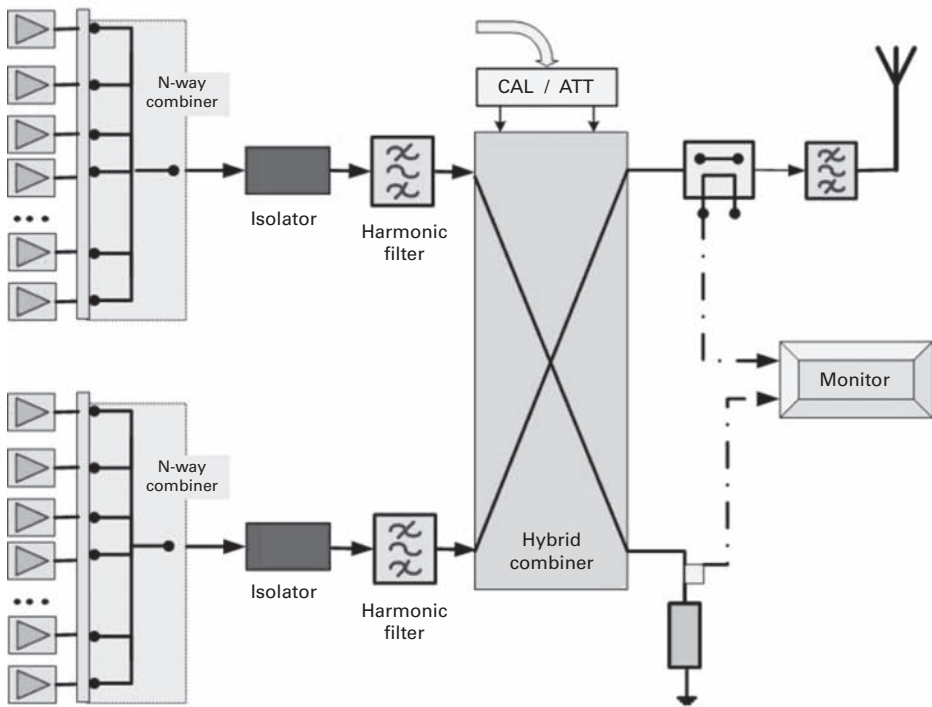
improved, PA. In this case, instead of using power combiners after the power modules, only the signals from the different modulators are previously combined to compose a multi-standard signal to drive the PA. This approach, which can be implemented in analog or in digital, reduces the number of RF components. However, transmitters for digital broadcasting can range in power from tens of watts to some kilowatts. For low or medium transmit powers, the structures of [Chapter 5](#) could be employed. Nevertheless, for higher power transmitters the conventional amplifier structure has to be used.

[Figure 4.1](#) shows some subsystems within a high power amplifier, including the blocks for the refrigerator and the power supply (AC-DC power converter). Notice the organization of the PAs is like a “fractal” structure: the lined-up power modules are also composed of lined-up power amplifiers. For the sake of simplicity in the figure, the driver amplifiers have been omitted, and the power splitters are assumed to be in one stage (N-way), as are the power combiners. However, depending on the transmit power, these splitting and combining stages may be multiple.

In [Figure 4.2](#) the most common alternatives to prevent the power amplifiers from reflected waves are presented. Isolators and harmonic filters, shown in [Figure 4.2\(b\)](#), may be also used along with the circulators ([Figure 4.2\(a\)](#)), albeit to find high power isolators operative at high frequencies is an issue. An overview of some different options in combining signals may be seen in [19].



(a) Circulators



(b) Hybrid combiner and isolators

Figure 4.2 Power amplifiers protection.

4.6 Conclusions

This chapter has given a technological view of mobile digital broadcasting systems that comprises a review of the main standards and proprietary solutions involved in the modulation and demodulation of the signals together with the efficient coding of the audiovisual sources. A perspective on the key elements in the digital front-end has been also provided.

To understand the variety of approaches presented in this [chapter one](#) should have in mind the evolutionary nature of digital broadcasting systems. The markets addressed by these systems have been changing constantly, trying to cover more and more potential consumers in a challenging trade-off between the market expectation and technical challenges and costs. The first technological solution, ATSC, was exclusively focused on video broadcasting to the home. Later on, the DVB-T produced a standard that could cope with receiver mobility installed in trains, buses, cars, etc. The newest standard, the Chinese DMTB, is addressed to any type of consumer including fixed home, vehicular mobile, and personal handheld devices. Having this evolution in mind it is easy to understand that newer standards, revisions, and proprietary alternatives have appeared in order to cover the potential market of broadcasting to mobile receivers. Today, different systems coexist showing local dominance in particular geographical areas (DVB-H in Europe, FloTV in USA, DMTB in China, etc.). Anyway, it is important to remark that the technology for mobile TV broadcasting is there and that many of these solutions have been in the market for a long time. Commercial products are available in many countries and product demonstrations have produced satisfactory results from the technical point of view.

However, the impact of mobile television has been limited and is being questioned by some companies. Its main competitors are smartphones and portable multimedia devices that may offer quality TV as a podcast or low-mid quality live TV using video streaming in 3G data networks. Podcasting has been proven to be a very efficient procedure for downloading TV series, news, etc. and all those contents that do not require the ‘live’ urgency. The quality that may be achieved by podcasting is satisfactory for personal devices and the downloading software can automatically manage up-loading and deletion of viewed programs in the terminal while the batteries are being recharged, without the participation of the user.

Therefore, mobile TV can only compete in those cases where quality live TV is a must. In fact, the main target for some mobile broadcasting divisions have been the transmission of live sports and many international and national events have already been transmitted in the mobile Pay-TV format. However, the commercial impact of these transmissions is still low due to a number of reasons that are under discussion.

List of abbreviations ([Chapter 4](#))

3G:	Third Generation (cellular telephony)
8-VSB:	8 level Vestigial Sideband
AAC:	Advanced Audio Coding

AAC-LC:	Low Complexity AAC
ACI:	Adjacent Channel Interferences
ADC:	Analog-to-Digital Converter
ATSC:	Advanced Television system Committee
BM-SC:	Broadcast Multicast Service Center
CCI:	Co-channel Interference
CDMA:	Code Division Multiple Access
CGC:	Complementary Ground Component
COFDM:	Coded Orthogonal Frequency Division Multiplexing
DAB:	Digital Audio Broadcasting
DAC:	Digital-to-Analog Converter
DBS:	Direct Broadcast Satellite
DFE:	Digital Front-End
DMB-T/H:	Digital Multimedia Broadcast for Terrestrial and Handheld
DMTB:	Digital Multimedia Terrestrial Broadcast
DQPSK:	Differential Quadrature Phase Shift Keying
DUC:	Digital Up-Converter
DVB-C:	Digital Video Broadcasting-Cable
DVB-H:	Digital Video Broadcasting-Handheld
DVB-RCS:	Digital Video Broadcasting – Return Channel via Satellite
DVB-RCT:	DVB-Return Channel Terrestrial
DVB-S:	Digital Video Broadcasting-Satellite
DVB-SH:	Digital Video Broadcasting – Satellite to Handheld
DVB-T:	Digital Video broadcasting – Terrestrial
DVB-T2:	Digital Video Broadcasting – Second Generation Terrestrial
EIRP:	Equivalent Isotropic Radiated Power
ETSI:	European Telecommunications Standards Institute
FACH:	Forward Access Channel
FCC:	Federal Communications Commission
FEC:	Forward Error Correction
GaN:	Gallium Nitride
GERAN:	GSM/EDGE Radio Access Network
GPS:	Global Positioning System
H.264/MVC:	MultiView Coding
HD:	High Definition
HDTV:	High Definition Television
HE-AAC:	High Efficiency AAC codec
ICI:	Inter Carrier Interference
IMR:	Image Rejection
ISDB-T:	Integrated Services Digital Broadcasting – Terrestrial
ITU:	International Telecommunication Union
JVT:	Joint Video Team
LDMOS:	Lateral Diffused Metal Oxide Semiconductors

LDPC:	Low Density Parity Check Code
MBMS:	Multimedia Broadcast/Multicast Service
MediaFlo:	Media Forward Link Only
MFN:	Multiple Frequency Network
MPEG:	Moving Picture Experts Group
MPEG-2 TS:	MPEG-2 Transport Stream
MPEG-4/AVC:	MPEG-4 Advanced Video Coding
MTCH:	Multimedia Traffic Channel
MUSE:	Multiple Sub-Nyquist Sampling Encoding System
OFDM:	Orthogonal Frequency Division Multiplexing
PA:	Power Amplifier
PAPR:	Peak-to-Average Power Ratio
PG:	Processing Gain
PLP:	Physical Layer Pipes
QAM:	Quadrature Amplitude Modulation
QCIF:	(Quarter Common Intermediate Format
QVGA:	Quality VCA
RS:	Reed-Solomon (codes)
SAR:	Successive Approximation Register
SC:	Satellite Component
S-CCPCH:	Secondary-Common Control Physical Channel
SD:	Standard Definition
SFN:	Single Frequency Network
SiC:	Silicon Carbide
SoC:	System on Chip
TDM:	Time Division Multiplex
T-DMB (or DMB-T):	Terrestrial Digital Multimedia Broadcasting
TDS-OFDM:	Time-Domain Synchronous OFDM
UMTS:	Universal Mobile Telecommunications System
UTRAN:	UMTS Terrestrial Radio Access Network
WCDMA:	Wireless CDMA

References

- [1] *Technical Specification Group Radio Access Network; Introduction of the Multimedia Broadcast Multicast Service (MBMS) in the Radio Access Network (RAN)*. 3GPP TS 25.346 V7.0.0 2006.
- [2] Citta, R. and Sgrignoli, G. "ATSC transmission system: 8-VSB tutorial," ITVS 1997 Montreux Symposium, 1997.
- [3] Davis, R. and Twitchell, E. "The Harris VSB exciter for digital ATV," NAB 1996 Engineering Conference, 1996.
- [4] ETSI DVB-H (EN 302 304), "Digital Video Broadcasting (DVB) Transmission System for Handheld Terminals (DVB-H)," Nov. 2004.

- [5] "Digital Video Broadcasting (DVB) DVB-H implementation guidelines." DVB document A09r3. 2009.
- [6] "Second Generation Terrestrial: The world most advanced digital terrestrial system." DVB Fact Sheet, July 2010.
- [7] ETSI DVB RCT (EN 301 958), "*Interaction Channel for Digital Terrestrial Television (RCT) Incorporating Multiple Access OFDM*." March 2001.
- [8] European Standard (Telecommunications series), *Digital Video Broadcasting (DVB); Framing Structure, Channel Coding and Modulation for Digital Terrestrial Television*. European Broadcasting Union, ETSI EN 300 744 V1.6.1 (2009–01).
- [9] European Standard (Telecommunications series). *Digital Video Broadcasting (DVB); Frame Structure Channel Coding and Modulation for a Second Generation Digital Terrestrial Television Broadcasting System (DVB-T2)* ETSI EN 302 755 V1.1.1 Sept. 2009.
- [10] Faria, G., Henriksson, J. A., Stare, E., and Talmola, P. "DVB-H: Digital broadcast services to handheld devices." *Proceedings of the IEEE*. vol. **94**. no. 1, Jan. 2006.
- [11] Fazel, K. and Kaiser, S. *Multi-Carrier and Spread Spectrum Systems*. John Wiley & Sons, 2008.
- [12] Furht, B. and Ahson, S. *Handbook of Mobile Broadcasting. DVB-H, DMB, ISDB-T, and MediaFLO*. CRC Press, 2008.
- [13] Holma, H. and Toskala, A., eds. *HSDPA/HSUPA for UMTS*, John Wiley & Sons, 2006.
- [14] *Guidelines for the Transition from Analogue to Digital Broadcasting*, International Telecommunication Union – ITU, 2010, www.itu.int/ITU-T/terrestrial/dvb/Guidelines-transition-analogue-digital%20%28Part%20D%29.pdf
- [15] Karamchedu, R. "Does China have the best digital television standard on the planet?", *IEEE Spectrum*, May 2009.
- [16] Kelley, P. and Rigal, C. "DVB-SH. Mobile digital TV in S-Band", *EBU Technical Review*. July 2007.
- [17] *FLO[™] Technology Overview*, Qualcomm, 2010.
- [18] Reimers, U. *DVB: the family of International Standards for Digital Video Broadcasting*, Second Edition, Springer, 2008.
- [19] Ramsden, T. "Selecting the right signal combiner." www.rfdesign.com. April 2007.
- [20] Shinagawa, M., Akazawa, Y., and Wakimoto, T. "Jitter analysis of high speed sampling systems. *IEEE Journal of Solid-State Circuits*, vol. **25**. no.1, Feb. 1990.

5

Digital front-end for multiple standards and multimode compatibility

Eduard Bertran

5.1 Introduction

Usually the development of a new standard for wireless communications does not leave the previous ones as obsolete, so it is necessary to share different complex technologies in both the user equipment (UE) and the base-station (base transceiver station, BTS). This implies a strong need for systems integration, especially in handset devices. Besides, the market tendency is to embed in the same UE a number of communications, location, and entertainment applications that some years ago were split among cell phones, PDAs, laptops and dedicated devices, allowing access to both cellular communications networks and WLANs (i.e, different versions of the IEEE 802.11 standard, commercialized as WiFi). Additionally, some WPAN (Wireless Personal Area Networks) applications, such as Bluetooth or Zigbee are also included in some UE, as well as mobile broadcast tuners/decoders (DVB-H, DMB-T, ISDB-T, MediaFLO) and GPS receivers for allowing users to utilize their personal devices as navigators. Also, GPS signals are used in some CDMA base-stations for clock synchronization and in broadcasting transmitters for synchronizing single frequency networks (SFN).

As a consequence, the equipment has to deal with heterogeneous wireless access networks, different in terms of coverage, access technologies, bandwidth, power, data rate, and latency. Besides, developments for new base-stations should support portions of evolving standards, at least the anticipation to modify hardware components and to reconfigure (or upgrade) the software. A multimodal device is a device capable of coping with a number of different standards and applications, supporting different modes operating at different access techniques, data rates, powers, sensitivities, modulations, and codes.

Even though some standards share the same or similar frequency bands for signal transmission, others are definite for transmission at separated frequencies. The International Telecommunications Union (ITU) has defined a set of possible frequency bands for standards deployment, and among these frequencies each national spectrum regulation body decides where to operate within their influence area, usually in coordination with other countries. As a consequence, the same standards may be deployed at different frequency bands around the world. This fact is an additional challenge for

mobile terminals designs looking at worldwide commercialization, forcing to consider multiband aspects in equipment design. In this way, *multiband* may be defined as the capability of the handheld devices or the base-stations (or transmitters in broadcasting) to simultaneously operate in multiple frequency bands.

Furthermore, apart from the operating band, another aspect also affecting multiband designs is the operational bandwidth, already defined in each standard. Some new standards have not defined a fixed value, as this bandwidth is *scalable*. While the first multiband approach, the one related to the transmission band, poses a challenge to be solved by means of physical RF devices, the second approach (scalable or multiple operation bandwidths) may be also dealt with by software-based methods.

Multimode-multiband capabilities could be provided from implementations embedding RF and SDR techniques. If the programmable blocks are made as flexible ASICs (i.e., including analog circuits digitally programmable or tunable), they allow the real-time adjustment of the behavior of some subsystems, as well as providing flexibility in designs. However, when it is possible to build them as software, flexibility, and adaptability are improved, and even some functions may be easily upgraded following market or standards evolution. Hardware reuse favors smaller printed circuits, as well as the returns from design investments.

In [23] and [9] some factors affecting coexistence of FDD (Frequency-division duplexing) and TDD (Time-division duplexing) systems are reported. Some prominent aspects are:

- *Transmitter out-of-band and spurious emission levels (transmitter linearity)*: non constant-envelope modulations and, moreover, OFDMA and CDMA access techniques require the use of linear power amplifiers, which are less power efficient. To increase the efficiency, it is necessary to sacrifice linearity, so the linearity needs to be balanced against power efficiency. This sacrifice, along with the intrinsic non-linear behaviour of the power transistors, produces both in-band distortion and out-of-band interference.
- *Receiver selectivity*: If it is not possible to have a sufficient suppression of out-of-band emissions in the transmitter, the receiver selectivity needs to be enough to evade these *out-of-band perturbations*. Besides, some parts of the spectrum are application-saturated (i.e., 2.4 GHz), so the receiver has to fit, as exactly as possible, the wanted frequency band.
- *Receiver blocking*: Signals from adjacent channels may reach the demodulator, reducing SNR (signal-to-noise ratio) and sensitivity. If the power received from adjacent channels is strong enough, then they can also block the input of the receiver from receiving the wanted signal.
- *Antenna discrimination*: When using multiple antennas in multiband transceivers, significant *coupling* between the antennas may be produced according to the layout of the antennas and their radiation diagram. What is beneficial is the use, when possible, of directional antennas and/or the use of different types of polarizations.

The rest of this chapter is organized into the following sections. In the [next section](#), an overview of the key structures of digital front-end for multiple bands and multiple modes

will be presented. [Section 5.3](#) is devoted to pointing out some considerations regarding the analog-to-digital conversion for multimode and multiband applications. [Section 5.4](#) outlines preselect filters and LNA (low noise amplifier) interface in multimode-multiband transmitters. Power amplifier architectures in multi-standards will be addressed in [Section 5.5](#). The [final section](#) of this chapter will give a summary and further discussions.

5.2 Receiver architectures

The structure of the receiver digital front-end (DFE) may be diverse [15][14], and it is not only related to the fact of having a multimode or multiband operation. In digital implementations, multimode affects mainly the baseband processing, while the multiband operation is closely related to the diplexers, combiners, filters, mixers, or oscillators. Filters in the DFE may be a wideband structure with tunable or parallel filters (in a number according to the frequency bands) [2][18].

Even though the transmitter section of the handsets, and specially the power amplifiers, are the most power consuming elements, during most of the time the handsets are working in reception, and the receiver consumes around 66 percent of the battery charge [21]. Besides, handset designs have to result in affordable products for worldwide commercial success. These facts emphasize the significance of an accurate selection of the basic receiver architecture.

Superheterodyne, Zero-IF (or Homodyne) and Low-IF receiver structures have been historically the most used to design radio terminals. The selection criteria depend on the manufacturer, and have to balance among cost, size, reliability, or usability of re-designs.

Superheterodyne receivers are good for sensitivity and selectivity, gain is distributed over several amplifiers (so stability is increased), and they make the hardware upgrade easier if the product has to be re-designed to other frequency bands, i.e., by inserting specific RFICs (radio frequency integrated circuits) according to different countries' operation. However some parts are difficult to integrate, as is the case with high-frequency filters (especially if they are selective *SAW* –surface acoustic wave- filters) and oscillators. Furthermore the number of components makes it difficult to reduce the power consumption and may increase spurious responses produced by local oscillators (LO).

The two-stage Superheterodyne (or dual band receiver) is a solution used in some SDR designs that provides good performance at the expense of slightly increased complexity and implementation cost.

The merit of the Zero-IF (Direct Conversion or Homodyne) receiver architecture is the reduction of hardware components. Neither image reject filter nor IF stages are necessary (this structure does not have image frequency interference), so becoming cheaper and easy to implement in a small circuit area. Additionally, the power consumption is low. The drawbacks are related to DC offsets and leakages, which complicates the implementation. The LO leakage, due to an imperfect isolation between LO and both input port of the mixer and the LNA, produces a feed-through to the LNA and the input port of the mixer, creating a self-mixing problem that generates DC offsets at the mixer output, so

producing saturation of the following amplification stages in the receiver chain. Besides, second-order distortion can be down-converted in the mixer, crossing the channel filter and appearing, placed around DC, at the ADC (analog-to-digital converter) input. In this way, balanced mixers are better than single-ended ones because of their improved linearity and isolation.

Another issue may be due to antenna couplings: as the LO oscillator frequency is the same as the carrier frequency, radiations from LO to antenna could interfere with the received signal. Additionally, the lack of an IF stage may result in an increase of the low-frequency flicker noise appearing at the centre of the band ($1/f$ noise).

Constant DC offset can be compensated by measuring it when the transceiver is in the transmit mode (no reception), or during special TDMA frames in some standards. This measurement is utilized to subtract the offset value during reception. In FDD operation, the pre-select filter can be used to attenuate the LO leakage before the down-conversion in order to reduce the stringent linearity requirements for the mixers [12] [5].

The Low-IF structure is a balance between the advantages of the Superheterodyne and those of the Zero-IF receiver architectures. The idea is to use a Low-IF frequency, instead of Zero-IF, in order to cope with the aforementioned DC offset problem while preserving a structure with lower complexity than the Superheterodyne. The signal path to the ADC is AC coupled, thus eliminating DC offset problems. Because the image rejection is made at Low-IF, after down-conversion, any change in the transmission frequency may be easily solved by changing accordingly the LO, thus keeping the same IF.

Like the Zero-IF, the Low-IF receiver implementation is facilitated by the fact that in most standards the transmit ACLR (adjacent channel leakage ratio) specifications don't allow strong signal levels in adjacent channels. Besides, operational bandwidths in cellular equipment and multimedia broadcasting range from hundreds of kHz to a few tens of MHz (up to 8 MHz in multimedia standards), so the bandwidth requirements in the posterior ADC are not very high. In Low-IF the adjacent channel image rejection is usually made in the DSP section, after the ADCs. However, some implementations relax the DSP burden by including Hartley or Weaver image-rejection schemes before the ADC.

Both the Zero-IF and the Low-IF are attractive for developing low-cost solutions, and consequently a number of RFICs with a high level of integration have been commercialized since the 1990s. Apart from ICs suited for different standards of cellular telephony, others are tailored for handheld convergence of multimedia and cellular radio. The DVB-H or ISDB-T (Japanese mobile digital TV) front-end designers can take advantage of direct-conversion IC tuners which cover part of the UHF spectrum. These ICs include LNA, RF variable-gain amplifiers, I & Q down-converting mixers, Low-IF amplifiers, and bandpass filters. For DMB-T, the RF front-end integrated circuits cover the VHF-III and the L-band, with very similar block diagrams to the devices for DVB-H or ISDB-T.

Sometimes confused with the Low-IF structure, the Wideband-IF receiver is based on a dual conversion architecture in which the tuned signal is first down-converted from RF to IF with a quadrature mixer (as in Low-IF), and later from IF to baseband by means of an image rejection mixer (IMR, a kind of quadrature mixer). In a simplified approach, we could say that in the Low-IF receiver the IMR is done (or partially done) after the ADCs,

while the Wideband-IF structure places the IMR mixer before the ADCs. Wideband-IF implementation also takes advantage of existing RFIC and ASIC solutions, along with SAW filters.

Direct Conversion (Zero-IF), Low-IF, and Wideband IF use image reject filters, which as the I&Q demodulators are also used in Heterodyne or Dual-Band receivers, must have well-compensated I and Q branches. In UTRA/FDD, gain and phase mismatches can be compensated taking advantage of the pilot symbol-assisted channel estimation scheme [20].

Despite having different antialiasing requirements, all the previous receiver architectures are candidates to be used for subsampling receivers, also named IF-sampling receivers even when the Direct Conversion structure is used. In [17] is presented a summary of sampling strategies and antialiasing filters complexity for various receiver architectures. For Low-IF and Direct Conversion, a simple low-pass filter may perform both the functionalities of blocking adjacent channels and the antialiasing filtering before the ADC converters. Also in [17] it is shown that if the distortion due to aliasing is specified to be less than 1/2 LSB of dynamic range, the required stop-band attenuation in the antialiasing filter becomes stronger as the number of bits in the ADC increases. Therefore, as IF sampling architectures operate with signals with slightly enlarged PAPR (Peak-to-Average Power Ratio), higher dynamic ranges on the ADC are required and, consequently, the attenuation requirements for the antialiasing filters are scarcely higher in receivers using IF-sampling techniques.

In a subsampling receiver the sampling frequency is much lower than the carrier RF frequency, so the operation of actual ADCs is possible. The most usual situation for subsampling is to use a one-stage Superheterodyne receiver, getting the samples at IF (IF-sampling). Subsampling means that the sampling rate can be much less than the IF carrier; however it has to be greater than twice the signal bandwidth to avoid destructive aliasing (Nyquist sampling criterion). In IF-sampling the I and Q components are extracted by using digital mixers, including advanced signal processing techniques for sampling bandpass signals. Hence, neither noise nor distortions are of concern in such mixers, except for the ADC quantization noise. Quantization noise is decreased as either the clock frequency or the number of bits are increased (better than the previous expression “number of bits”, it is more realistic to use the ENOB – effective number of bits- a parameter that only considers the number of bits producing quantization levels above the noise floor).

The detailed design of the abovementioned general structures has to follow the usual requirements for single-mode (not multimode) receivers, with few exceptions. The band select filter has to remove unwanted spectrum, the LNA has to be as noiseless as possible, showing high $IP3$ as well as the mixers, and local oscillators have to be as stable (jitter free) as possible. Related to the SNR and the distortion of the cascaded receiver subsystems it is the receiver sensitivity that determines the minimum detectable signal, or better, the minimum allowable signal that the receiver may detect while keeping the desired BER (bit error rate). In a multi-standard environment, the minimum sensitivity is imposed by the most sensitive standard.

From [8], [10], [16], [19], [21], and [22], and by also using diverse technical and commercial documentation from different manufacturers, as well as measured or simulated data presented in technical conferences and seminars, Table 5.1 has been put

Table 5.1 Overview of parameter dispersion among standards.

	LTE	UMTS	WiMAX (Mobile)	CDMA 2000	GSM	EDGE	BLUETOOTH	DVB-H	ZIGBEE	WLAN 802.11	GPS	GALILEO
Channel BW	1.25–20 MHz	3840 kHz	1.25– 20 MHz	1228 kHz	200 kHz	270 kHz	1 MHz	6 MHz (5–8)	5 MHz	22 MHz 20+20 in 802.11n	2.046 MHz	4.046 MHz
SNR (dB) in ADC		74		83	85	85						
Receiver SNR (dB)	5–30	–20	3–20		10		20	22		5–20	–30	
Sensitivity (dBm)	–123	–117	–100	–105	–102	–90	–70	–94	–85	–82	–150 –160	–150 –160
Max. Inp. Level (dBm)	–15	–19	–30, –45		–15		–20			–30, –10	–15 (–123)	
ADC sampling rate (Mps)	30.72	61.44 153.6 for OSR=40	22.4 140 (*) – 160(*)	30.72 76.8 for OSR=64	26 OSR=48	26 OSR=65	8	16 70 (*) –150 (*)	12	40 150(*) –230 (*)	8–64	
ENOB	10	8	10–12	14	12–14	12–14	6–12	8–11	12	8–14	7	7

(OSR: Oversampling ratio for using $\Delta\Sigma$ converters).

Dynamic ranges: GSM: 87 dB, 802.11: 50–70 dB, Bluetooth: 50 dB.

Note(*): results obtained by computer simulation in [16], considering signal interferers and for different ratios between signal and interferers.

together. The objective of this table is just to give, at a glance, an overview of parameter dispersion among standards likely to be embedded in a multimode-multiband handheld device. So, some values are just an indicator, such as the maximum input levels, sampling rates or the ENOB values. Other parameters, such as the SNR for applications supporting adaptive modulation, depend on the modulation format. Also, sensitivity depends on the wanted BER. Notice that, due to the dispersion of the documental sources used, some parameters (such as the sensitivity and the SNR in the detectors, or the desired SNR in ADCs and the ENOB) may be more or less unmatched.

5.3 Analog-to-digital converters

The ADCs have to cope with the required signal bandwidth and dynamic margin, and to work with enough ENOB to accomplish the SNR requirements. The programmable (or variable) gain amplifiers, PGA, coming before ADC in the signal acquisition chain, have to permit the operation of the multiple standards with different signal amplitudes, adapting these amplitudes to the fixed dynamic range of the ADCs. To operate with signal amplitudes not suited for the ADC range is equivalent to a loss in the number of bits or, conversely, it produces overload distortion, depending on the balance between maximum signal variation and ADC dynamic range. As an example, in WCDMA, TDD mode, the PGA has to permit up to 40 dB of variation.

In multimode-multiband, the dynamic range of the ADC has to be assessed more carefully than in single-mode, because of the *signal superposition* when two or more standards are received simultaneously. As a result, the PAPR is higher than that of each individual band (if the sampling is made at baseband, the PAPR is 3 dB less than that of the IF signal [17]). Apart from this, if the multiband bandwidths are significantly different, the channelization filters prior to the ADC should be consequent. These filters may result in a big bank if the number of standards is high. Hence, instead of a filter bank it is preferable, regarding circuit cost and size, to use a unique filter capable of coping with the highest bandwidth among the possible received bands. If this filter is not tunable, or only scarcely tunable, a number of received signals will cross it, leading to the paradox that, at a given moment, the sum of the energy of the signals from the unwanted standards, even though narrowband, could be higher than the energy due to the wanted signal (or simply, to be significantly elevated with regard to the wanted signal energy). This fact could produce overload distortion in the ADC. Besides, in mobile equipment the kind of fading is diverse depending on the variable physical environment. As a consequence, the received signal amplitudes fluctuate, so introducing additional challenges if fading speed is too high for adequate PGA operation.

From the aforementioned situation, one would be tempted to use ADCs with the dynamic margin as large as possible. Nevertheless, this idea has two pitfalls. One is the increase of the quantization noise, because only the used bits participate in the computation of the SNR due to quantization noise. To spoil some bits (quantization levels) for not matching closely enough the dynamic range is equivalent to operating with an ADC

having a reduced number of bits than it actually has. The other is power consumption, typically increased with the dynamic range (along with the clock frequency of the ADC).

If the Nyquist bandwidth exactly fits the system bandwidth, that is, the sampling frequency, f_s , is exactly twice the maximum signal frequency, f_m , then the signal to quantization noise ratio is

$$SNR = 6.02n + K(\text{in dB})$$

where n is the number of bits and K a parameter dependent on the input signal statistics. This SNR value is obtained assuming no jitter in the ADC clock. On the other hand, if the sampling frequency is augmented over the Nyquist frequency, this SNR is increased by the processing gain, defined as:

$$\text{processing gain } (PG) = 10 \cdot \log(f_s/2f_m).$$

So resulting in a SNR value of

$$SNR = 6.02n + K + PG.$$

Oversampling has not only benefits in reducing the quantization noise (by increasing the processing gain) but also it simplifies the implementation of the antialiasing filters whose transition band may be relaxed. However, oversampling has a counterpart in the power consumption: it implies to clock the ADCs at higher values, thus making these devices more power hungry.

In the previous SNR equations what has been supposed is a constant amplitude signal, fitted to the ADC dynamic range. If the PAPR of the signal is not 0 dB (non constant envelope signals), then the SNR becomes, approximately,

$$SNR = 6.02n + K + PG - \alpha$$

where α is the PAPR.

The crest factor reduction techniques presented in [Chapter 11](#) of this book reduce the PAPR, so they also have the advantages of fitting the signal variation to the ADC dynamic margin. The price may be a loss in data throughput because of the need to transmit additional information to the receiver (i.e., partial transmit sequences), to add redundancy in the number of bits (i.e., codewords), or to distort the signal (i.e., clipping) with the consequent increase in the symbol error rate (SER).

Among the different kinds of ADC converters the ones based on parallel, pipelined, and folding architectures, as well as the delta-sigma converters, are the most used in SDR designs. Conversely, the successive approximation architecture, very usual in low-frequency designs, is not suitable for SDR because of its low conversion speed. Pipelined converters allow sampling rates up to hundreds of MHz. Folding and interpolating converters allow conversion rates up to 3.6 Gsps, with typical ENOB values of 8 bit (using a 0.18 μm CMOS process, a manufacturer has recently presented a 12 bits, 3.6 Gsps converter).

Delta-sigma ($\Delta\Sigma$) converters take advantage of the beneficial oversampling effects on the quantization noise, apart from other structural effects capable of displacing the noise to other bands. Although they may be built at low cost, with acceptable power

consumption, they are relatively less used in digital radios using the IF-sampling structure, mainly because of some dynamic range restrictions and the necessary high-frequency clocks. Unlike other kinds of converters, such as the pipelined, in $\Delta\Sigma$ converters the oversampling process is not just an added value for increasing the SNR. Instead, in $\Delta\Sigma$, oversampling is also mandatory to compensate for the reduced number of quantification levels inside the converter loop (which may be as low as one), thus becoming competitive with other kind of converters. The augmentation of the sampling frequency leads to lower signal variation between consecutive samples of a band-limited signal, so it is possible to operate with a reduced number of quantization levels. On the other hand, this reduced number of quantization levels, along with the sampling frequency, may produce slope overload errors in the converter.

Some high-frequency $\Delta\Sigma$ converters have been reported [3], and current nm-CMOS technologies [6], as well as SiGe BiCMOS, allow converters [7] at GHz frequencies. However they have still to benefit from undergoing research, and most of the commercialized $\Delta\Sigma$ converters operate below 10 Msps, so their main applications are in low-frequency instrumentation and in audio-frequency equipment. In the last few years, some continuous-time $\Delta\Sigma$ converters (capable of operating at sampling rates up to 50 Msps and beyond) have appeared on the market, so allowing their use in IF-sampling receivers suited for modern wireless communication standards.

5.4 Preselect filters and LNA interface for multiband receivers

The diplexers (or duplexers, if the frequency band is the same for both transmission and reception), along with the architecture of the RF preselect filters, are the topological receiver aspects more related to the multimode-multiband reception. Indirectly, the preselect filter organization also affects some arrangement conditions of the down-converters [18] [2].

Diplexer filters for multiband receivers include band-select filters that, once isolated from the transmit signal, split the output signal among different branches with each one being fitted to a different band. This process is not trivial because it supposes to simultaneously manage both filter selectivity and antenna efficiency. The use of the same antenna means that the different received standards operate in closer frequency bands (in a spectrum as wide as the antenna allows). The design of very frequency-selective filters (high Q) is difficult in passive filter implementations (because of the resulting size), and it is power-consuming in active filters. To allocate a single multiband filter after the antenna (Figure 5.1(a)) makes the LNA design challenging because, apart from the basic need for low noise figures, it has to operate in a huge frequency band, hence risking the amplifier stability.

An alternative is the use of a tunable filter (Figure 5.1(b)) after the antenna. This is realistic if the transmission bands of the different standards are not very separated along the spectrum. Conversely it would be necessary to change, accordingly to the band, not only the filter adjustable parameters but also the filter technology, in order to keep the filter's selectivity (i.e. for 800 MHz the filter technology is not necessarily the same as

that for 5 GHz). The typical use of good lumped-elements filters is up to 1 GHz, helical filters up to 2.5 GHz, SAW filters up to 3.5 GHz, ceramic filters up to 5 GHz, and cavity filters up to tens of GHz. Microstrip filters operate beyond hundreds of MHz to tens of GHz. Obviously, tunable filters are not practical if it is necessary to change among different technologies.

A bulky alternative is the use of a bank of switched filters (Figure 5.1(c)), with the appropriate frequency response and technology in each branch. This solution permits the separated design of the filters, and the switch may be made by means of PIN diodes, HEMT switches, or by using new MEMS (micro-electro-mechanical systems) devices. With regard to Figure 5.1(c), if the carrier frequencies are not close (i.e., GPS and IEEE 802.16n) and the unique antenna in the figure is not wideband enough, it has to be separated in different antennas. The switch is removed in the scheme of Figure 5.1(d), where a unique broadband LNA amplifies the signal prior to its separation in different frequency branches. Besides, in separate multiband developments it would be necessary to make a different number of down-conversions along the branches, selecting the most suitable filter technology after each down-converter. A typical situation may be seen in receivers covering three bands (2.3/2.4–2.5/2.5–2.7 GHz, 3.5–3.7 GHz and 5.13–5.35 GHz). This scenario may correspond to a receiver for IEEE802.11a/b/g (or n). In such designs, the VCO output (i.e. around 3.6 GHz), may be separated in two branches, one of them with the VCO frequency divided by two. The outputs of the three band select filters are translated to a common IF, by using one or two down-conversion stages according to the band. A similar situation happens in receivers with GPS and UMTS coexistence, or GSM and WiFi among other combinations that may also include DVB-H, ISDB-T or DMB-T.

Figure 5.1(e) shows an alternative that translates the signal filtering to IF. It is easier and more suitable for low-size (or low-cost) developments; nevertheless the image frequency, as well as the interchannel interference (ICI) may present a problem due to the lack of the pre-select filter.

For Low-IF single conversion tuners, the multimode-multiband architecture of Figure 5.1(f) has been commercialized in integrated CMOS devices for multimedia receptors including, in several bands, reception of digital multimedia broadcasting (DMB) (terrestrial and/or satellite), dual-band digital audio broadcasting (DAB), and FM. The building blocks of these ICs include LNAs, a PGA, a down-conversion mixer, a bandwidth-adjustable image rejection bandpass filter, an IF programmable gain amplifier, and a VCO.

Before the IF mixer or the DAC converter, depending on the selected receiver structure (Superheterodyne, Zero-IF . . .), it could be necessary to have an adjustable gain amplifier to fit the different signal amplitudes to the mixers or DACs dynamic margins, as has been stated above. The received signal amplitude not only depends on the standard, but also on the transmit power, the transmitter distance, or the preselect filter structure. In fact, if the filters allow the pass of a relatively wide range of frequencies, the signals in adjacent bands act as interferers to the wanted signal (enlarging the PAPR). This fact may increase intermodulation in mixers, saturation in amplifiers, overload distortion in ADCs, or, even, to prevent (*blocking*) the receiver from receiving the wanted signal.

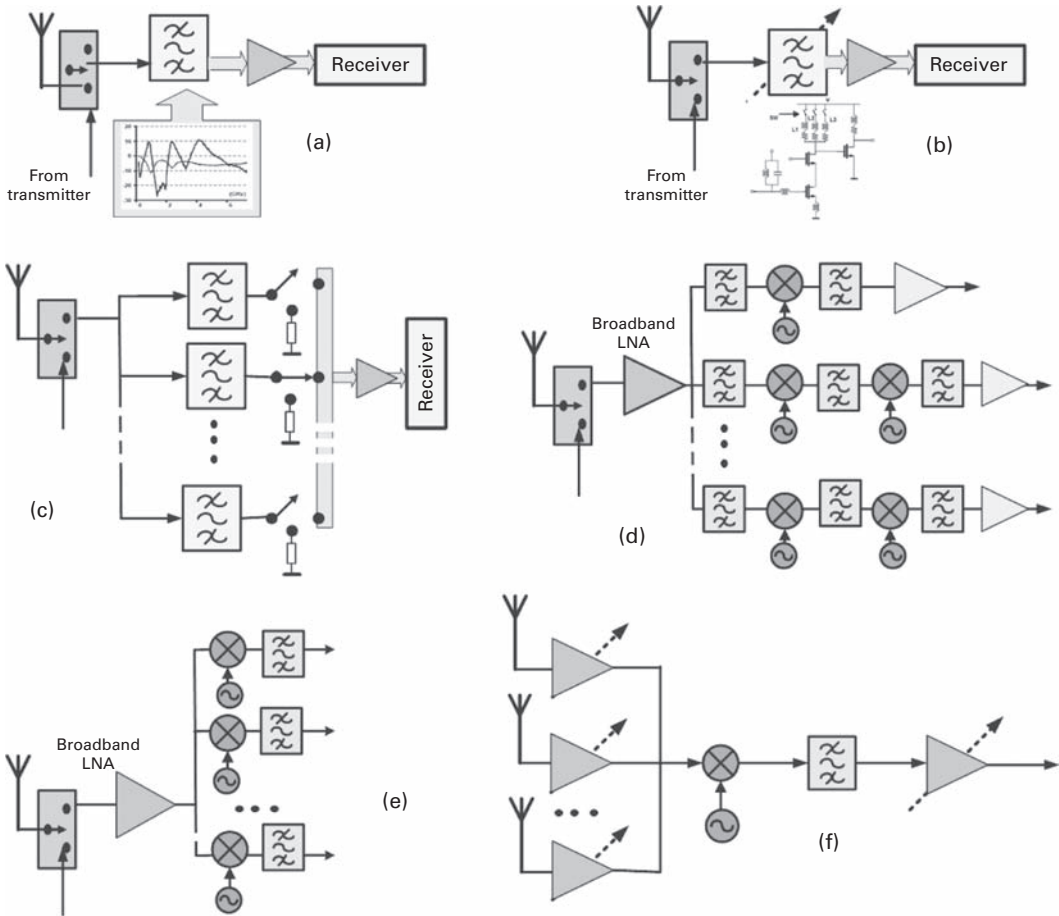


Figure 5.1 Preselect filter structures.

5.5 Power amplifier architectures in multimode-multiband transmitters

Among the transmitter's subsystems, the unavoidable power amplifier (PA) is the most sensitive subsystem for an affordable development of multimode-multiband designs. Apart from not being cheap, the PA is the most power consuming element in the transmitter chain, and the most critical one for transmission linearity. While in handheld devices, battery operated, the PA power consumption is a key factor for the autonomy time (a critical aspect for commercial success), in base-stations or in broadcasting transmitters the PA efficiency is the most significant factor in the electrical bill, including the feeding of both transceivers and air conditioners.

In previous chapters what has been introduced is some PA issues regarding the PAPR produced by non constant envelope modulations and, moreover, for the CDMA and OFDM (or SC-FDMA) access techniques. On one side, the non-constant amplitude

signals make impractical the use of efficient switched power amplifier structures, even when the signal bandwidth allows their usage (the use of polar transmitters in new standards is still addressed by undergoing research). Class AB are the usual PAs, although recently more efficient Doherty structures have been commercialized. On the other side, the high PAPR values (in the order of 15 dB for a high number of carriers in LTE, or for multi-carrier CDMA) make it impossible to bias the “linear” PA at power-efficient back-off levels.

Conversely, if the PA is forced to have low back-off operation, then the out-of-band emission (defined in standards as the ACLR, adjacent channel leakage ratio) increases, so creating two drawbacks. First, the transmit power could become not compliant with the mandatory spectrum mask for product certification, and second, a band transmitting with a bad PA within a multiband scenario may become a blocker to reception of other bands. Apart from the out-of-band emissions, also in-band distortions, measured as the EVM (error vector magnitude), appear in low backed-off PAs. As a result, the signal constellation is degenerated and the data throughput shrinks.

The previous chapters of this book have also introduced the merits of the crest factor reduction techniques and, moreover, of the PA *linearizers*, the latter not requiring a sacrifice in data throughput. In multimode transmitters there is no conceptual or structural difference with regard to the linearization techniques, only some quantitative matters are different. Because of the augmented number of simultaneous signals to be transmitted from the different standards, in the case of transmitters using an unique PA (i.e., a broadband PA), the PAPR might be more increased, so magnifying the linearization challenges. And if the transmissions are not made simultaneously (i.e., sequential), then the PA has to manage very diverse PAPR values. Such a management could shrink the PA linearity or the power efficiency.

5.5.1 PA structures for multimode-multiband handsets

Transmit power in handsets for mobile communications depends on the communication standard and the number of Power Classes implemented in the user terminal. Maximum power in GSM handsets is 39 dBm for Class 2, 37 dBm for Class 3, 33 dBm for Class 4 (typical), and 29 dBm for Class 5. In DCS1800 power ranges between 30 and 36 dBm. The UMTS maximum transmit powers, are 33 dBm (Class 3), 27 dBm (Class 2), 24 dBm (Class 3), and 21 dBm (Class 4). In Mobile WiMAX and LTE, the specified power is 23 dBm. Other applications transmitting from the handheld antenna may be Bluetooth, whose power modes depend on the distance, i.e. 20 dBm (100 m), 4 dBm or 0 dBm (10 cm), and WiFi. In the case of standard IEEE 802.11 g, the maximum transmission powers are 16 dBm (OFDM mode) and 18 dBm (CCK).

Some commercialized RFICs provide most of the blocks within the transmitter front-end schematics, including PA drivers, but keeping open the choice for selecting the kind of PA [25]. For multiband handsets, the different PA design alternatives may be grouped in three main strategies:

- *To switch among different matching networks*, each one tuned at a different frequency band. It reduces the development cost because the power transistor is unique, but it does

not allow simultaneous transmission in different bands. The PA linearization may be based on DPDs (Digital Predistorters). The main challenges of this structure derive from the control of the spurious emission due to matching network switches, as well as the compliance with transient masks introduced in new standards in addition to the classical spectrum mask. Apart from PIN diodes and HEMT switches, what recently has appeared is MEMS and what is under research is the use of metamaterials (artificial materials engineered to provide special response to electromagnetic waves) to develop switchable directional couplers [11]. The use of power varicaps (variable capacitance diodes) should permit a smooth variation of the matching network, without switching transients.

- *To embed independent PAs in the same chip.* Currently this is the most widely commercialized solution for handheld devices being common RFICs for different WiFi bands (2.4 and 5 GHz), for GSM900 and DCS1800, or GSM and UMTS [1][4][13]. Due to the independent operation of the PAs, it is not especially complicated to design them, as well as their linearizers. And regarding the abovementioned transmit power levels for handhelds the size of the power device is not large, despite the two embedded PAs. A challenge for the manufacturer of the integrated circuits is to avoid coupling and intermodulations between the different PAs buried in the same semiconductor substrate. The spurious emission depends on the used duplexers, circulators, power combiners (i.e., Chireix) or power switches, the last not allowing simultaneous transmission.
- *To use a wideband PA and to separate the different applications by input and output filtering.* The linearization of this structure is more difficult because, unlike the first case, the filters operate simultaneously and the amplifier operation is for multiple bands at the same time. Besides, the PAPR values depend on the number of concurrent active bands driving the amplifier. The wide bandwidth operation, necessary for multiband amplification, makes stability an issue and, consequently, gain has to be sacrificed for PA stability. The advantages of this structure are the use of an unique power transistor device and to avoid switches in load matching.

5.5.2 Multi-amplifier arrangements in power transmitters

With the general name of power transmitters, we include both the base-stations for mobile telephony and broadcasting transmitters, despite their different transmit power.

Typical maximum transmit power in UMTS, Mobile Wimax, and LTE base-stations is 20 W (43 dBm), although some units allow up to 40 W (46 dBm). The GSM base-stations Classes 1 to 3 (from 320 W to 80 W) are not normally implemented, being usually Class 5 (20 W) and even Class 4 (40 W). Class 6 is 10 W, used in special cells. On the other hand, multimedia broadcasting transmitters, including different standards (DVB-T, DVB-H, DAB, ISDB-T, ...) operate at transmit powers ranging typically from 50 to 800 W.

According to these power values, the design on the PA for base-stations is less challenging than the design for broadcasting transmitters. Both have in common the need for efficient schemes, combining the power efficiency with the linearity

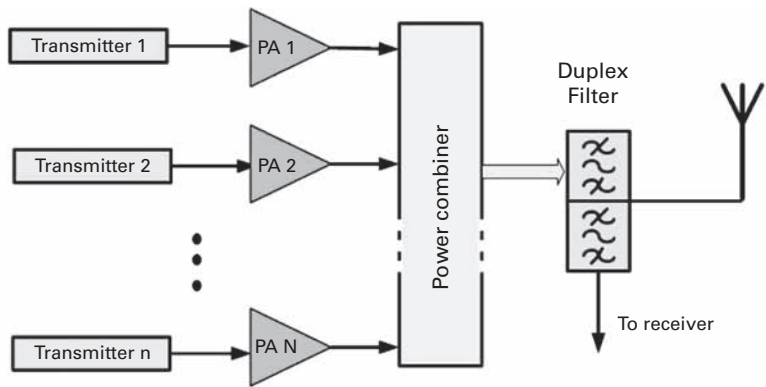
requirements as introduced in previous sections. But in broadcast transmitters the need for efficient power combiners is higher because of the higher transmit power. Furthermore, the ACLR requirements in mobile broadcasting standards are also severe. As an example, in ISDB-T, the Japanese norm ARIB Std B31 specifies an ACLR of -54.4 dBc for a difference of 3 MHz from the center frequency, increased to -74.4 dB/10 kHz if the transmit power is 2.5 W. Regarding the PA bandwidth, transmitters for different standards of mobile broadcasting have to cope with a maximum bandwidth of 8 MHz, which may be relaxed in DAB or DMB-T. Hence, the most stringent situations regarding PA bandwidths are in transmitters for LTE or Mobile WiMAX.

Introducing some power values, achieving a quasi-linear power of the order of 400–600 W (as an example), requires a number of PAs (operating at high back-off levels, for the sake of linearity) whose outputs have to be power combined. This bank of lined-up PAs may be driven by 25–30 W amplifiers, pre-driven by other 3 W amplifiers. Obviously, power detectors and temperature controllers are important in such structures. By using traditional *power MOSFETs*, the number of transistors is around 20. Nowadays it is possible to obtain in the market devices based on Doherty PAs already connectorized (ready to use) with peak powers over 100 W, or to use amplifiers whose RF line-up is composed by GaN (Gallium Nitride) devices, capable of operating from 500 MHz to 8 GHz (some samples can reach higher frequencies) and to deliver 50–100 W with gains around 30 dB, or to provide 20–50 W (enough for a common base-station) with gains around 50 dB. Current power efficiency of GaN amplifiers is up to 60%.

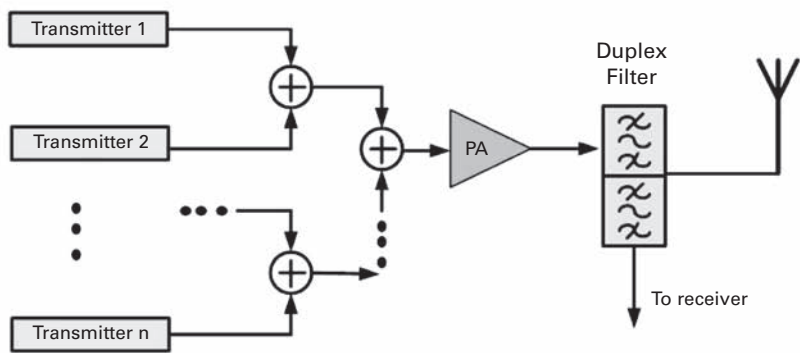
Normally the power transistors of the line-up operate in class AB, showing efficiencies lower than 20 percent, and needing big air- or liquid-cooled heat sinks. Cavity combiners are normally preferred to hybrid combiners due to the lower losses, even though combiner losses are dismissed as secondary in comparison with the inefficiency in the PAs.

The combiners referred above are employed for getting enough linear output power from the combination of individual amplifiers. This use is just at the PA design level, and these power combiners can be transparent if the PA is bought in an already connectorized closed box. Another application of the combiners, at the system level, is to send out several transceivers through a *single* antenna. For instance, in a GSM BTS, the number of transceivers may range, typically, between 1 and 16. Even, in multimode systems operating at different frequency bands, the common antenna may be multiband in order to reduce the cost and the environmental impact. In BTS for mobile telephony dual (GSM/DCS) or tri-band (GSM/DCS/UMTS) antennas are commercialized.

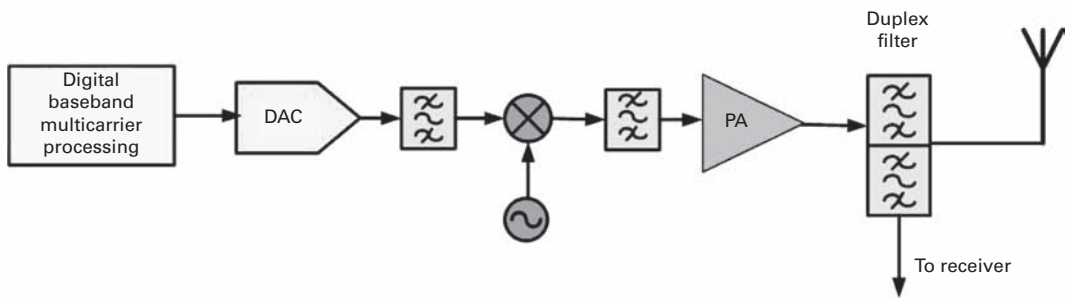
Power combiners are also used in broadcast transmitters with multimode transmission of different standards, such as DVB-T, DVB-H or ISDB-T, as well as in multi-carrier transceivers for base-stations. Indeed, in GSM systems each carrier is divided into eight time slots (TDMA), so it may allocate eight full-rate users. In the base-stations these carriers are separated 200 kHz (FDMA), thus allocating 124 carriers in a 25 MHz bandwidth. Multi-carrier transceivers allow several carriers to be processed by a single power amplifier in the downlink direction and by a single wideband receiver in the uplink direction. But multi-carrier transceivers are not exclusive to GSM systems, i.e. W-CDMA base-stations also uses multi-carrier in order to provide different services and applications, to increase user's coverage or mobility or to separate types of traffic. This is made by means of the joint transmission



(a) Traditional architecture in a BTS



(b) Traditional post-amplification architecture in a BTS



(c) Modern multicarrier architecture

Figure 5.2 Power combinations.

of different data packs, each occupying a 3.84 MHz bandwidth within different 5 MHz spaced channels.

In the first BTS transmitter designs, the output from the different PAs was power combined (Figure 5.2(a)), a solution showing important losses and heat dissipation requirements. Apart, good isolation between PAs through hybrid combiners was an issue to avoid intermodulation distortion (IMD).

Later, the need of amplifiers capable of coping with more efficient modulation formats, with increased PAPR values, increased the cost of the PA units. This fact boosted an alternative structure consisting of a previous low-power combination and a posterior amplification by using a single, “linear,” PA (Figure 5.2(b)). Current structures in multi carrier base-stations take advantage of digital signal processing (Figure 5.2(c)) to construct the multi-carrier signal at baseband, then extracting the composite digital signal to a DAC whose output, once up-converted, drives a unique broadband PA [24].

In the structures of Figures 5.2(b) and 5.2(c) most of the analog RF components are needed only once per antenna instead of once per carrier (Figure 5.2(a)). Besides, the loss of energy in the passive combiners is avoided because the carriers are combined at low power levels (Figure 5.2(b)) or at baseband by using digital processors (Figure 5.2(c)).

The digital architecture of Figure 5.2(c) needs a DAC. The same considerations about the use of the crest factor reduction techniques that were mentioned in Section 5.3 regarding the ADC converters may be repeated here for the DAC. Hence, designers have to consider in every new development if the PAPR due to the multimode operation could become large enough to justify the loss in data throughput.

The structures shown in Figures 5.2(b) and 5.2(c) have to amplify the combination of multiple signals with fluctuating envelopes. To overcome the resulting linearity-efficiency trade-off due to the increased PAPR, some linearizing structures were tested in the past, such as the Feed-Forward amplifiers. Nowadays the current state of the art in DSP and FPGA devices makes it preferable to use digital linearization techniques, such as the LINC (which requires two PAs) or the digital predistortion techniques presented in other related chapters of this book.

5.6 Conclusions

Portable and mobile communications equipment need to cope with mass market business to be commercially successful. Hence, its design has the additional challenge of coping with multi-standard operation in tight circuits embedded in the same user equipment, including capability for multimode-multiband functionality. This is different than multi-standard fixed equipment (i.e., base-stations), where a set of independent subsystems may operate in parallel, each one for a different standard, and sharing some common resources.

While in the receiver section of the mobile equipment Low-IF and Zero IF seem to be the most promising structures because of their simplicity and cost, the band-select filters are an active area of research, looking for smaller and highly selective tunable filters. Also research for cost-competitive high sampling frequency ADC and DAC is continuously ongoing.

The main challenge in the transmitter section is a consequence of the increasing need to achieve the ubiquity objective. This means that devices have to be battery operated, and their commercial success depends on batteries' lifespan, as well as the data-rate manageable by the devices. The lifetime of the batteries is very sensitive to the PA, whose design has to face the trade-off between power efficiency and linearity (closely related with the maximum data-rate). Hence, active research is covering from the basic technology level, developing new power transistors made of enhanced materials, to the system level. Hence, PA architectures, such as Doherty or Polar transmitters are being revitalized, as well as some linearizing structures.

Since DSP and FPGA are no longer seldom-used devices in communications equipment, digital linearizing solutions, such as the ones that will be detailed in the following chapters, may take advantage of the presence of these devices, overcoming some implementation restrictions usual in classical analog linearizers.

List of abbreviations (Chapter 5)

ACLR:	Adjacent Channel Leakage Ratio
ADC:	Analog-to-Digital Converter
ASIC:	Application-Specific Integrated cCircuit
BER:	Bit Error Rate
BiCMOS:	Bipolar-CMOS
BTS:	Base Transceiver Station
CCK:	Complementary Code Keying
CDMA:	Code Division Multiple Access
CMOS:	Complementary Metal-Oxide Semiconductor
DAB:	Digital Audio Broadcasting
DAC:	Digital-to-Analog Converter
DCS:	Digital Cellular Service
DFE:	Digital Front-End
DMB-T:	Digital Multimedia Broadcasting – Terrestrial
DPD:	Digital Predistorter
DSP:	Digital Signal Processor
DVB-H:	Digital Video Broadcasting Handheld
ENOB:	Effective Number of Bits
EVM:	Error Vector Magnitude
FDD:	Frequency Division Duplexing
FDMA:	Frequency Division Multiple Access
FM:	Frequency Modulation
FPGA:	Field Programmable Gate Array
GaN:	Gallium Nitride (Semiconductor)
GPS:	Global Positioning System
HEMT:	High Electron Mobility Transistor
IC:	Integrated Circuit

ICI:	Interchannel Interference
IF:	Intermediate Frequency
IMD:	Intermodulation Distortion
IMR:	Image Rejection Mixer
ISDB-T:	Integrated Services Digital Broadcasting – Terrestrial
ITU:	International Telecommunications Union
LNA:	Low-Noise Amplifier
LO:	Local Oscillator
LTE:	Long Term Evolution
MEMS:	Micro-Electro-Mechanical System
MOSFET:	Metal Oxide Semiconductor Field Effect Transistor
OFDMA:	Orthogonal Frequency-Division Multiple Access
OSR:	Oversampling Ratio
PA:	Power Amplifier
PAPR:	Peak-to-Average Power Ratio
PGA:	Programmable Gain Amplifiers
RFIC:	Radio Frequency Integrated Circuit
SAW:	Surface Acoustic Wave (filter)
SC-FDMA:	Single-carrier FDMA
SDR:	Software Defined Radio
SER:	Symbol Error Rate
SFN:	Single Frequency Network
SiGe:	Silicon-Germanium (Semiconductor)
SNR:	Signal-to-Noise Ratio
TDD:	Time Division Duplexing
TDMA:	Time Division Multiple Access
UE:	User Equipment
UMTS:	Universal Mobile Telecommunications System
VCO:	Voltage Controlled Oscillator
WiMAX:	Worldwide Interoperability for Microwave Access (Industry association)
WLAN:	Wireless Local Area Network
WPAN:	Wireless Personal Area Networks)
$\Delta\Sigma$:	Delta-Sigma converter

References

- [1] “Agilent Multi-mode Handset Manufacturing. Challenges and Solutions.” Agilent Application Note, 2005.
- [2] Bauernschmitt, U., Block, C., Hagn, P. *et al.*, “Concepts for RF Front-Ends for Multi-Mode, Multi-Band Cellular Phones,” *10th European Conference on Wireless Technology*, October 2007, Munich.
- [3] Bolatkale, M., Breems, L., and Makinwa, K., “High-Speed Sigma-Delta Converters,” *Proceedings ProRISC*, Nov. 2008.

- [4] Gard, K., Barnett, K., Dunworth, J., Segoria, T., Walker, B., *et al.*, "Direct Conversion Dual-Band SiGe BiCMOS Transmitter and Receive PLL IC for CDMA/WCDMA/AMPS/GPS Applications," *IEEE International Solid-State Circuits Conference*, 2003.
- [5] Gharpurey R., Yanduru N., Dantoni F., Litmanen P. *et al.*, "A Direct Conversion Receiver for the 3G WCDMA Standard," *IEEE Custom Integrated Circuits Conference*, May 2002.
- [6] Le Guillou, Y., "Analyzing sigma-delta ADCs in deep-submicron CMOS technologies," www.rfdesign.com. February 2005.
- [7] Hart, A. and Voinigescu, S. P., "A 1 GHz Bandwidth Low-pass $\Delta\Sigma$ ADC with 20 GHz to 50 GHz Adjustable Sampling Rate," *IEEE Radio Frequency Integrated Circuits Symposium, 2008. RFIC 2008*. April 2008.
- [8] Hueber, G., Stuhlberger, R., Holm, A., and Springer, A., "Multi-Mode Receiver Design for Wireless Terminals," *10th European Conference on Wireless Technology*, Munich, October 2007.
- [9] "Coexistence between IMT-2000 Time Division Duplex and Frequency Division Duplex Terrestrial Radio Interface Technologies around 2600 MHz Operating in Adjacent Bands and in the same Geographical Area." ITU-R M.2030 Report, 2003.
- [10] Kenington, P. B, *RF and Baseband Techniques for Software Defined Radio*. Artech House, 2005.
- [11] Lehmann, T., Hettstedt, F., and Knöchel, F., "Reconfigurable PA Networks using Switchable Directional Couplers as RF Switch." *10th European Conference on Wireless Technology*. October 2007.
- [12] Lie, D. Y. C., Kennedy J., Livezey D. *et al.*, "A Direct-Conversion W-CDMA Front-End SiGe Receiver Chip," *IEEE Radio Frequency Integrated Circuits Symposium*, May 2002.
- [13] Magoon, R., Molnar, A., Zachan, J., Hatcher, G., and Rhee, W., "A Single-Chip Quad-Band (850/900/1800/1900 MHz) Direct Conversion GSM/GPRS RF Transceiver with Integrated VCOs and Fractional-N Synthesizer," *IEEE Journal of Solid-State Circuits*, vol. **37**, no. 12, December 2002.
- [14] Mehta, J. L., "Transceiver Architectures for Wireless ICs," www.rfdesign.com. February 2001.
- [15] Razavi, B., "Design Considerations for Direct-Conversion Receivers," *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, vol. **44**, no. 6, June 1997.
- [16] Rodriguez-Parera, S., Bourdoux, A., Horlin, F., Carrabina, J., and Van der Perre, J., "Front-End ADC Requirements for Uniform Bandpass Sampling in SDR," *IEEE 65th Vehicular Technology Conference*, 2007.
- [17] Roupael, T. J., *RF and Digital Signal Processing for Software-Defined Radio. A Multi-Standard Multi-Mode Approach*. Elsevier, 2009.
- [18] Ryyänen, J., Lindfors, S., Stadius K., and Halonen, K. A. I., "Integrated Circuits for Multi-Band Multi-Mode Receivers," *IEEE Circuits and Systems Magazine*, Second Quarter, 2006.
- [19] Silva, A., Guilherme A. J., and Horta, N., "Reconfigurable Multi-mode Sigma-Delta Modulator for 4G Mobile Terminals," *Integration, the VLSI Journal*, vol. **42**, 2009.
- [20] Springer, A., Maurer, L., and Weigel, R., "RF System Concepts for Highly Integrated RFICs for W-CDMA Mobile Radio Terminals," *IEEE Transactions on Microwave Theory and Techniques (MTT)*, vol. **50**, no. 1, Jan. 2002.
- [21] Tuttlebee, W., *Software Defined Radio. Enabling Technologies*. John Wiley & Sons, 2002.
- [22] Van Veldhoven, R. H. M., "A Triple-Mode Continuous-Time $\Sigma\Delta$ Modulator with Switched-Capacitor Feedback DAC for a GSM-EDGE/CDMA2000/UMTS Receiver," *IEEE Journal of Solid-State Circuits*, vol. **38**, No. 12, December 2003.
- [23] WiMAX Forum, "Service Recommendations to Support Technology Neutral Allocations FDD/TDD Coexistence." 10 April 2007.

-
- [24] Wright, A., Nesper O., “Multi-Carrier WCDMA Basestation Design Considerations – Amplifier Linearization and Crest Factor Control,” PMC-Sierra, Technology White Paper PMC-2021396. August, 2002.
 - [25] Zannoth, M., and Rühlicke, Klepser B.U. “A Highly Integrated Dual-Band Multimode Wireless LAN Transceiver,” *IEEE Journal of Solid-State Circuits*, vol. **39**, no. 7, July 2004.

Part II

DPD and CFR

6 General principles and design overview of digital predistortion

R. Neil Braithwaite

6.1 Introduction

Wireless communication systems use radio frequency (RF) signals to transmit data between base stations and mobile users. The RF power amplifier (PA) is located within the transmitter and is a key component of the down-link connecting the base station to the mobile. Power amplifiers tend to be either linear or efficient, but not both. Fortunately, an efficient power amplifier may be used within a digital transmitter if the nonlinear behavior of the PA is compensated using digital predistortion (DPD).

This chapter discusses digital predistortion techniques suitable for use in a digital transmitter. [Section 6.2](#) reviews the nonlinear behavior of a power amplifier and its effect on the output spectrum. [Section 6.3](#) provides an overview of digital predistortion avoiding equations for the most part. Details of the basic algorithms used appear in [Section 6.4](#). [Section 6.5](#) discusses some advanced topics in DPD.

6.2 Power amplifier nonlinearity

In the RF transmission of digital information, sampled data sequences are converted from digital to RF and amplified, as shown in [Figure 6.1](#). The digital input is denoted by $x(k)$ and the RF signals at the input and output of the power amplifier are denoted by $x_{RF}(t)$ and $y_{RF}(t)$, respectively. For a linear system most of the power transmitted by the base-station is confined to an allocated bandwidth, which is referred to as a channel. However, efficient power amplifiers operating close to saturation exhibit an undesired nonlinear behavior, generating a distortion spectrum several times wider than the allocated channel. This broadening of the transmitted spectrum is referred to as spectral regrowth.

The interaction of the input signal with the PA nonlinearity generates both harmonic and intermodulation distortion (IMD). Harmonics are polynomial functions of the carrier frequency and IMDs are polynomial functions of the digital modulation located near the carrier frequency. In most RF transmitters, a filter is placed after the power amplifier to limit the transmit bandwidth. The passband spans, typically, several channels: for

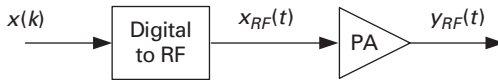


Figure 6.1 Digital transmitter with power amplifier (PA).

example, in a wideband code division multiple access system (WCDMA, [1]) the transmit band can be 60 MHz wide whereas the individual channels are 5 MHz wide. As a result, the transmit filter can be relied upon to remove the harmonics of the carrier frequency, but not the intermodulation products of the signal modulation. The latter needs to be compensated using a linearization approach. A popular choice is predistortion applied at the digital baseband.

The out-of-band and in-band distortions are regulated by specifications associated with the modulation format of the input signal. Out-of-band distortion is measured by the amount of distortion power falling into adjacent channels neighboring the allocated channel. This may be described as the adjacent channel power or the adjacent channel leakage. The ratio of the adjacent and allocated channel powers is referred to as the adjacent channel power ratio (ACPR) or the adjacent channel leakage ratio (ACLR). In general, the specifications make out-of-band distortion a key measurement of power amplifier linearity particularly for CDMA-based modulation formats.

In-band distortion increases the error vector magnitude (EVM). The EVM is the root mean square (RMS) difference between the demodulated signal prior to quantization and the desired digital signal. It is a key measurement within Orthogonal Frequency Division Multiplexing (OFDM) systems, particularly when the individual subcarriers are modulated using 64-QAM (quadrature amplitude modulation), and is degraded by impairments including power amplifier nonlinearities.

To illustrate IMD distortion, consider the PA response to a Gaussian noise input signal that is band limited to fill the allocated channel. The output spectrum is the sum of the linear component and several intermodulation modes (the third and fifth order IMDs are shown in Figure 6.2). The linear component, also referred to as a carrier signal, remains within the allocated channel whereas the distortion spreads into the adjacent channels. The IMD bandwidth increases with the polynomial order. However, a significant portion of the distortion lies within the allocated channel.

The RF transmitter may transmit several carrier signals occupying nearby channels. In such cases, the intermodulation distortion comprises components originating from the intermodulation within individual carriers as well as intermodulation between carriers. The effect of the latter is to spread the distortion over a much larger bandwidth, as shown in Figure 6.3.

The PA nonlinearity can be classified as memoryless or memory-based. For the former, the nonlinear modes are functions of the instantaneous input only. For the latter, the nonlinear modes are also affected by past input values. The PA memory alters the frequency response of the distortion generated and increases the number of coefficients needed to model the IMD. The memory effects exhibited by power amplifiers tend to be more discernible for wide bandwidth input signals, such as those comprising multiple carrier signals. In particular, dramatic memory effects are common with the Doherty

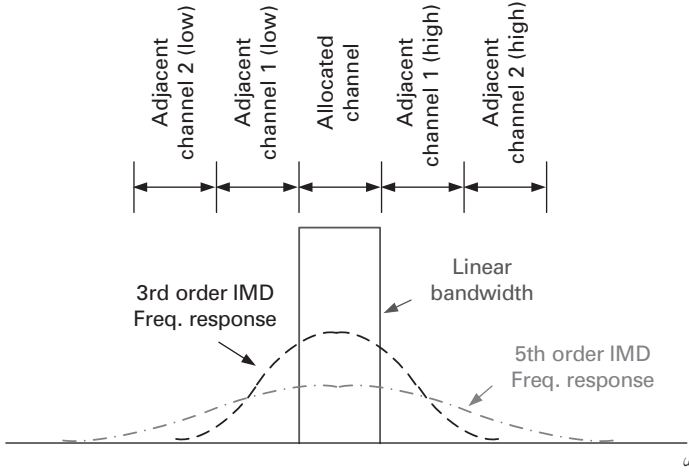


Figure 6.2 Output spectrum comprising the linear signal component, as well as the third- and fifth-order intermodulation distortion (IMD). Also shown are the allocated and adjacent channels. Note the increase in IMD bandwidth with the polynomial order.

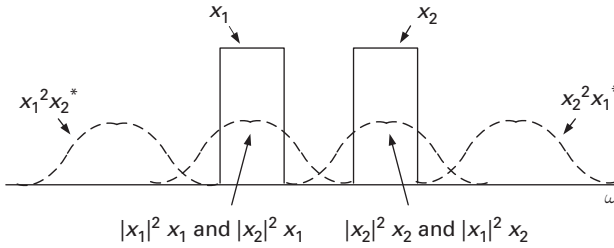


Figure 6.3 Linear and third-order IMD spectra for two carrier signals, x_1 and x_2 . The third-order IMD of the two carriers has six components: $|x_1|^2 x_1$, $|x_2|^2 x_2$, $|x_1|^2 x_2$, $|x_2|^2 x_1$, $x_1^2 x_2^*$, and $x_2^2 x_1^*$. The first two components are IMDs within the individual carriers, the others are IMDs between carriers. Note the increased bandwidth of the distortion spectrum arising from $x_1^2 x_2^*$ and $x_2^2 x_1^*$. Adapted from [2] © IEEE 2008.

amplifier configuration [3], which is popular in cellular basestations due to its high efficiency when operating at power levels near saturation.

The output of the PA is modeled as a nonlinear function of the input $x_{RF}(t)$:

$$y_{RF}(t) = F_{PA}\{x_{RF}(t)\}$$

where $F_{PA}\{\}$ is a nonlinear function whose response is controllable using adjustable complex coefficients b_i . The coefficients of the PA model $F_{PA}\{\}$ are selected, typically, to minimize the difference between the actual output $y_{RF}(t)$ and the modeled output $F_{PA}\{x_{RF}(t)\}$, as shown in Figure 6.4. That is, coefficient selection minimizes the expected squared error,

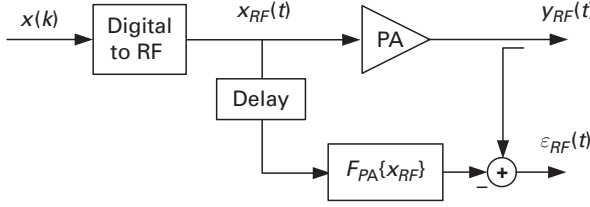


Figure 6.4 Modeling the RF output of a power amplifier $y_{RF}(t)$ using a nonlinear function of the RF input $x_{RF}(t)$.

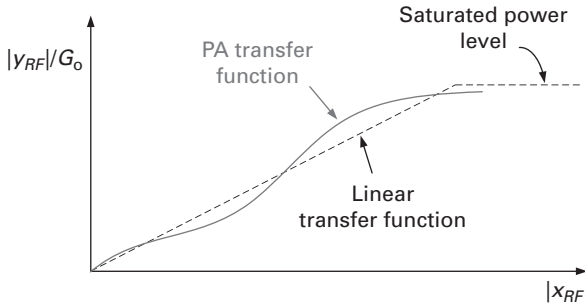


Figure 6.5 Typical power transfer function for a class AB biased power amplifier and the ideal linear transfer function.

$$E[|\varepsilon_{RF}(t)|^2] = E[|y_{RF}(t) - F_{PA}\{x_{RF}(t)\}|^2],$$

where $E[\]$ denotes expected value. Note that a nominal delay through the power amplifier is present. It is compensated typically using a bulk delay external to the nonlinear mapping between the input and output, as shown in Figure 6.4.

In terms of power amplifier models, the simplest is the memoryless model which is a nonlinear mapping of the instantaneous input and output signals. The magnitude component of the memoryless mapping is referred to as the power transfer function. The power transfer function of a class AB biased amplifier is shown in Figure 6.5. The most notable feature is the saturated power, which is the level where additional input drive will not increase the output power. The linear response for the transfer function is also shown in Figure 6.5. The output magnitude is normalized by the desired gain of the PA, denoted by G_o , so that the linear transfer function has a slope of unity.

The memoryless nonlinearity of the power amplifier can also be modeled by specifying the complex gain as a function of the input signal envelope $|x_{RF}|$. The output of the PA becomes

$$y_{RF}(t) = G_{PA}(|x_{RF}|) \cdot x_{RF}(t)$$

where $G_{PA}(|x_{RF}|)$ denotes the complex gain of the PA. The amplitude and phase components of the gain are referred to as the AM-AM and AM-PM curves. These curves get their name from the fact the amplitude modulation (AM) within the input signal results in

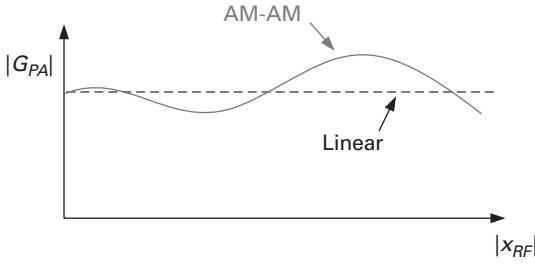


Figure 6.6 AM-AM curve for a class AB biased power amplifier representing the amplitude component of the gain as a function of the input signal amplitude.

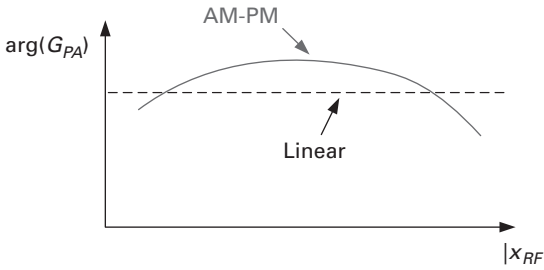


Figure 6.7 AM-PM curve for a class AB biased power amplifier representing the phase component of the gain as a function of the input signal amplitude.

amplitude and phase modulation (AM and PM) in the power amplifier gain. The AM-AM and AM-PM curves for a class AB biased amplifier are shown in [Figure 6.6](#) and [Figure 6.7](#), respectively. The linear responses for the AM-AM and AM-PM curves are constant across the range of input envelopes, $|x_{RF}|$.

It is possible to build the AM-AM and AM-PM curves by sampling the input and output RF signals, computing the gain ($G_{PA} = y_{RF}/x_{RF}$), and marking each coordinate location, $(|x_{RF}|, |G_{PA}|)$ or $(|x_{RF}|, \arg\{G_{PA}\})$, on the respective plots, as shown in [Figure 6.8](#) and [Figure 6.9](#). The memoryless curve represents an average of the sampled gain, where the integration window used for the averaging spans a small interval of horizontal axis $|x_{RF}|$ (RF input envelope). This interval is referred to later as a bin. The effect of memory is visible as dispersion about the memoryless curves. Note that the nominal delay shown in [Figure 6.4](#) is selected to minimize the above-mentioned dispersion.

Memory within a power amplifier can be modeled by applying filtering before and after a memoryless nonlinear block and feedback around it, as shown in [Figure 6.10](#). Each filter introduces frequency-dependent behavior. However, they need to be considered separately due to the presence of the nonlinear block and the fact that multiplicative blocks do not commute with nonlinearities. Simplified memory models using only one of the filter blocks can often be used successfully when the nonlinearity or memory is modest.

Nonlinear memory can also be caused by an unwanted gain modulation that is a function of the present and past envelope values. This is due in part to time constants

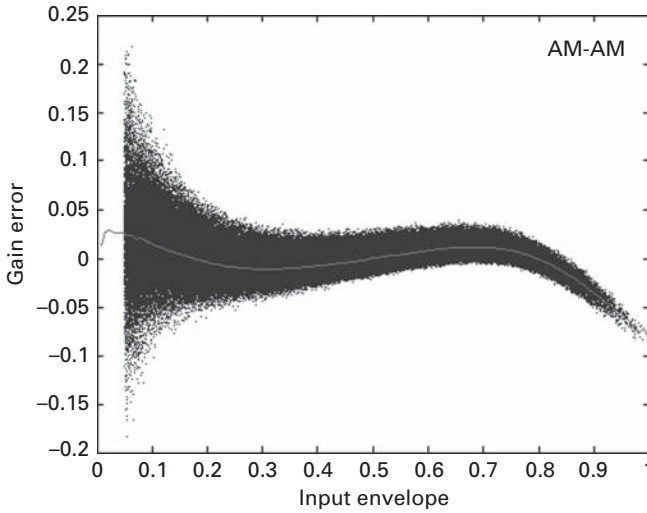


Figure 6.8 Measured gain magnitude and the estimated AM-AM curve. The AM-AM curve is an average of the measured gain magnitude.

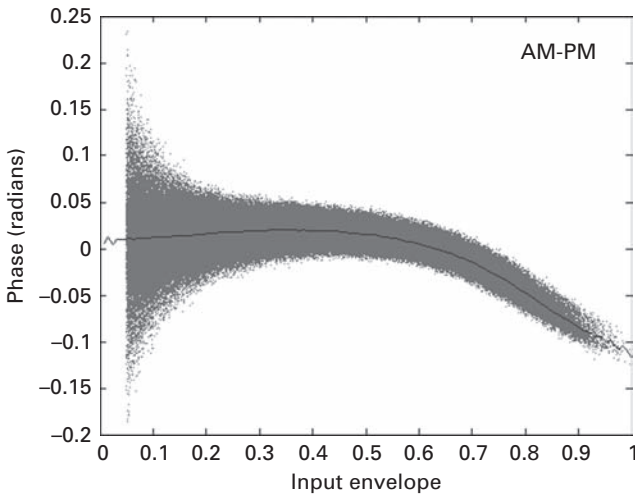


Figure 6.9 Measured phase and the estimated AM-PM curve. The AM-PM curve is an average of the measured phase.

associated with currents in the bias network. Slower gain variations associated with thermal effects may be due to changes in the average input power and time constants associated with thermal dissipation by the heat sinking around the transistor.

The memoryless and memory power amplifier models capture the effects induced by variations in the input signal envelope $|x_{RF}(t)|$. The behavior of a power amplifier is also affected by slow-varying conditions such as ambient temperature, carrier frequency,

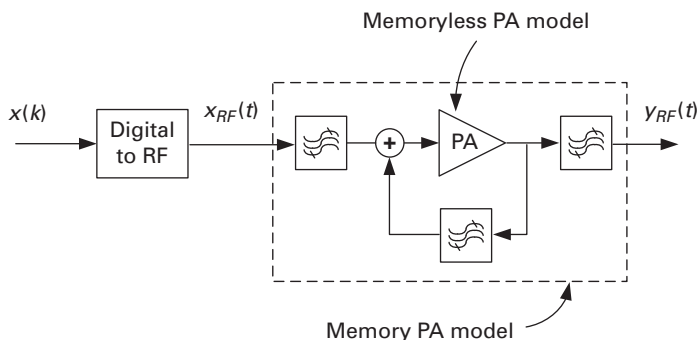


Figure 6.10 Memory model of a power amplifier.

number of carriers used, and modulation format. In this chapter, these non-envelope parameters are referred to collectively as the operating condition. Some envelope-related measurements, such as the average input power, may be considered part of the operating condition if they are varying slowly and not included in the original power amplifier model.

A nonlinear model is valid for a particular operating condition. While it is possible to create one large model of the power amplifier accounting for all possible operating conditions, it is not done in practice. In most cases, the nonlinear model is made as simple as possible to meet the accuracy desired at a given operating condition and new models are created as needed within the multi-dimensional space defined by the operating conditions.

It should be apparent that nonlinearities can be modeled as input–output mappings or by the variations in the gain of the power amplifier. The model chosen for the PA is often used for the predistortion function (discussed in [Section 6.3](#)) as well. The author’s preference is the latter because it allows the problem of linearization to be defined as gain regulation, which is a well-studied problem in the field of adaptive control theory [4]. All of the available control theories and adaptive techniques can be exploited to make high performing, robust linearization designs. The gain regulation view of power amplifier linearization requires the input and linear component of the output signal to be time aligned and equalized in their own modules separate from that of the DPD module.

The input–output mapping model is popular because it allows predistortion to be viewed as the inverse mapping. In most cases, the inverse mapping is obtained directly by measuring the output-to-input mapping. This is wonderful for a static predistortion design based on exhaustive experimental measurements made offline. Unfortunately, this simplicity in the inverse mapping is misleading because it does not acknowledge many of the difficulties associated with real-time adaptive systems that use least mean square (LMS) estimation techniques. The most significant of these problems is the dependence on the input signal excitation to make the nonlinear modes in the mapping observable. Be wary of predistortion papers describing such problems as new discoveries because in reality these problems were identified and addressed decades before within the adaptive control field.

6.3 Digital predistortion overview

An overview of digital predistortion is provided without discussing the details of the algorithms used. Digital predistortion and the challenges faced when implementing it are summarized. Adaptive systems, where the performance is optimized by measuring the output signal and adjusting DPD coefficients iteratively to reduce the residual nonlinearities, are presented. Learning structures are discussed that augment adaptive systems to improve the rate of convergence.

6.3.1 Digital predistortion challenges

The linearization of a power amplifier using digital predistortion involves applying a nonlinear transformation to the digital signal that opposes the subsequent nonlinearity within the power amplifier, as shown in Figure 6.11. The predistorted signal, denoted by $x_{DPD}(k)$, is

$$x_{DPD}(k) = F_{DPD}\{x(k)\}$$

where $F_{DPD}\{\}$ is the nonlinear predistortion function. The predistorted signal may also be described as

$$x_{DPD}(k) = G_{DPD}(|x(k)|) \cdot x(k)$$

where $G_{DPD}(|x|)$ is the nonlinear gain of the DPD. The DPD nonlinearity is controlled by a set of complex coefficients denoted by a_i . Selection of the coefficients may be done offline based on experimental measurements of the PA nonlinearity or adaptively based on real-time measurements of the PA output signal (see Subsection 6.3.2).

Without predistortion, the power amplifier design is a trade-off between linearity and efficiency. Operating the amplifier at a power level near saturation improves efficiency but subjects the signal to more severe nonlinearities. Digital predistortion is used to compensate for the nonlinear behavior allowing for more efficient operation. The trade-off now becomes one of efficiency versus DPD complexity. It is important to understand the goal when attempting to balance PA efficiency and DPD complexity. If the goal is to reduce the thermal build-up in the transistor, only the final stage efficiency of the power amplifier is important and DPD complexity is secondary. However, the DPD consumes power and the consumption is related to the size of the hardware needed to implement the DPD. Thus, DPD complexity must be weighted higher if the goal of efficiency is to reduce overall power consumption of the transmitter.

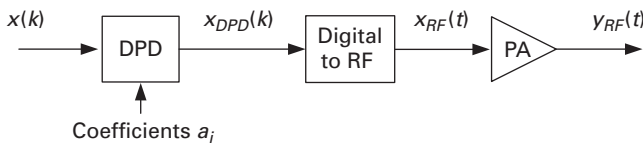


Figure 6.11 Digital transmitter with digital predistortion.

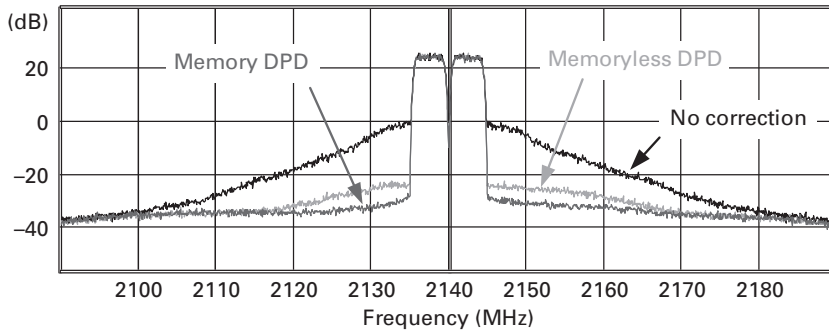


Figure 6.12 Output spectra for a two-carrier WCDMA signal amplified by a Doherty PA. Digital predistortion reduces the ACPR. The memoryless DPD shows a large improvement over the uncorrected PA. However, increasing the complexity of the DPD by including memory correction results in a much smaller incremental improvement. Adapted from [5] © IEEE 2009.

Although digital predistortion is a simple concept, many challenges are encountered when implementing a DPD system. In general, the predistorted signal $x_{DPD}(k)$ has a wider bandwidth than the input signal $x(k)$ because of the presence of the distortion compensation signals. As a result the digital sampling rate of $x_{DPD}(k)$ needs to be several times higher than the Nyquist rate associated with $x(k)$. This in turn means that higher speed digital-to-analog converters (DACs) are needed within the digital-to-RF conversion stage. The DAC speed limitations determine the maximum bandwidth of the input signal that can be predistorted digitally.

Thermal effects make digital predistortion more challenging. The temperature of the transistor, which affects the nonlinear behavior of the power amplifier, changes with the average input power level for high efficiency amplifiers. The problem is most severe for time domain duplexed (TDD) systems, such as WiMax, where the power amplifier is turned on and off in a duty cycle spanning several milliseconds. The digital predistortion must model the heating and cooling of the transistor as well as the changes in the power amplifier characteristics induced by the temperature cycle.

Digital predistortion suffers from diminishing returns. A large initial improvement is possible by predistorting the power amplifier using a simple memoryless DPD model. An example for the case of a two-carrier WCDMA signal amplified by a Doherty PA is shown in Figure 6.12. The memoryless DPD reduces the ACPR by more than 20 dB, but the use of a more complex memory DPD provides an incremental improvement of 5 dB only. That is, the DPD model requires more coefficients for incremental improvement in distortion cancellation. This causes the coefficient set size to expand exponentially. In practice, this is not a serious problem for traditional amplifier bias conditions, such as class AB, because the input power level can be backed off, reducing the amount of correction needed to match the amount of DPD correction provided. However, this may not hold true for some of the high efficiency amplifier configurations, such as those based on switching or drain modulation. Thus, the power amplifier must be designed to ensure

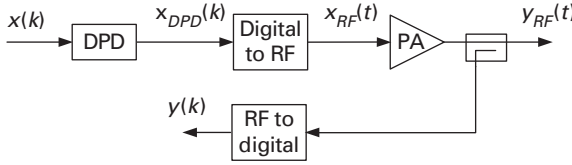


Figure 6.13 Digitally predistorted transmitter with a digital observation path from the PA output.

that the nonlinearities created in pursuit of efficiency can be predistorted by the available DPD function. This is sometimes referred to as “predistortability.”

The input signal influences the DPD performance. Signals spanning a wide bandwidth stimulate memory effects in the power amplifier. As a result, the DPD model requires more coefficients to obtain the necessary correction for a wide bandwidth signal compared to a narrow bandwidth signal applied to the same power amplifier. In addition, signals with high peak-to-average power ratios (PAPR) are difficult for the DPD because they tend to push the power amplifier deeper into saturation. This stimulates higher-order nonlinear modes in the power amplifier, thereby increasing the order of the DPD model. As a result, crest factor reduction (CFR) is often used in conjunction with the DPD. However, CFR increases the EVM in order to reduce the ACPR. This may be an acceptable trade-off for CDMA-based signals but a poor one for OFDM-based signals. Although CFR is useful in a digital transmitter, it is outside of the scope of this chapter.

Digital predistortion often increases the PAPR of the signal entering the PA in order to compensate for gain reductions at peak power levels. It also expands the dynamic range of the envelope voltage $|x_{DPD}(k)|$ compared to $|x(k)|$. Thus, one must reserve headroom for the DPD both in terms of PA saturation and digital saturation. This is done by backing off the average power level of the input signal $x(k)$ by more than the original PAPR (after CFR) from the two saturation levels. However, rogue signal peaks occurring infrequently, such as a probability of less than 10^{-4} , can often be clipped by the saturation without discernible degradation to the ACPR or spectral mask. Thus, the headroom offset is often applied relative to the power level where the complementary cumulative density function (CCDF) of $|x(k)|^2$ is 10^{-4} .

It is convenient to describe the PA output signal as a digital baseband signal, as shown in Figure 6.13. The down-converted and digitized version of $y_{RF}(t)$ is referred to as an output observation signal, and is denoted by $y(k)$. This allows us to describe the PA nonlinearity and the PA gain as functions of the digital signals $x_{DPD}(k)$ and $y(k)$ instead of using the RF signals $x_{RF}(t)$ and $y_{RF}(t)$. The digitized output $y(k)$ is described by

$$y(k) = F_{PA}\{x_{DPD}(k)\}$$

or

$$y(k) = G_{PA}(|x_{DPD}(k)|) \cdot x_{DPD}(k).$$

It should be apparent that the transformation from $y_{RF}(t)$ to $y(k)$ must be linear, or least 10 dB better than the desired linearity of the system, to avoid introducing distortion into $y_{RF}(t)$. A predistorted system that is linear with respect to $x(k)$ and $y(k)$ may appear

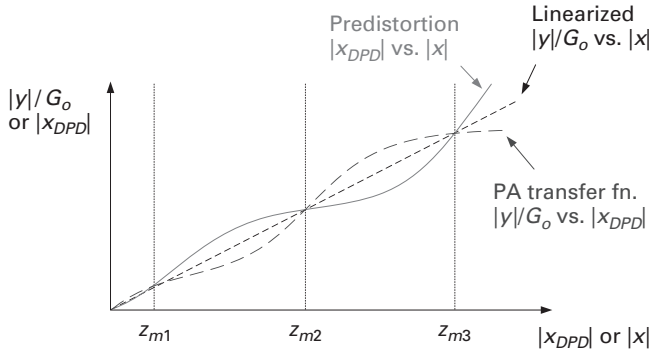


Figure 6.14 Transfer function for a memoryless PA nonlinearity, the ideal linear transfer function, and the predistortion function needed to obtain the linear response.

nonlinear with respect to $x(k)$ and $y_{RF}(t)$, and that nonlinearity will be the inverse of the nonlinearity between $y_{RF}(t)$ and $y(k)$.

Consider a memoryless PA nonlinearity. The desired DPD function F_{DPD} is approximately equal to the inverse of the PA response. The magnitude portion of the DPD is obtained by flipping the PA transfer function about the 45 degree line, which corresponds to the linear response assuming that the PA output magnitude $|y(k)|$ has been normalized by the desired gain G_o . The transfer functions for the PA, DPD, and linear responses are shown in Figure 6.14. The horizontal and vertical axes for the DPD case are $|x(k)|$ and $|x_{DPD}(k)|$, respectively, whereas for the PA case they are $|x_{DPD}(k)|$ and $|y(k)|$. The input envelope levels where $|x(k)| = |x_{DPD}(k)| = |y(k)|/G_o$ are of interest because they are the points where the linear, DPD, and PA transfer functions cross. These points are referred to as zero crossings. The DPD function will be unity at the zero crossings and will expand or compress the DPD gain as needed between the zero crossings.

The PA transfer function must be monotonic over the range of $|x_{DPD}(k)|$ in order for a valid inverse to exist (that is, to be unique and finite). There must be a one-to-one continuous mapping between $|x_{DPD}(k)|$ and $|y(k)|$, and $d|y|/d|x_{DPD}| > 0$. The problem occurs at saturation for a typical PA where $d|y|/d|x_{DPD}| = 0$. As a result, the predistortion function should not attempt to linearize the PA for input envelope values beyond a selected level of $|x(k)|$. That is, the DPD function $F_{DPD}\{x(k)\}$ is selected such that the predistorted output (the cascade of the DPD and PA functions) follows the desired linear response up to the saturation level.

The phase of the DPD function is approximately equal to the negative of the AM-PM curve of the PA, as shown in Figure 6.15. The horizontal axis for the DPD AM-PM curve will stretch and compress slightly between zero crossings because $|x(k)|$ replaces $|x_{DPD}(k)|$.

For memory-based DPD, the coefficients a_i are selected to minimize the difference between the input $x(k)$ and the normalized output $y(k)/G_o$. The optimal coefficients a_i minimize

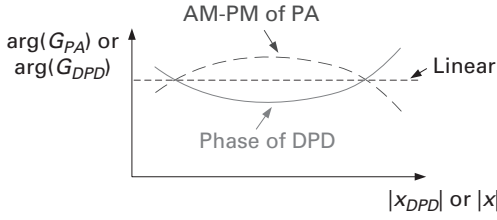


Figure 6.15 Phase curves for a memoryless PA nonlinearity, the ideal linear response, and the predistortion function needed to obtain the linear response.

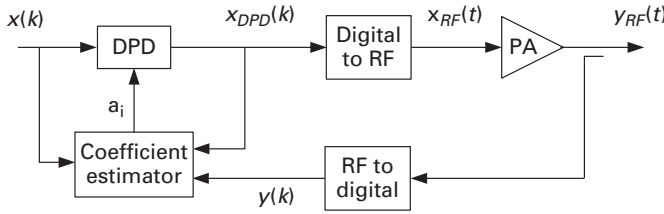


Figure 6.16 RF power amplifier with a coefficient estimator allowing for adaptive digital predistortion (DPD).

$$E[|\varepsilon(k)|^2] = E[|G_o^{-1} \cdot y(k) - x(k)|^2].$$

There are cases where it is advantageous to filter $\varepsilon(t)$, thereby biasing the coefficient optimization so that the mean square error is minimized for specific frequency bands. For example, reducing the out-of-band distortion is more important than the in-band distortion for CDMA modulation formats, so minimizing

$$E[|L\{\varepsilon\}|^2] = E[|G_o^{-1} \cdot L\{y\} - L\{x\}|^2],$$

where $L\{\}$ indicates a filtering operation, may allow the specification to be met using a less complex DPD function.

6.3.2 Adaptive digital predistortion

Because digital predistortion (DPD) models the inverse nonlinearity of a power amplifier (PA) using a small set of coefficients, adequate results are obtained only over a narrow range of operating conditions. The range is extended by making the DPD adaptive. An adaptive system estimates the residual nonlinearity and adjusts the predistortion coefficients in an iterative fashion to minimize it.

The adaptive DPD system in its general form is shown in Figure 6.16. A coefficient estimator is added along with an observation path that digitizes the RF output signal so that nonlinearities are measured based on the input and output signals, as opposed to offline using experimental data obtained from probing signals. The corrections to the DPD coefficients needed to reduce the residual nonlinearities further are computed.

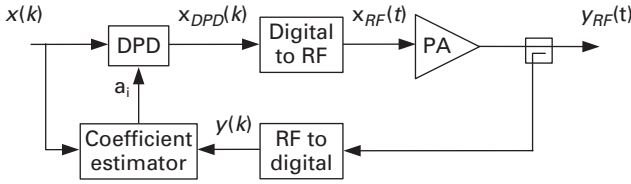


Figure 6.17 RF power amplifier with adaptive DPD based on the MRAS structure. The coefficient estimator compares $x(k)$ and $y(k)$. It is considered closed loop because the estimation includes the effect of the DPD function.

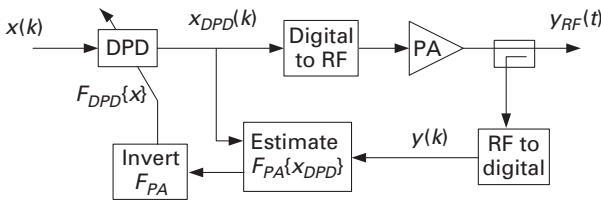


Figure 6.18 RF power amplifier with adaptive DPD based on the STR structure. The PA nonlinearity is estimated then inverted to obtain the DPD function. It is considered open loop because the DPD block is outside of the estimation loop.

Adaptation of the coefficients adds feedback to the digital predistortion system. This can lead to confusion when discussing PA linearization because techniques based on signal feedback exist which have little in common with adaptive digital predistortion. The DPD coefficients tend to vary slowly at a rate similar to the variations in the operating condition. In contrast, the correction of signal-based feedback must be at least as fast as the envelope variations. Examples of signal-based feedback are RF feedback, envelope feedback, and Cartesian feedback [3]. None of these signal-based feedback methods are discussed further in this chapter.

Two examples of adaptive DPD systems are presented [6]. The difference between the two systems is how the coefficients a_i are estimated. Figures 6.17 and 6.18 depict the “model reference adaptive system (MRAS)” and the “self-tuning regulator (STR)” architectures, respectively [4]. The MRAS uses a closed loop estimation where the optimization of the coefficients is done by comparing the input signal $x(k)$ and output signal $y(k)$. The STR estimates the PA nonlinearity $F_{PA}\{x_{DPD}(k)\}$ based on $x_{DPD}(k)$ and $y(k)$. Once converged it is inverted to obtain the predistortion function $F_{DPD}(x(k))$. The STR is open loop in the sense that the DPD function does not effect the estimation. That is, the DPD block is outside of the estimation loop.

Consider the STR structure. The DPD correction is based on the inversion of the estimated PA nonlinearity. The inverse calculation becomes difficult when the STR includes memory correction. Inaccuracies in the inversion degrade distortion cancellation because the predistortion is applied open loop. As a result, computing F_{DPD} from the inversion of F_{PA} is best limited to the memoryless case.

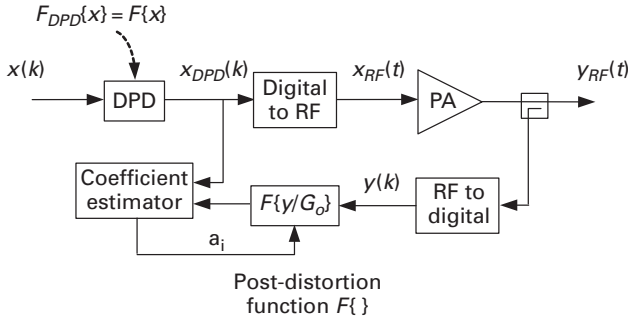


Figure 6.19 RF power amplifier with adaptive DPD based on the STR structure. The coefficient estimator compares $x_{DPD}(k)$ to the post-distorted output $F\{y(k)/G_o\}$. It is considered open loop because the DPD block is outside of the estimation loop.

The inversion problem for the STR is avoided by estimating the inverse PA non-linearity directly (referred to as indirect learning in [7]). The normalized output signal $y(k)/G_o$ is post-distorted by an adjustable nonlinear function, $F\{y(k)/G_o\}$, with the goal of making $F\{y(k)/G_o\} = x_{DPD}(k)$, as shown in Figure 6.19. The optimal predistortion function $F_{DPD}\{x(k)\}$ is assumed to be equal to the optimal post-distortion function $F\{y(k)/G_o\}$. This solves the inversion problem but creates others. For example, the inverse function is undefined near saturation. To compensate, some researchers define a region of validity with respect to the input envelope to allow the adaptation to converge. Another drawback of this approach is that the comparison, $F\{y(k)/G_o\} = x_{DPD}(k)$, has a significantly larger bandwidth and requires more dynamic range than the input signal $x(k)$. It is shown in [8] that the steady-state error increases when the observation output signal $y(k)$ is sub-sampled or band limited with respect to the predistorted signal $x_{DPD}(k)$.

Another consideration is that the strict inverse of the PA nonlinearity may not be desired. Linearization requirements are specified as the maximum allowable distortion generated within adjacent channels, as opposed to the deviation from a linear gain. It can be advantageous to prevent the predistorted signal $x_{DPD}(k)$ from saturating the PA. This is done by clipping the output of $F_{DPD}\{x(k)\}$ or implementing some smoother digital saturation. The digital saturation produces IMDs but this may be acceptable if the level is below the specification and the generation of higher-order distortion by the power amplifier is avoided. The clipped version of $F_{DPD}\{x(k)\}$ is simple to implement because the STR estimation is open loop which allows one to deviate from $F\{y(k)/G_o\}$ as needed without affecting the estimation process.

The MRAS, being a closed-loop process, is well suited for memory correction. Errors in the DPD coefficients are eliminated by the integral action of the control loop (as shown later in Section 6.4). During the optimization, the coefficients a_i within $F_{DPD}\{x(k)\}$ are adjusted with the goal of making $y(k)/G_o = x(k)$. The bandwidth of $y(k)$ approaches that of $x(k)$ as the predistortion converges. As a result, the estimation process is more robust to sub-sampling of $y(k)$ (see [8]). Finally, the inversion of the PA nonlinearity, mentioned in the STR approach, is not required because it is part of the iterative process which uses the first order of a series expansion as an approximation. However, the approach works best if the residual distortion in $y(k)$ is small because the adaptation is based on a first-order

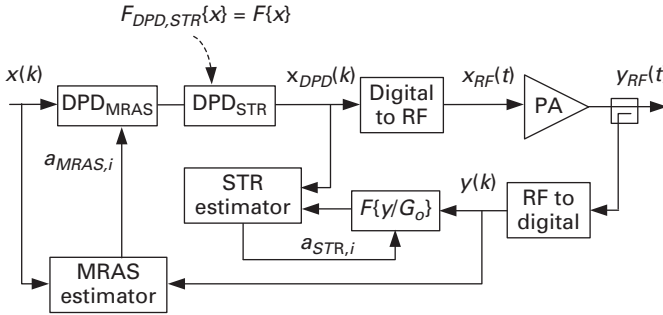


Figure 6.20 RF power amplifier with adaptive DPD using a series combination of the MRAS and STR estimators.

approximation. In addition, clipping within the power amplifier can cause the coefficient adaptation to diverge because the estimator will attempt repeatedly to increase the peak output power beyond saturation without success.

A series combination of the two approaches [6] is shown in Figure 6.20. It solves several problems that arise when implementing either the STR or MRAS in isolation. The presence of the STR predistortion solves both deficiencies of the MRAS mentioned above. The bulk of the predistortion correction can be achieved using a memoryless STR predistortion making the residual distortion seen by the MRAS small. In addition, the STR can be implemented to include a clipping function, and when clipping is detected it marks the present data capture being accumulated for the memory correction as invalid. This exception processing inhibits the estimator, preventing the MRAS coefficients from diverging. Thus the series combination of a memoryless STR with clipping and a memory-based MRAS is more robust than either the STR or MRAS in isolation.

The STR and MRAS estimation stages are run independently; however, the structure shown in Figure 6.20 avoids unstable interactions. Interactions between the DPD stages are stable because the memoryless STR uses an open-loop estimation based on $x_{\text{DPD}}(k)$ and $y(k)$, whereas the MRAS predistortion is a closed-loop estimation based on $x(k)$ and $y(k)$. As a result, the STR coefficients are unaffected by the MRAS correction; however, the MRAS correction tracks changes in the STR coefficients.

There are alternative implementations where the PA nonlinearities are measured without down-converting and digitizing $y_{\text{RF}}(t)$ to create $y(k)$. The gain of the PA can be measured directly using a commercially available differential gain detector circuit, such as Analog Devices' AD8302. A STR implementation is shown in Figure 6.21. The benefit of using a gain detector instead of $y(k)$ is that the data acquisition rate of the analog-to-digital conversion (ADC) is reduced. However, the bandwidth of the gain detector is modest compared to the down-converted output $y(k)$. In the author's opinion, the gain detector method is best suited for analog predistortion using the MRAS structure, as described in [9]. However, analog predistortion is outside of the scope of this chapter.

Another method of measuring the residual nonlinear gain of a predistorted PA is described in [10]. Deformations of a known probing signal are used to measure mild nonlinearities. The inexpensive data acquisition system comprises a bridge circuit and a

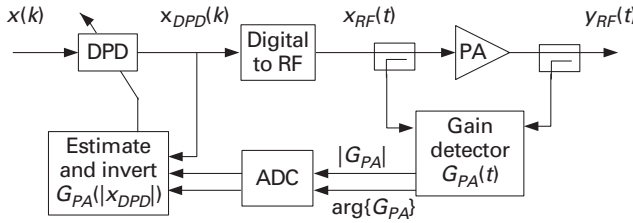


Figure 6.21 RF power amplifier with adaptive DPD based on the STR structure. The gain of the PA is measured directly using a commercially available differential gain detector instead of measuring the PA output $y(k)$.

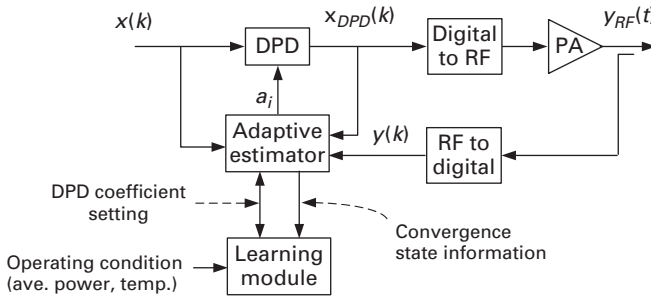


Figure 6.22 RF power amplifier with adaptive DPD and a learning module.

square-law detector. What makes this method interesting is the use of selective sampling to convert a WCDMA input signal into the desired probing signal. A full description of the method can be found in [10]. The main point being made by mentioning the system is to show that $y(k)$ is not required to perform adaptive DPD. However, $y(k)$ is used commonly in adaptive DPD systems and is effective.

6.3.3 Learning-based digital predistortion

As mentioned earlier, adequate DPD results are obtained over a narrow range of operating conditions and that range is extended by making the DPD adaptive. However, the performance of the predistortion is degraded while the estimator is converging to a steady state. A good initial estimate of the coefficient setting is desired to minimize the duration of this transient degradation.

A learning module can reduce the coefficient settling (convergence) time of the adaptive estimator. Past successful steady-state coefficient settings are remembered along with their corresponding operating conditions, and restored subsequently when the operating condition returns. The restored coefficients need not be exact to be useful; they need only be better than the present coefficient setting to reduce the convergence time. After restoring the coefficients, the learning module is inhibited to allow the adaptive estimator to refine the coefficient setting. A DPD system with an adaptive estimator and a learning module is shown in Figure 6.22.

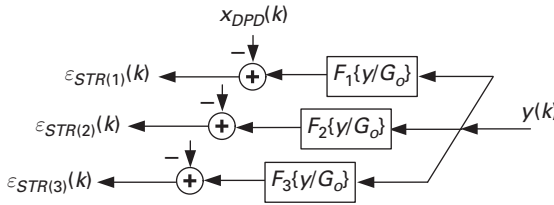


Figure 6.23 Method of arbitrating between the coefficient setting being used currently by the adaptive estimator and those stored in the learning database. Multiple candidate DPD functions from the learning database are tested concurrently using the STR-based post-distortion. Lowest LMS error is preferred.

The key to learning is the data structure which correlates past successful coefficient settings with different operating conditions. The data structure used should be compact (no repeated entries) and current (new entries favored). Successful coefficient settings may be generated online by the adaptive estimator or offline from experimental data. The latter method is more accurately referred to as knowledge-based DPD; however, the underlying method is the same as if the coefficients were generated by the adaptive estimator.

In designing a learning module, several problems must be addressed. Attributes of the operating condition that affect the PA nonlinearity must be identified and measured. An arbitration method is needed to determine when past coefficient settings from the learning module are better than the present coefficient settings generated by the adaptive estimator. In addition, a method is needed for adding new entries to the data structure and pruning redundant ones.

Let us consider the problem of arbitration. One advantage of the post-distortion function used within the STR architecture is that several candidate inverse nonlinearity models, $F_n\{y(k)/G_o\}$, can be applied to the output signal $y(k)$ simultaneously, as shown in Figure 6.23. The quality of the model fit for each $F_n\{y(k)/G_o\}$ is judged by the residual squared error, denoted by $E[\varepsilon_{STR(n)}^2]$. If the present adaptive solution is not adequate, the candidate solution with the lowest residual squared error is selected. Thus a means of arbitrating between the adaptive estimator and several promising functions stored in the memory of the learning module is provided. The MRAS adaptive structure can also use the STR-based test to rank candidate functions, $F_n\{y(k)/G_o\}$, as potential replacements for $F_{DPD}\{x(k)\}$.

An alternative method of arbitrating between the adaptive estimator and candidate functions from the learned database is described in [11]. The test involves inserting promising functions directly into $F_{DPD}\{x(k)\}$ when the adaptive solution is not sufficient. This trial-and-error approach disturbs the DPD function; however, this may not be considered significant if the DPD is already failing to meet the linearity specification. This method is well suited for adaptive DPD systems that do not measure $y(k)$.

Learning is discussed further in Section 6.5. At this point it is important to know that the main advantage of learning is the minimization of the transient degradations associated with abrupt changes in the operating conditions. The secondary advantage is that

learning allows the order of the DPD model to be reduced because the operating condition range over which the model is accurate can be narrowed.

6.4 Digital predistortion basics

Details of basic algorithms used in adaptive digital predistortion are discussed. Three methods of implementing digital predistortion are reviewed: look-up-tables (LUTs), basis waveforms, and cascaded structures. Prior to reviewing the DPD methods, aspects of the DPD implementation and the data acquisition used for coefficient estimation will be discussed.

In practice, the DPD and the estimator are implemented using different hardware. The DPD uses a field programmable gate array (FPGA), typically, to perform the necessary computations in real-time. In contrast, the coefficient estimator can be implemented using a digital signal processor (DSP) or other type of general processor because it has the luxury of accumulating measurements of $x(k)$, $x_{DPD}(k)$, and $y(k)$ in batches before estimating the residual nonlinearities.

Adaptation is based on estimation of the DPD coefficient errors from data captures of $x(k)$, $x_{DPD}(k)$, and $y(k)$. After the estimation, the DPD coefficients are updated with the intention of reducing the coefficient error. The process of capturing data, estimating the coefficient error, and updating the DPD coefficients is repeated indefinitely, giving the system the ability to adapt to variations in the operating condition. The time line for the data capture, estimation, coefficient update, and settling time before the next data capture is shown in Figure 6.24.

The size of the data capture should be large enough that the batch estimates of the input signal statistics do not vary between captures, assuming the input signal is stationary. It is assumed that the data capture length is $K = 16k$ samples per signal, unless noted otherwise. This topic is discussed further in Section 6.5.

It is important to remember that there is a delay between when the signal $x(k)$ first appears in transmitter and when its replica appears at the output $y(k)$. The data captures of $x(k)$ and $y(k)$ must be time-aligned to remove any temporal offsets, such as those shown in Figure 6.25. A cross-correlation between larger size data captures of $x(k)$ and $y(k)$ may be needed to obtain the initial alignment, where the coarse time offset is indicated by the lag

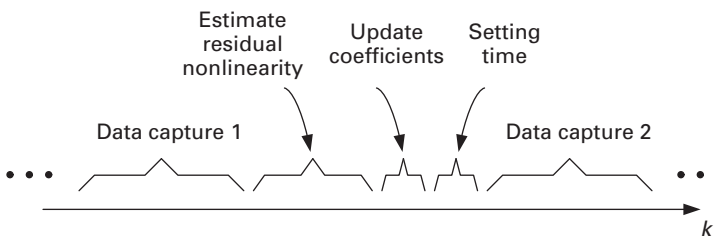


Figure 6.24 The DPD coefficient estimator time line. The repeated pattern includes the data capture, estimation, coefficient update, and settling time before the next data capture.

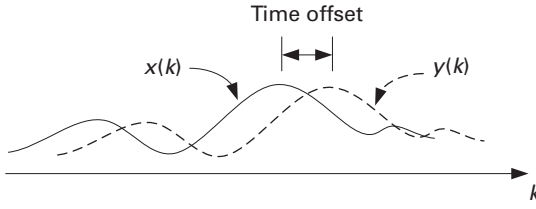


Figure 6.25 Temporal offset between the input signal $x(k)$ and the output signal $y(k)$.

associated with the cross-correlation peak. A delay equalization filter may be applied to achieve sub-sample alignment between $x(k)$ and $y(k)$. Some of the coefficient estimation approaches (see [Subsection 6.4.3](#)) are less sensitive to time misalignment; however, it is due to the fact that the delay equalization function is included within the DPD model.

The DPD system must be mindful of other adaptive processes within the transmitter. For example, an equalizer may be present to regulate the gain flatness of the transmitter. A similar equalizer may be in the observation path. In general, only one of the equalizers can be made adaptive on the basis of measurements of the output $y(k)$. The equalizer on the other path is calibrated offline typically, and its coefficients remain static thereafter. This static path is sometimes referred to as the golden path. Other adaptive processes within a transmitter include automatic gain control (AGC) [12] and corrections of modulator imbalance [13] as well as carrier leakage. The rate of adaptation of each process should be different from that of the DPD system, preferably static, if the DPD model overlaps the other processes.

6.4.1 LUT implementations of DPD

In the following, various implementations of the LUT-based DPD are discussed [14][15][16][17]. Adaptive memoryless DPD is implemented using STR and MRAS estimators. Memory correction is implemented using a multi-dimensional LUT (see [Subsection 6.4.2](#)). For convenience, only the MRAS structure using a gain-based DPD model is considered for the memory correction even though other structures are possible.

For LUT-based memoryless DPD, the predistorted signal $x_{DPD}(k)$ is

$$x_{DPD}(k) = G_{DPD}(|x(k)|) \cdot x(k),$$

where the predistortion gain G_{DPD} is represented by a LUT comprising a finite number of bins (denoted by N_{bins}). Each bin spans a small interval of $|x|$ whose length is denoted by η . Each input envelope sample $|x(k)|$ entering the predistorter is converted into a bin index n using

$$n(k) = \text{round}\left\{\frac{|x(k)|}{\eta}\right\}$$

where

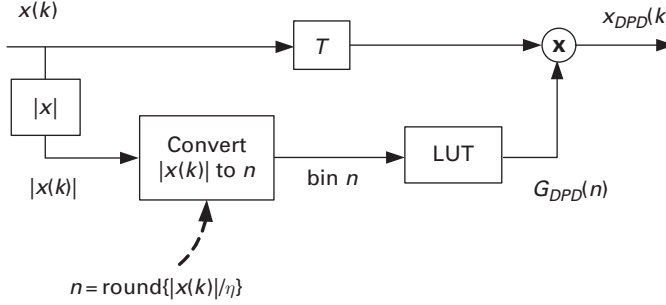


Figure 6.26 DPD function implemented as a gain-based LUT.

$$\eta = \frac{\max\{|x(k)|\}}{N_{bins}}.$$

The DPD gains $G_{DPD}(n)$ associated with bins n are applied to time sequence $x(k)$ in real-time to provide the gain variations needed to predistort the system, as shown in Figure 6.26. A typical value for N_{bins} is 128.

An adaptive system is used to estimate the $G_{DPD}(n)$ LUT based on past measurements of the output $y(k)$. Two types of estimators are considered for the memoryless system. The STR case is discussed first, followed by the MRAS case.

The STR uses past data captures to estimate the PA gain LUT, G_{PA} , as a function of $|x_{DPD}|$ then inverts it to obtain the predistortion gain LUT, G_{DPD} . The PA gain G_{PA} measured at time sample k is

$$G_{PA}(k) = \frac{y(k) \cdot x_{DPD}^*(k)}{|x_{DPD}(k)|^2}.$$

At first glance it may seem puzzling why one needs to estimate the G_{PA} LUT when the PA gain can be measured. However, $G_{PA}(k)$ is based on $y(k)$, which as mentioned earlier, is delayed compared to $x(k)$ or $x_{DPD}(k)$. Thus, to obtain an estimate of the DPD gain that is time-aligned with $|x_{DPD}(k)|$, the former must be predicted from latter using the estimated G_{PA} LUT.

The LUT implementation of G_{PA} quantizes $|x_{DPD}(k)|$ into bins, converting time samples k into bin indices n_D using

$$n_D(k) = \text{round}\left\{\frac{|x_{DPD}(k)|}{\eta_D}\right\}$$

where

$$\eta_D = \frac{\max\{|x_{DPD}(k)|\}}{N_{bins}}.$$

Note that the quantization of $|x_{DPD}|$ is different than that of $|x|$ used in the $G_{DPD}(n)$ LUT. However, the difference is small.

Statistical moments of the measured PA gain, obtained by accumulating samples from the most recent data capture into bins n_D , are

$$L_0(n_D) = \sum_k \beta_{n,D}(k)$$

$$L_1(n_D) = \sum_k \beta_{n,D}(k) \cdot G_{PA}(k)$$

$$L_2(n_D) = \sum_k \beta_{n,D}(k) \cdot |G_{PA}(k)|^2$$

where the quantization of $|x_{DPD}(k)|$ is

$$\beta_{n,D}(k) = \begin{cases} 1 & \text{when } \left| \frac{|x_{DPD}(k)|}{\eta_D} - n_D \right| < 0.5 \\ 0 & \text{otherwise} \end{cases}.$$

The mean of $G_{PA}(n_D)$, used to form the AM-AM and AM-PM curves for the PA, is

$$E[G_{PA}(n_D)] = \frac{L_1(n_D)}{L_0(n_D)}.$$

The bin variance is

$$\sigma^2(n_D) = \frac{L_2(n_D) \cdot L_0(n_D) - [L_1(n_D)]^2}{[L_0(n_D)]^2}.$$

The variance is a measure of the dispersion in the AM-AM and AM-PM curves shown earlier in [Figure 6.8](#) and [Figure 6.9](#).

The DPD gain $G_{DPD}(n)$ is obtained by inverting the normalized PA gain $G_{PA}(n_D)/G_o$. Unfortunately, the estimated PA gain curve can be jagged due to measurement noise. In such cases the estimated $G_{PA}(n_D)$ is altered to ensure that the PA transfer function is monotonic and that a valid inverse exists: that is,

$$n_D \cdot |G_{PA}(n_D)| > (n_D - 1) \cdot |G_{PA}(n_D - 1)|.$$

Smoothing across bins and averaging over time improves the situation, typically. However, in some cases the magnitude of the estimated PA gain must be increased for some bin indices to meet the monotonic condition.

Smoothing across bins should be applied to the statistical moments, L_0 and L_1 , rather than the estimated PA gain $G_{PA}(n_D)$. The reason for doing this is that the quantization process may be uneven resulting in variations in the number of samples within neighboring bins. In general, the estimated PA gain $G_{PA}(n_D)$ is more accurate for bins containing many time samples k . Thus bins with larger values of $L_0(n_D)$ should receive more weight when smoothing across neighboring bins. Computing $G_{PA}(n_D)$ from smoothed values of L_0 and L_1 achieves this goal with the least amount of effort. Smoothing is performed in several iterations using the following diffusion equation:

$$L_0(n_D)|_{s+1} = L_0(n_D)|_s + \mu_s \cdot [L_0(n_D - 1)|_s + L_0(n_D + 1)|_s]$$

where $0 < \mu_s < 1$ and $L_0(n_D)|_s$ is the smoothed L_0 for bin n_D after s iterations. The same smoothing is applied to the bins of L_1 .

Time averaging of L_0 , L_1 , and L_2 is applied over successive data captures. The time averaged values of L_0 are

$$L_0(n_D)|_{t+1} = L_0(n_D) + \mu_t \cdot L_0(n_D)|_t$$

where $0 < \mu_t < 1$ and $L_0(n_D)|_t$ is the time averaged L_0 for bin n_D after t data captures. The term μ_t represents a forgetting factor that discounts measurements obtained from older data captures. The same time averaging is applied to L_1 and L_2 . The order in which smoothing and time averaging is applied is a matter of personal preference. The author prefers to time average first, then smooth across bins prior to inverting the $G_{PA}(n_D)$ LUT to obtain the $G_{DPD}(n)$ LUT.

In some instances, it is desirable to convert the estimated AM-AM and AM-PM curves into a polynomial function. This can be done instead of bin smoothing, to simplify the inversion of G_{PA} , or to allow G_{DPD} to be implemented as a polynomial function (see [Subsection 6.4.3](#)). The transformation is based on a weighted LMS fit which produces polynomial coefficients a_i minimizing

$$J_{LUT} = \sum_{n=1}^{N_{bins}} w(n_D) \left| G_{PA}(n_D) - \sum_i a_i \cdot |\eta_D \cdot n_D|^i \right|^2$$

where $w(n_D)$ is the bin weight.

There are several different weighting schemes: uniform, histogram, or stochastic. The uniform weighting, $w(n_D) = 1$, results in the poorest ACPR performance. This is unfortunate because the novice engineer will almost always use it. The histogram weighting sets $w(n_D)$ equal to $L_0(n_D)$. This provides good ACPR performance and is the same as the LMS estimation using basis waveforms described in [Subsection 6.4.3](#). The stochastic weighting adjusts the weighting of the histogram based on the bin variance making $w(n_D) = L_0(n_D)/\sigma^2(n_D)$. The stochastic weighting provides the best ACPR performance; however, the improvement over the histogram weighting is not enough to cause concern. The distributions of three weighting schemes for a two-carrier WCDMA signal are shown in [Figure 6.27](#). The stochastic distribution is biased towards higher bins (higher values of $|x_{DPD}|$) because σ^2 is larger typically for lower values of $|x_{DPD}|$, as was shown earlier in [Figure 6.8](#) and [Figure 6.9](#).

Now consider the MRAS implementation of the estimator. The MRAS estimates the residual gain error of the predistorted system from past data captures and uses it to update the DPD gain $G_{DPD}(n)$. The system gain error ΔG_{sys} at time sample k is

$$\Delta G_{sys}(k) = \frac{[y(k) - G_o \cdot x(k)] \cdot x^*(k)}{|x(k)|^2}.$$

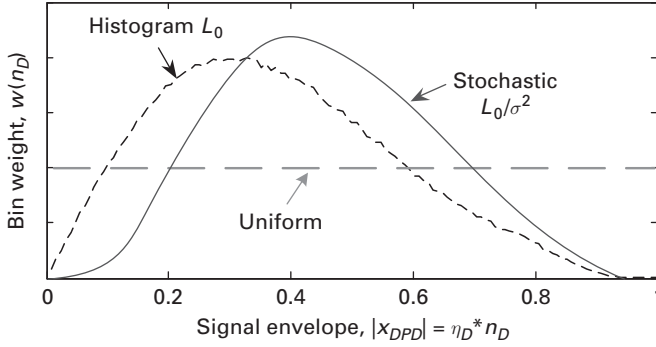


Figure 6.27 Distributions of bin weights $w(n_D)$ for the uniform, histogram, and stochastic weighting schemes. The input is a two-carrier WCDMA signal whose spectrum is shown in Figure 6.12. Adapted from [6] © IEEE 2009.

The LUT implementation of ΔG_{sys} quantizes the input envelope $|x(k)|$, converting it into bin indices n . Note that the LUTs representing ΔG_{sys} and G_{DPD} share the same bin indices, which is convenient.

Statistical moments of the residual gain error ΔG_{sys} , obtained by accumulating samples from the most recent data capture into bins n , are

$$L_0(n) = \sum_k \beta_n(k)$$

$$L_1(n) = \sum_k \beta_n(k) \cdot \Delta G_{sys}(k)$$

$$L_2(n) = \sum_k \beta_n(k) \cdot |\Delta G_{sys}(k)|^2$$

where the quantization of $|x(k)|$ is

$$\beta_n(k) = \begin{cases} 1 & \text{when } \left| \frac{|x(k)|}{\eta} - n \right| < 0.5 \\ 0 & \text{otherwise} \end{cases}$$

The mean of the residual gain error $\Delta G_{sys}(n)$ is

$$E[\Delta G_{sys}(n)] = \frac{L_1(n)}{L_0(n)}.$$

The bin variance is the same as in the STR case.

Bin smoothing can be used for L_0 and L_1 prior to computing $\Delta G_{sys}(n)$, as in the STR case. However, time averaging is replaced by an update equation for $G_{DPD}(n)$:

$$G_{DPD}(n)|_{t+1} = G_{DPD}(n)|_t - \mu_\Delta \cdot \Delta G_{sys}(n)$$

where $G_{DPD}(n)|_t$ is the estimate of the DPD gain after t data captures and $0 < \mu_\Delta < 1$. The term μ_Δ is a convergence constant. It can also be useful to add a small constant to $L_0(n)$ before computing $\Delta G_{sys}(n)$ so that the residual gain error is underestimated when $L_0(n)$ is small. This reduces the sensitivity to noise but extends the convergence time, much like reducing the convergence constant μ_Δ (see [section 6.5](#)).

6.4.2 LUT memory DPD

Digital predistortion capable of correcting PA memory can be implemented using a multi-dimensional LUT [18]. Let the predistorted signal $x_{DPD}(k)$ be defined by

$$x_{DPD}(k) = G_{DPD}(|x(k)|, |x(k-1)|) \cdot x(k)$$

where the DPD gain G_{DPD} is a function of the present sample of the input envelope $|x(k)|$ as well as the previous sample, $|x(k-1)|$. Digital predistortion using a two-dimensional LUT is not unlike the 1D case. Both $|x(k)|$ and $|x(k-1)|$ are quantized to obtain bin indices n_0 and n_1 : that is,

$$n_0(k) = \text{round}\left\{\frac{|x(k)|}{\eta}\right\}$$

$$n_1(k) = n_0(k-1).$$

The DPD gains $G_{DPD}(n_0, n_1)$ associated with bins (n_0, n_1) are applied to time sequence $x(k)$ in real-time to provide the gain variations needed to predistort the system.

The MRAS estimator for the 2D case is discussed below. As in the 1D case, the residual gain error LUT, $\Delta G_{sys}(n_0, n_1)$, is estimated from the most recent data capture and used to update the $G_{DPD}(n_0, n_1)$ LUT. The statistical moments for each bin (n_0, n_1) are

$$L_0(n_0, n_1) = \sum_k \beta_{n_0}(k) \cdot \beta_{n_1}(k-1)$$

$$L_1(n_0, n_1) = \sum_k \beta_{n_0}(k) \cdot \beta_{n_1}(k-1) \cdot \Delta G_{sys}(k)$$

$$L_2(n_0, n_1) = \sum_k \beta_{n_0}(k) \cdot \beta_{n_1}(k-1) \cdot |\Delta G_{sys}(k)|^2.$$

The mean of the residual gain error $\Delta G_{sys}(n_0, n_1)$ is

$$E[\Delta G_{sys}(n_0, n_1)] = \frac{L_1(n_0, n_1)}{L_0(n_0, n_1)}$$

and the bin variance is

$$\sigma^2(n_0, n_1) = \frac{L_2(n_0, n_1) \cdot L_0(n_0, n_1) - [L_1(n_0, n_1)]^2}{[L_0(n_0, n_1)]^2}.$$

The updated DPD gain is

$$G_{DPD}(n_0, n_1)|_{t+1} = G_{DPD}(n_0, n_1)|_t - \mu_\Delta \cdot \Delta G_{sys}(n_0, n_1).$$

It should be apparent that the LUT implementation can be extended to include additional dimensions.

The primary problem with the multi-dimensional LUT is that the number of bin elements grows exponentially with the dimension (memory depth). Another problem is related to the update of the bin elements. Some of the bin elements will have a low histogram count $L_0(n_0, n_1)$ because changes between $|x(k)|$ and $|x(k-1)|$ are limited by the bandwidth of the input signal.

An alternative 2D LUT is

$$x_{DPD}(k) = G_{DPD}(|x(k)|, \Delta|x(k)|) \cdot x(k)$$

where

$$\Delta|x(k)| = |x(k)| - |x(k-1)|.$$

Both $|x(k)|$ and $\Delta|x(k)|$ are quantized to obtain indices n_0 and n_1 , the moments $L_0(n_0, n_1)$, $L_1(n_0, n_1)$, and $L_2(n_0, n_1)$ are computed from the most recent data capture, and the mean and variance for each bin element (n_0, n_1) are estimated. The key difference is that the number of bins needed to represent $\Delta|x(k)|$ is reduced compared to $|x(k-1)|$.

The problem of a low histogram count is reduced in severity by bin smoothing. The 2D diffusion equation becomes

$$L_0(n_0, n_1)|_{s+1} = L_0(n_0, n_1)|_s + D_2 + \mu_1 \cdot L_0(n_0)$$

where

$$D_2 = D_{2,0} + D_{2,1}$$

$$D_{2,0} = \mu_{2,0} \cdot [L_0(n_0 - 1, n_1) + L_0(n_0 + 1, n_1)]$$

$$D_{2,1} = \mu_{2,1} \cdot [L_0(n_0, n_1 - 1) + L_0(n_0, n_1 + 1)]$$

and μ_1 , $\mu_{2,0}$, and $\mu_{2,1}$ are constants less than unity. Note that the recursive application of the 2D diffusion equation smoothes across the bins of the 2D LUT using D_2 . The diffusion equation shown above also applies bin smoothing across the memory depth by including $L_0(n_0)$ from the memoryless DPD. The value of μ_1 is set proportional to $\sigma^2(n_0, n_1)/\sigma^2(n_0)$ so that the memoryless estimation contributes less if the 2D bin variance is small.

The motivation for increasing the memory depth is to improve the model accuracy, which implies that the memory correction should reduce the 2D bin variance $\sigma^2(n_0, n_1)$ compared to the 1D value of $\sigma^2(n_0)$. If the bin variance is unchanged from the memoryless estimate, the increase in memory depth, by splitting the accumulated samples of bin n_0 amongst sub-bins n_1 , was likely unnecessary. In such cases $G_{DPD}(n_0)$ is a good estimate of $G_{DPD}(n_0, n_1)$.

6.4.3 DPD using basis waveforms

Another approach to DPD involves generating a set of basis waveforms from nonlinear transformations of the input signal $x(k)$ then creating the desired predistorted signal $x_{DPD}(k)$ from a weighted sum of these waveforms. The predistorted signal is

$$x_{DPD}(k) = F_{DPD}\{x(k)\} = \sum_{i=1}^N a_i \cdot \gamma_i\{x(k)\}$$

where $\gamma_i(k)$ are basis waveforms that are nonlinear functions of input $x(k)$, a_i are the DPD coefficients, and N is the number of basis waveforms used in the DPD model. The constant N is also referred to as the model order. Any set of nonlinear transformations may be used to generate the basis waveforms; however, polynomial functions of $x(k)$ are the most popular.

Let us look at some basis waveform sets used commonly in DPD systems. The discussion starts with the memoryless polynomial. Several memory basis sets are reviewed including the discrete Volterra series and memory polynomial. Gain-based implementations of the memoryless and memory polynomials are also discussed. Suitable coefficient estimators for the basis waveform approach are presented.

The simplest DPD model is the memoryless predistorter whose basis waveforms are

$$\gamma_i(k) = |x(k)|^{i-1} \cdot x(k).$$

The predistorted signal becomes

$$x_{DPD}(k) = \sum_{i=1}^N a_i \cdot |x(k)|^{i-1} \cdot x(k).$$

The DPD function is referred to as a memoryless polynomial and is shown in [Figure 6.28](#).

Despite the simplicity of the memoryless polynomial, there are varying opinions amongst researchers on which waveforms should be included in the basis set. The DPD model comprises odd- and even-order terms. The odd-order terms are $x(k)$, $|x(k)|^2 x(k)$, $|x(k)|^4 x(k)$, $\dots$, $|x(k)|^{2n} x(k)$. The even-order terms are $|x(k)|x(k)$, $|x(k)|^3 x(k)$, $\dots$,

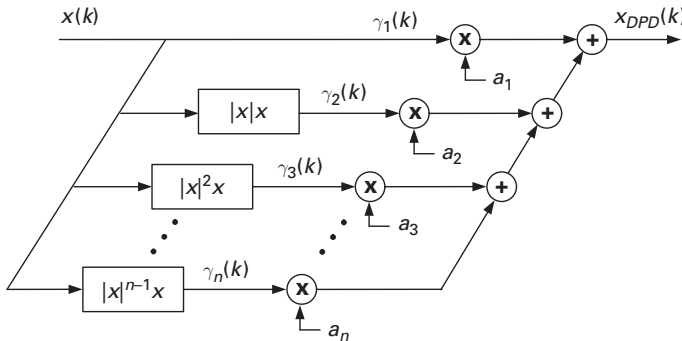


Figure 6.28 DPD function implemented as a memoryless polynomial.

$|x(k)|^{2n+1}x(k)$. Many researchers omit the even-order terms from the basis set. The reasoning given is that the even-order distortion products generated by the RF signal passing through the PA are suppressed by the output filter. However, it is shown in [7] that including even-order terms in the DPD model improves the ACPR performance.

There are reasons why even-order terms may be beneficial in the DPD. Nonlinearities can occur in the baseband analog portion of the transmitter, prior to the up-conversion to RF. The even-order modes from such nonlinearities appear near the carrier frequency and are not suppressed by the output filter. A second explanation is that the even-order terms within the DPD function are compensating for higher-order odd terms within the PA [19]. One advantage of using both odd- and even-order terms in the DPD rather than including additional higher-order odd terms in the basis set is that lower-order terms require less dynamic range to generate the waveforms digitally. On the other hand, if using an odd-order basis set provides sufficient linearization, then use it.

The discussion of odd- and even-order terms is part of a broader topic: basis waveform selection. The desire for accuracy from the DPD function encourages the use of many basis waveforms. However, additional basis waveforms require more FPGA resources to implement, which is undesirable for cost-sensitive applications. Another reason to limit the number of basis waveforms is that over-specifying the DPD model causes the coefficient estimation to be ill-conditioned. It is shown later in Section 6.5 that poor conditioning in the coefficient estimation degrades the ACPR performance of the pre-distorted system. As a result, there is a strong motivation to identify and select basis waveforms similar to the IMD distortion generated by the interaction of the PA non-linearity and the input signal $x(k)$, so that adequate predistortion is obtained using the fewest basis waveforms possible.

It is necessary to expand the basis set to include memory correction when the memoryless DPD model is not sufficient. One possible basis set with memory is the discrete Volterra series. The basis waveforms for polynomial order $p = 3$ and memory depth of M are

$$\gamma_i(k) = x(k - m_1) \cdot x(k - m_2) \cdot x^*(k - m_3)$$

where $-M \leq m_1, m_2, m_3 \leq M$ and $i = (m_3 + M) \cdot (2M + 1)^2 + (m_2 + M) \cdot (2M + 1) + (m_1 + M + 1)$. The basis waveforms for odd polynomial orders $p \geq 5$ are

$$\gamma_i(k) = A(m_1 : m_{0.5(N+1)}) \cdot A(m_{0.5(N+3)} : m_N)^*$$

where

$$A(m_l : m_j) = x(k - m_l) \cdot x(k - m_{l+1}) \cdots x(k - m_j).$$

The drawback of the Volterra series is that the number of basis waveforms created is enormous. The model order for $p = 3$ is $(2M + 1)^3$, which is already large for $M > 1$, and becomes exponentially larger as higher values of p are included in the basis set. In addition, many of the basis waveforms are duplicated. As a result, pruned versions of the Volterra series are used for memory correction in practical implementations [19][20][21][22].

With respect to pruning the Volterra basis set, one might question why both positive and negative delays (denoted by m_i above) are included. Although physical delays through the PA are positive, it is important to remember that the nominal delay has been removed. As a result, a negative delay is actually a physical delay that is less than the nominal value. In addition, the memory terms within the DPD model can be viewed as a delay equalization filter for the nonlinear modes. Delay equalization of band-limited signals often uses finite impulse response (FIR) kernels whose taps surround the intended range of delays. For these reasons both positive and negative delays are included in the basis sets within this chapter unless noted otherwise.

A popular method of pruning the Volterra series is to keep only the basis waveforms formed by delaying the modes of a memoryless polynomial. The retained basis waveforms are

$$\gamma_{p,m}(k) = |x(k-m)|^{p-1} \cdot x(k-m)$$

where $-M \leq m \leq M$ and $1 \leq p \leq P$. Note that the subscript of the basis waveform $\gamma_{p,m}$ has been split into the polynomial order p and the delay term m , for convenience. The model order of the predistortion is $P*(2M+1)$ assuming the FIR size, $2M+1$, is the same for each polynomial order p . The model order N for this pruned basis set is significantly less than the Volterra series for $M > 1$.

The predistortion signal $x_{DPD}(k)$ is

$$x_{DPD}(k) = \sum_{p=1}^P \sum_{m=-M}^M a_{p,m} \cdot |x(k-m)|^{p-1} \cdot x(k-m).$$

The resulting DPD function, referred to as a memory polynomial [7][23], implements the predistortion as a set of nonlinear FIR structures with adaptable coefficients, as shown in Figure 6.29 and Figure 6.30. Note the each FIR filter associated with a polynomial order p can have a different frequency response, depending on the selection of the DPD coefficients $a_{p,m}$.

The MRAS and STR estimators suitable for the memoryless polynomial, memory polynomial, and Volterra series are described below. The MRAS adapts the DPD

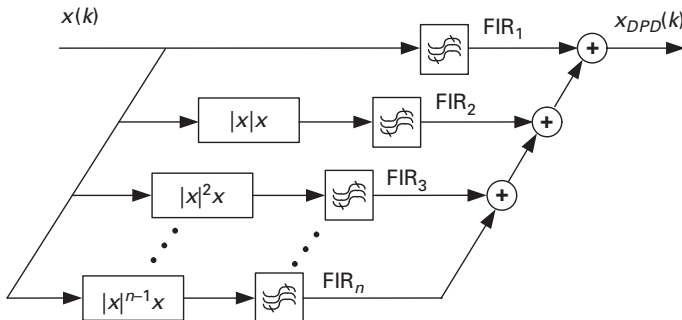


Figure 6.29 DPD function implemented as a memory polynomial.

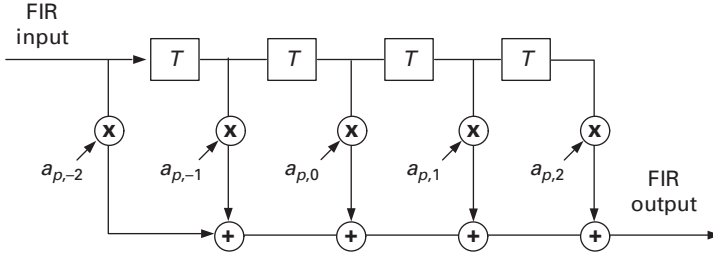


Figure 6.30 Finite impulse response (FIR) filter used in the memory polynomial. The DPD coefficients $a_{p,m}$ are adaptable.

coefficients a_i based on time-aligned data captures of the input and output signals: $\underline{x} = [x(1) \dots x(K)]^T$ and $\underline{y} = [y(1) \dots y(K)]^T$. The MRAS estimator minimizes

$$J_{MRAS} = \sum_k \left| \varepsilon_{MRAS}(k) - \sum_i \Delta a_i \cdot \gamma_i\{x(k)\} \right|^2$$

where Δa_i are the coefficient errors and

$$\varepsilon_{MRAS}(k) = G_o^{-1} \cdot y(k) - x(k).$$

The LMS estimate of the coefficient errors Δa_i is

$$\underline{\Delta a} = Q^{-1} \cdot Z^H \cdot \underline{\varepsilon}_{MRAS}$$

where

$$Q = Z^H Z,$$

$$Z = [\underline{\gamma}_1 \quad \dots \quad \underline{\gamma}_N],$$

and $(\cdot)^H$ denotes conjugate transpose, $\underline{\Delta a} = [\Delta a_1 \dots \Delta a_N]^T$, $\underline{\varepsilon}_{MRAS} = [\varepsilon_{MRAS}(1) \dots \varepsilon_{MRAS}(K)]^T$, and $\underline{\gamma}_i = [\gamma_i(1) \dots \gamma_i(K)]^T$. The DPD coefficients are updated recursively using

$$a_i(t_{l+1}) = a_i(t_l) - \mu_A \cdot \Delta a_i$$

where t_l indicates the iteration number.

The matrix Q is dependent on the basis waveforms γ_i , which in turn are dependant on the statistics of the input signal $x(k)$. The conditioning of Q affects the accuracy of the estimation process. Condition problems indicate that the basis set has been selected poorly for the input signal $x(k)$ being transmitted. Methods for avoiding conditioning problems are discussed in [Section 6.5](#).

The STR adapts DPD coefficients by comparing the post-distorted output signal $F\{y(k)/G_o\}$ with the predistortion signal $x_{DPD}(k)$. The STR estimate of the coefficient errors Δa_i minimizes

$$J_{STR} = \sum_k \left| \varepsilon_{STR}(k) - \sum_i \Delta a_i \cdot \gamma_i \{G_o^{-1} \cdot y(k)\} \right|^2$$

where

$$\varepsilon_{STR}(k) = F\{G_o^{-1} \cdot y(k)\} - x_{DPD}(k).$$

The estimation of the DPD coefficient errors and the update of the DPD coefficients are the same as the MRAS case.

As mentioned earlier, an alternative method of predistortion is to apply a nonlinear gain $G_{DPD}(|x(k)|)$ to the signal $x(k)$. The predistorted signal for the gain-based implementation of the memoryless polynomial is

$$x_{DPD}(k) = G_{DPD}(|x|) \cdot x(k) = \left[1 + \sum_{i=2}^N a_i \cdot |x(k)|^{i-1}\right] \cdot x(k).$$

The gain-based memoryless polynomial, shown in Figure 6.31, differs from the standard implementation because the correction is multiplied instead of added to $x(k)$. In addition, the gain-based version of the DPD does not control the linear component of the gain. It is assumed that the automatic gain control (AGC) function is performed elsewhere within the transmitter. However, from the point of view of the MRAS estimator, the basis waveforms are still defined by

$$\gamma_i(k) = |x(k)|^{i-1} \cdot x(k).$$

The gain-based memory polynomial, shown in Figure 6.32, delays the nonlinear modes of the DPD gain $G_{DPD}(|x|)$ only, but not the linear signal path $x(k)$. The basis waveforms are

$$\gamma_{p,m}(k) = |x(k-m)|^{p-1} \cdot x(k)$$

and the predistortion signal $x_{DPD}(k)$ becomes

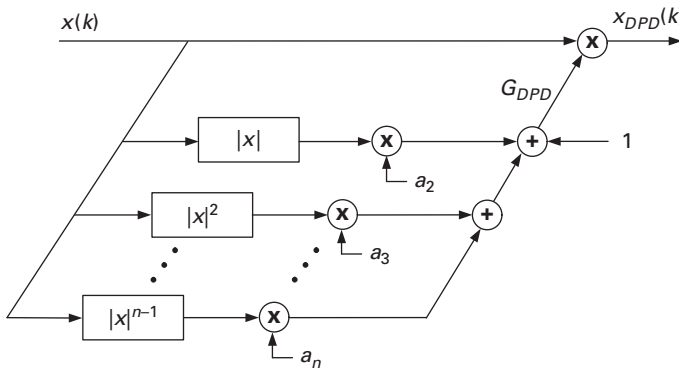


Figure 6.31 DPD function implemented as a gain-based memoryless polynomial. Note that the correction is multiplied instead of added to $x(k)$.

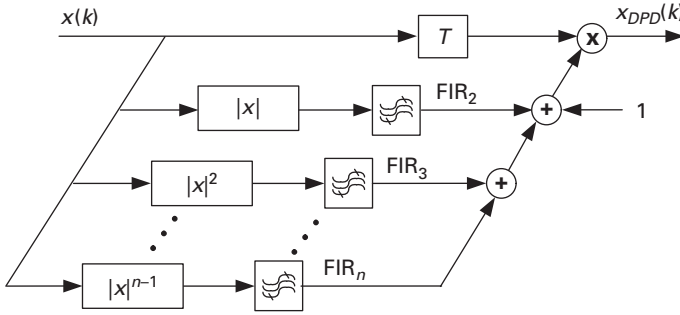


Figure 6.32 DPD function implemented as a gain-based memory polynomial.

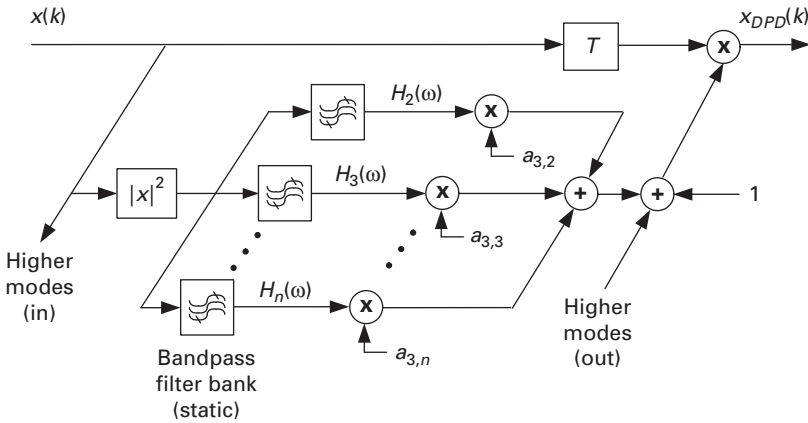


Figure 6.33 Gain-based memory polynomial where the memory is implemented using a bank of static filters.

$$x_{DPD}(k) = [1 + \sum_{p=2}^P \sum_{m=-M}^M a_{p,m} \cdot |x(k-m)|^{p-1}] \cdot x(k).$$

As in the memoryless case, it is assumed that AGC and equalization of the linear components are performed elsewhere in the transmitter.

The gain-based memory polynomial uses a different multi-tap FIR filter for each nonlinear mode of the DPD gain $G_{DPD}(|x|)$. It is possible to implement the filtering in other ways such as using a bank of static filters to separate the nonlinear modes into several frequency bands, as shown in Figure 6.33. Complex coefficients weight the summation of the bandpass filtered signals, adjusting the frequency response of each nonlinear mode of $G_{DPD}(|x|)$ directly. The static frequency responses within Figure 6.33 are denoted by $H_n(\omega)$.

At first glance it may appear that the static filter bank requires significantly more FPGA resources to implement than a single adaptive FIR filter based on weighted delay taps. This is probably true. However, the static filter bank can be implemented efficiently using

infinite impulse response (IIR) structures [24] or by re-using the static filter bank for higher-order modes [25].

The re-use of the static filter bank for higher-order modes is shown in Figure 6.34. The squared envelope of the input signal $|x(k)|^2$ is filtered by a set of kernels whose frequency responses are denoted by $H_i(\omega)$. The resulting signals, denoted by $g_i(k)$, are used in all odd-order branches. The basis waveform for an odd polynomial order p and filter $H_i(\omega)$ is

$$\gamma_{p,i}(k) = g_i(k) \cdot |x(k)|^{p-3} \cdot x(k)$$

where

$$g_i(k) = \sum_{m=-M}^M |x(k-m)|^2 \cdot h_i(m)$$

and $h_i(m)$ are the static weights of the FIR filter used to implement $H_i(\omega)$. The predistortion signal $x_{DPD}(k)$ becomes

$$x_{DPD}(k) = [1 + \sum_{p=3, \text{odd}}^P \sum_i a_{p,i} \cdot g_i(k) \cdot |x(k)|^{p-3}] \cdot x(k).$$

The estimation of the DPD coefficients for a gain-based polynomial, either memoryless or memory-based, benefits from spectral weighting. Signals entering the estimator, such as the input signal $x(k)$, output signal $y(k)$, and basis waveforms $\gamma_i(k)$, should be filtered to notch the in-band frequencies. This prevents linear errors from biasing the DPD coefficient estimation and improves the distortion cancellation outside of the bandwidth of $x(k)$. The estimation filter $L\{\}$ is chosen so that $L\{x(k)\}$ is small. Note that the filter $L\{\}$ is applied only to the signals within the estimator, and not those signals used in the DPD function to create $x_{DPD}(k)$.

A MRAS estimator that includes an estimation filter $L\{\}$ minimizes

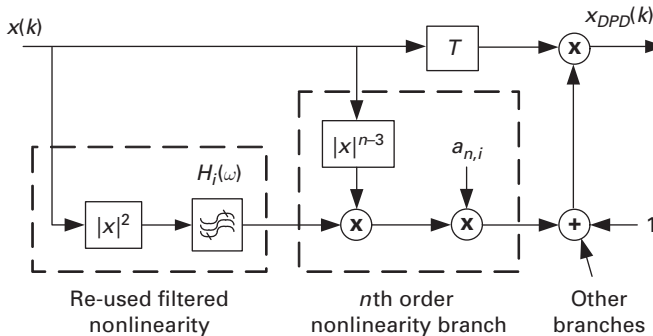


Figure 6.34 Gain-based memory correction where the static filter bank is re-used in each nonlinear mode of the DPD gain $G_{DPD}(|x(k)|)$.

$$J_{MRAS} = \sum_k \left| L\{\varepsilon_{MRAS}(k)\} - \sum_{i=1}^N \Delta a_i \cdot L\{\gamma_i(k)\} \right|^2.$$

The weighted LMS estimate of Δa_i is

$$\underline{\Delta a} = Q_L^{-1} \cdot Z_L^H \cdot L\{\varepsilon_{MRAS}\}$$

where

$$Q_L = Z_L^H Z_L,$$

$$Z_L = [L\{\gamma_1\} \quad \cdots \quad L\{\gamma_N\}].$$

The matrix Q_L is dependant on the estimation filter $L\{\}$ and the statistics of the input signal $x(k)$. Methods for avoiding conditioning problems in Q_L are discussed in [Section 6.5](#). The estimation filter $L\{\}$ can also be used to compensate for bandwidth limitation within the observation path, as will also be shown in [Section 6.5](#).

6.4.4 Cascaded DPD structures

Cascaded DPD structures are characterized by two or more separate adaptive blocks in series. Two examples of a memoryless nonlinearity and an adaptive filter in cascade are presented. The filter follows the nonlinearity for the Hammerstein model and precedes the nonlinearity for the Wiener model.

The Hammerstein model is shown in [Figure 6.35](#). The predistorted signal $x_{DPD}(k)$ is

$$x_{DPD}(k) = \sum_{m=-M}^M x_{MLN}(k-m) \cdot h_{post}(m)$$

where $h_{post}(m)$ are the coefficients of the adaptive post-filter and the memoryless nonlinearity (MLN) is

$$x_{MLN}(k) = \sum_{i=1}^N a_i \cdot |x(k)|^{i-1} \cdot x(k).$$

This structure differs from a memory polynomial because the same FIR filter is applied to each nonlinear mode. That is, the Hammerstein model is a sub-set of the memory

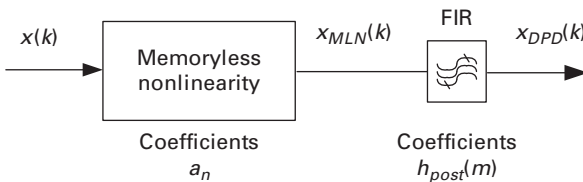


Figure 6.35 Hammerstein model of the DPD function.

polynomial. Another way that the Hammerstein model differs from the memory polynomial is that it is possible to implement the memoryless nonlinearity as a LUT [26].

Since it is possible to write the predistorted signal as

$$x_{DPD}(k) = \sum_{m=-M}^M \sum_{p=1}^P a_p \cdot h_{post}(m) \cdot |x(k-m)|^{p-1} \cdot x(k-m),$$

one might be tempted to define the basis waveforms as

$$\gamma_{p,m}(k) = |x(k-m)|^{p-1} \cdot x(k-m)$$

and use the MRAS estimator from the [previous subsection](#) to estimate the cascaded coefficient product $a_p h_{post}(m)$. However, decomposing the cascaded estimate into separate coefficients, a_p and $h_{post}(m)$, is not straightforward. It is much easier to use two STRs to estimate the memoryless polynomial coefficients a_p first, then estimate the coefficients of the adaptive post-filter, $h_{post}(m)$, as shown in [Figure 6.36](#).

The Wiener model is shown in [Figure 6.37](#). The predistorted signal $x_{DPD}(k)$ is

$$x_{DPD}(k) = \sum_{i=1}^N a_i \cdot |x_h(k)|^{i-1} \cdot x_h(k)$$

where

$$x_h(k) = \sum_{m=-M}^M x(k-m) \cdot h_{pre}(m)$$

and $h_{pre}(m)$ are the coefficients of the adaptive pre-filter. Unlike the Hammerstein model, the Wiener model is not a sub-set of the memory polynomial, but is a sub-set of the

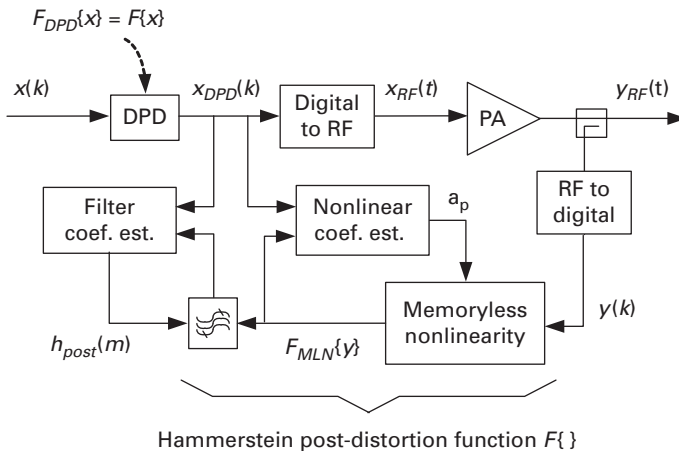


Figure 6.36 Adaptive Hammerstein DPD implemented using two cascaded STR estimators with post-distortion functions.

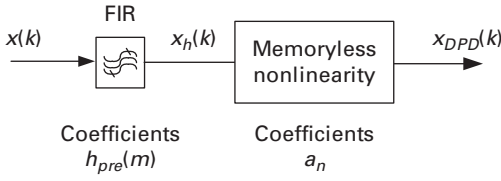


Figure 6.37 Wiener model of the DPD function.

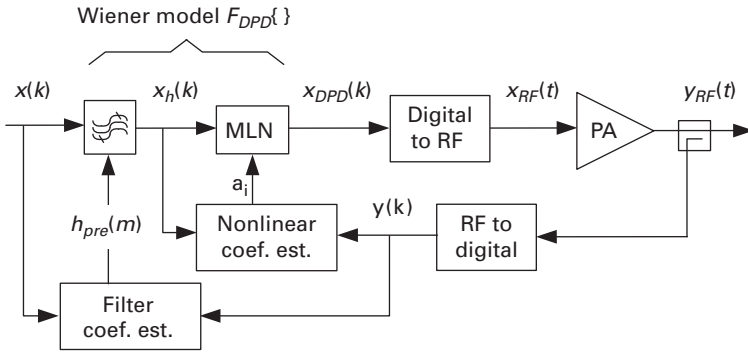


Figure 6.38 Adaptive Wiener DPD implemented using two cascaded MRAS estimators.

Volterra series. The coefficients of the adaptive pre-filter $h_{pre}(m)$ and the memoryless nonlinearity a_i are estimated using two MRAS structures, as shown in Figure 6.38.

The Hammerstein model for the DPD assumes that the PA response can be described accurately by a Wiener model. That is, we would like the PA nonlinear block to be followed immediately by the nonlinear block within the post-distortion function used in the STR-Hammerstein approach. Similarly, the Wiener model for the DPD assumes that the PA response can be described accurately by a Hammerstein model. We want the DPD nonlinear block used in the MRAS-Wiener approach to be followed immediately by the PA nonlinear block.

6.5 Advanced topics in DPD

The [previous section](#) described many basic aspects of an adaptive DPD system. In the following, advanced topics are discussed: model order selection, the use of learning to implement temperature compensation, and convergence properties.

6.5.1 Model order selection

Over-specifying the model order of the DPD for the input signal presented degrades the output spectrum. It is shown that the severity of degradation is reduced by selecting the appropriate estimation rank and increasing the estimation bandwidth.

The selection of the basis waveform set is important for the predistortion to achieve the required distortion reduction. A rich basis set that describes accurately the physical properties of the power amplifier provides good distortion cancellation, but only if the coefficients are selected correctly. Unfortunately, the input signal often lacks the statistical content in terms of amplitude variation and spectrum to excite fully the space spanned by the basis set. This causes the correlation matrix Q or Q_L to be rank deficient. Thus, the appropriate model order for an adaptive DPD system depends on the input signal $x(k)$ as much as the power amplifier nonlinearity.

The difference in the model orders of the DPD module and the coefficient estimator can be resolved using an eigenvalue decomposition of Q or Q_L . When the input signal $x(k)$ is not rich enough to provide a full rank correlation matrix, the number of independent coefficient adjustments is reduced to equal the number of non-zero eigenvalues present in Q or Q_L .

Let the eigenvectors of Q_L be denoted by v_j , where the index j is listed in order from the largest to smallest eigenvalue λ_j . The number of eigenvectors used, denoted by N_{MRAS} , represents the rank of the estimation. The rank reduced estimate of the coefficient error, denoted by $\Delta\alpha_i$, is

$$\underline{\Delta\alpha} = \Lambda_{MRAS}^{-1} \cdot V_{MRAS}^H \cdot Z_L^H \cdot L\{\underline{\varepsilon}_{MRAS}\}$$

where $\underline{\Delta\alpha} = [\Delta\alpha_1 \dots \Delta\alpha_{N_{MRAS}}]^T$, $\Lambda_{MRAS} = \text{diag}\{\lambda_1 \dots \lambda_{N_{MRAS}}\}$, and

$$V_{MRAS} = [v_1 \dots v_{N_{MRAS}}].$$

The DPD coefficients are updated recursively using

$$\underline{a}(t_{l+1}) = \underline{a}(t_l) - \mu_{\Delta} \cdot V_{MRAS} \cdot \underline{\Delta\alpha}$$

where $\underline{a} = [a_1 \dots a_N]^T$. Since only dominant eigenvectors are used, the number of adaptive coefficients estimated is reduced and condition problems are avoided. Eigenvalue decomposition may also be used to reduce the model order for an estimator using Q as the correlation matrix.

It is possible to redefine the basis set using the eigenvectors of Q_L . The new basis waveforms become

$$\gamma_{i,new} = v_i^T \cdot Z_L.$$

The resulting basis set will be orthogonal with respect to the input signal $x(k)$ and estimation filter $L\{\}$. This transformation improves the conditioning of the estimation by reducing the number of basis waveforms used, but it does not reduce the computational load on the FPGA, in general. However, it can be helpful when the digital domain has limited dynamic range.

The above-mentioned estimation assumes that the bandwidth of the output observation path is larger than the bandwidth of predistorted signal $x_{DPD}(k)$. This is not always the case when the polynomial order of the basis set is high. Consider an observation receiver that uses intermediate frequency (IF) sampling as shown in Figure 6.39. The IF filter in the observation path limits the bandwidth of $y(k)$.

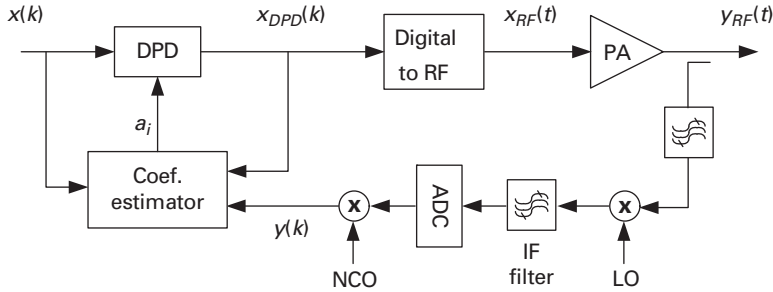


Figure 6.39 Digital transmitter with DPD and MRAS estimator where the observation path is implemented using an IF receiver with a narrow bandwidth IF filter.

Fortunately it is not necessary to measure the full spectrum of the IMD waveform to adapt the predistortion coefficients. Zhu [27] has shown that the output of a nonlinear function in response to a band limited input signal need only be sampled at the Nyquist rate of the input signal rather than the wider bandwidth of the output signal if the nonlinear function is a one-to-one continuous mapping. With respect to DPD applications, this means that the observation bandwidth must be as wide as the linear signal bandwidth to estimate the coefficients if the predistortion module is memoryless.

In general a coefficient estimate based on the correlation matrix Q will have poor IMD cancellation over a wide bandwidth because the observation filter is interpreted incorrectly as being memory within the PA. Any memory correction available within the predistortion module will degrade the distortion cancellation compared to a memoryless model. The deleterious effects of the limited observation bandwidth are minimized by using the Q_L estimator.

The basis waveforms within the estimator are filtered so that the comparison between the observed PA output and the basis waveforms is made over the identical frequency band. That is, the filters applied to the basis set and the output $y(k)$ must have the same frequency responses after accounting for the observation filter. The basis set used by the DPD module is not altered, making it possible to extrapolate the distortion cancellation beyond the observed bandwidth. However, there is no guarantee that the unseen IMDs will be reduced by the narrow bandwidth optimization. In such cases, the model order of the estimation affects the overall distortion cancellation performance. The model order N_{MRAS} should be selected to improve distortion cancellation within the observed bandwidth without over-fitting so that the IMDs outside are at least as low as those achievable using memoryless predistortion.

Consider the following example [28]. The input signal $x(k)$ is formed by two 5 MHz bands of noise separated by 15 MHz and amplified by a Doherty PA. The DPD is implemented using a gain memory polynomial having 76 coefficients ($N=76$) where the highest polynomial order is $p=5$. The estimation rank N_{MRAS} is reduced using eigenvalue decomposition and the bandwidth (BW) of the estimation filter is varied. The output spectra for a reduced rank of $N_{MRAS}=15$ and the bandwidths of 50 MHz and

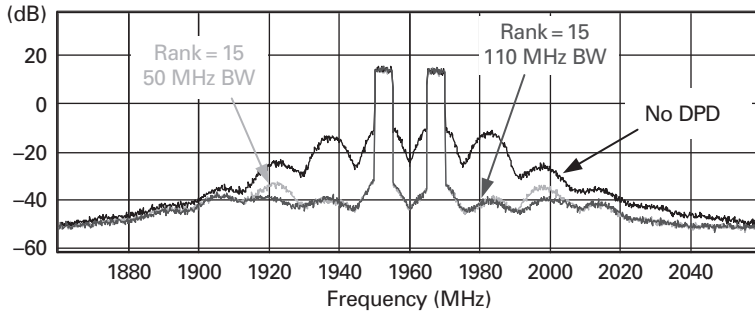


Figure 6.40 Spectrum of a Doherty PA with DPD applied using an estimation rank of $N_{MRAS} = 15$. The estimator bandwidths are 50 MHz and 110 MHz. The uncorrected spectrum is shown in black. Adapted from [28 © IEEE 2008].

110 MHz are shown in Figure 6.40. The ACPR rises outside of the observation band for the 50 MHz case because the rank is too high for the estimation bandwidth.

The rank of the estimation, N_{MRAS} , should be reduced if it is not possible to increase the bandwidth. The ACPR measured at 15 MHz and 30 MHz offsets from the nearest carrier are listed in Table 6.1 for ranks of 5 and 15 for the 50 MHz BW case and a rank of 15 for the 110 MHz BW case. The rank $N_{MRAS} = 15$, BW = 50 MHz case shows a degraded ACPR at the 30 MHz offset. The 30 MHz offset ACPR is improved by reducing the rank N_{MRAS} to 5 or by increasing the estimation BW to 110 MHz.

An alternative method of improving the condition of Q_L when the observation path has limited bandwidth is to measure several disjoint bands before updating the DPD coefficients. The LO frequency of the IF receiver is adjusted to change the center frequency of the observed bandwidth. The weighted LMS estimate of Δa_i becomes

$$\underline{\Delta a} = \left[\sum_n Q_L(\omega_n)^{-1} \right] \cdot \left[\sum_n Z_L(\omega_n)^H L\{\underline{\varepsilon}_{MRAS}\}(\omega_n) \right]$$

where $Q_L(\omega_n)$, $Z_L(\omega_n)$, and $L\{\underline{\varepsilon}_{MRAS}\}(\omega_n)$ are obtained using an estimation filter $L\{\}$ whose bandpass response is centered at frequency ω_n . The rank of $\sum Q_L(\omega_n)$ determines the condition of the estimation.

Table 6.1 ACPR for a Doherty PA DPD-corrected using different estimation ranks and estimation bandwidths (from [28] © IEEE 2008).

Adjacent channel offset	Rank = 5 50 MHz BW	Rank = 15 50 MHz BW	Rank = 15 110 MHz BW
−15 MHz	−51.9 dB	−55.3 dB	−54.1 dB
+15 MHz	−51.6 dB	−55.1 dB	−55.0 dB
−30 MHz	−51.1 dB	−48.4 dB	−54.4 dB
+30 MHz	−53.2 dB	−49.7 dB	−54.5 dB

Adding several narrow observation bands can be challenging because of the need to time align data captures of $x(k)$ and $y(k)$. Ensuring that each band overlaps partially with the input signal $x(k)$ makes time alignment easier. Higher-order basis waveforms derived from $x(k)$ may be used for time alignment when the observation band is outside of the linear signal bandwidth. However, this works only when the output signal $y(k)$ contains residual distortion. Note the time alignment is performed prior to applying the estimation filter $L\{\}$.

The widest estimation bandwidth required is determined in part by the output filtering present between the PA and the antenna. This may be in the form of an output filter or a duplexer. For example, if the transmit band is limited to 60 MHz by an output filter, the spectral regrowth appearing outside of the 60 MHz band is not a serious concern. As a result, an estimation bandwidth wider than the output filter bandwidth is not necessary from an ACPR point-of-view.

6.5.2 Learning and temperature compensation

The discussion of the learning module, introduced in [Section 6.3](#), is continued, providing details of the implementation. As mentioned earlier, learning is best used to reduce the convergence time in response to abrupt changes in the operating condition. The learning database can also be used for temperature compensation. A known temperature compensation implementation is reviewed and an alternative approach using basis waveforms and the learning database is proposed.

In the following, several aspects of the learning module are detailed. First, the composition of the entries within the learning database is defined. Next, a method of identifying and pruning redundant entries is presented. Finally, it is shown how entries within the learning database whose past operating conditions are most similar to the present condition are identified. The selected entries are referred to as nearest neighbors in [\[11\]](#).

Entries in the learning module must contain the information needed to correlate the operating condition to a successful DPD coefficient setting. Thus, the learning module requires measurements of the operation conditions (OCs) affecting the PA nonlinearity. In general, this will include several parameters creating a multi-dimensional OC vector, denoted by $\underline{u}$. Examples of the OC parameters include the average input power and the ambient temperature. The model order of the coefficient estimation N_{MRAS} can also be considered an OC parameter. Each database entry $C(n)$ is defined by its DPD coefficients $a_i(n)$ and the corresponding OC parameters $u_i(n)$: that is,

$$C(n) = [\underline{a}^T(n) \quad \underline{u}^T(n) \quad t(n) \quad \rho(n)]$$

where $t(n)$ is a time stamp and $\rho(n)$ is a group number. A new entry is added to the learning database when the estimate of the DPD coefficients a_i converges.

It is necessary to prune the learning database by removing redundant entries. Let the OC distance between two entries $C(n_1)$ and $C(n_2)$ be

$$d_{OC}(n_1, n_2) = [\underline{u}(n_1) - \underline{u}(n_2)]^T \cdot W \cdot [\underline{u}(n_1) - \underline{u}(n_2)]$$

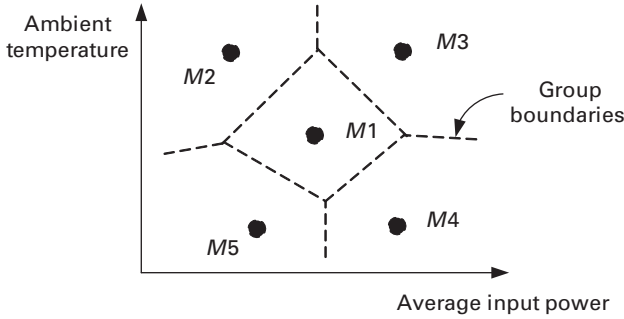


Figure 6.41 Markers placed within the operating condition (OC) space to define disjoint groups p for the learning database.

where W is a matrix whose diagonal elements are weights of the OC dimensions. Database entries separated by a small distance and belonging to the same group are considered redundant. The older entry is deleted.

The group number p is assigned to prevent excessive pruning. Markers denoted by M_j are placed throughout the OC space, as shown in Figure 6.41, based on the anticipated variation in the coefficient settings a_i . The group number of an entry indicates the index of the closest marker, as measured by $d_{OC}(n, M_j)$. Each group retains at least one entry after pruning.

It is necessary to identify the database entries nearest to the current OC parameters. Assume the DPD coefficients used currently in the adaptive controller and the current OC vector are stored in the database as entry $C(0)$. The distances $d_{OC}(0, n)$ between $C(0)$ and the entries $C(n)$ within the learning database are measured. The entries having small distances are considered nearest neighbors and are used as candidates for replacing the adaptive DPD coefficients. The STR-based test shown in Figure 6.23 can be used to determine the best of the candidates from the learning database.

As mentioned earlier in Section 6.3, the coefficients from the nearest neighbor entry can be inserted directly into the DPD function $F_{DPD}\{x(k)\}$ when the adaptive solution is not sufficient. This approach would be used if the STR-based test is not feasible. However, after retrieving the nearest neighbor, the learning database must be inhibited until the current operating condition changes enough that a new entry is identified as the nearest neighbor. This allows the adaptive controller to converge without interruption when no new information is available from the learning module.

Now consider temperature compensation. An effective method is demonstrated in [29]. The approach uses two LUTs to control the DPD gain G_{DPD} : a nominal LUT which is the same as a memoryless DPD function, and a delta LUT which represents the sensitivity of the DPD gain to thermal variations. An external device is used to correlate self-heating of the transistor with the input signal power and to compute a scale term for weighting the contribution of the delta LUT to the DPD gain G_{DPD} .

The experimental results presented in [29] demonstrate the value of thermal compensation and the effectiveness of the approach. For example, failing to correct for temperature variations will often result in an undesired gain droop when the input signal power

risks and holds for an extended period. However, the delta LUT in [29] is obtained experimentally and few details are provided on how this could be implemented within an adaptive framework. In addition, the external device used to correlate the input signal power to the self-heating of the transistor is not discussed; however, the example provided describes the PA's thermal model as an eighth-order Butterworth low-pass filter with a time constant of 2 ms.

In the following, an alternative temperature compensation method is proposed (untested) which uses the basis waveform representation of the DPD function $F_{DPD}\{x(k)\}$ so that shorter-term memory is also corrected. The computation of the sensitivity of the DPD to thermal variations is made adaptive using the past coefficients settings and operating conditions stored in the learning database. A model for predicting the heat accumulated within the transistor is also provided.

Assume the transfer of heat from the transistor to the sink is described by time constants that are known or measured experimentally. The temperature variation in the transistor is modeled as a function of the input signal power, as shown in Figure 6.42. The input power $|x(k)|^2$ is low-pass filtered to match the thermal time constants and an offset, denoted by $P_{ambient}$, is included to account for changes in the ambient temperature [30]. The thermal flow model becomes

$$P_{thermal}(k) = P_{ambient} + \sum_m |x(k-m)|^2 \cdot h_{thermal}(m)$$

where $h_{thermal}$ are the coefficients of the low-pass filter. In most cases the lowpass filter would be implemented as a weighted sum of low-order IIR filters, one for each discernible time constant, instead of using the FIR kernel $h_{thermal}$ listed above. The latter was used for notational convenience.

Temperature compensation must also determine the sensitivity of the DPD function $F_{DPD}\{x(k)\}$ to thermal variations: that is,

$$\frac{\Delta F_{DPD}}{\Delta P} = \sum_{i=1}^N \delta_{thermal(i)} \cdot \gamma_i(k)$$

where $\delta_{thermal(i)}$ are the thermal coefficients. The estimate of the thermal coefficients is based on several neighboring entries $C(n)$ retrieved from the learning database:

$$\delta_{thermal(i)} = \frac{\sum_n [a_i(n) - a_i(0)] \cdot (P_n - P_0)}{\sum_n (P_n - P_0)^2}$$

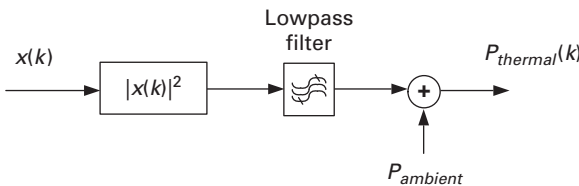


Figure 6.42 Thermal flow model representing heat generated by the transistor and dissipated by the heat sink.

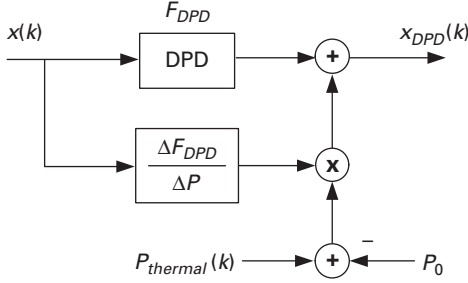


Figure 6.43 DPD with thermal compensation.

where $a_i(n)$ and P_n are the DPD coefficients and the average input power for database entry $C(n)$, respectively, and $a_i(0)$ and P_0 are those associated with the nominal predistortion function $F_{DPD}\{x(k)\}$. The value P_n is a long-term average spanning at least one data capture interval. The integration interval is assumed to be significantly longer than the time constants used to obtain $P_{thermal}(k)$.

The DPD function with temperature compensation becomes

$$x_{DPD}(k) = F_{DPD}\{x(k)\} + \Delta F_{DPD}\{x(k)\}$$

where

$$\Delta F_{DPD}(k) = \frac{\Delta F_{DPD}}{\Delta P} \cdot [P_{thermal}(k) - P_0].$$

The block diagram is shown in Figure 6.43. Temperature compensation can be important for input signals $x(k)$ containing bursts, such as HSDPA or TDD signals. However, temperature compensation for such signal dynamics is performed in an open-loop manner. Its use, in practice, is considered only when the DPD system can not meet the specifications without it.

6.5.3 Comments on convergence properties

Convergence properties of the coefficient estimation algorithms presented earlier are discussed. A comparison to alternative methods that update the DPD coefficients a_i after each time sample k , instead of after each data capture, is made.

Recall that the DPD coefficient update is

$$\underline{a}(t_1) = \underline{a}(t_0) - \mu_\Delta \cdot Q^{-1} \cdot Z^H \cdot \underline{\varepsilon}.$$

There are three components to the update: the correlation matrix Q , the measurement vector $Z^H \underline{\varepsilon}$, and the convergence constant μ_Δ . Each affects the convergence properties of the DPD coefficient estimation.

Let the basis waveforms sampled at time k be denoted by

$$\psi(k) = [\gamma_1(k) \quad \cdots \quad \gamma_N(k)].$$

We can rewrite Q and $Z^H \underline{\varepsilon}$ as

$$\underline{Q}_{1:K} = \sum_{k=1}^K \psi^H(k) \cdot \psi(k)$$

$$\underline{Z}^H \cdot \underline{\varepsilon} = \sum_{k=1}^K \psi^H(k) \cdot \varepsilon(k)$$

where $\underline{Q}_{1:K}$ indicates a correlation matrix computed from a data capture spanning the time interval $[1:K]$. The value of K affects the rate of convergence and the steady-state coefficient error variance. The selected K should be large enough to provide averaging needed to obtain a good estimate of $\underline{Q}$ and suppress measurement noise in $\underline{Z}^H \underline{\varepsilon}$. However, an excessive K will result in slow convergence. It can be beneficial to use a small value of K initially while $|\Delta a_i|$ is large and increase K as the DPD coefficients converge.

Let us look at the effect of approximating $\underline{Q}^{-1}$ on the convergence. Assume the approximation of $\underline{Q}^{-1}$ is denoted by $\underline{S}$. The DPD coefficient update becomes

$$\underline{a}(t_1) = \underline{a}(t_0) - \mu_\Delta \cdot \underline{S} \cdot \underline{Z}^H \cdot \underline{\varepsilon}.$$

The coefficient error $\underline{\Delta a}$ at a time t_0 can be defined as

$$\underline{\Delta a}(t_0) = \sum_i \Delta a_i(t_0) \cdot \underline{v}_{SQ(i)}$$

where $\underline{v}_{SQ(i)}$ are the eigenvectors of the matrix product $\underline{S}\underline{Q}$ and $\Delta a_i(t_0)$ are complex weights at time t_0 . After n iterations, we have

$$\Delta a_i(t_n) = \Delta a_i(t_0) \cdot (1 - \mu_\Delta \cdot \lambda_{SQ(i)})^n$$

where $\lambda_{SQ(i)}$ is the i th largest eigenvalue of $\underline{S}\underline{Q}$. If $\mu_\Delta = 1/\lambda_{SQ(1)}$, the weights Δa_i associated with the dominant eigenvectors will converge quickly towards zero whereas Δa_i for the remaining eigenvectors of $\underline{S}\underline{Q}$ will converge slowly because $\mu_\Delta \lambda_{SQ(i)} \ll 1$. Thus, it is important for convergence that $\underline{S}$ be selected to minimize the spread of the eigenvalues of $\underline{S}\underline{Q}$. Note that for a reduced-order estimation (see [Subsection 6.5.1](#)), only the spread of the N_{MRAS} largest eigenvalues is important.

The selection of K affects the accuracy of $\underline{Z}^H \underline{\varepsilon}$. Noise in the measurement of $\varepsilon(t)$ contributes to the variance in $\underline{Z}^H \underline{\varepsilon}$ as well as the variance in the steady-state estimate of the DPD coefficients a_i . The averaging provided by K reduces the steady-state variance.

Let us now look at the role of the convergence constant μ_Δ . The minimum requirement is that $\mu_\Delta < 2/\lambda_{SQ(1)}$. Anything higher will cause instability in the coefficient estimator. The coefficient estimation sequence will exhibit an oscillatory behavior with a stable decay along the dominant eigenvectors when $1/\lambda_{SQ(1)} > \mu_\Delta > 2/\lambda_{SQ(1)}$. A stable, well-damped decay is ensured along all eigenvectors when $\mu_\Delta < 1/\lambda_{SQ(1)}$.

Once stability is established, the choice of μ_Δ is a trade-off between the rate of convergence and the steady-state variance. Convergence is defined by the rate at which the coefficient error $|\Delta a_i|$ decays. The decay rate is at least $(1 - \mu_\Delta \lambda_{SQ(N,MRAS)})$ per iteration, where $\lambda_{SQ(N,MRAS)}$ is the smallest eigenvalue of importance for an estimator using a reduced order of N_{MRAS} . The steady-state variance is determined by the variance

in $\mu_\Delta Z^H \underline{\varepsilon}$. Thus, selecting a low μ_Δ will reduce the steady-state variance, but will cause the convergence rate to be slow.

It is possible to update the DPD coefficients a_i after each time sample k instead of waiting for the complete data capture. If Q is measured using a sliding window spanning $[k-K+1:k]$, the update for time sample k becomes

$$\underline{a}(k+1) = \underline{a}(k) - \mu_\Delta \cdot Q_{k-K+1:k}^{-1} \cdot \psi^H(k) \cdot \varepsilon(k).$$

The DPD coefficients a_i will converge following an exponential decay instead of being reduced in discrete steps after each data capture. However, the convergence properties remain largely unchanged because the increase in the update rate by a factor of K has been offset by a reduction in the measurement vector $Z^H \underline{\varepsilon}$ by a factor of $1/K$.

Instead of using a sliding window, Q can be estimated using a recursive method:

$$Q_{2:K+1} = \mu_Q \cdot Q_{1:K} + \psi^H(K+1) \cdot \psi(K+1)$$

where

$$\mu_Q = \frac{K-1}{K}.$$

This reduces the storage requirements, compared to using data captures of $K = 16k$ samples, making it suitable for a FPGA implementation of the estimator. However, it should be apparent that a recursive estimation of Q does not improve the rate of convergence over the sliding window estimate. The value of K is still the most important parameter.

Some researchers use a recursive least squares (RLS) method that is similar to above except that the recursive equation is applied to Q^{-1} directly instead of Q . Comparisons of the rate of convergence are made frequently between the RLS and LMS approaches where the superiority of the former is almost always claimed. However, in most cases, the definition being used for LMS is actually the steepest descent approximation [31] whose DPD coefficient update is

$$\underline{a}(k+1) = \underline{a}(k) - \mu_\Delta \cdot \psi^H(k) \cdot \varepsilon(k).$$

That is, Q^{-1} is approximated by the identity matrix I . The constant μ_Δ must be reduced by a factor of $1/\lambda_{Q(1)}$ to compensate for inaccuracies in the approximation, where $\lambda_{Q(1)}$ is the largest eigenvalue of Q (which is proportional to K). The only way by which the steepest descent will provide good convergence is if the basis waveforms are redefined (orthogonalized) using the eigenvectors of the original Q , as described earlier in [Subsection 6.5.1](#), and normalized by their respective eigenvalues.

The superior convergence properties of the RLS implementation is a side-effect of the fact that the initial estimate of Q is often selected conservatively (too small) and $\mu_Q = 1$. This is equivalent to making K small initially while $|\Delta a_i|$ is large and increasing K over

time. This will show favorable convergence properties such as a fast initial rate and low steady-state variance. That is, it will show favorable convergence properties until the operating condition changes abruptly. The system will not adapt to the change unless either Q is reset or reduced, or $\mu_Q < 1$. Fortunately, this can be done easily within the RLS framework.

In summary, the RLS algorithm is useful when it is understood by the user and implemented correctly. However, the same can be said for the steepest descent implementation. How useful is a comparison of the convergence rates for RLS and steepest descent implementations when the eigenvalue spread of the latter can be increased arbitrarily by scaling the input signal $x(k)$ prior to computing the basis waveforms? If both the RLS and steepest descent are optimized to their fullest extent, their performances are equivalent. Despite this, many researchers see value in the further study of the fact that a sub-optimal implementation of the RLS algorithm performs better than a sub-optimal implementation of the steepest descent algorithm.

6.6 Concluding comments

Within this concluding section, important DPD related topics omitted from the chapter are mentioned, along with a list of key researchers and companies in the DPD field. Thoughts of the author regarding future research directions and a summary of the chapter are also provided.

6.6.1 Other work and key researchers

There are several important DPD-related topics not covered in this chapter due to space limitations. Some of the omitted topics are listed below with references. In addition, several researchers and companies are mentioned whose names can be used for productive Internet searches for DPD-related papers.

Artificial neural networks are being used to implement nonlinear DPD functions which include FIR feedback, as shown in [Figure 6.10](#). A recent example can be found in [\[32\]](#). The effects of linear degradations within the transmitter, such as modulator imbalances and equalization errors, on the DPD performance have been shown in [\[29\]](#) and [\[33\]](#). Digital predistortion systems specific to power amplifiers based on envelope tracking are available [\[34\]](#).

There are many researchers studying digital predistortion of power amplifiers. Professor J. Cavers published much of the early work. Professor J. S. Kenney and Professor F. Ghannouchi have led successful DPD and PA research groups at their respective universities for several years. Professor P. Asbeck's group is well known for DPD and PA research related to envelope tracking. Some of the newer names supervising DPD research efforts include Professor S. Boumaiza and Professor A. Zhu. In the related area of power amplifier modeling, searches for papers co-authored by Dr. J. Wood or Professor J. C. Pedro are recommended. Professor S. C. Cripps is a required reference for anything related to power amplifiers [\[3\]](#).

Many DPD systems have been developed by companies such as Texas Instruments, Optichron, PMC Sierra, and Intersil. Vendors such as Nijura have produced DPD systems to support sales of their core product line relating to envelope modulation. Many FPGA vendors, such as Xilinx and Alterra, have developed DPD algorithms optimized for their products as a form of sales support. However, all of these companies face a difficult challenge in having to design a DPD system without knowing the characteristics of the power amplifier or input signal format in advance.

6.6.2 Future work

At present, few DPD systems are intended to be used for input bandwidths wider than 20 MHz. However, LTE-Advanced and neutral host applications can have wider bandwidths, such as 100 MHz. Wider bandwidths require faster digital components and more detailed DPD models, in general. However, an output filter placed after the PA attenuates a portion of the intermodulation distortion as the bandwidths become wider. This may result in a relaxation of the requirements on the DPD system. An understanding of the sampling requirements for representing nonlinear modes in the digital domain becomes important.

Sampling requirements for nonlinear modes have been discussed briefly in [2]. Aliasing of nonlinear modes is an interesting phenomenon that deserves more of our attention. In particular, the effect of an aliased nonlinear mode on the adaptation of the estimator is not understood fully, but clearly has the potential for creating problems for the DPD system. To date, having the luxury of over-sampling the digital domain has reduced the practical impact of such problems.

Coefficient estimation techniques using narrow bandwidth observation signals, such as those described in [28] and [Subsection 6.5.1](#), should prove useful for those seeking to increase the bandwidth of a DPD system.

6.6.3 Conclusion

Many aspects of digital predistortion have been discussed: the DPD module, the coefficient estimator, and the use of learning. Within each of these topics, several choices are available. For the DPD, does one use a nonlinear mapping $F_{DPD}\{x(k)\}$ or a gain-based function $G_{DPD}(|x(k)|)$? Should a memoryless or memory model be used and what is the appropriate DPD model order N ? Which is better: LUT, basis waveforms, or cascaded structures? For the estimator, does one use the MRAS or STR structure, or a combination? Should an estimation filter be used and what is its frequency response? What is the appropriate rank N_{MRAS} of the estimator? Should learning be used to improve convergence? Is temperature compensation needed? Even with all the possible combinations, this list is not complete.

To answer the question “which is best?,” it is important to remember that the goal of digital predistortion is to meet the specification of the modulation format used to create the input signal $x(k)$. There is no prize for “destroying” the ACPR specification by 10 dB.

Whatever combination of DPD function and estimator structure passes the specification with sufficient margin for the manufacturing process is acceptable. After the pass/fail criterion, DPD implementations are judged by the amount of computing resources required and the increase in efficiency of the power amplifier made possible.

In practice, most people will not have a choice in the design of the DPD system. Instead, they will inherit a legacy or commercial system that may require modifications when the input signal format changes or when a new power amplifier design is needed. In such cases, it is important to be able to determine the structure of the DPD system so that extensions can be made effectively should the existing system fail to meet the specifications.

References

- [1] “Universal Mobile Telecommunications System (UMTS); Base station (BS) radio transmission and reception (FDD) (3GPP TS 25.104 ver. 6.3.0 Rel. 6),” ETSI TS 125 104 V6.3.0 (2003–09).
- [2] R. N. Braithwaite, “Adaptive digital predistortion of nonlinear power amplifiers using reduced order memory correction,” presented at workshop on “Highly efficient linear power transmitters for wireless applications based on switching mode amplifiers,” 2008 *IEEE MTT-S International Microwave Symposium*, Atlanta, GA, June 15–20, 2008.
- [3] S. C. Cripps, *RF Power Amplifiers for Wireless Communications*, Norwood, MA: Artech House, 1999.
- [4] K. J. Astrom, “Adaptive feedback control,” *Proceedings of the IEEE*, vol. **75**, no. 2, pp. 185–217, Feb. 1987.
- [5] R. N. Braithwaite and S. Carichner, “An improved Doherty amplifier using cascaded digital predistortion and digital gate voltage enhancement,” 2009 *IEEE MTT-S International Microwave Symposium*, Boston, MA, June 7–12, 2009, pp. 1073–1076.
- [6] R. N. Braithwaite and S. Carichner, “An improved Doherty amplifier using cascaded digital predistortion and digital gate voltage enhancement,” *IEEE Transactions on Microwave Theory and Techniques*, vol. **57**, no. 12, pp. 3118–3126, Dec. 2009.
- [7] L. Ding, G. T. Zhou, D. R. Morgan, *et al.* “A robust digital baseband predistorter constructed using memory polynomials,” *IEEE Trans. Comm.*, vol. **52**, no. 1, pp. 159–165, January 2004.
- [8] L. Rexberg “Digital linearization of power amplifiers – a control problem,” 2002. Available: www.s3.kth.se/contorl/reglermote/papers/rexberg.pdf
- [9] R. N. Braithwaite, “Low cost, low delay UMTS power amplifier using digital-controlled adaptive analog predistortion”, in *Proceedings 36th European Microwave Conference*, Manchester, UK, Sep. 10–15, 2006, pp. 1641–1644.
- [10] R. N. Braithwaite, “Measurement and correction of residual nonlinearities in a digitally predistorted power amplifier,” 75th *ARFTG Microwave Measurement Conference*, Anaheim, CA, May 28, 2010, pp. 14–17.
- [11] R. N. Braithwaite, “A self-generating coefficient list for machine learning in RF power amplifiers using adaptive predistortion,” *IEEE 36th European Microwave Conference*, Manchester, UK, Sept. 10–15, 2006, pp. 1229–1232.

- [12] C. Luke, I. Johnson, M. Cope, A. Mansell, and S. A. Wood, "System and method for forward path gain control in a digital predistortion linearized transmitter," US patent 7,590,190, Sep. 15, 2009.
- [13] J. K. Cavers, "New Methods for Adaptation of Quadrature Modulators and Demodulators in Amplifier Linearization Circuits", *IEEE Transactions on Vehicular Technology*, vol. **46**, no. 3, pp 707–716, August 1997.
- [14] J. K. Cavers, "A linearizing predistorter with fast adaptation," *Proc IEEE Vehicular Technology Conference*, pp. 41–47, Orlando, May 1990.
- [15] M. Faulkner, T. Mattsson and W. Yates, W., "Adaptive linearisation using pre-distortion," *Proc. IEEE Vehicular Technology Conference*, pp. 35–40, Orlando, May 1990.
- [16] J. K. Cavers, "Amplifier linearization using a digital predistorter with fast adaptation and low memory requirements," *IEEE Transactions on Vehicular Technology*, vol. **39**, no. 4, pp. 374–382, November 1990.
- [17] J. K. Cavers, "Optimum table spacing in predistorting amplifier linearizers," *IEEE Transactions on Vehicular Technology*, vol. **48**, no. 5, pp. 1699–1704, September, 1999.
- [18] O. Hammi, F. Ghannouchi, S. Boumaiza, and B. Vassilakis, "A data-based nested LUT model for RF power amplifiers exhibiting memory effects," *IEEE Microwave and Wireless Components Letters*, vol. **17**, no. 10, pp. 712–714, October 2007.
- [19] E. G. Lima, T. R. Cunha, H. M. Teixeira, M. Pirola, and J. C. Pedro, "Base-band derived Volterra series for power amplifier modeling," *2009 IEEE MTT-S International Microwave Symposium*, Boston, MA, June 7–12, 2009, pp. 1361–1364.
- [20] A. Zhu, J. C. Pedro, and T. R. Cunha, "Pruning the Volterra series for behavioral modeling of power amplifiers using physical knowledge," *IEEE Transactions on Microwave Theory and Techniques*, vol. **55**, no. 5, pp. 813–821, May 2007.
- [21] A. Zhu, P. J. Draxler, J. J. Yan, *et al.* "Open-Loop digital predistorter for RF power amplifiers using dynamic deviation reduction-based Volterra series," *IEEE Transactions on Microwave Theory and Techniques*, vol. **56**, no. 7, pp. 1524–1534, July 2008.
- [22] J. Staudinger, J. Nanan, and J. Wood, "Memory fading Volterra series model for high power infrastructure amplifiers," *2010 IEEE Radio Wireless Symposium*, New Orleans, LA, Jan. 10–14, 2010, pp. 184–187.
- [23] J. Kim and K. Konstantinou, "Digital predistortion of wideband signals based on power amplifier model with memory," *Electronics Letters*, vol. **37**, no. 23, pp. 1417–1418, Nov. 2001.
- [24] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, 4th edn., Upper Saddle River, NJ: Pearson Prentice Hall, 2007. (Figure 9.2.5, page 571).
- [25] R. N. Braithwaite, "Memory correction of a Doherty power amplifier with a WCDMA input using digital predistortion," *2006 IEEE MTT-S International Microwave Symposium*, San Francisco, CA, June 11–16, 2006, pp. 1526–1529.
- [26] E. G. Jeckeln, H. Shih, E. Martony, and M. Eron, "Method for modeling amplitude and bandwidth dependent distortion in nonlinear RF devices," *2003 IEEE MTT-S International Microwave Symposium*, Philadelphia, PA, June 8–13, 2003, pp. 1733–1736.
- [27] Y.-M. Zhu, "Generalized sampling theorem," *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, vol. **39**, no. 8, pp. 587–588, Aug. 1992.
- [28] R. N. Braithwaite, "Wide bandwidth adaptive digital predistortion of power amplifiers using reduced order memory correction," *2008 IEEE MTT-S International Microwave Symposium*, Atlanta, GA, June 15–20, 2008, pp. 1517–1520.

- [29] Intersil application note AN1022, “Operation and performance of the ISL5239 predistortion linearizer,” July 2002.
- [30] S. A. Wood, Personal communications, 2004.
- [31] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd edn., Reading, MA: Addison-Wesley, 1984.
- [32] F. Mkadem, M. B. Ayed, S. Boumaiza, J. Wood, and P. Anen, “Behavioral modeling and digital predistortion of power amplifiers with memory using two hidden layers artificial neural networks,” *2010 IEEE MTT-S International Microwave Symposium*, Anaheim, CA, May 23–28, 2010, pp. 656–659.
- [33] J. K. Cavers, “The effect of quadrature modulator and demodulator errors on adaptive digital predistorters for amplifier linearization,” *IEEE Transactions on Vehicular Technology*, vol. **46**, no. 2, pp. 456–466, May 1997.
- [34] A. Zhu, P. J. Draxler, H. Chin, *et al.* “Digital predistortion for envelope-tracking power amplifiers using decomposed piecewise Volterra series,” *IEEE Transactions on Microwave Theory and Techniques*, vol. **56**, no. 10, pp. 2237–2247, Oct. 2008.

7 Power amplifier nonlinear modeling for digital predistortion

Gabriel Montoro and Pere L. Gilabert

7.1 Introduction

Nowadays, one of the main common objectives in all Electrical Engineering research areas consists of reducing energy consumption by enhancing power efficiency. It is well known that the power amplifier (PA) is one of the most power hungry devices in radiocommunications. Therefore, to amplify non-constant envelope modulated signals, the use of linear Class-A PAs operating at high-power back-off levels to guarantee the desired linearity is no longer a desirable solution since it results in power inefficiency. In a classical Cartesian I-Q transmitter with static supply, the PA has to linearly amplify a carrier signal which is both phase and amplitude modulated and usually showing high peak-to-average power ratios (PAPRs), which implies that for having linear amplification it is necessary to use extremely inefficient class-A or class-AB PAs. Power amplifier system level linearizers, such as digital predistortion (DPD), extend the linear range of power amplifiers which, properly combined with crest factor reduction (CFR) techniques [1], enable PAs to be driven harder into compression (thus more efficient) while meeting linearity requirements.

Thanks to the intensive processing capabilities offered by the “always faster” digital signal processors, some power supply control architectures with great potential for high-efficiency operation have been revived. The PA drain supply modulation is carried out using techniques such as envelope elimination and restoration (EE&R) [2] and envelope tracking (ET) [3],[4] in conjunction with DPD. Therefore, the use of linearizers, and more precisely DPD, becomes an essential solution to mitigate nonlinear distortion effects arising from the use of more efficient but highly nonlinear PAs (Class D, E, F switched PAs) in both Cartesian and Polar transmitter architectures.

Typical and commonly used in several communication applications Class AB PAs show efficiencies ranging from 5 to 10 percent when operated with significant back-off levels. However, introducing CFR and adaptive DPD techniques, the efficiency can be improved by a factor of 3 to 5. In addition, by using more efficient topologies such as Doherty amplifiers, or even Class AB amplifiers with dynamic ET in combination with DPD, the efficiency can be improved up to 50 percent [5].

Moreover, due to the fact that modern communication standards use high-speed envelope signals exhibiting significant bandwidths, memory effects have to be taken into account. Memory effects do not affect equally all linearization techniques, for example, feedback or feedforward linearizers are less sensitive than DPD. Digital predistortion linearization is quite sensitive to memory effects, which can be a drawback in cancelling distortion in wideband signals, since its linearization performance is reduced [6]. Thanks to the envelope filtering technique the inverse memory effects that are generated inside the PA can be reproduced aiming at a later cancellation. For this reason, in the following sections an overview on PA behavioral models capable of reproducing PA nonlinear behavior and memory effects will be presented. It is important to emphasize that the selection of a proper behavioral model is a crucial initial step towards DPD.

Therefore, the rest of this chapter is organized into three sections. [Section 7.2](#) is devoted to explaining the basic principles behind the predistortion linearization technique. In [Section 7.3](#), we present an overview on some of the most common PA behavioral models published in the literature, because they are the starting point to the development of advanced DPD algorithms. Finally, [Section 7.4](#) summarizes the main conclusions.

7.2 Principles of predistortion linearization

Conceptually, non-adaptive memoryless predistortion is a linearization technique quite easy to understand since it consists of preceding the PA with a device called predistorter in order to counteract the nonlinear characteristic of the PA. The objective of the predistorter is then to ideally reproduce the inverse PA nonlinear behavior and as a result have linear amplification at the output of the PA.

[Figure 7.1](#) shows the basic principles of an open-loop predistorter. Several solutions have been developed to realize the predistorter, from digital baseband processing to processing the signal directly at radiofrequency using diodes as nonlinear devices. Moreover, most current predistortion solutions already introduce some kind of feedback mechanism in order to allow *adaptive predistortion* that permits a more robust operation of the linearizer. Therefore all the different solutions proposed for the realization of the

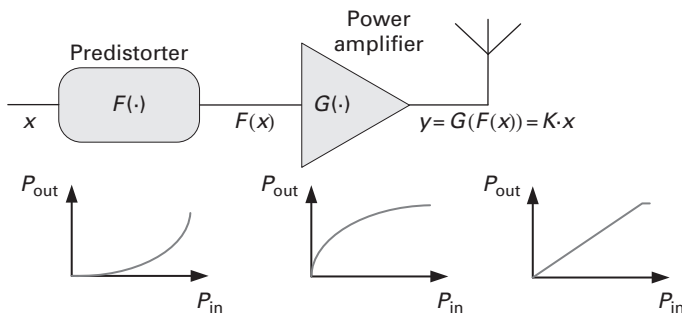


Figure 7.1 Fundamentals of predistortion linearization.

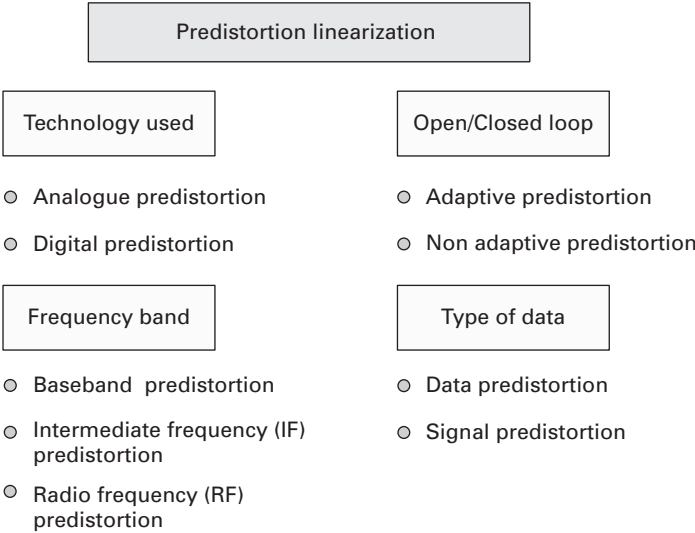


Figure 7.2 Classification of predistortion linearizers.

predistorter can be somehow classified according to some specific categories. An example of possible alternatives in the realization of the predistorter is depicted in Figure 7.2.

According to the position of the predistorter in the complete transmitter, predistortion can be carried out at *radiofrequency* (RF), *intermediate frequency* (IF) or *baseband* (BB). Predistortion at IF and BB have the advantage of being independent of the final frequency band of operation, as well as more robust concerning environmental parameters. Besides, the cost of ADC (analog-to-digital converters) and DAC (digital-to-analog converters) decreases at low frequency operation. One drawback regarding the predistortion compensation at IF or BB is the increasing linearity requirements because the up-converters can introduce additional distortion. However, up-converters (or I&Q modulators) can be avoided operating at IF and using software radio techniques, such as the so-called *IF sampling*. Another significant issue is nonlinear distortion in the feedback loop (e.g. introduced by down-converters) when considering adaptive predistortion. This nonlinear distortion does not have to be compensated but can mask the open loop linearity performance and produce unwanted nonlinear compensations.

One example of *RF predistortion* using analog devices is the cubic predistorter shown in Figure 7.3. The cubic RF predistorter is aimed at cancelling third-order intermodulation distortion (IMD) by adding a properly corrected in amplitude and phase cubic component to the input signal. In the case of a bandpass system only the third-order IMD products are usually reduced. When a higher linearity improvement is required, predistortion is usually combined with feedforward architectures. Figure 7.3 shows the block diagram of a typical cubic predistorter. The input signal is split into two paths. In order to ensure a perfect match in the recombination process of both split signals, a time-delay element is added in the upper path. The lower path is formed by a cubic nonlinearity performing the nonlinear predistortion, a gain (variable attenuator) and phase (variable phase-shifter)

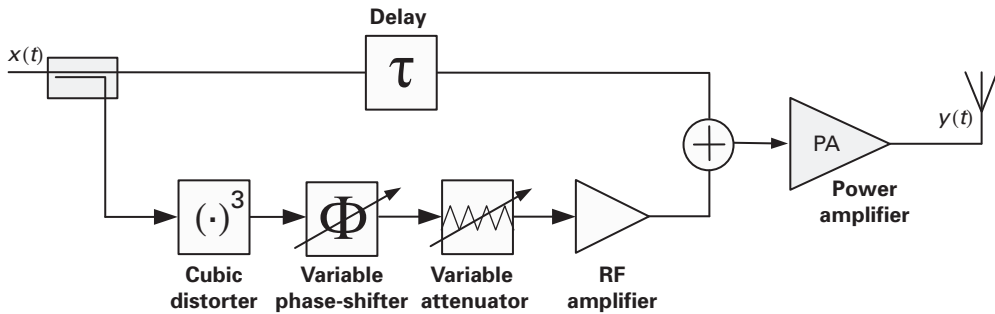


Figure 7.3 Block diagram of a RF cubic predistorter.

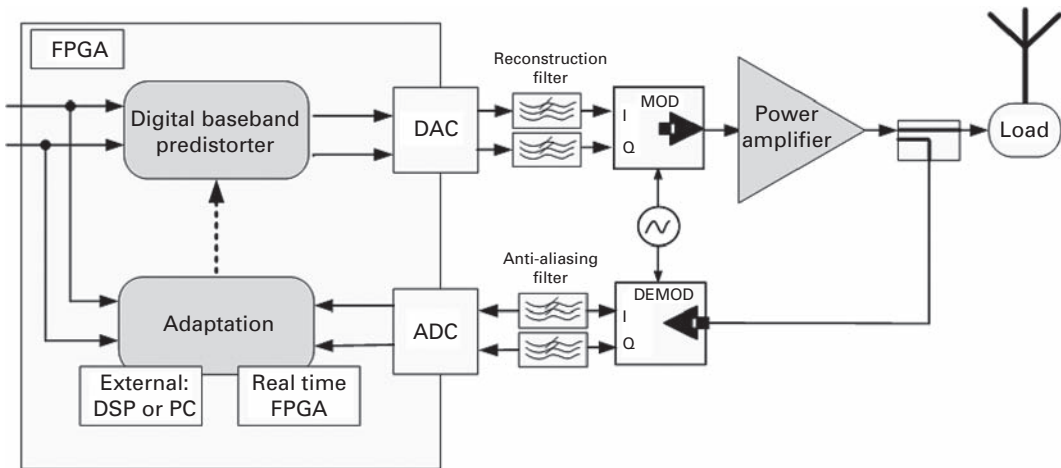


Figure 7.4 Simplified block diagram of a transmitter with baseband adaptive digital predistortion linearization.

controllers that ensure a correct match at the combiner and finally, a post-distortion amplifier that buffers and amplifies the resulting signal. The amplifier in this path is considered to be a small-signal device that does not contribute to the overall nonlinear distortion. Finally the predistorted signal in the lower path is subtracted from the signal in the upper one and fed to the PA.

The nonlinearity in the predistorter is usually created making use of the nonlinear characteristics of diodes: single diode, anti-parallel diode, varactor-diode [7],[8], and also FET transistors [9].

On the other hand, digital predistortion makes use of digital processing devices such as digital signal processors (DSPs) or FPGAs. Figure 7.4 shows the simplified block diagram of a transmitter containing an adaptive digital predistortion module at BB. Two main approaches can be found in digital predistortion: *signal* and *data predistortion*. The main difference between signal and data predistortion regards the position in the transmitter chain where the predistortion is carried out [10]. The digital *data predistortion* technique is custom tailored to specific digital modulation formats. This means

that the predistortion function of data predistorters aims at compensating the data vector space (constellation). Therefore, the predistorter coefficients are optimized by minimizing the error vector magnitude (EVM), i.e. compensating the in-band distortion introduced by the PA. However, digital data predistortion does not directly compensate the out-band distortion. In addition, data predistortion is not transparent to all different modulation formats, so it is suitable for specific applications but unsuitable from a more versatile point of view. On the other hand, digital *signal predistortion* is aimed at cancelling both in-band and out-band distortion, as long as the saturation level of the PA permits it. Moreover, digital signal predistortion is totally independent from the PA type, band of operation, amplifier class, power technology, and the signal modulation.

7.3 Power amplifier behavioral models for digital predistortion

The digital baseband predistortion linearization technique is independent of the final frequency band of operation and avoids the complexity of the RF hardware, always operating at higher frequencies. This makes the DPD linearizer more versatile and it is also a huge field of research for engineers for optimizing digital signal processing architectures and algorithms. However, the cancellation performance of DPD can be seriously degraded due to its sensitivity to memory effects generated in RF power amplifiers. The envelope filtering technique is a memory effects compensation technique aimed at reproducing the inverse memory effects that are generated inside the PA [6].

Therefore, when designing adaptive digital predistorters at baseband, some key aspects have to be taken into account:

- the need for PA behavioral models capable of meeting the accuracy requirements (in terms of nonlinear and memory effects modeling)
- an efficient procedure for the identification and adaptation of the baseband DPD based on the previous PA behavioral model
- an efficient implementation of the DPD function in a digital processing device (DSPs or FPGAs) without introducing an excessive computational cost.

The PA models can be classified according to the type of data needed for their extraction in *physical* and *empirical* models [11],[12]. *Physical models* require the knowledge of the electronic elements that comprise the PA, their constitutive relations and the theoretical rules describing their interactions. These models make use of a nonlinear modeling of the PA active device and other passive components to form a set of nonlinear equations relating terminal voltages and currents. The equivalent circuit description of the PA permits accurate results which are very suitable for *circuit-level* simulation. However, this kind of method often provide excessive details and thus is very time (and computational resources) consuming for system level linearization purposes. On the other hand, *empirical or behavioral models* do not need an *a priori* knowledge of the PA internal composition (very useful for traveling-wave-tube (TWT) PAs for example). For this reason they are also known as *black-box models*. Their extraction just relies on a set of

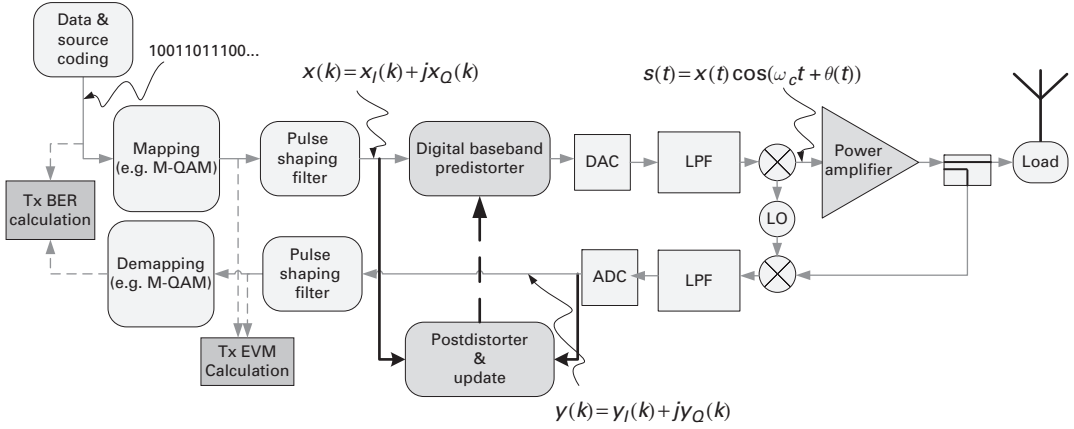


Figure 7.5 Block diagram of a transmitter with digital baseband predistortion.

input-output observations. Therefore, their accuracy is highly sensitive to the adopted model structure and the parameter extraction procedure.

Power amplifier behavioral (or black-box) models at system level are single-input single-output (SISO) systems. The extraction of a PA behavioral model for DPD linearization purposes is carried out by means of input and output *complex envelope* signal observations. Figure 7.5 shows a simplified transmitter block diagram presenting digital baseband predistortion linearization. Notice the different type of signals (bandpass, baseband, analog, digital) present after each of the main transmitter blocks.

A bandpass modulation can be defined as the process whereby the amplitude, frequency, or phase of an RF carrier, or a combination of them, is varied in accordance with the information to be transmitted. An amplitude and phase modulated *bandpass signal* can be described as

$$s(t) = A(t) \cos(\omega_c t + \theta(t)) \quad (7.1)$$

where $\omega_c = 2\pi f_c$ is the angular carrier frequency, $A(t)$ and $\theta(t)$ are the time-varying amplitude and phase, respectively. This bandpass signal can also be expressed in a polar form as

$$s(t) = \text{Re}\{x(t)e^{j\omega_c t}\} \quad (7.2)$$

with $x(t)$ being the complex envelope defined as

$$x(t) = A(t)e^{j\theta(t)} = A(t) \cos(\theta(t)) + jA(t) \sin(\theta(t)) = x_I(t) + jx_Q(t) \quad (7.3)$$

where $x_I(t)$ and $x_Q(t)$ are the in-phase (I) and quadrature (Q) components of the complex envelope, respectively. Using trigonometric identities the bandpass signal can be also expressed by means of their Cartesian I and Q components:

$$s(t) = \text{Re}\{(x_I(t) + jx_Q(t))(\cos(\omega_c t) + j\sin(\omega_c t))\} = x_I(t) \cos(\omega_c t) - x_Q(t) \sin(\omega_c t). \quad (7.4)$$

Therefore, considering the complex envelope $x(t)$ as a baseband signal described by a sequence of ideal pulses $x_I(t)$ and $x_Q(t)$ appearing at discrete times ($k=1,2,3 \dots$) in a digital baseband environment, from (7.3) the complex envelope in the discrete form can be expressed as

$$x(k) = x(t)|_{t=kT_s} = x_I(kT_s) + jx_Q(kT_s) = x_I(k) + jx_Q(k). \quad (7.5)$$

Thus, in order to extract the PA behavioral models that will allow us to design the digital predistorter at baseband, it is necessary to have input ($x(k)$) and output ($y(k)$) discrete complex envelope signals, as schematically shown in Figure 7.5.

An extended description of the existing PA behavioral models as well as several aspects related to the extraction, validation, and comparison of those behavioral models can be found in [12]. However, since DPD nonlinear cancellation performance depends on the ability of these models to accurately reproduce PA nonlinear and memory effects (envelope filtering technique), some of the most significant PA models, in terms of DPD implementation, will be presented and discussed in the following.

7.3.1 Memoryless power amplifier behavioral models

In memoryless models, the output signal is assumed to be a nonlinear function of the instantaneous input signal only (Figure 7.6). The envelope frequency of the input signal is much smaller than the envelope bandwidth of the amplifier. Consequently, the amount of amplitude and phase distortion depends only on the input signal level at the corresponding respective time instant.

Some of the most representative memoryless PA behavioral models that can be found in the literature are: the Saleh or modified Saleh model, the Bessel-Fourier model, the Hetrakul and Taylor model, the Berman and Mahle model, and the Wiener expansion. As an example, the Saleh and the memoryless polynomial models are presented in the following.

Saleh model

Saleh proposed two general functions to approximate the AM-AM (amplitude modulation) and AM-PM (phase modulation) envelope characteristics, initially applied to traveling wave tube (TWT) amplifier models [13], but later also used to characterize

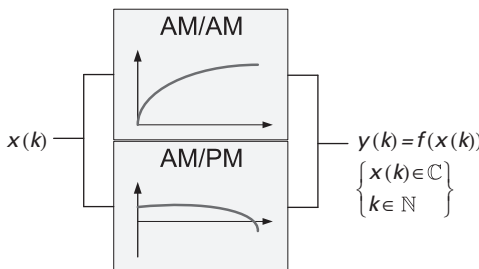


Figure 7.6 Memoryless behavioral model.

solid state power amplifier (SSPA) models. The polar representation corresponds directly to AM-AM and AM-PM characterization and can be expressed as

$$f_A(|x(k)|) = \frac{\alpha_A |x(k)|}{1 + \beta_A |x(k)|^2} \quad (7.6)$$

$$f_\phi(|x(k)|) = \frac{\alpha_\phi |x(k)|^2}{1 + \beta_\phi |x(k)|^2} \quad (7.7)$$

where α_A , β_A , α_ϕ , β_ϕ are constant parameters chosen to approximate the real PA characteristics.

Memoryless Polynomial Model

Together with Saleh's models the memoryless polynomial model is one of the best known and most used models to describe the PA nonlinear static behavior. A general form of p^{th} order power series approximation can be expressed by:

$$y(k) = \sum_{p=0}^P \gamma_p \cdot x(k) \cdot |x(k)|^p \quad (7.8)$$

where $x(k)$ and $y(k)$ are the PA input and output complex envelope at sample k , P is the order of the polynomial function and γ_p are complex coefficients that have to be identified in order to characterize the PA memoryless nonlinear behavior.

7.3.2 Power amplifier behavioral models with memory effects

The precise gain in RF power amplifiers presenting memory effects is not only a function of the input signal amplitude at the same instant, but also dependent on the recent history of the input–output signals as well (see [Figure 7.7](#)).

Coping with high-speed envelope signals (presenting significant bandwidths) makes engineers reconsider the degradation resulting from memory effects. Time responses are convolved by the impulse response of the system and thus memory effects are no longer irrelevant when predistortion type linearization is employed to cancel out the intermodulation sidebands.

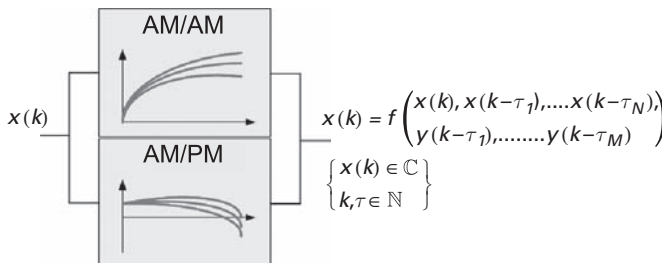


Figure 7.7 Behavioral models with memory.

To develop amplifier models that include memory effects, the amplifier has to be characterized using dynamic measurement systems. The most common *dynamic non-linear models* considered in the literature for characterizing the PA behavior are neural networks (NNs), Volterra series based models, and the so-called *modular* or *two/three-box* models. In the following, some of these most common dynamic structures used in PA behavioral modeling are presented and discussed.

Artificial neural networks: time delay neural networks

An artificial neural network (ANN) is a system composed of a large number of basic elements arranged in layers that are highly interconnected [14]. In general, the structure is formed by several inputs and outputs, which may be trained to react to the inputs stimulus in a particular desired way. These systems pretend to emulate the human brain by learning how to behave on the basis of the former knowledge of the environment's problem.

Multi-layer perceptrons (MLP) is one of the most popular structures of ANN models. They follow a general class of feedforward ANN structure (because the flow of data is from inputs to outputs) and have the capability of producing the general approximation of any function [15]. The MLP model consists of a finite number of units called *neurons* or *perceptrons*, where each unit of each layer is connected to each unit of the subsequent/previous layer. These connections are called links or synapses and they only exist between neurons in subsequent layers. Connections among neurons within a layer do not exist, nor are there any direct connections across layers. The first layer is called the input layer, followed by intermediate groups of neurons called hidden layers, and the result of the network is obtained in the output layer. The number of neurons in the hidden layer is usually determined by trial-and-error, to find the trade-off that provides the simplest network with acceptable performance.

In the last few years NNs have been also used to model the dynamic phenomena, i.e. for power amplifier (PA) modeling [16],[17], proposing the use of a special type of MLP model, the time-delayed neural networks (TDNN).

A TDNN is based on the feedforward MLP NN with the addition of tapped delay lines which generate delayed samples of the input variables. They are used to add the history of the input signals to the model, required to model memory effects. The TDNN entries include not only the current value of the input signal, but also its previous values, as shown in the TDNN neuron model in Figure 7.8 and described in (7.9).

$$y_j(k) = \psi \left(\sum_{i=0}^N w_{ij} x(k - \tau_i) \pm \theta_j \right). \quad (7.9)$$

The TDNN structure consists of two sections, namely a linear time invariant (LTI) system and a nonlinear memoryless system (MLP), as shown in Figure 7.9. The LTI part provides the TDNN with the capability of performing dynamic mappings which depend on past input values, making them suitable for time series prediction, nonlinear system identification, and signal processing applications. In this case, the most popular type are

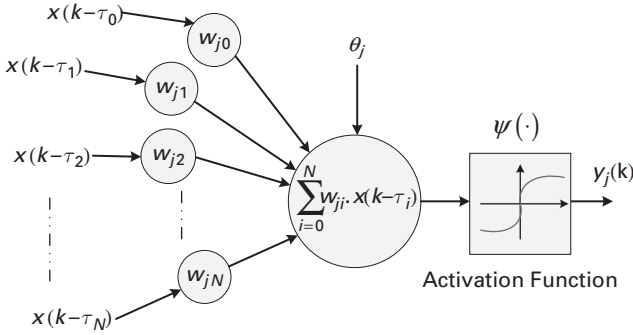


Figure 7.8 TDNN neuron model.

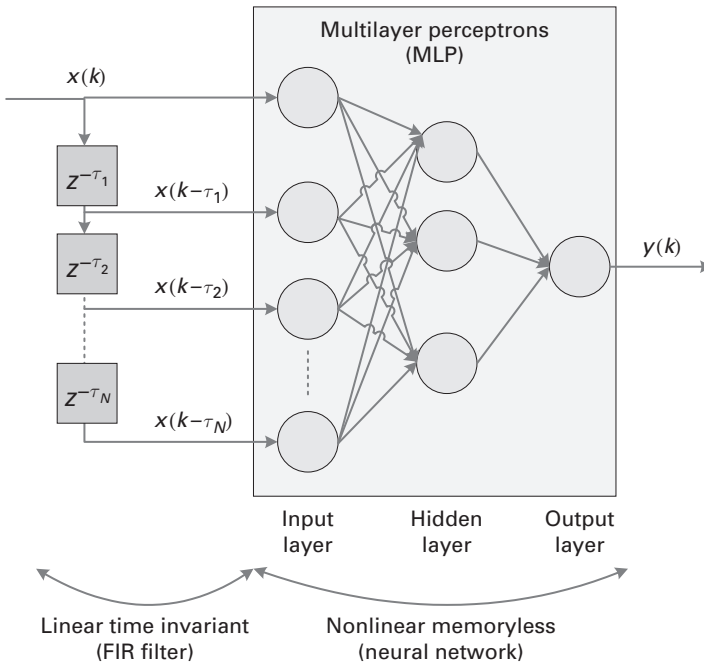


Figure 7.9 Time-delayed neural network (TDNN).

finite impulse response (FIR) neural networks which are obtained by replacing input with FIR filters. The memory depth of the element or system analyzed is reflected in the length of the taps.

Comparing the NN approach with the polynomial one, contrary to the inherent local approximating properties of polynomials, NNs can behave as global approximations: a potential advantage when modeling strongly nonlinear systems. Besides, in principle, NNs work better than the classical models in extrapolating beyond the zone exploited for parameter extraction.

Volterra series based models

Vito Volterra in [18] was the first to study the *functional series*, considered after him as an extension of the Taylor series expansion [19]. Volterra kernels are reliable descriptors of the system's function, covering a large number of possible system states. For that reason they can be considered to model nonlinear systems with memory such as PAs. Therefore, considering $x(k)$ and $y(k)$ as the complex envelope input–output relationship in the PA, the discrete-time complex Volterra series expansion can be expressed as

$$\begin{aligned}
 y(k) = & \sum_{i=0}^{M_1-1} h_1(i) \cdot x(k-i) + \\
 & + \sum_{m=0}^{M_3-1} \sum_{n=0}^{M_3-1} \sum_{p=0}^{M_3-1} h_3(m, n, p) \cdot x(k-m) \cdot x(k-n) \cdot x^*(k-p) + \\
 & \vdots \\
 & + \sum_{m=0}^{M_Q-1} \sum_{n=0}^{M_Q-1} \cdots \sum_{q=0}^{M_Q-1} h_Q(m, n, \dots, q) \cdot x(k-m) \cdot x(k-n) \cdots x^*(k-q)
 \end{aligned} \tag{7.10}$$

where $h_1, h_3, \dots, h_Q$ are the so-called low-pass equivalent Volterra kernels, k denotes discrete time, and $M_1, M_3, \dots, M_Q$ are the number of delays considered in each kernel respectively.

The number of parameters in Volterra series grow exponentially when considering higher-order kernel extractions, which can add unnecessary computational complexity when considering them for identification and linearization purposes. The number of coefficients (n_c) that we have to extract when considering the discrete-time Volterra series is described by the following equation:

$$n_c = \sum_{i=1}^{\frac{N_k+1}{2}} D^{2i-1} \quad (M_1 = M_3 = \dots = M_Q = D) \tag{7.11}$$

with N_k being the highest-order kernel considered (for the sake of simplicity only odd kernels have been considered) and assuming the same memory depth for each kernel ($M_1 = M_3 = \dots = M_Q = D$).

On the one hand, a common solution to avoid the computational complexity problem of Volterra series in modeling dynamic systems (such as the PA) is the use of pruning techniques [20]; or alternative configurations to reproduce truncated and simplified versions of Volterra kernels as for example: the Laguerre-Volterra model [21], the Kautz-Volterra model [22], or the Wiener-Bose model [23]. Further details on Volterra series and their alternative configurations can be found in [12] and [19].

On the other hand, more simplified deviations of the general Volterra series are the so-called *modular approaches*, consisting of structures of linear time-invariant filters and memoryless nonlinearities. Some of the most important published models based on a modular approach will be discussed in the following.

Memory polynomial: nonlinear moving average behavioral models

Memory polynomials or nonlinear moving average (NMA) models are some of the most simple models that take into account the nonlinear dynamic behavior of a PA and have been used for predistortion applications [24], [25]. The block scheme of an NMA model is depicted in Figure 7.10.

The input–output relationship of an NMA model can be expressed as

$$y_{NMA}(k) = \sum_{n=0}^N f_n(x(k - \tau_n)) \quad (7.12)$$

where the nonlinear function $f_n(\cdot)$ is expressed by a polynomial function in (7.13) and where τ 's are the most suitable non-consecutive delays for describing the PA model ($\tau_0 = 0$):

$$f_n(x(k - \tau_n)) = \sum_{p=0}^P \alpha_{pn} x(k - \tau_n) |x(k - \tau_n)|^p. \quad (7.13)$$

Rewriting (7.13) in a more compact matrix notation,

$$y_{NMA}(k) = \lambda^H \Phi \quad (7.14)$$

where the superindex H denotes Hermitian (conjugate transpose) and where $\lambda = [\alpha_{00}, \alpha_{10}, \dots, \alpha_{P0}, \alpha_{01}, \dots, \alpha_{P1}, \dots, \alpha_{0N}, \dots, \alpha_{PN}]^H$ and

$$\Phi = \begin{bmatrix} x(k), x(k)|x(k)|, \dots, x(k)|x(k)|^P, x(k - \tau_1), \dots, \\ x(k - \tau_1)|x(k - \tau_1)|^P, \dots, x(k - \tau_N), \dots, x(k - \tau_N)|x(k - \tau_N)|^P \end{bmatrix}^T.$$

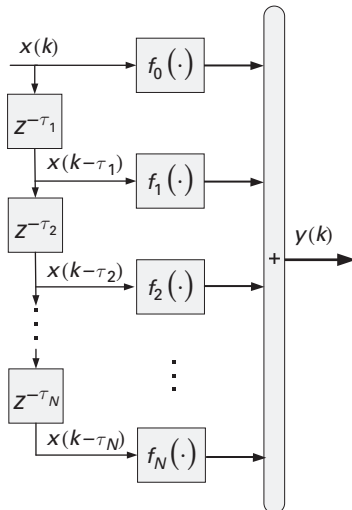


Figure 7.10 Block diagram of a Nonlinear Moving Average (memory polynomial).

The cost function to be minimized in order to extract the PA model parameters is the *Normalized Square Error* (NSE), defined in (7.15)

$$J(k) = |\bar{e}(k)|^2 = |(y(k) - \lambda^H \Phi) / y(k)|^2. \quad (7.15)$$

The error is defined as the difference between the PA measured output and the model estimated output, divided by the measured output.

Augmented nonlinear moving average

The augmented NMA PA behavioral model presented in [26] is an extension of the NMA model that introduces the main p-order nonlinear products of pairs of delayed samples of the input ($x(k - \tau_i), x(k - \tau_j)$) in order to improve nonlinear memory modeling.

The augmented NMA model structure is depicted in Figure 7.11 and its input–output relationship can be expressed as

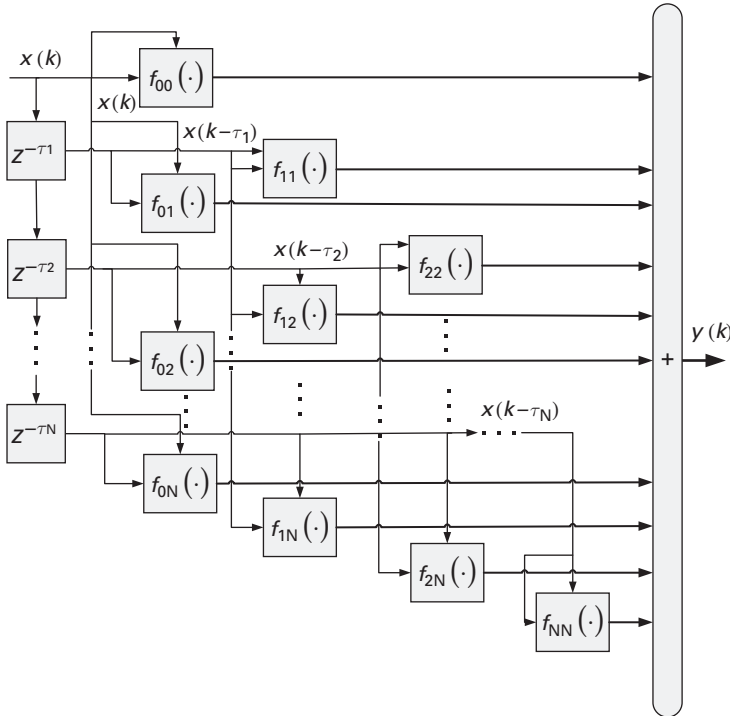


Figure 7.11 Block diagram of an Augmented Nonlinear Moving Average model.

$$y_{AugNMA}(k) = \sum_{i=0}^N \sum_{j=i}^N f_{ij}(x(k - \tau_i), x(k - \tau_j)) \quad (7.16)$$

where τ 's are the most suitable delays for describing the PA model and $f_{ij}(\cdot)$ are static nonlinear functions implemented using polynomials

$$f_{ij}(x(k - \tau_i), x(k - \tau_j)) = \sum_{p=0}^P \alpha_{ijp} x(k - \tau_i) |x(k - \tau_j)|^p. \quad (7.17)$$

Expressing (7.16) in a more compact matrix notation:

$$y_{AugNMA}(k) = \lambda^H \Phi \quad (7.18)$$

where $\lambda = [\alpha_{00}, \dots, \alpha_{0N}, \alpha_{11}, \dots, \alpha_{1N}, \dots, \alpha_{(N-1)N}, \alpha_{NN}]^H$,

$$\Phi = \begin{bmatrix} f_{00}(x(k), x(k)), \dots, f_{0N}(x(k), x(k - \tau_N)), \\ f_{11}(x(k - \tau_1), x(k - \tau_1)), \dots, f_{1N}(x(k - \tau_1), x(k - \tau_N)), \dots, \\ f_{(N-1)N}(x(k - \tau_{N-1}), x(k - \tau_N)), f_{NN}(x(k - \tau_N), x(k - \tau_N)) \end{bmatrix}^T$$

and assuming $\tau_0 = 0$.

The cost function to be minimized in order to extract the coefficients defining the augmented NMA model is the same as the one defined in (7.15).

Hammerstein model

Hammerstein models are composed by a memoryless nonlinearity followed by a linear time-invariant (LTI) system, as shown in Figure 7.12.

For the sake of simplicity, the LTI block is implemented with a FIR filter,

$$y_{Ham}(k) = \sum_{n=0}^N \alpha_n f_0(x(k - \tau_n)). \quad (7.19)$$

Assuming that the memoryless nonlinearity is implemented with polynomials:

$$y_{Ham}(k) = \sum_{n=0}^N \alpha_n \sum_{p=0}^P \gamma_p x(k - \tau_n) |x(k - \tau_n)|^p. \quad (7.20)$$

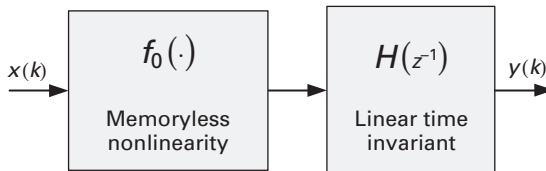


Figure 7.12 Block diagram of a Hammerstein model.

In order to rewrite (7.20) in a matrix notation that allows the extraction of the Hammerstein model coefficients we need to assume that $d_{np} = \alpha_n \gamma_p$, thus the Hammerstein model output can be expressed as

$$y_{Ham}(k) = \lambda^H \Phi \quad (7.21)$$

where $\lambda = [d_{00}, \dots, d_{0P}, d_{10}, \dots, d_{1P}, \dots, d_{N0}, \dots, d_{NP}]^H$, with $\tau_0 = 0$ and

$$\Phi = \begin{bmatrix} x(k), \dots, x(k)|x(k)|^P, x(k - \tau_1), \dots, x(k - \tau_1)|x(k - \tau_1)|^P, \dots \\ x(k - \tau_N), \dots, x(k - \tau_N)|x(k - \tau_N)|^P \end{bmatrix}^T.$$

Analogously, its cost function is defined as in (7.15).

Once we have extracted all d_{np} parameters, the two-stage identification algorithm proposed in [27] can be used to identify α_n and γ_p separately with a singular value decomposition (SVD).

The Hammerstein structure has been considered to identify PA behavioral models that later have been used to predistort the PA device. Moreover, infinite impulse response (IIR) filters have also been considered for implementing the LTI block of the Hammerstein scheme, showing better performance than using only FIR filters but having the risk of introducing instabilities, as presented in [28],[29].

More elaborated structures can be constructed by adding single Hammerstein structures in parallel in order to achieve a more refined PA behavioral model identification. This is the case of the *parallel-Hammerstein* model depicted in Figure 7.13.

The mathematical expression for the input–output relationship in the parallel-Hammerstein model, assuming again FIR filters for the LTI blocks, can be expressed as

$$y_{P-H}(k) = \sum_{r=1}^R \sum_{n=0}^{N_r} a_{nr} f_r(x(k - \tau_n)) \quad (7.22)$$

where R is the number of branches considered in the parallel-Hammerstein model, and with N_r ($r = 1, \dots, R$) being the memory depth of each of the FIR blocks considered in each branch respectively. Results showing the use of the parallel-Hammerstein model to identify PA behavioral models can be found in [30].

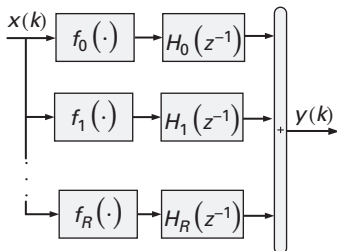


Figure 7.13 Block diagram of a parallel-Hammerstein model.

Wiener model

The Wiener model, as the Hammerstein model, can be seen as reduction and simplification of the general Volterra series expansion. The Wiener model is composed by a LTI system followed by a memoryless nonlinearity, as shown in Figure 7.14.

By considering again a FIR filter for the LTI block, the input–output relationship in a Wiener model can be defined as

$$y_{Wien}(k) = f_0 \left(\sum_{n=0}^N \alpha_n x(k - \tau_n) \right) \quad (7.23)$$

where $\tau_0 = 0$. Assuming that the memoryless nonlinearity is implemented with polynomials,

$$y_{Wien}(k) = \sum_{p=0}^P \gamma_p \left(\sum_{n=0}^N \alpha_n x(k - \tau_n) \right) \left| \sum_{n=0}^N \alpha_n x(k - \tau_n) \right|^p. \quad (7.24)$$

As deduced from (7.24), the coefficients of the filter cannot be easily extracted because the whole FIR filter expression is raised to the p th power within the power series. In order to solve this problem, we need to first estimate an intermediate variable $v(k)$ as proposed in [31]. Then, the estimation problem is divided into two steps.

To estimate the intermediate variable, the following assumptions are considered:

- the linear subsystem (the FIR filter) is stable,
- the nonlinear function (power series) is invertible, and
- there is no noise in the system.

With these assumptions, we can calculate the intermediate variable $v(k)$ in (7.25). Therefore to proceed with the two-step identification it is first necessary to rewrite (7.23) by means of an intermediate variable $v(k)$.

$$\begin{aligned} v(k) &= \sum_{n=0}^N \alpha_n x(k - \tau_n) \\ y_{Wien}(k) &= f_0(v(k)). \end{aligned} \quad (7.25)$$

Then, we can proceed with the Wiener model estimation in two steps as described in [31]. Some examples using the Wiener model to identify PA behavioral models are presented in [32],[33].

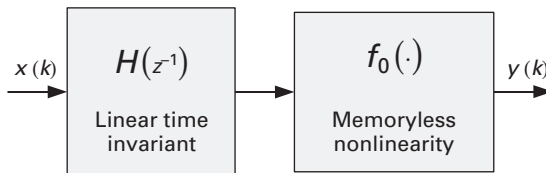


Figure 7.14 Block diagram of a Wiener model.

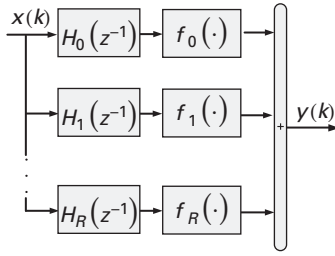


Figure 7.15 Block diagram of a parallel-Wiener model.

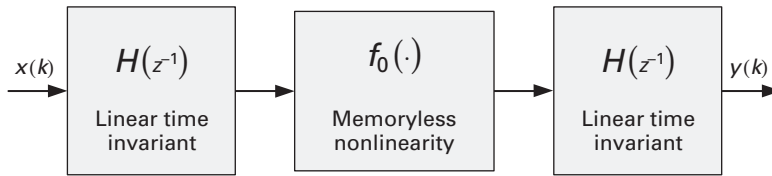


Figure 7.16 Block diagram of a Wiener-Hammerstein cascade model.

Similarly to parallel-Hammerstein models, *parallel-Wiener* models can be implemented by adding single Wiener structures in parallel in order to achieve a more refined PA behavioral model identification. A general parallel-Wiener structure is depicted in Figure 7.15.

The mathematical expression for the input–output relationship of a parallel-Wiener model, assuming again FIR filters for the LTI blocks, can be expressed as

$$y_{P-W}(k) = \sum_{r=1}^R f_r \left(\sum_{n=0}^{N_r} a_{nr} x(k - \tau_n) \right) \quad (7.26)$$

where R is the number of selected branches in the parallel-Wiener model with N_r ($r = 1, \dots, R$) being the maximum number of delays (memory depth) of each of the FIR blocks considered in each branch respectively. Results showing the use of the parallel-Wiener model to identify PA behavioral models can be found in [34].

Wiener-Hammerstein cascade: three-box modeling

The two-box model methodology (Hammerstein or Wiener) can be extended to the three-box model topology. If we consider a Wiener model, an additional filter is included at the output of the memoryless nonlinearity to yield a filter–nonlinearity–filter cascade configuration. The Wiener-Hammerstein cascade model is a three-box model consisting of the cascade interconnection among a LTI block followed by a memoryless nonlinearity block and finally followed by another LTI block, as schematically shown in Figure 7.16.

If we consider both LTI blocks implemented as FIR filters, the input–output relationship in a Wiener-Hammerstein cascade model can be defined as

$$y_{Wien-Hamm}(k) = \sum_{j=0}^M \beta_j f_0 \left(\sum_{i=0}^N \alpha_i x(k - \tau_i - \tau_j) \right). \quad (7.27)$$

In order to use this model for PA behavioral modeling purposes, it is necessary to use the two-step identification technique described for the Wiener model, identifying first intermediate variables that permit the final extraction of the coefficients of this three-box model.

Such a configuration is commonly used for satellite communication channels, where the PA at the satellite transponder is driven near saturation to exploit the maximum power efficiency for the downlink. The Wiener–Hammerstien configuration has been used for modeling and predistortion linearization purposes in [24] and [35], respectively.

Nonlinear auto-regressive moving average models

The advantage of using a nonlinear auto-regressive moving average (NARMA) model is the introduction of a nonlinear feedback path (IIR terms) that may permit relaxing the number of delayed samples considered to model the PA. Reducing the complexity of the PA model acquires importance when, for example, the PA model is considered for linearization purposes, such as in DPD linearization. However, one of the main weaknesses of the NARMA model is its stability, since the use of nonlinear feedback paths can result in overall system instability. Therefore, in order to guarantee the stability of a NARMA PA model, a stability test based on the small-gain theorem has to be first conducted, as explained in [36].

The NARMA model can be seen as an extension of a simpler NMA model that incorporates a nonlinear auto regressive block. The block diagram describing the NARMA model is depicted in Figure 7.17.

The general input–output mathematical expression in a NARMA model can be described as

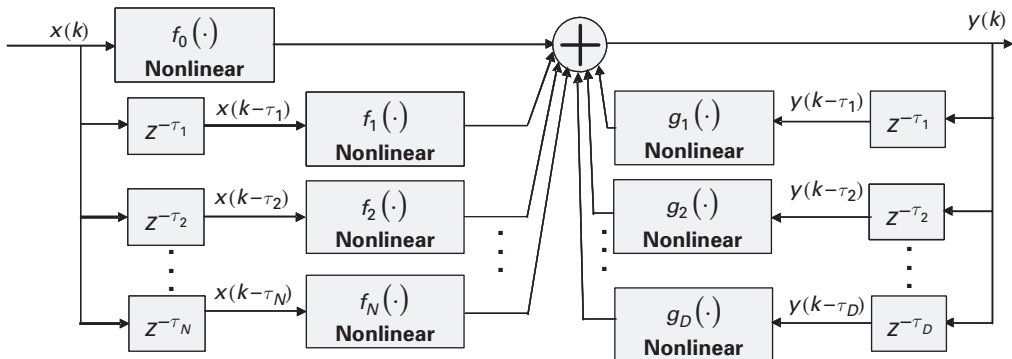


Figure 7.17 Block diagram of the Nonlinear Auto Regressive Moving Average (NARMA) model.

$$y_{NARMA}(k) = y(k) = \sum_{i=0}^N f_i(x(k - \tau_i)) - \sum_{j=1}^D g_j(y(k - \tau_j)) \quad (7.28)$$

with $f_i(\cdot)$ and $g_j(\cdot)$ being nonlinear memoryless functions, and where τ_i and τ_j ($\tau \in \mathbb{N}$; and with $\tau_0 = 0$) are the most significant sparse delays of the input and the output respectively, contributing to the description of the PA memory effects.

The present output sample ($y(k)$) depends on the sum of different static nonlinearities related to the present sample of the input ($x(k)$), and both input ($x(k - \tau_i)$) and output ($y(k - \tau_j)$) past samples ($i = 1, 2 \dots N; j = 1, 2 \dots D$), as shown in Figure 7.17. By considering polynomials to implement the $f_i(\cdot)$ and $g_j(\cdot)$ nonlinear functions, we can rewrite (7.28) as

$$y_{NARMA}(k) = y(k) = \sum_{i=0}^N \alpha_i \left(\sum_{p=0}^P \gamma_p x(k - \tau_i) |x(k - \tau_i)|^p \right) - \sum_{j=1}^D \beta_j \left(\sum_{p=0}^P \delta_p y(k - \tau_j) |y(k - \tau_j)|^p \right), \quad (7.29)$$

where P is the order of the memoryless polynomial and N and D are the number of the delayed input and output samples (memory depth) considered for modeling the PA dynamics.

We can rewrite (7.29) in a more compact matrix notation assuming that $a_{ip} = \alpha_i \cdot \gamma_p$ and $b_{jp} = \beta_j \cdot \delta_p$. Then (7.29) results in

$$y_{NARMA}(k) = y(k) = \lambda^H \Phi \quad (7.30)$$

$$\text{where } \lambda = \begin{bmatrix} a_{00}, a_{01}, \dots, a_{0P}, a_{10}, \dots, a_{1P}, \dots, a_{N0}, \dots, a_{NP}, \\ -b_{10}, \dots, -b_{1P}, \dots, -b_{D0}, \dots, -b_{DP} \end{bmatrix}^H,$$

$$\Phi = \begin{bmatrix} x(k), x(k)|x(k)|, \dots, x(k)|x(k)|^P, x(k - \tau_1), \dots, \\ x(k - \tau_1)|x(k - \tau_1)|^P, \dots, x(k - \tau_N), \dots, x(k - \tau_N)|x(k - \tau_N)|^P, \\ y(k - \tau_1), \dots, y(k - \tau_1)|y(k - \tau_1)|^P, \dots, y(k - \tau_D), \dots, \\ y(k - \tau_D)|y(k - \tau_D)|^P \end{bmatrix}^T$$

and its cost function is defined as in (7.15).

7.4 Discussion

The PA is a key device within the transmitter because it is responsible for a significant part of the overall power efficiency and linearity performance. The PA inherent nonlinear behavior has been confronted from two different approaches: (a) an accurate circuit-level design aiming at obtaining highly linear PAs (e.g. Doherty Amplifiers) and when

necessary making use of linearization solutions at device level; and (b) system-level linearization techniques, a broad approach that considers the PA as a nonlinear black-box to be linearized.

This chapter has shown the principles of predistortion linearization emphasizing digital predistortion solutions carried out at baseband. To counteract the nonlinear behavior of PAs means that it is crucially important to find PA behavioral models capable of reproducing not only its nonlinear static behavior, but also its dynamics.

In this chapter, an overview of some of the most used PA behavioral models have been presented. In order to be used for DPD purposes, PA behavioral models have to deal with the trade-off between computational complexity and accuracy in the reproduction of the PA nonlinear behavior and memory effects. Some of these models can reproduce the PA nonlinear behavior but not take into account possible PA dynamics; these are the so-called memoryless behavioral models. On the other hand, there are behavioral models capable of reproducing PA memory effects as well as its nonlinear behavior and thus suitable for applying the envelope filtering technique in DPD linearization. Among these behavioral models, NNs and Volterra series are very accurate descriptors of nonlinear dynamics since they cover a large sum of possible states and their nonlinear combinations. However, since the optimum could be better than needed, these models add computational complexity to DPD algorithms which can result in an inefficient power consumption of the DSP device. For that reason several simplified models derived from Volterra series (two and three box modeling) have been proposed as possible candidates for implementing the DPD function in a DSP. The NARMA model presents a good trade-off between accuracy and complexity, since its modular nature allows scalability. Therefore, in case of necessity it is possible to incorporate several branches in parallel to characterize nonlinear memory effects. Moreover, its recursive architecture can relax the number of required delays to characterize nonlinear memory effects.

References

- [1] P.L. Gilabert, M.E. Gadringer, G. Montoro, *et al.* An Efficient Combination of Digital Predistortion and OFDM Clipping for Power Amplifiers. *International Journal of RF and Microwave Computer-Aided Engineering*, **19**:583–591, September 2009.
- [2] P.M. Cabral, J.C. Pedro, J.A. Garcia, and L. Cabria. A Linearized Polar Transmitter for Wireless Applications. In *Proc. IEEE MTT-S International Microwave Symposium Digest*, pages 935–938, 2008.
- [3] D. Y. C. Lie, J. Lopez, and Yan Li. Low-power RF Wideband Polar Transmitter Design using the Envelope-tracking Technique. In *Proc. 9th International Conference Solid-State and Integrated-Circuit Technology ICSICT 2008*, pages 1536–1543, 2008.
- [4] D.F. Kimball, Jinho Jeong, Chin Hsia, *et al.* High-efficiency Envelope-tracking W-CDMA Base-station Amplifier using Gan Hfets. *IEEE Transactions on Microwave Theory and Techniques*, **54**(11):3848–3856, 2006.
- [5] H. Gandhi. Digital Predistortion Linearizes Broadband PAs. *Microwaves and RF*, pages 1–4, 2008.
- [6] J. Vuolevi. *Distortion in RF Power Amplifiers*. Artech House, 2003.

- [7] K. Yamauchi, K. Mori, M. Nakayama, *et al.* A Novel Series Diode Linearizer for Mobile Radio Power Amplifiers. In *Proc. IEEE MTT-S International Microwave Symposium Digest*, volume 2, pages 831–834, June 1996.
- [8] C. S. Yu, W. S. Chan, and W.-L. Chan. 1.9 GHz Low Loss Varactor Diode Pre-Distorter. *Electronics Letters*, **35**:1681–1682, September 1999.
- [9] T. S. Nielsen and S. Lindfors. A 2.4 GHz MOSFET Predistorter for Dual-Mode Blue-tooth/IEEE 802.11b Transmitter. In *Proc. of Norchip Conference*, pages 15–20, November 2002.
- [10] S. Andreoli, H. G. McClure, P. Banelli, and S. Cacapardi. Digital Linearizer for RF Amplifiers. *IEEE Transactions on Broadcasting*, **43**(1):12–19, March 1997.
- [11] J. C. Pedro and S. A. Maas. A Comparative Overview of Microwave and Wireless Power Amplifier Behavioral Modeling Approaches. *IEEE Transactions on Microwave Theory and Techniques*, **53**(4):1150–1163, April 2005.
- [12] D. Scheurs, M. O'Droma, A. A. Goacher, and M. Gadringer, editors. *RF Power Amplifier Behavioural Modeling*. Cambridge University Press, 2009.
- [13] A. Saleh. Frequency-Independent and Frequency-Dependent Nonlinear Models of TWT Amplifiers. *IEEE Transactions on Communications*, **29**(11):1715–1720, November 1981.
- [14] Q. J. Zhang and K. C. Gupta. *Neural Networks for RF and Microwaves Design*. Artech House, 2000.
- [15] F. Scarselli and A. C. Tsoi. Universal Approximation Using Feedforward Neural Networks: A Survey of Some Exciting Methods, and Some New Results. *Neural Networks*, **11**:15–37, 1998.
- [16] J. O'Toole and T. J. Brazil. Novel Neural Network-Bessel Transform for Behavioural Modelling of a Power Amplifier. In *High Frequency Postgraduate Student Colloquium, 2000*, pages 45–50, September 2000.
- [17] J. Xu, M. C. E. Yagoub, R. Ding, and Q.-J. Zhang. Neural-Based Dynamic Modeling of Nonlinear Microwave Circuits. *IEEE Transactions on Microwave Theory and Techniques*, **50**(12):2769–2780, December 2002.
- [18] V. Volterra. *Theory of Functionals and of Integral and Integro-Differential Equations*. Dover Phoenix Editions, 1959.
- [19] V. J. Mathews and G. L. Sicuranza. *Polynomial Signal Processing*. John Willey & Sons, 2000.
- [20] A. Zhu, J. C. Pedro, and T. R. Cunha. Pruning the Volterra Series for Behavioral Modeling of Power Amplifiers Using Physical Knowledge. *IEEE Transactions on Microwave Theory and Techniques*, **55**(5):813–821, May 2007.
- [21] Anding Zhu and T. J. Brazil. Rf power amplifier behavioral modeling using volterra expansion with laguerre functions. In *Proc. IEEE MTT-S International Microwave Symposium Digest*, 2005.
- [22] M. Isaksson and D. Ronnow. A Kautz-Volterra Behavioral Model for RF Power Amplifiers. In *Proc. IEEE MTT-S International Microwave Symposium Digest*, pages 485–488, 2006.
- [23] D. D. Silveira and G. Magerl. Extraction and Improvements of a Behavioral Model Based on the Wiener-Bose Structure Used for Baseband Volterra Kernels Estimation. In *Proc. IEEE MTT-S International Microwave Symposium Digest*, pages 2007–2010, 3–8 June 2007.
- [24] J. Kim and K. Konstantinou. Digital Predistortion of Wideband Signals Based on Power Amplifier Model with Memory. *Electronics Letters*, **37**:1417–1418, November 2001.

- [25] W. Bosch and G. Gatti. Measurement and Simulation of Memory Effects in Predistortion Linearizers. *IEEE Transactions on Microwave Theory and Techniques*, **37**(12):1885–1890, December 1989.
- [26] P.L. Gilabert, D.D. Silveira, G. Montoro, M.E. Gadringer, and E. Bertran. Heuristic Algorithms for Power Amplifier Behavioral Modeling. *IEEE Microwave and Wireless Components Letters*, **17**(10):715–717, Oct. 2007.
- [27] Er-Wei Bai. An Optimal Two Stage Identification Algorithm for Hammerstein-Wiener Nonlinear Systems. In *Proc. American Control Conference*, volume 5, pages 2756–2760, June 1998.
- [28] P.L. Gilabert, D.D. Silveira, G. Montoro, and G. Magerl. RF-Power Amplifier Modeling and Predistortion Based on a Modular Approach. In *Proc. IEEE European Microwave Integrated Circuits Conference (EuMIC'06)*, pages 265–268, September 2006.
- [29] T. Liu, S. Boumaiza, and F.M. Ghannouchi. Augmented Hammerstein Predistorter for Linearization of Broad-Band Wireless Transmitters. *IEEE Transactions on Microwave Theory and Techniques*, **54**(4):1340–1349, June 2006.
- [30] D.D. Silveira, H. Arthaber, P.L. Gilabert, G. Magerl, and E. Bertran. Application of Optimal Delays Selection on Parallel Cascade Hammerstein Models for the Prediction of RF-Power Amplifier Behavior. In *Proc. IEEE Asia-Pacific Microwave Conference (APMC'06)*, volume 1, pages 283–286, December 2006.
- [31] A. Hagenblad. Aspects of the Identification of Wiener Models. Ph.D. thesis, Linköpings Universitet, Sweden, 1999.
- [32] P.L. Gilabert, E. Bertran, and G. Montoro. Multiple Lookup Table Predistortion for Adaptive Modulation. *Proc. European Microwave Association*, pages 304–311, December 2005.
- [33] P.L. Gilabert, G. Montoro, and E. Bertran. On the Wiener and Hammerstein Models for Power Amplifier Predistortion. In *Proc. IEEE Asia-Pacific Microwave Conference (APMC'05)*, volume 2, December. 2005.
- [34] D. Silveira, M. Gadringer, H. Arthaber, and G. Magerl. RF-Power Amplifier Characteristics Determination Using Parallel Cascade Wiener Models and Pseudo-Inverse Techniques. In *Proc. IEEE Asia-Pacific Microwave Conference Proceedings (APMC'05)*, volume 1, December. 2005.
- [35] S. Chang and E.J. Powers. A Simplified Predistorter for Compensation of Nonlinear Distortion in OFDM Systems. In *Proc. IEEE Global Telecommunications Conference (GLOBECOM '01)*, volume 5, pages 3080–3084, November. 2001.
- [36] G. Montoro, P.L. Gilabert, E. Bertran, A. Cesari, and D.D. Silveira. A New Digital Predictive Predistorter for Behavioral Power Amplifier Linearization. *IEEE Microwave and Wireless Components Letters*, **17**(6):448–450, June 2007.

8

Look-up table based digital predistortion schemes and implementation

Pere L. Gilabert and Gabriel Montoro

8.1 Introduction

Many years of research have dealt with predistortion techniques for memoryless PA. Recently, several solutions have included memory effects compensation, since those effects seem to be of significant concern when considering high bandwidths with multi-level and multi-carrier modulation formats. Digital predistortion solutions, usually based on a particular PA behavioral model, have to be designed to be later implemented in a digital signal processor.

An efficient way to implement the predistortion function without introducing an excessive computational cost is by using look-up tables (LUTs). An LUT is a data structure used to replace a runtime computation with a simpler array indexing operation. Therefore, computational complexity and processing time is reduced by using LUTs, since retrieving a value from memory is in general faster than running the algorithm required to generate this value. In addition, LUT-based DPD has shown better performance than using other low-order parametric models such as polynomials. An LUT-based DPD consists, among other blocks, of a memory block that contains a representation of the inverse characteristic of the amplifier and an address calculator to index the memory block. Then, as we will discuss in the following, according to the type of LUT architecture considered it will incorporate several real or complex adders and multipliers to perform the predistortion of the input complex data signal.

In order to map the predistortion function into an LUT, some considerations regarding LUT organization have to be taken into account, such as

- The LUT architecture (one-dimensional [1-D] or two-dimensional [2-D] LUTs) to deal with the discrete complex signal's envelope.
- The optimum size of the LUT (trade-off between accuracy and memory size).
- The LUT indexing and spacing between entries within the LUT.

Another issue that has to be taken into account, despite not being exclusively an issue related to LUT organization, is the complexity of the adaptation algorithm since it is closely related to the frequency of the LUT updates. In the following subsections, some of these LUT organization issues will be discussed.

Among the different existing behavioral models, the nonlinear auto-regressive moving average (NARMA) model is capable of reproducing both PA nonlinear distortion and dynamics. In addition, the NARMA model introduces a nonlinear feedback path in its architecture (nonlinear infinite impulse response (IIR) terms) that may permit relaxing the complexity of the PA model, for example, in comparison with a behavioral model that uses only finite impulse response (FIR) terms to model the PA dynamics. Finally, a DPD based on a NARMA model is well suited for being easily implemented using only a set of look-up tables (LUTs).

In the following sections, we will take as an illustrative example the NARMA-based DPD, and we will define the most critical issues in designing an adaptive DPD to be finally implemented in a hardware digital platform.

Therefore, the rest of this chapter is organized into three sections. In [Section 8.2](#), we present some considerations regarding LUT organization for the design of LUT-based DPDs. [Section 8.3](#) deals with the implementation in an FPGA device of a NARMA-based digital adaptive predistorter designed as the combination of a set of LUTs. In the [final section](#), some discussions are presented. For simplicity, this chapter uses the same notations and abbreviations as those in [Chapter 7](#) of this book.

8.2 Look-up table based schemes

8.2.1 LUT schemes

Digital baseband predistortion handles the complex envelope of the PA input and output RF signals. Therefore, the LUT architecture depends on the way this complex envelope is treated in order to be predistorted. Look-up table based predistorters may be classified by its LUT approach in [\[1\]](#):

- mapping predistorters (two 2-D LUTs)
- polar predistorters (two 1-D LUTs)
- complex gain based predistorters (one 1-D LUT).

In *Mapping Predistortion* the complex input signal is represented by its Cartesian in-phase (x_I) and phase-quadrature (x_Q) components. By using two 2-D LUTs (one for the I and another for the Q) both input Cartesian components are mapped to a new constellation of Cartesian components: $y_I = x_I + f_I(x_I, x_Q)$ and $y_Q = x_Q + f_Q(x_I, x_Q)$. [Figure 8.1](#) shows the block scheme of a mapping predistorter architecture. Some results using mapping predistortion are reported in [\[2\],\[3\],\[4\]](#). The major drawback with the mapping predistorter is the size of the two 2-D LUTs, which results in long adaptation times and computational complexity.

Polar Predistortion was presented by Faulkner *et al.* in [\[5\]](#) and uses two 1-D LUTs containing magnitude gain and phase rotation, respectively. The principle is illustrated in [Figure 8.2](#). The input signal amplitude, R_{in} , is used to point an address of the LUT containing an amplitude gain factor $R_{out} = F_R(R_{in})$. Then, this factor is used to multiply the original input signal amplitude (amplitude correction). Moreover, this gain factor is

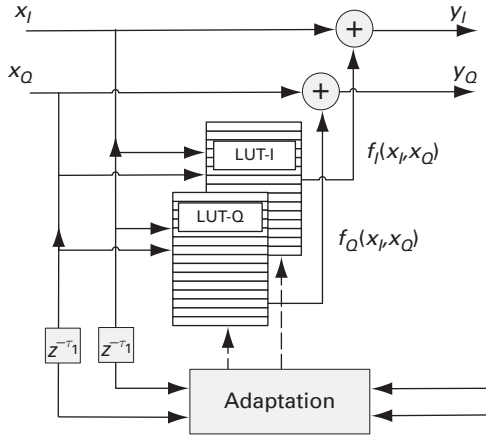


Figure 8.1 Mapping predistortion.

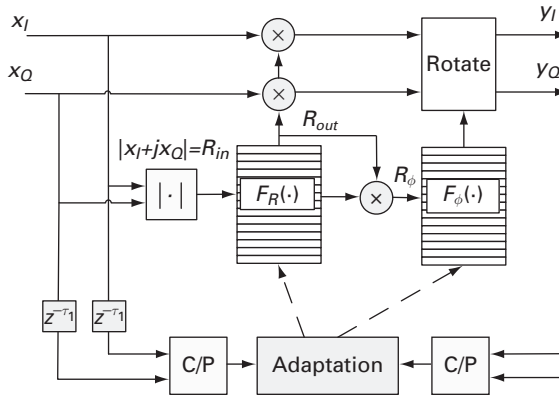


Figure 8.2 Polar predistortion.

also used to multiply the input amplitude and then addressing a second table containing the predistortion phase, $\Phi_{out} = F_\Phi(R_\Phi)$. This phase is used to rotate the signal previously predistorted in amplitude (phase correction). Both tables are 1-D, so the access time and the memory requirements are reduced respect to the previous Mapping Predistortion.

Complex Gain Predistortion was presented by Cavers in [6],[7] and later has been used by many other authors, such as in [44],[45],[46],[47]. Instead of having two 1-D tables with amplitude gain and phase rotation, this approach (see Figure 8.3) has only an LUT containing complex-valued gain factors given in Cartesian form. The complex gain predistorter uses the power of the input signal ($R = |x|^2$) to point to a unique LUT address containing the complex gain of a predistortion function, $G_{LUT}(R)$. The complex gain ($G_{LUT}(R)$) that results from the LUT addressing is used to predistort the input signal x by computing the complex product $y = x \cdot G_{LUT}(R)$. With only one 1-D table the complex gain predistortion reduces the complexity and adaptation time in comparison with mapping predistortion.

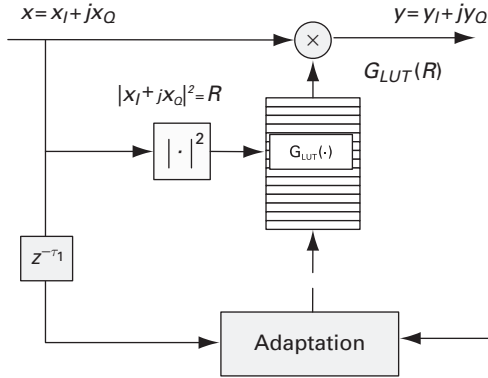


Figure 8.3 Complex Gain Predistortion.

8.2.2 LUT size and word length

As discussed before, digital predistortion schemes use 2-D or 1-D LUTs. The 2-D table method (mapping predistortion) builds a table that is indexed by the I and Q inputs of the predistorter and stores the appropriate predistorter output. The disadvantage of 2-D tables is their large memory requirement and, in adaptive systems, the large number of samples required before the table is fully updated. The advantage of the 2-D table is that no polar-to-Cartesian or Cartesian-to-polar conversions are necessary. Complex gain or polar methods use two 1-D tables (amplitude-phase or I-Q gains) to correct amplitude and phase distortion. A disadvantage of the 1-D table is that it often requires conversions between Cartesian and polar representations. Another drawback is that only phase-invariant nonlinearities can be corrected without adding additional LUTs in parallel. If the number of bits used to quantize the signal is represented by n :

- a full two-dimensional table memory requires $(2^n)^2$ entries, while
- a one-dimensional table requires two tables of 2^n values.

The *table size* has an inverse relationship with the adjacent channel interference. Each doubling of the table size decreases the adjacent channel power ratio (ACPR) by 6 dB up to a limit after which increasing the table size no longer reduces the adjacent channel interference, and may indeed degrade performances [12]. Furthermore, in adaptive systems, LUT size has an inverse relationship with the convergence speed of the adaptation, so the LUT should be kept as small as possible.

The *word length* of parameters stored in memory is related to the noise and accuracy of the whole system. Since a reduction in the adjacent channel interference is desired, then the dominant noise must not be due to quantization. The desired level of adjacent-channel-interference suppression will set the minimum signal to quantization noise ratio (SNR_q) for the table entries; SNR_q first depends on word length and then on the ratio of the peak amplitude to the root mean square (RMS) amplitude of the signal. Further details on this topic can be found in [13].

8.2.3 LUT spacing

The organization of the LUT spacing has been an interesting topic of discussion for several years [14], [15], [16], since a uniform or non-uniform spacing of the LUT is closely related to the linearization performance achieved by DPD linearizers. The so called *companding function* is responsible for deriving the spacing of the input levels in the LUT. It performs a processing of input data for pointing the LUT in different resolution ranges (e.g. concentrating most of the memory registers for predistortion operation near to the PA compression point or the other way round, concentrating them in the cut-off region).

Figure 8.4 shows the basic structure of an LUT-based complex gain DPD with a companding function $s(\cdot)$ that is responsible for the uniform or non-uniform spacing of the LUT entries.

Therefore, if an LUT of N_t entries (see Figure 8.4) is considered, the width of the LUT bins (d) in the y domain will be:

$$d = \frac{1}{N_t \cdot s'(r_m)} \quad (8.1)$$

where r_m is defined as $r_m = |x_T|$.

The most common companding functions reported in the literature for DPD purposes are:

- Amplitude:

$$s(r_m) = r_m \rightarrow s'(r_m) = 1. \quad (8.2)$$

- Power:

$$s(r_m) = r_m^2 \rightarrow s'(r_m) = 2 \cdot r_m. \quad (8.3)$$

- μ -law (as the one used in American companders for voice telephony):

$$s(r_m) = \frac{\ln(1 + \mu \cdot r_m)}{\ln(1 + \mu)} \rightarrow s'(r_m) = \frac{\mu}{\ln(1 + \mu)} \cdot \frac{1}{1 + \mu \cdot r_m} \quad (8.4)$$

where $\mu = 255$ (8 bits) in the North American and Japanese standards.

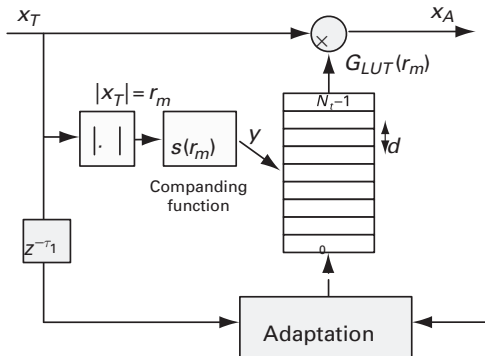


Figure 8.4 Structure of an LUT-based predistorter with companding function.

- Cavers' optimum indexing. Let the optimum companding function be $s_{opt}(\cdot)$, then the derivative can be expressed as [15]:

$$s'(r_m) = \left(\int_0^1 (w(r_m))^{\frac{1}{3}} dr_m \right)^{-1} \cdot (w(r_m))^{\frac{1}{3}} \quad (8.5)$$

where $w(r_m)$ is the nonnegative weighting function defined as

$$w(r_m) = \frac{r_m^2}{12 \cdot N_t^2} \cdot \frac{|g'(r_m)|^2}{|g'(r_m)|^4} \cdot p_r(r_m)$$

where $g(r_m)$ is the complex gain of the PA and $p_r(r_m)$ the signal amplitude *probability density function* (pdf). The condition to be satisfied by the optimum DPD gain ($f_o(r_m)$) is: $f_o(r_m) \cdot g(r_m) = K$, where K is the ideal linear gain. The optimum companding function takes into account both the derivative of the amplifier gain as well as the weighted PDF of the signal amplitude.

- Sub-optimum indexing [16]. Considering Cavers' optimum indexing in (8.5) but defining $w(r_m) = r_m^2 \cdot p_r(r_m)$.

The best linearity performance is achieved with Cavers' *optimum companding function*. However, due to its computational complexity and its dependence on the signal's pdf, PA nonlinearities and back-off can make it unsuitable. On the other hand, results reported advert that spacing by *μ -law* and *power* suffer from significant intermodulation power generation at high and low signal levels, respectively, because their LUT entries are unnecessary concentrated at the other end of the amplitude range. Cavers concludes in [15] that *amplitude* spacing (uniform spacing) despite being suboptimal, is an excellent choice for DPD purposes. It is simple, independent on the amplifier, the modulation format or the operation back-off and, moreover, its performance is very close to the limit defined by optimum spacing.

8.3 FPGA implementation of an LUT-based DPD

In order to implement the predistortion function in a general digital processor, it is important to have in mind the trade-off between accuracy (in the predistortion performance) and computational complexity. Therefore, despite some of the presented behavioral models being very good descriptors of the PA behavior, their complexity must be reduced in order to be used for DPD purposes. Such is the case of Volterra-based models for example, that first have to be simplified in order to be implemented as a set of LUTs in an FPGA [53]. On the other hand, the intrinsic nature of PA behavioral models based on a modular approach, make them suitable to be implemented as a set of (scalable in complexity) LUTs. Moreover, another important aspect when designing reconfigurable subsystems is to provide the DPD with scalability, since it provides versatility and flexibility when faced with possible changes in the transmitter.

Among the different PA behavioral models presented in previous sections, the NARMA model is capable of reproducing both PA nonlinear distortion and dynamics. Moreover, it has the advantage of introducing a nonlinear feedback path that may permit

relaxing the complexity of the PA model, in comparison with a model using only FIR terms. However an additional test on its stability is a mandatory previous step that has to be performed in order to prevent the whole system from instabilities.

This section is therefore going to be devoted to the design of a *digital baseband adaptive predistorter* based on a NARMA structure, whose parameters can be easily obtained from a closely related NARMA behavioral model of the PA. This kind of model embeds both linear and nonlinear parts, thus avoiding a cascaded Linear–Nonlinear decomposition as in Hammerstein or Wiener models. In addition, and very importantly, the NARMA DPD is well suited for being implemented by using a set of simple LUTs. The adaptation process of the proposed DPD relies just on the PA NARMA model, and its stability, despite its nonlinear feedback structure, may be assessed and ensured.

8.3.1 Identification approaches for a NARMA-based DPD

Behavioral models describe the PA nonlinear dynamic behavior but, in order to linearize the PA, it is necessary to obtain its inverse characteristic (envelope filtering technique). In this subsection two identification approaches oriented at obtaining the PA inverse characteristic are presented. The first one, named *postdistortion and translation method*, is commonly used in the literature to extract the inverse function (DPD function) of the PA in two steps. The second one is a new identification method, the *predictive predistortion method*, oriented at overcoming some formal issues that present in the former one.

Indirect learning: postdistortion and translation method

The indirect learning or post-distortion and translation method is a commonly used technique to identify the predistortion function [18],[19],[20],[21],[10]. Its basic functioning is schematically depicted in Figure 8.5. In the indirect learning approach, a first postdistortion function is estimated using the power amplifier's input $x_A(k)$ and output $y_A(k)$ baseband modulated data. Once the coefficients defining the postdistortion function are estimated, they are copied to an identical model that is used to predistort the input signal that will later feed to the PA.

The main advantage of this identification approach consists of the fact that the postdistortion function is obtained from direct input and output PA observations,

$$F_{Post}(G(x_T)) = K \cdot x_T \quad (8.7)$$

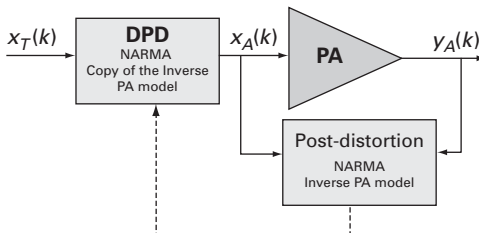


Figure 8.5 Block diagram of the indirect learning: postdistortion and translation method.

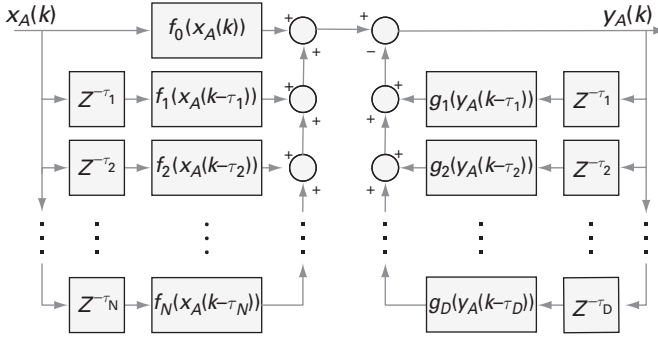


Figure 8.6 General block diagram of a NARMA structure.

with $F_{Post}(\cdot)$ and $G(\cdot)$ being the postdistortion and the PA nonlinear functions respectively, and x_T the discrete complex envelope of the signal to be transmitted.

From now on, the NARMA model depicted again in Figure 8.6 will be considered for the post and predistortion purposes. Following the notation shown in Figure 8.6, the general NARMA expression can be defined again as

$$\hat{y}_A(k) = \sum_{i=0}^N \hat{f}_i(x_A(k - \tau_i)) - \sum_{j=1}^D \hat{g}_j(y_A(k - \tau_j)). \quad (8.8)$$

Therefore, the postdistortion function in the indirect learning approach (see Figure 8.5) can be expressed as the inverse model:

$$\hat{x}_A(k) = \sum_{i=0}^N \hat{f}_i(y_A(k - \tau_i)) - \sum_{j=1}^D \hat{g}_j(x_A(k - \tau_j)) \quad (8.9)$$

where $\hat{f}_i$ and $\hat{g}_j$ are estimated nonlinear functions that can be implemented with polynomials or by using look-up tables (LUTs). In addition, $\tau_0 = 0$ while τ_i and τ_j ($\tau \subset N$) are the most significant sparse delays of the input and the output respectively, contributing to the description of the PA memory effects. As explained in [22], to estimate these optimal delays we can use heuristic search algorithms, however they are computationally demanding which is clearly a drawback in real-time applications.

To implement the predistorter function in the FPGA it is necessary to map nonlinear functions $\hat{f}_i$ and $\hat{g}_j$ into a set of LUTs. But first these nonlinear functions defining the postdistorter and predistorter nonlinear operation have to be calculated. For that reason, these static nonlinear functions in (8.9) are implemented with polynomials,

$$\begin{aligned} \hat{f}_i(y_A(k - \tau_i)) &= \sum_{p=0}^P \alpha_{pi} \cdot y_A(k - \tau_i) |y_A(k - \tau_i)|^p \\ \hat{g}_j(x_A(k - \tau_j)) &= \sum_{p=0}^P \beta_{pj} \cdot x_A(k - \tau_j) |x_A(k - \tau_j)|^p \end{aligned} \quad (8.10)$$

with P being the polynomial order, α_{pi} and β_{pj} their complex coefficients respectively, $i = (0, 1, \dots, N)$, $j = (1, \dots, D)$ and $\tau_0 = 0$.

Expanding (8.9) with (8.10) and expressing it in a more compact matrix notation, it is possible to rewrite (8.9) as

$$\hat{x}_A(k) = \lambda^H \theta \quad (8.11)$$

$$\text{where } \lambda = \begin{pmatrix} \alpha_{00}, \alpha_{10}, \dots, \alpha_{P0}, \alpha_{01}, \dots, \alpha_{P1}, \dots, \alpha_{0N}, \dots, \alpha_{PN}, \\ -\beta_{01}, \dots, -\beta_{P1}, \dots, -\beta_{0D}, \dots, -\beta_{PD} \end{pmatrix}^H,$$

$$\theta = \begin{pmatrix} y_A(k), y_A(k)|y_A(k)|, \dots, y_A(k)|y_A(k)|^P, y_A(k - \tau_1), \dots, \\ y_A(k - \tau_1)|y_A(k - \tau_1)|^P, \dots, y_A(k - \tau_N), \dots, y_A(k - \tau_N)|y_A(k - \tau_N)|^P, \\ x_A(k - \tau_1), \dots, x_A(k - \tau_1)|x_A(k - \tau_1)|^P, \dots, x_A(k - \tau_D), \dots, \\ x_A(k - \tau_D)|x_A(k - \tau_D)|^P \end{pmatrix}^T$$

and where the superindex H denotes Hermitian.

The cost function to be minimized in order to extract the α_{pi} and β_{pj} complex coefficients describing the postdistorter (based on a NARMA architecture) nonlinear function is defined by the following equation:

$$J(k) = |e(k)|^2 = |x_A(k) - \hat{x}_A(k)|^2 = |x_A(k) - \lambda^H \theta|^2. \quad (8.12)$$

The estimation error is defined as the difference between the original PA input data and the estimated output data of the postdistorter. Different identification algorithms can be used to extract the postdistorter parameters, such as *gradient algorithms* (Newton method, Steepest Descent, LMS, Fast Kalman) or *parametric estimation methods* (Least Squares, Recursive LS, Extended LS).

Once we have the postdistorter function identified, the predistortion function can be mapped in the FPGA. To fulfill this objective it is first necessary to express the predistortion function as the combination of several LUTs. With x_T being the predistorter input and x_A the predistorter output (see Figure 8.5), the predistortion function remains as an exact copy of the estimated postdistortion function:

$$x_A(k) = \sum_{i=0}^N \hat{f}_i(x_T(k - \tau_i)) - \sum_{j=1}^D \hat{g}_j(x_A(k - \tau_j)). \quad (8.13)$$

Next we need to set up the input–output relationship of the predistortion function as a Cartesian complex product between the input–output sample ($x_T(k), x_A(k)$) and a complex gain (G_{LUT}) that depends on the envelope of the signal. Therefore we can rewrite the static nonlinear functions in (8.10) as

$$\begin{aligned} \hat{f}_i(x_T(k - \tau_i)) &= x_T(k - \tau_i) \cdot G_{LUT-\hat{f}_i}^*(|x_T(k - \tau_i)|) \\ \hat{g}_j(x_A(k - \tau_j)) &= x_A(k - \tau_j) \cdot G_{LUT-\hat{g}_j}^*(|x_A(k - \tau_j)|) \end{aligned} \quad (8.14)$$

where the asterisk (*) denotes the complex conjugate of the gain (G_{LUT}). This generalization is the crucial step towards an FPGA practical DPD implementation, since it

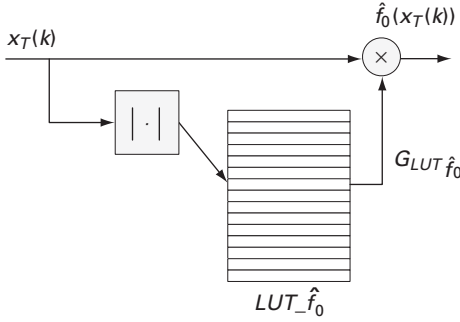


Figure 8.7 Basic Predistortion Cell for FPGA implementation.

enables the DPD to be stated in terms of a *Basic Predistortion Cell* (BPC). A simplified block diagram of a BPC based on a complex gain DPD with *uniform spacing* is shown in Figure 8.7.

Therefore, a BPC is the fundamental building block from which DPD functions derived from the NARMA model (or also Volterra derivations) can be quickly mapped into an FPGA device.

Finally the input–output relationship of the predistorter in (8.13) can be expressed as the combination of a set of LUTs,

$$\begin{aligned}
 x_A(k) = & x_T(k) \cdot \underbrace{\left[\alpha_{00} + \alpha_{10}|x_T(k)| + \cdots + \alpha_{P0}|x_T(k)|^P \right]^*}_{LUT_{\hat{f}_0} = G_{LUT_{\hat{f}_0}}^* (|x_T(k)|)} + \\
 & x_T(k - \tau_1) \cdot \underbrace{\left[\alpha_{01} + \alpha_{11}|x_T(k - \tau_1)| + \cdots + \alpha_{P1}|x_T(k - \tau_1)|^P \right]^*}_{LUT_{\hat{f}_1} = G_{LUT_{\hat{f}_1}}^* (|x_T(k - \tau_1)|)} + \\
 & \vdots \\
 & x_T(k - \tau_N) \cdot \underbrace{\left[\alpha_{0N} + \alpha_{1N}|x_T(k - \tau_N)| + \cdots + \alpha_{PN}|x_T(k - \tau_N)|^P \right]^*}_{LUT_{\hat{f}_N} = G_{LUT_{\hat{f}_N}}^* (|x_T(k - \tau_N)|)} - \\
 & x_A(k - \tau_1) \cdot \underbrace{\left[\beta_{01} + \beta_{11}|x_A(k - \tau_1)| + \cdots + \beta_{P1}|x_A(k - \tau_1)|^P \right]^*}_{LUT_{\hat{g}_1} = G_{LUT_{\hat{g}_1}}^* (|x_A(k - \tau_1)|)} - \\
 & \vdots \\
 & x_A(k - \tau_D) \cdot \underbrace{\left[\beta_{0D} + \beta_{1D}|x_A(k - \tau_D)| + \cdots + \beta_{PD}|x_A(k - \tau_D)|^P \right]^*}_{LUT_{\hat{g}_D} = G_{LUT_{\hat{g}_D}}^* (|x_A(k - \tau_D)|)}. \quad (8.15)
 \end{aligned}$$

The predistorter input–output relationship expressed in (8.15) by means of BPCs is graphically shown in Figure 8.8. The implementation of the DPD function in the FPGA is therefore possible thanks to this multi-LUT architecture, consisting of the combination of a set of BPCs [10].

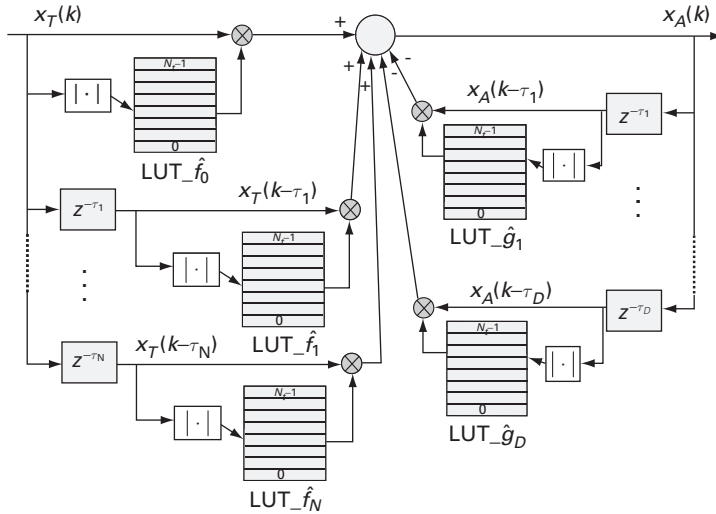


Figure 8.8 Multi-LUT implementation of a NARMA-based predistorter with the indirect learning approach.

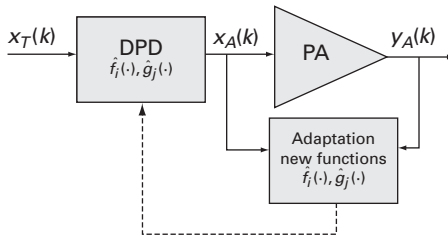


Figure 8.9 Block diagram of the Predictive Predistorter approach.

Predictive predistortion method

Despite the good results shown by the indirect learning approach (regarding linearization performance), the postdistortion and translation method assumes the commutative property for cascading nonlinear systems, which is not mathematically rigorous at all. For that reason in [11] an alternative method for identifying and adapting the predistortion function was proposed, the *predictive predistortion method*. This method permits an identification and later adaptation of the predistortion function that relies just on the PA NARMA model. The stability of the NARMA model, as has been discussed previously, may be in any case assessed and ensured.

The *digital predictive predistortion* approach follows the block diagram shown in Figure 8.9, where digital predistortion linearization is carried out at baseband by adaptively forcing the PA to behave as a linear device.

The functioning of the DPD is quite intuitive: first, it is necessary to perform an identification of the low-pass complex envelope PA behavioral model; for example, the NARMA model described in Equation (8.8). Then, considering the NARMA model, a first set of nonlinear functions ($\hat{f}_i$ and $\hat{g}_j$) defining the PA behavior are calculated at baseband using the PA input x_A and output y_A discrete complex envelope data. Once $\hat{f}_i$ and $\hat{g}_j$ are

identified, y_D is now considered as the desired linearized PA output. This desired output is defined as the signal to be transmitted (x_T) multiplied by a linear and real gain (G_{linear}),

$$y_D(k) = x_T(k) \cdot G_{linear}. \quad (8.16)$$

As shown in Figure 8.9, $x_A(k) = x_T(k)$ if no baseband predistortion is considered. From the PA NARMA model expression in (8.8) we can obtain the following equation

$$\hat{f}_0(x_A(k)) = y_A(k) - \sum_{i=1}^N \hat{f}_i(x_A(k - \tau_i)) + \sum_{j=1}^D \hat{g}_j(y_A(k - \tau_j)) \quad (8.17)$$

and solving (8.17) we can further obtain the necessary amplifier input $\hat{x}_A(k)$ in order to achieve a certain output $y_A(k)$. Since the desired output $y_D(k)$ is evaluated a priori (see (8.16)), then in (8.17) y_A is replaced by y_D (and the same for all delayed output samples). In other words, the desired output $y_D(k)$ is considered as a prediction of the future value of $y_A(k)$ (current output) and consequently, the input value of the PA ($\hat{x}_A(k)$) that permits achieving the desired output ($y_A(k) = y_D(k)$) is calculated. Finally, the digital PD output $\hat{x}_A(k)$ (and PA input) can be expressed as

$$\hat{x}_A(k) = \hat{f}_0^{-1} \left(y_D(k) - \sum_{i=1}^N \hat{f}_i(\hat{x}_A(k - \tau_i)) + \sum_{j=1}^D \hat{g}_j(y_D(k - \tau_j)) \right). \quad (8.18)$$

In order to implement the predistortion function in an FPGA, it is necessary to express it with a combination of BPCs. For that reason, similarly to what has been done previously, we now consider the NARMA-based PA behavioral model expressed in (8.8), expanded by means of their polynomial expression. Hence it results in the following

$$\begin{aligned} \hat{y}_A(k) = & x_A(k) \cdot \underbrace{\left[\alpha_{00} + \alpha_{10}|x_A(k)| + \dots + \alpha_{p0}|x_A(k)|^p \right]^*}_{LUT-\hat{f}_0=G_{LUT-\hat{f}_0}^*(|x_A(k)|)} + \\ & x_A(k - \tau_1) \cdot \underbrace{\left[\alpha_{01} + \alpha_{11}|x_A(k - \tau_1)| + \dots + \alpha_{p1}|x_A(k - \tau_1)|^p \right]^*}_{LUT-\hat{f}_1=G_{LUT-\hat{f}_1}^*(|x_A(k - \tau_1)|)} + \\ & \vdots \\ & x_A(k - \tau_N) \cdot \underbrace{\left[\alpha_{0N} + \alpha_{1N}|x_A(k - \tau_N)| + \dots + \alpha_{pN}|x_A(k - \tau_N)|^p \right]^*}_{LUT-\hat{f}_N=G_{LUT-\hat{f}_N}^*(|x_A(k - \tau_N)|)} - \\ & y_A(k - \tau_1) \cdot \underbrace{\left[\beta_{01} + \beta_{11}|y_A(k - \tau_1)| + \dots + \beta_{p1}|y_A(k - \tau_1)|^p \right]^*}_{LUT-\hat{g}_1=G_{LUT-\hat{g}_1}^*(|y_A(k - \tau_1)|)} - \\ & \vdots \\ & y_A(k - \tau_D) \cdot \underbrace{\left[\beta_{0D} + \beta_{1D}|y_A(k - \tau_D)| + \dots + \beta_{pD}|y_A(k - \tau_D)|^p \right]^*}_{LUT-\hat{g}_D=G_{LUT-\hat{g}_D}^*(|y_A(k - \tau_D)|)}. \end{aligned} \quad (8.19)$$

Therefore, in a similar manner as in the indirect learning approach, we can rewrite the predistortion function in (8.18) in a more convenient DPD expression, in terms of the (delayed) complex inputs and outputs multiplied by their corresponding complex gain (G_{LUT}),

$$\hat{x}_A(k) = z(k) \cdot G_{LUT-\hat{f}_0}^*(|z(k)|) \quad (8.20)$$

where $z(k)$ is defined as

$$\begin{aligned} z(k) = & y_D(k) + \sum_{j=1}^D y_D(k - \tau_j) \cdot G_{LUT-\hat{g}_j}^*(|y_D(k - \tau_j)|) - \\ & - \sum_{i=1}^N \hat{x}_A(k - \tau_i) \cdot G_{LUT-\hat{f}_i}^*(|\hat{x}_A(k - \tau_i)|) \end{aligned} \quad (8.21)$$

Thanks to (8.20) and (8.21), the DPD can be mapped into the FPGA as a set of parallel and cascade BPCs, as depicted in Figure 8.10. Furthermore, suitable DPD operation is obtained by just downloading the appropriate complex gain values into each BPC LUT.

8.3.2 Implementation of the adaptive DPD function in an FPGA

As presented in the previous subsection, the NARMA-based predistortion function can be described by a Multi-LUT architecture, as depicted in Figure 8.8 when applying the indirect learning method or, alternatively, in Figure 8.10 when applying the predictive NARMA predistortion method. In both cases, the hardware implementation of the predistortion function is carried out in an FPGA board in charge of the DPD processing at the actual sampling rate, and thus allowing high data throughput. The VHDL (VHSIC hardware description language; VHSIC: very-high-speed integrated circuit) is the general-purpose language used to design the FPGA. Alternatively, we can use the Xilinx Inc. System Generator [23] which consists of a Simulink library called the Xilinx Blockset and software to translate a Simulink model into a hardware realization of the model. It automatically produces command files for FPGA synthesis, HDL simulation, and implementation tools, so that the user can work entirely in graphical environments in going from system specification to hardware realization.

The adaptation process, consisting of periodically calculating new predistortion values to fill the LUT bins of each BPC, can be performed in both a host PC running Matlab or in the same FPGA in charge of the real-time predistortion. Therefore, provided that the DPD function is carried out in an FPGA device, two possible configurations can be considered to provide the DPD with adaptativity:

- External adaptation; where the algorithms that will provide new updated LUT contents to the DPD function are programmed in an external device (PC or DSP).
- Near real-time FPGA adaptation; where the LUT contents forming the BPC are continuously updated in the same FPGA board without interrupting the transmission (taking advantage of the parallel processing capabilities of this device).

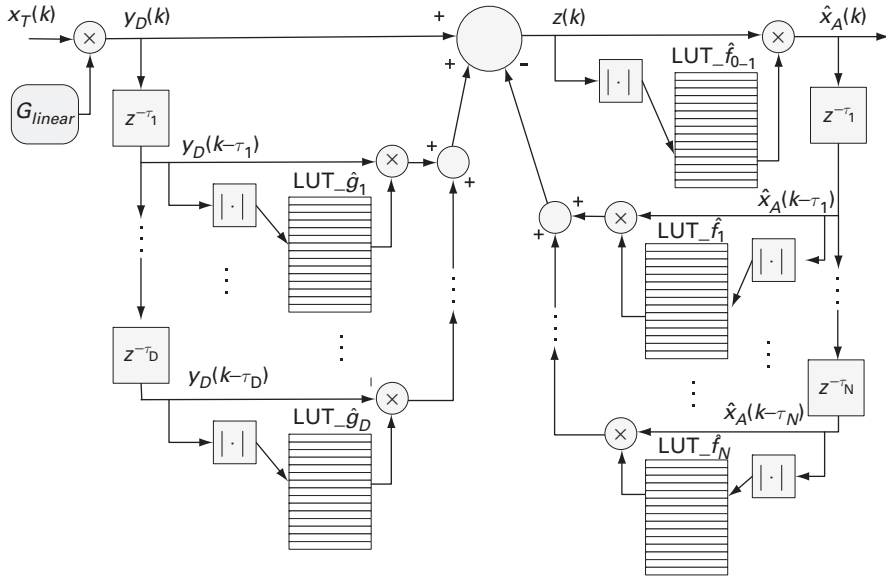


Figure 8.10 Predictive digital predistorter implementation with multi-LUTs.

In the following, some details of the DPD FPGA implementation by means of System Generator description blocks will be presented. In addition both configurations (external and real-time) to perform closed-loop adaption will be presented and discussed.

External adaptation

In this kind of configuration, while predistortion is carried out in an FPGA, the more complex adaptation algorithms are computed in an external device, such as a DSP board as in [24],[25], and thus will have more relaxed adaptation time constants. For simplicity and to enhance flexibility during the prototyping procedures, the DSP device can be replaced by an external host PC in which Matlab is in charge of the adaptation. Figure 8.11 shows a block diagram of an adaptive DPD based on a NARMA PA behavioral model where the adaptation (update of the LUT contents forming the BPCs) is carried out in an external device.

The predistortion of the input signal is performed real-time in the FPGA, while the LUT contents are being updated periodically. The DPD function follows a NARMA architecture that corresponds to the one described in (8.20) and (8.21), obtained through the predictive predistortion method. Figure 8.12 shows the block diagram of the NARMA-based predistorter composed of several direct and recursive BPCs contributing in parallel. In general a BPC is composed of the following hardware blocks: a complex multiplier, a random access memory (RAM) acting as an LUT and an address calculator to select the position of memory within the LUT. However, in order to access the RAM from an external device (such as when using a PC or a DSP to update the LUT contents), a Shared Memory is required.

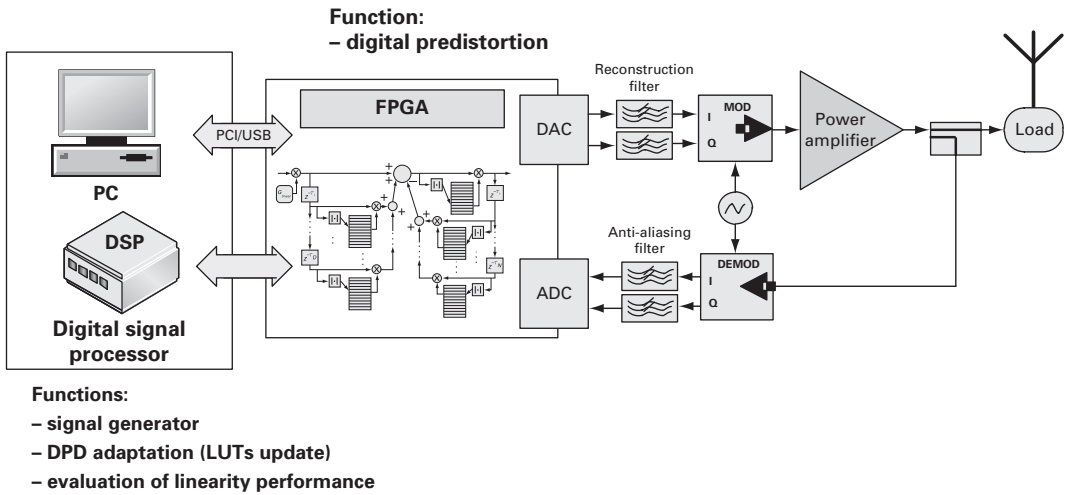


Figure 8.11 Block diagram of a DPD with external adaptation, carried out in a PC or DSP board.

Figure 8.13 shows the implementation of a BPC with external adaptation using System Generator blocks. As can be observed, the shared memory block has three default inputs: *address* (addr), *data in* (din), and *write enable* (we) and one default output: *data output* (dout). In addition, by setting access protection to the mode *lockable* causes two additional ports to appear: an input port *request* (req) and an output port *grant*. Driving the we port with 1 indicates that the value on the din port should be written to the memory address pointed to by port addr, whose value has to be included in the range between 0 and $L-1$, L being the memory depth of the LUT (in Figure 8.13 L is equal to $2^{11} = 2048$). When access protection is set to *lockable*, the req and grant ports are used to control access to the memory. Before a read or write can occur, a request must first be made by setting req to 1. When grant becomes 1, the request for access has been allowed and read or write operations can proceed. As can be observed in Figure 8.13, the din port is set to 0, because the data contents to update the LUT are provided from an external device, i.e. a PC.

Figure 8.14 shows the implementation of a complex multiplier (including the complex conjugate of the DPD gain) with System Generator blocks. By means of the System Generator BitBasher block the 28 bits of unsigned complex data are split into two signed 14.13 (14 bits with the binary point located at the 13th bit) I/Q data components. Similarly, the complex gains within the LUT are split into two signed 14.12 real and imaginary gains. Then, it is possible to perform the complex multiplication consisting of four single (or real) multiplications and two additions. Finally the resulting signed 14.13 I/Q data components of the predistorted signal are again merged into unsigned 28.0 complex data by means of a BitBasher.

Among the different companding functions to address the LUT, the amplitude or uniform spacing companding function has shown a good trade-off between accuracy

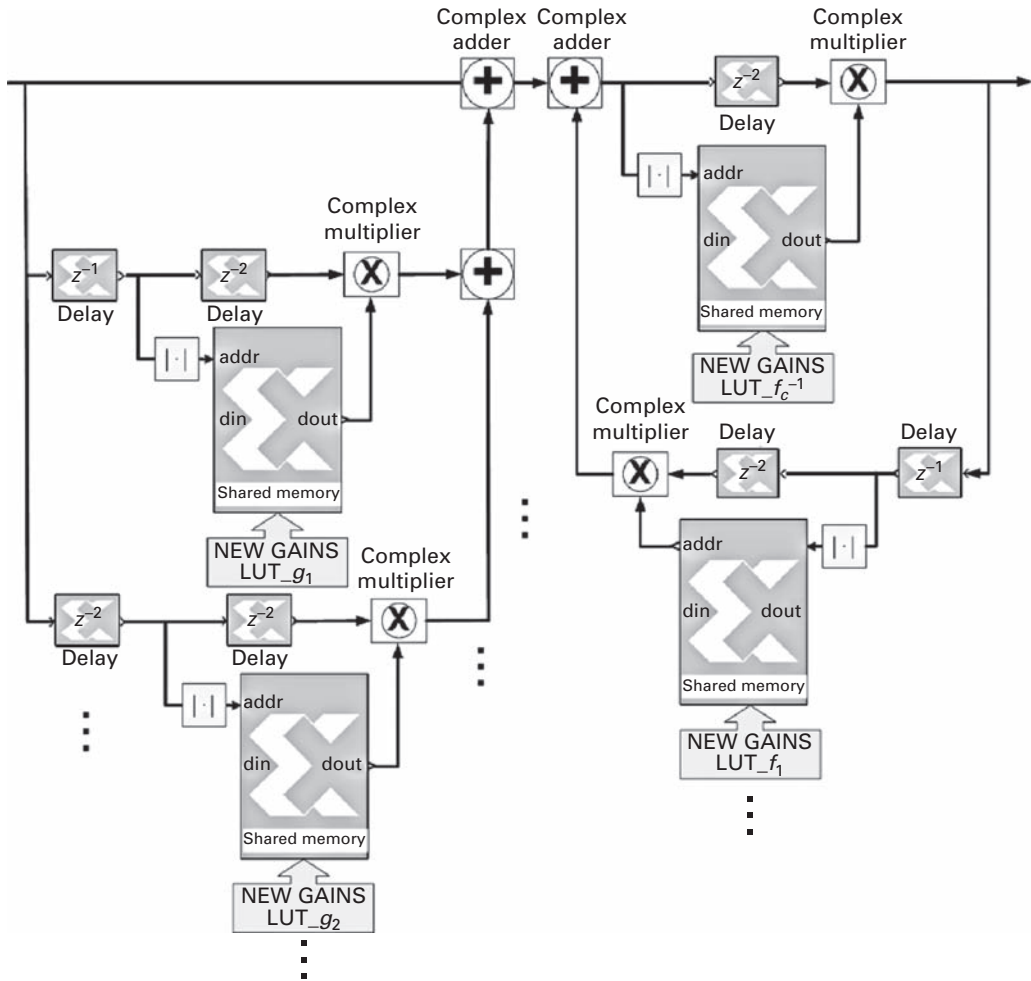


Figure 8.12 NARMA DPD with external adaptation implemented with System Generator.

and complexity. Therefore, when considering a uniform spacing it is necessary to calculate the absolute value (modulus) of the complex envelope of the signal. The modulus block to address the LUT implemented with System Generator blocks is shown in Figure 8.15. The square root function can be implemented with a CORDIC (COordinate Rotation DIgital Computer) square root block. However, the CORDIC block is more complex and introduces a higher latency (31 clock periods) than mapping the square root function in a RAM block (1 clock period of latency). As shown in Figure 8.15, once the squared absolute value of the complex envelope of the input signal is obtained (ranging from 0 to 1), then it is multiplied by L , L being the memory depth of the LUT (RAM) containing the L values of the square root function (the LUT contents are obtained by evaluating the square root of L real numbers ranging between 0 and 1, in steps of $1/L$).

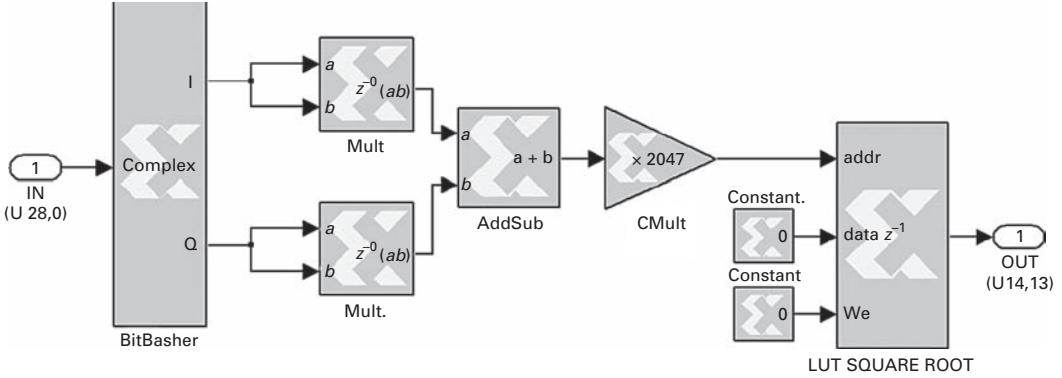


Figure 8.15 Modulus implemented with System Generator.

As shown in Figure 8.11, to provide the DPD with adaptation capabilities a closed-loop processing is necessary (a feedback loop from the PA output towards the FPGA, through the demodulator and A/D converters). In this particular implementation, the FPGA provides the external host PC with buffers (through Shared Memories) of predistorted and PA output data of $L = 2^{11}$ I/Q samples each, from which the NARMA-based DPD function is derived. The interface between the development kit board (containing the FPGA, A/D and D/A converters, clocks, etc.) and the host PC can be either PCI or USB. The D/A and A/D blocks in System Generator must be driven by a 14-bit fixed-point signal, with the binary point at position 13.

The update of the LUT contents describing the DPD relies just on the PA NARMA model defined in (8.8), since the DPD nonlinear functions of the Predictive-NARMA DPD in (8.18) are the same as in (8.8) except for $\hat{f}_0^{-1}$. The PA model identification procedure is performed using the least squares (LS) algorithm, since the constraint of computational complexity has disappeared when operating in a PC running Matlab.

Therefore, considering x_A the complex data vector at the DPD output (PA input – see Figure 8.9) and y_A the corresponding time-aligned complex data vector of the PA output (and normalized by the linear PA gain to allow signals comparison), both vectors of L samples length, we define

$$x_A^{pi}(k) = x_A(k - \tau_i) |x_A(k - \tau_i)|^p \quad (8.22)$$

$$y_A^{pj}(k) = y_A(k - \tau_j) |y_A(k - \tau_j)|^p. \quad (8.23)$$

Then, the NARMA input–output relationship in (8.8) can be expressed in a matrix notation as

$$y_A = Q\hat{\delta} \quad (8.24)$$

where $y_A = [y_A(0), \dots, y_A(L-1)]^T$,

$$Q = [x_A^{00}, \dots, x_A^{p0}, \dots, x_A^{0N}, \dots, x_A^{PN}, y_A^{01}, \dots, y_A^{p1}, \dots, y_A^{0D}, \dots, y_A^{PD}],$$

$$x_A^{pi} = [x_A^{pi}(0), \dots, x_A^{pi}(L-1)]^T, \quad y_A^{pj} = [y_A^{pj}(0), \dots, y_A^{pj}(L-1)]^T \quad \text{and}$$

$$\delta = \begin{bmatrix} \alpha_{00}, \dots, \alpha_{P0}, \dots, \alpha_{0N}, \dots, \alpha_{PN}, \\ -\beta_{01}, \dots, -\beta_{P1}, \dots, -\beta_{0D}, \dots, -\beta_{PD} \end{bmatrix}^T.$$

Hence, the Least Square (LS) solution for (8.24) is:

$$\hat{\delta} = (Q^H Q)^{-1} Q^H y_A \quad (8.25)$$

where superindex H denotes complex conjugate transpose.

The adaptive process performed in the host PC is schematically depicted in Figure 8.16 and described by the following steps:

- In the first step, identify by means of the LS algorithm in (8.25), new $\hat{f}_i$ and $\hat{g}_j$ nonlinear functions (described in (8.8)) by monitoring current PA input (x_A) and output (y_A) data vectors, as shown in Figure 8.9.
- Second step: test the stability of the resulting PA NARMA model.
- The third step consists of inverting the $\hat{f}_0$ memoryless nonlinear function to obtain the digital predistorter output as shown in (8.18).
- The last step consists of generating all necessary LUTs to implement the predistorter function described in (8.18). From (8.19), (8.20), and (8.21) complex gains (G_{LUT}) are computed and fed into the FPGA in the BPC convenient LUT form. Then, back to step one.

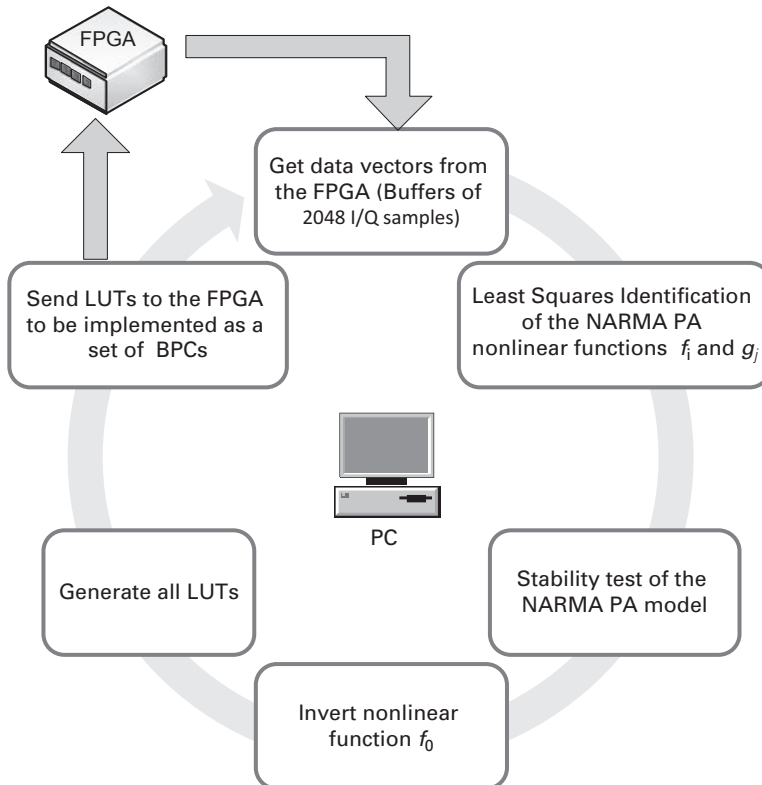


Figure 8.16 Flow diagram of the external adaptation carried out in the host PC.

At every adaptation routine, all LUT contents (containing the BPC complex gains) are recalculated, even if they have never been addressed. The computational complexity introduced by the LS algorithm, handling matrices of 2048×2048 complex data, makes a real-time adaptation unfeasible.

External adaptation results

As an example, in the following some results of the performance of a NARMA-based DPD with external adaptation implemented in an FPGA will be shown and discussed. The DPD function capable of compensating PA memory effects was implemented in the FPGA, while the adaptation of the LUT contents was carried out in a PC running Matlab. For the sake of simplicity a PA model with memory effects was also designed in the same FPGA. This PA model was based on a Hammerstein configuration, and it was extracted from input–output records of a PA based on a Freescale MRF21170 MOSFET transistor. By adjusting the PA quiescent point two PA modes of operation were considered: class AB and class B operation. The test signal used to evaluate the performance of the adaptive DPD was a multi-carrier (OFDM-based) 16-QAM signal presenting a significant PAPR (up to 14 dB).

Figures 8.17 and 8.18 show the AM-AM characteristics for both PA modes of operation, class AB and class B, respectively. Similarly, Figures 8.19 and 8.20 show the output power spectra, and Figures 8.21 and 8.22 the 16-QAM constellation at the PA after demodulation. Since no crest factor reduction techniques were considered, with an OFDM-based test signal having high PAPR there is a trade-off between the allowed amount of signal clipping after linearization (which introduces signal distortion) and the predistortion gain (a higher gain implies better power efficiency). For example, the spectral regrowth (out-of-band distortion) compensation in Figure 8.20 is lower (around 10 dB of ACPR) than the one achieved in Figure 8.19; the reason can be observed in Figure 8.18, where the linearized AM-AM characteristic presents some clipping. The price to pay for avoiding clipping and its consequent linearity degradation is a loss of gain

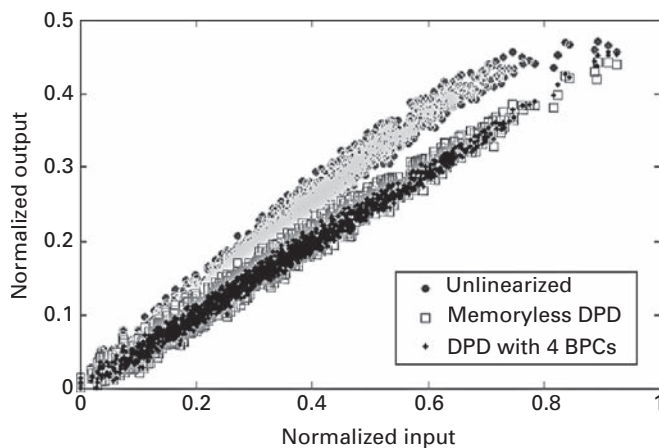


Figure 8.17 AM-AM characteristic for a class AB PA.

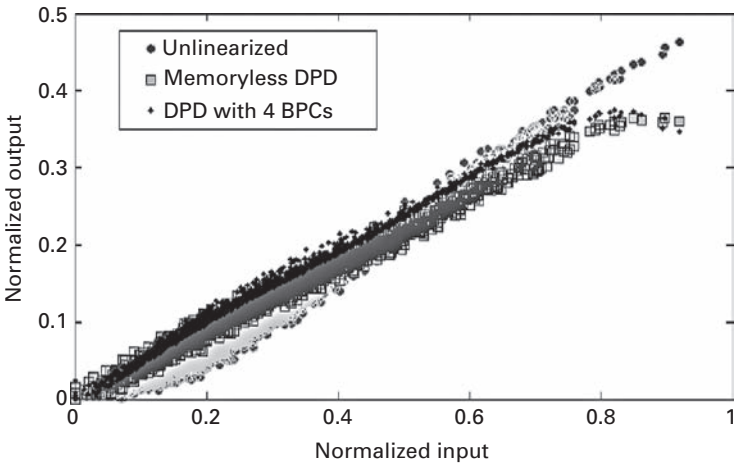


Figure 8.18 AM-AM characteristic for a class B PA.

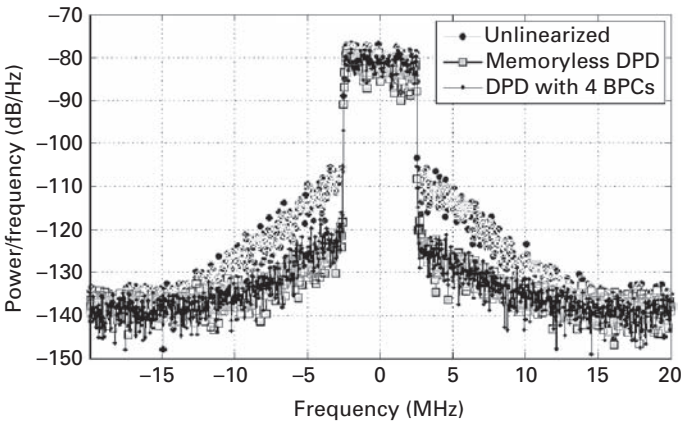


Figure 8.19 Output power spectra for a class AB PA.

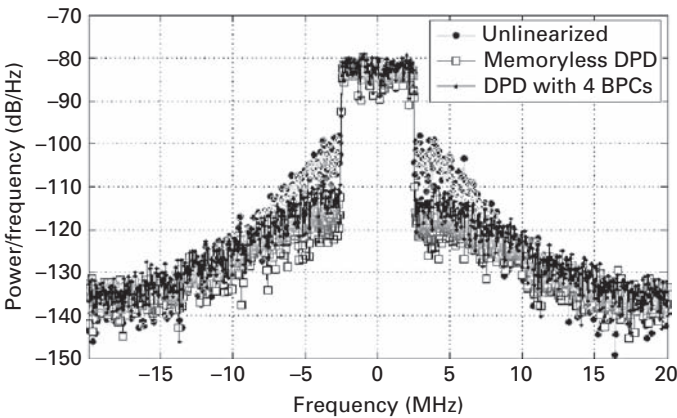


Figure 8.20 Output power spectra for a class B PA.

and thus the power efficiency (see Figure 8.19). Finally, the advantage of compensating PA memory effects is mainly reflected in the in-band distortion compensation as observed in Figure 8.21 and Figure 8.22. The use of the NARMA architecture with memory effects compensation capabilities improves the EVM figures with respect to using only memoryless DPD.

Experimental results of the implementation of this NARMA-based adaptive DPD can be found in [26].

Near real-time adaptation

A more compact DPD configuration consists of performing the adaptation in the same FPGA, as schematically depicted in Figure 8.23. The adaptation can be performed in the same FPGA by considering the use of an iterative algorithm capable of updating the

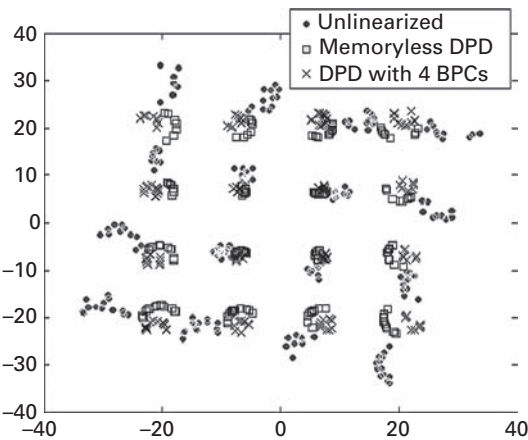


Figure 8.21 16-QAM constellation for a class AB PA.

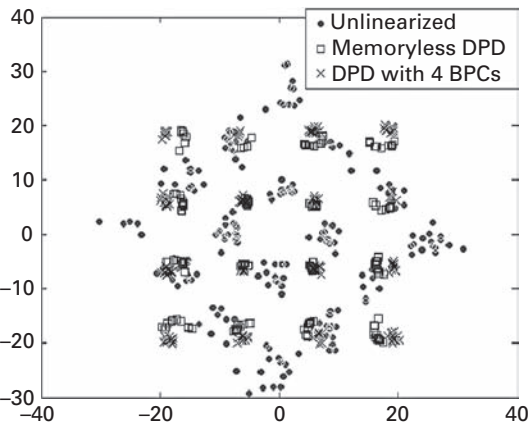


Figure 8.22 16-QAM constellation for a class B PA.

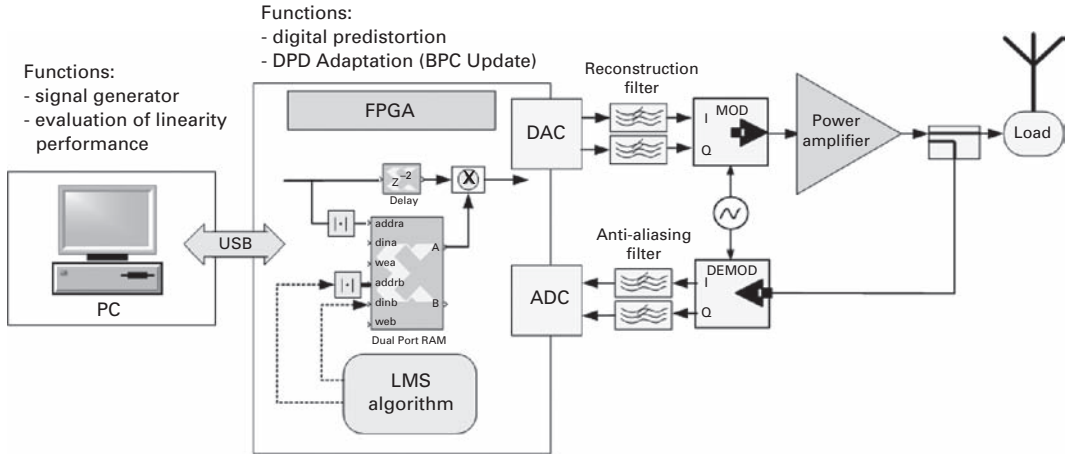


Figure 8.23 Block diagram of a DPD with almost real-time adaptation carried out in the FPGA.

individual predistortion complex gains (G_{LUT}) stored in BPCs. For example, taking advantage of the parallel computing capabilities of the FPGA, the least mean square (LMS) algorithm [27] is used to simultaneously update all gains involved in a predistortion execution. The complex gains involved in the predistortion are being updated at every clock period (near real-time). Obviously, there is a small delay between the instant the complex gain has been used for predistortion and the instant that the same gain is being updated (around 40 clock periods). This small delay is due to the feedback loop (the PA output signal has to be downconverted at baseband) and the LMS algorithm computing latency. Therefore, unlike external adaptation where all BPCs complex gains were recalculated at every adaptation execution, the FPGA device can perform near real-time updates without the need of a DSP or any other kind of advanced coprocessor.

To show the functioning of the adaptive DPD fully implemented in an FPGA, let us focus on the functioning of a predistortion function consisting of a single BPC, with this being capable of only compensating for memoryless nonlinear distortion. By observing the block diagram of the indirect learning approach in Figure 8.5, it is then possible to define the memoryless DPD function stated in terms of a single BPC as

$$x_A(k) = x_T(k) \cdot G_{LUT-\hat{f}_0}^*(|x_T(k)|) \quad (8.26)$$

with $x_T(k)$ being the original signal to be transmitted (Tx Data in Figure 8.24), $x_A(k)$ the predistorted signal (Tx DPD in Figure 8.24), and $(G_{LUT-\hat{f}_0})^*$ the complex gain stored in the LUT of the single BPC defining the predistortion function. Figure 8.24 shows the BPC responsible for the DPD function. The main difference with the BPC presented in Figure 8.13 is the Dual-Port RAM, consisting of a block that has two independent sets of ports for simultaneous reading and writing. Independent address, data, and write enable ports allow shared access to a single memory space. By default, each port set has one output port and three input ports for address, input data, and write enable. The use of a Dual-Port RAM allows the LUT complex gains to be updated continuously without interrupting the normal data transmission.

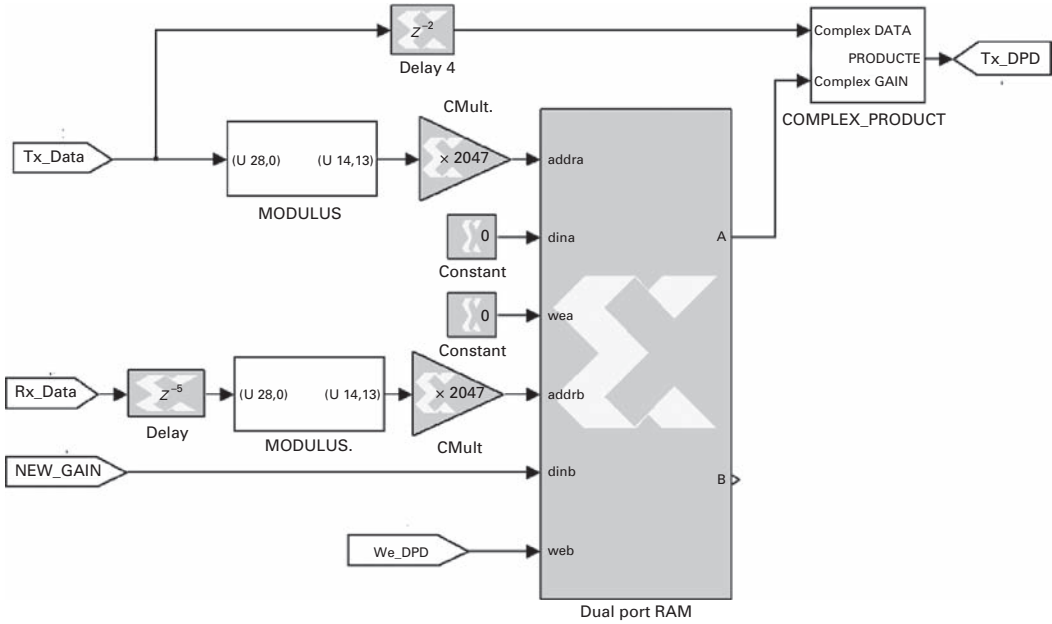


Figure 8.24 Predistortion BPC with a Dual-Port RAM for LMS adaptation.

The new gains that will update the predistortion function or, more precisely, the LUT contents of the predistortion BPC in Figure 8.24, are calculated in the postdistortion block as schematically depicted in the indirect learning block diagram in Figure 8.5. The postdistortion BPC implemented with System Generator blocks is depicted in Figure 8.25, where again a Dual-Port RAM block is used to allow shared access to a single memory space. The postdistortion gains are calculated to be later copied in the predistortion function. The new postdistortion (and predistortion) gains are estimated using the LMS algorithm:

$$G_{LUT-\hat{f}_0}^{(NEW)}(|y_A(k)|) = G_{LUT-\hat{f}_0}^{(OLD)}(|y_A(k)|) + \mu \cdot y_A(k) \cdot e^*(k) \quad (8.27)$$

with $y_A(k)$ being the PA output signal (Rx Data in Figure 8.25), μ being the step size that determines the learning rate of the LMS algorithm, and $e^*(k)$ the complex conjugate of the identification error, defined as

$$e(k) = x_A(k) - \hat{x}_A(k) \quad (8.28)$$

where $x_A(k)$ is the predistorted signal (Tx DPD in Figure 8.25) and $\hat{x}_A(k)$ is the estimated predistorted signal (Tx DPD MOD in Figure 8.25) that can be defined as

$$\hat{x}_A(k) = y_A(k) \cdot G_{LUT-\hat{f}_0}^* (|y_A(k)|). \quad (8.29)$$

Therefore, after a transient period, in which all LUT gains are continuously being updated (in both the postdistortion BPC and the predistortion BPC), the PA output

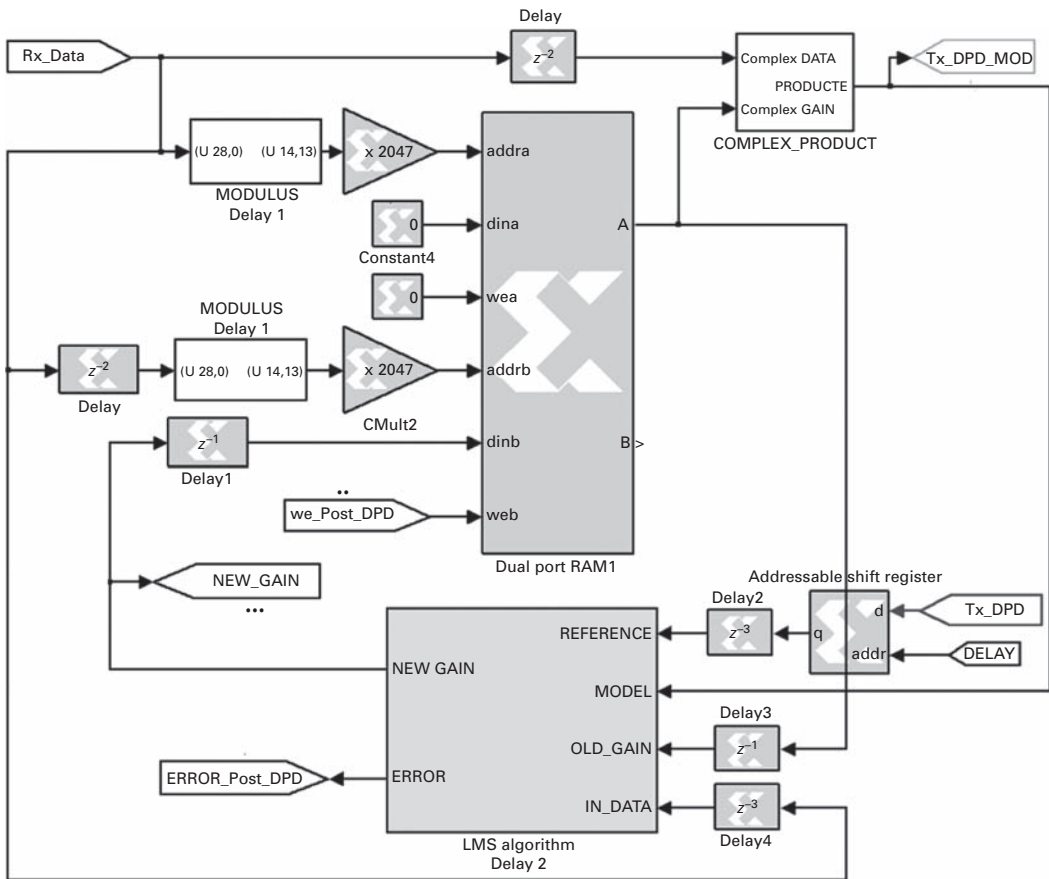


Figure 8.25 Postdistortion BPC with a Dual-Port RAM and a LMS block to perform online adaptation.

converges to the desired output (having again the original signal to be transmitted), by achieving then the desired linear amplification.

Near real-time adaptation results

In the following some results showing the performance of the adaptive memoryless DPD with near real-time adaptation performed in the same FPGA will be presented and discussed. In this case, a class AB memoryless PA model was used to test the adaptive DPD functioning, since only a BPC was considered to perform DPD. Again, the test signal used to evaluate the performance of the adaptive DPD was a multi-carrier (OFDM-based) 16-QAM signal presenting a significant PAPR.

Figure 8.26 shows the AM-AM characteristics considering a memoryless class AB PA, while Figure 8.27 and Figure 8.28 show the output power spectra and the 16-QAM constellation, respectively. In this case, the AM-AM characteristic in Figure 8.26 shows no clipping, but instead there is a loss of gain that can be observed in both output power spectra and the 16-QAM constellation. The identification error converged to a residual

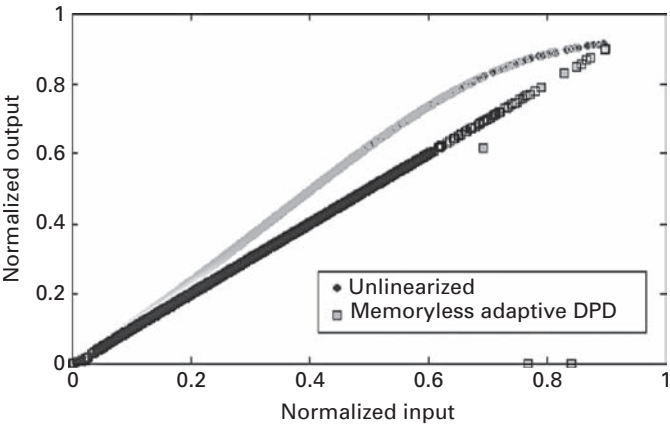


Figure 8.26 AM-AM characteristic for a memoryless class AB PA.

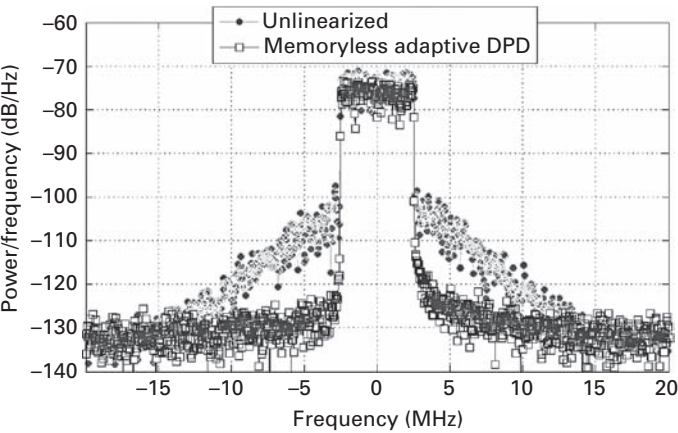


Figure 8.27 Output power spectra for a memoryless class AB PA.

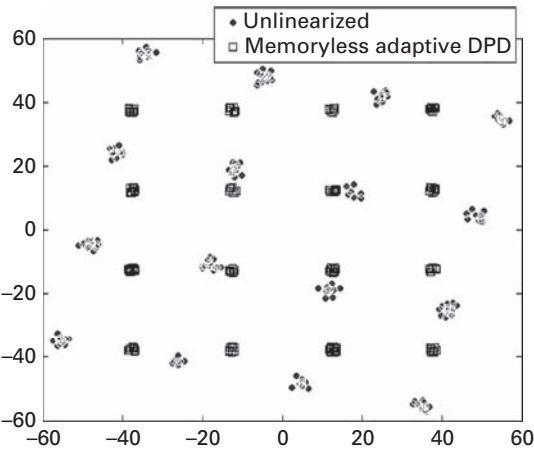


Figure 8.28 16-QAM constellation for a memoryless class AB PA.

value of 0.05 in less than 4 milliseconds. However, as observed in Figure 8.26 (there is a sample located outside the main linearized straight line) the perfect convergence of all complex gains to their desired value may take a little bit longer than 4 milliseconds. This delay is due to the fact that the convergence speed depends on both the μ value of the LMS algorithm and the frequency with which a particular complex gain is addressed.

8.4 Discussion

The use of LUTs to design the predistortion function provides an efficient solution to be later implemented in a digital signal processor without introducing an excessive computational cost. This chapter has discussed several considerations regarding the organization of the LUTs. In particular, we have followed a complex gain LUT architecture with uniform spacing to design the NARMA-based DPD in terms of BPCs. Moreover, as the conducting example, the principles of the extraction of the NARMA-based DPD function by means of the predictive predistortion approach have been described, as well as its later derivation into a set of LUTs (or BPCs) for its FPGA implementation.

Practical design issues and real-time DPD hardware implementation topics have been also tackled in this chapter. More precisely, we have introduced the design, using the Xilinx Inc. System Generator, of the fundamental building blocks defining a BPC. The replica of several BPCs have the flexible and scalable nature of the NARMA-based digital adaptive predistorter. In addition, two types of identification and adaptation processes have been presented: (a) one performed by an external device that processes (by means of the LS algorithm) offline blocks of input–output data records to estimate the new complex gains of the BPCs forming the predistortion function; and (b) a near real-time adaptation (LMS algorithm) carried out sample by sample in the same FPGA.

As shown throughout this chapter, linearization techniques such as DPD can help to mitigate nonlinear distortion in PAs. However, the resulting power efficiency achieved with linearization techniques applied to PAs operating as current sources is still in the range of 15 to 30 percent [28]. To overcome this power efficiency limitation, currently, highly efficient switching mode (class D, E, F) RF power amplifiers are being considered. The conventional Cartesian transmitter architectures are being modified or adapted to ensure optimal system-level amplification with highly efficient switching mode RF PAs. System-level architectures (some including power supply control) with great potential for high-efficiency operation such as: linear amplification with nonlinear components (LINC) [29]; envelope tracking (ET) PAs [30], [31]; envelope elimination and restoration (EE&R) [32], [33]; or polar transmitters (PTs) [34]; have been revived thanks to current high-speed digital signal processing platforms which substitute their analog counterparts, subjected to tolerances and periodic adjustments. These emergent software defined radio (SDR) transmitter architectures are implemented in versatile digital signal processing devices that allow the introduction of correcting mechanisms and reconfigurability.

Therefore, one of the future challenges consists of redesigning the BPC architecture of the predistortion function (initially designed to be used in Cartesian transmitters dealing with the In-phase and Quadrature-phase components of the baseband complex signal), to

be capable of compensating the nonlinear distortion generated by these highly efficient transmitter architectures. These architectures with dynamic supply require the generation of the polar (module and phase) components of the transmitted signal and thus, the BPC forming the predistortion function has to be redesigned according to this. Moreover, these highly efficient PA architectures are also highly sensitive to misalignments (i.e. delay mismatches). Hence, there is still plenty of room for proposing new adaptive/correcting mechanisms to compensate for all sources of linear and nonlinear distortion that arise due to the intrinsic configuration of these new transmitter architectures.

References

- [1] L. Sundström. Digital RF Power Amplifier Linearizers. Analysis and Design. Ph.D. thesis, Lund University, August 1995.
- [2] Y. Nagata. Linear Amplification Technique for Digital Mobile Communications. In *Proc. IEEE Vehicular Technology Conference (VTC'89)*, volume 1, pages 159–164, May 1989.
- [3] M. Minowa, M. Onoda, E. Fukuda, and Y. Daido. Backoff Improvement of an 800-MHz GaAs FET Amplifier for a QPSK Transmitter Using an Adaptive Nonlinear Distortion Canceller. In *Proc. IEEE Vehicular Technology Conference (VTC'90)*, pages 542–546, May 1990.
- [4] A. Mansell and A. Bateman. Practical Implementation Issues for Adaptive Predistortion Transmitter Linearisation. In *IEE Colloquium on Linear RF Amplifiers and Transmitters*, April 1994.
- [5] M. Faulkner and M. Johansson. Adaptive linearization using predistortion-experimental results. *IEEE Transactions on Vehicular Technology*, **43**(2):323–332, May 1994.
- [6] J. K. Cavers. A Linearizing Predistorter with Fast Adaptation. In *Proc. IEEE Vehicular Technology Conference (VTC'90)*, pages 41–47, May 1990.
- [7] J. K. Cavers. Amplifier Linearization Using a Digital Predistorter with Fast Adaptation and Low Memory Requirements. *IEEE Transactions on Vehicular Technology*, **39**(4):374–382, November 1990.
- [8] W.-J. Kim, S. P. Stapleton, Kim J. H., and C. Edelman. Digital Predistortion Linearizes Wireless Power Amplifiers. *IEEE Microwave Magazine*, **6**(3):54–61, September 2005.
- [9] M. Helaoui, S. Boumaiza, A. Ghazel, and F. M. Ghannouchi. Power and Efficiency Enhancement of 3G Multicarrier Amplifiers Using Digital Signal Processing with Experimental Validation. *IEEE Transactions on Microwave Theory and Techniques*, **54**(4):1396–1404, June 2006.
- [10] P. L. Gilabert, G. Montoro, and A. Cesari. A Recursive Digital Predistorter for Linearizing RF Power Amplifiers with Memory Effects. In *Proc. IEEE Asia-Pacific Microwave Conference Proceedings (APMC'06)*, volume 2, pages 1043–1046, Dec. 2006.
- [11] G. Montoro, P. L. Gilabert, E. Bertran, A. Cesari, and D. D. Silveira. A New Digital Predictive Predistorter for Behavioral Power Amplifier Linearization. *IEEE Microwave and Wireless Components Letters*, **17**(6):448–450, June 2007.
- [12] Bo Shi, W. Shan, and L. Sundstrom. Effects of Look-Up Table Size on Adaptive Predistortion Linearizer Systems with Error Sources. In *Proc. IEEE European Microwave Conference (EuMC'03)*, volume 3, pages 1231–1234, October 2003.
- [13] L. Sundstrom, M. Faulkner, and M. Johansson. Quantization Analysis and Design of a Digital Predistortion Linearizer for RF Power Amplifiers. *IEEE Transactions on Vehicular Technology*, **45**(4):707–719, November 1996.

- [14] J.K. Cavers. Optimum Indexing in Predistorting Amplifier linearizers. In *Proc. IEEE Vehicular Technology Conference (VTC'97)*, volume 2, pages 676–680, May 1997.
- [15] J.K. Cavers. Optimum Table Spacing in Predistorting Amplifier Linearizers. *IEEE Transactions on Vehicular Technology*, **48**(5):1699–1705, September. 1999.
- [16] K.J. Muhonen, M. Kavehrad, and R. Krishnamoorthy. Adaptive Baseband Predistortion Techniques for Amplifier Linearization. In *Proc. Asilomar Conference on Signals, Systems, and Computers*, volume 2, pages 888–892, October 1999.
- [17] L. Guan and A. Zhu. Low-cost FPGA Implementation of Volterra Series-based Digital Predistorter for RF Power Amplifiers. *IEEE Transactions on Microwave Theory and Techniques* **58**(4):866–872, 2010.
- [18] R. Marsalek, P. Jardin, and G. Baudoin. From Post-Distortion to Pre-Distortion for Power Amplifiers Linearization. *IEEE Communications Letters*, **7**(7):308–310, July 2003.
- [19] W.-J. Kim, K.-J. Cho, S. P. Stapleton, and J.-H. Kim. Piecewise Pre-Equalized Linearization of the Wireless Transmitter With a Doherty Amplifier. *IEEE Transactions on Microwave Theory and Techniques*, **54**(9):3469–3478, September 2006.
- [20] L. Ding, G. T. Zhou, D. R. Morgan, *et al.* A Robust Digital Baseband Predistorter Constructed Using Memory Polynomials. *IEEE Transactions on Communications*, **52**(1):159–165, January 2004.
- [21] K.-J. Cho, W.-J. Kim, J.-H. Kim, and S. P. Stapleton. Linearity Optimization of a High Power Doherty Amplifier Based on Post-Distortion Compensation. *IEEE Microwave and Wireless Components Letters*, **15**(11):748–750, November 2005.
- [22] P. L. Gilabert, D. D. Silveira, G. Montoro, M. E. Gadringer, and E. Bertran. Heuristic Algorithms for Power Amplifier Behavioral Modeling. *IEEE Microwave and Wireless Components Letters*, **17**(10):715–717, Oct. 2007.
- [23] *Xilinx System Generator v2.1 Reference Guide* (www.xilinx.com/support/).
- [24] S. Vázquez. Implementació en DSP's d'un Predistorsionador Digital per a la Linealització d'Amplificadors de Potència. Master's thesis, Technical University of Catalonia (UPC), 2006.
- [25] H. Ben Nasr, S. Boumaiza, M. Helaoui, A. Ghazel, and F. M. Ghannouchi. On the Critical Issues of DSP/FPGA Mixed Digital Predistorter Implementation. In *Proc. IEEE Asia-Pacific Microwave Conference (APMC'05)*, volume 5, December 2005.
- [26] P. L. Gilabert, A. Cesari, G. Montoro, E. Bertran, and J. M. Dilhac. Multi Look-Up Table FPGA Implementation of a Digital Adaptive Predistorter for Linearizing RF Power Amplifiers with Memory Effects. *IEEE Transactions on Microwave Theory and Techniques*, page submitted for revision, 2007.
- [27] S. Haykin. *Adaptive Filter Theory*. Prentice Hall, 1991.
- [28] H. Gandhi. Digital Predistortion Linearizes Broadband PAs. *Microwaves and RF*, pages 1–4, 2008.
- [29] S. Chung, P. A. Godoy, T. W. Barton, D. J. Perreault, and J. L. Dawson. Asymmetric Multilevel Outphasing Transmitter using Class-e PAs with Discrete Pulse Width Modulation. In *Proc. IEEE MTT-S International Microwave Symposium Digest (MTT)*, pages 264–267, 2010.
- [30] D. Y. C. Lie, J. Lopez, and Y. Li. Design of Highly-efficient Wideband RF Polar Transmitters using the Envelope-tracking Technique. In *Proc. IEEE Bipolar/BiCMOS Circuits and Technology Meeting BCTM 2008*, pages 57–64, 2008.
- [31] J. Hoversten and Z. Popovic. Envelope Tracking Transmitter System Analysis Method. In *Proc. IEEE Radio and Wireless Symposium (RWS)*, pages 180–183, 2010.

-
- [32] I. Kim, Y. Y. Woo, J. Kim, *et al.* High-efficiency Hybrid EER Transmitter using Optimized Power Amplifier. *IEEE Transactions on Microwave Theory and Techniques* **56**(11):2582–2593, 2008.
 - [33] F. Wang, D. Kimball, J. Popp, *et al.* Wideband Envelope Elimination and Restoration Power Amplifier with High Efficiency Wideband Envelope Amplifier for WLAN 802.11g Applications. In *Proc. IEEE MTT-S International Microwave Symposium Digest*, 2005.
 - [34] P. M. Cabral, J. C. Pedro, J. A. Garcia, and L. Cabria. A Linearized Polar Transmitter for Wireless Applications. In *Proc. IEEE MTT-S International Microwave Symposium Digest*, pages 935–938, 2008.

9 Digital predistortion and its combination with crest factor reduction

by Anding Zhu

In wireless communications, it is often desirable to transmit a signal as efficiently as possible to achieve low power dissipation. In the meantime, it is necessary to keep signal distortion small. Unfortunately, these two desirable features contradict each other. For example, it is well known that the radio frequency (RF) power amplifier (PA) in wireless transmitters is inherently nonlinear, and when operated near saturation, it causes intermodulation products that interfere with adjacent channels. To reduce distortion, a common method is to “back-off” the output power of the PA to ensure that signal peaks do not exceed the saturation point of the amplifier to permit distortion-free transmission. This “back-off” has not been a big concern in narrowband systems, e.g. GSM/EDGE systems, where the modulation exhibits a modest few dB peak-to-average power ratio (PAPR), which degrades efficiencies by around 30 percent that is deemed acceptable. However, the “back-off” could dramatically degrade the PA efficiency in wideband systems because, for example, 3GPP LTE (Long Term Evolution) signals often exhibit PAPRs in excess of 10 dB. Backing off an amplifier to operate within its linear region with these waveforms rapidly forces highly inefficient operation, which can lead to typical efficiencies of less than 10%. That is not acceptable in practical situations.

Digital predistortion (DPD) is proposed to use digital signal processing techniques to compensate for the nonlinear distortion in the RF PA, thereby allowing it to be operated at higher drive levels for higher efficiency [1]. The attraction of this approach is that the nonlinear PA can be linearized by a standalone add-on digital block, freeing vendors from the burden and complexity of manufacturing complex analog/RF circuits. Digital predistortion has become one of the most popular and feasible linearization techniques in modern wireless communication systems.

This chapter introduces basic principles of digital predistortion techniques and addresses many major issues related to the state of the art of DPD development for RF power amplifiers in modern wireless communication systems. The chapter is organized into the following eight sections. The first section, [Section 9.1](#), will outline the principles of operation of a digital predistortion system. Various DPD models, covering both static nonlinearities and memory effects, will be presented in [Section 9.2](#) and DPD model extraction structures will be given in [Section 9.3](#). We will then introduce the bandwidth and sampling rate requirements and DPD system characterization procedures in

Section 9.4 and 9.5, respectively. Section 9.6 will discuss DPD implementation issues and Section 9.7 will introduce some of the widely used crest factor reduction techniques, which are proposed to reduce the peaks of the transmit signal to a satisfactory level before digital predistortion to allow the PA to be operated at higher average power to achieve higher power efficiency. In the final section, Section 9.8, we will summarize the chapter and discuss some challenges that DPD may face in the future.

9.1 Principles of operation

The principle of digital predistortion is simple: a nonlinear distortion function is built up within the digital domain that is the inverse of the distortion function exhibited by the power amplifier [1]. The basic block diagram of a predistortion scheme is shown in Figure 9.1. Typically, in a linear RF power amplifier, e.g., Class A/AB, the gain of the PA is flat in the small signal region and tends to decrease when the input power level increases. In order to compensate for this gain compression, a linearization block, called predistorter, can be inserted in the signal path prior to the PA, to predistort the signal by making a gain expansion in the high power level, namely, inverting the gain characteristic of the PA. A linear amplification can then be achieved when these two nonlinear systems are serially connected.

By employing a DPD, the linear range of the PA can be extended, so that the PA can be operated at a higher drive level to deliver higher output power without introducing severe distortion, which leads to higher power efficiency. This is illustrated in Figure 9.2 [2]. Without DPD, the relationship between the input power and output power of a PA is nonlinear. If the amplifier is operating in compression, the output versus input curve falls below the linear curve, and the actual output power of the PA is lower than that for linear operation. For instance, when driven with P_{in} , the PA only can produce P_{out} at point A rather than at point B, which is the linear response. This nonlinear response causes distortion to the signal, which forces that the PA must be backed off to its linear region and thus produces very low average power. However, through the introduction of DPD,

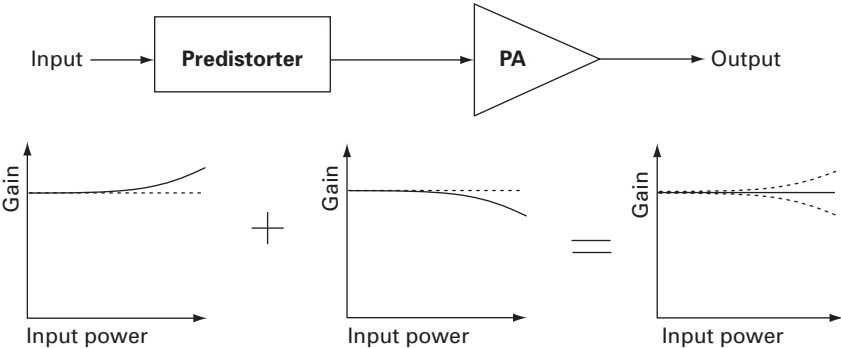


Figure 9.1 Predistortion concept.

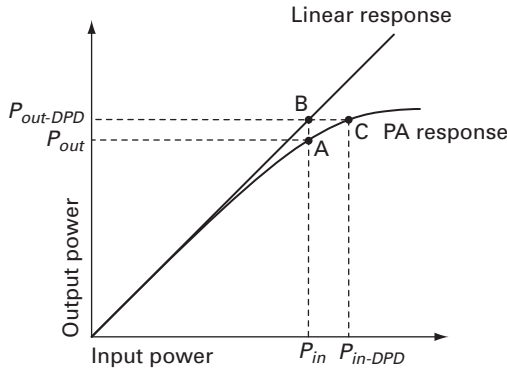


Figure 9.2 PA output vs. input.

the P_{out} versus P_{in} curve can be forced to have a linear response over a high input power level. For instance, with DPD, P_{in} can be increased to P_{in-DPD} to produce power at point C, coinciding with the linear response at point B. Although distortion is induced by the PA, it is compensated by the DPD, and thus the overall system is still linear. Operating a PA to the nonlinear compression region leads to much higher efficiency.

Most predistorters today are implemented in the digital baseband domain. In these systems, the original I/Q (in-phase/quadrature) baseband signals are pre-processed, e.g., adjusted by the inverse characteristics of the PA, to generate the predistorted signals. The predistorted signals are then passed through digital-to-analog converters (DACs), modulated and up-converted to the RF frequency, and finally sent to the PA. In order to extract and update the coefficients of the DPD, a small fraction of the transmit signal is fed back and transferred to baseband via a down-converter and an analog-to-digital converter (ADC). The model parameter extraction unit compares the captured input and output data, and extracts the coefficients for the DPD. This model extraction unit is normally only operational during the initial setup of the system or whenever the characteristics of the system are changed. The block diagram of a baseband digital predistortion system is illustrated in Figure 9.3.

9.2 DPD models

In recent years, various DPD systems have been proposed, with the main differences usually lying either in the implementation of predistortion architectures and models or in parameter extraction algorithms.

9.2.1 Memoryless system

In narrow band systems, distortion is mainly caused by static nonlinearities in RF power amplifiers. This means that only memoryless DPDs are required for compensating these

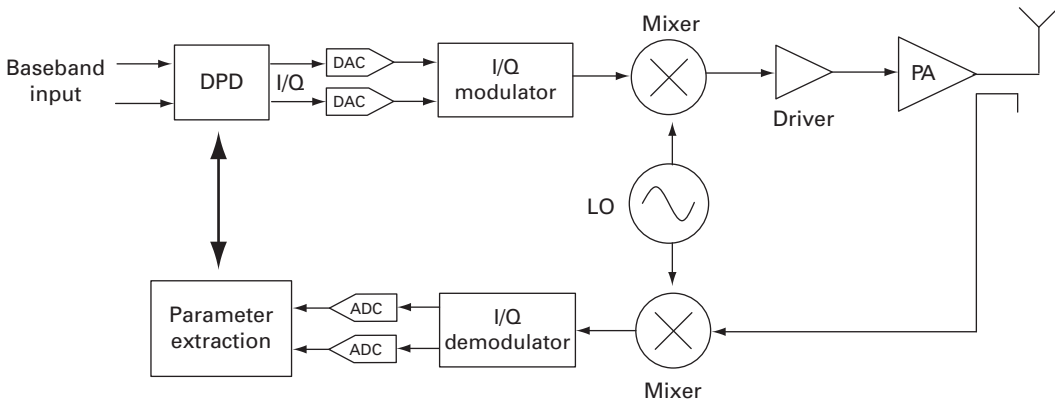


Figure 9.3 Block diagram of DPD.

distortions. There are two distinct DPD structures in memoryless systems: (i) look-up tables (LUTs) and (ii) polynomial functions.

Look-up table

The LUT-based DPD has been widely used because it is relatively simple and easily implemented. In this system, the predistortion data or coefficients are stored in a look-up table. The transmit signals are predistorted by processing with these coefficients in the table, and then converted to RF frequency and fed to the amplifier. There are mainly two methods of look-up table indexing: direct mapping and complex gain indexing.

The direct mapping method is the simplest solution for building the LUT. In this approach, a measure of the output signal is down-converted to baseband and compared to the original input signal to produce the errors, which are used to map input constellation points to desired output locations on the complex Cartesian (I/Q) plane using a two-dimensional table. The original input I/Q signal in Cartesian form is directly converted into a predistorted signal by using the mapping in the table and then up-converted and amplified. The direct mapping process allows the compensation for any memoryless distortion, including errors produced by the modulation process. The method however has a number of disadvantages. For instance, the memory requirement is huge because every point on the complex plane must be mapped. The convergence of the predistorter is slow because every region on the Cartesian plane must be accessed before convergence is complete [3]. An improvement can be made by converting the I/Q Cartesian form to a polar form with magnitude and phase, so that the two-dimensional tables can be converted to one-dimensional ones. However, the phase compensation is too complicated to implement [4].

The complex gain-based method uses the magnitude level of the signal as an index to produce the required complex gain output for a range of input power levels. This effectively avoids storing one-to-one mappings or magnitude and phase errors for each data point, and thus significantly reduces the system complexity [5]. In this approach, the table stores the complex gain values based on the magnitude level of the signal only. The

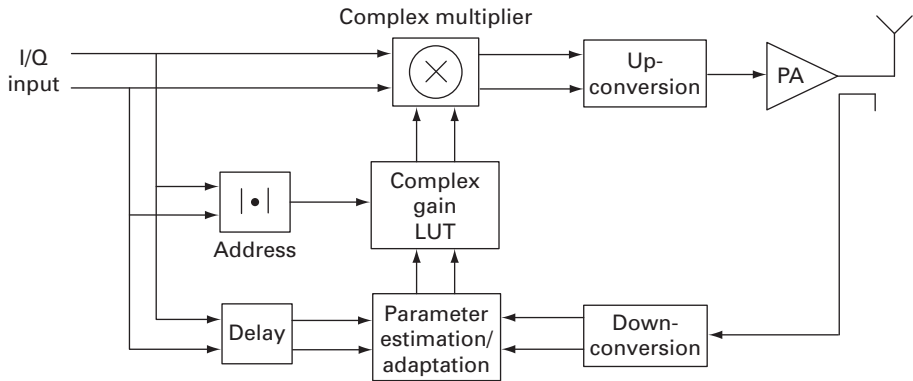


Figure 9.4 Complex Gain LUT DPD.

address to the table can be calculated by using CORDICs [6] or directly as the squared magnitude of the input I/Q signal. The maximum magnitude value can be normalized to unity and divided into a limited number of intervals, which are then used to build the LUTs for storing the complex gain values. Interpolation can be used to calculate intermediate gain values. The input signal is predistorted by a single complex multiplication between the original signal and the complex gain retrieved from the table. The block diagram of this system is shown in Figure 9.4. The memory requirement of this method is considerably smaller than that required by the direct mapping-based method. The convergence time is also significantly shorter, and re-convergence time is negligible. However, this approach is only suitable for reducing distortion caused by AM/AM (amplitude to amplitude) and AM/PM (amplitude to phase) conversions, because these complex gain values are dependent upon the signal envelope (amplitude) only. And the parameter estimation and computation required by this method is more complex than that of the mapping predistorter, and accurate control of the Cartesian modulator is required, as the method cannot correct errors due to the modulation process.

The LUT-based DPD is simple and easily implemented, and it can be fitted to any type of characteristics. However, while wireless systems are migrating to wideband operations, the bandwidth and dynamic range of the signals to be transmitted are expanded dramatically. In order to cover all of the operating regions, the LUT requires a large memory to store sufficient data, which significantly increases the system complexity.

Polynomial Functions

The analytical function-based DPDs use mathematical equations to represent the characteristics of the PA and the DPD. In this approach, each value of the predistorted output is directly calculated with a small number of pre-decided parameters when it is needed, which avoids storing all possible values and thus requires much less memory storage than the LUT-based methods. A simple way of describing the DPD behavior is by treating the output signal as a polynomial function of the input, which can be expressed as,

$$\tilde{y}(n) = \sum_{k=0}^{\frac{P-1}{2}} c_{2k+1} |\tilde{x}(n)|^{2k} \tilde{x}(n) \quad (9.1)$$

where $\tilde{x}(n)$ and $\tilde{y}(n)$ represents the baseband input and output signal, respectively, c_{2k+1} is the complex coefficient, and P is the nonlinear order. Note that only odd order terms are included in the equation above since the effects from even order terms are falling out of the first zone of the carrier frequency band and do not affect the modulated baseband signal.

9.2.2 System with memory

As wireless communication evolves towards high data-rate and broadband services, RF power amplifiers may exhibit frequency- or history-dependent behavior that have short and long time constants compared to the RF carrier signal or to its slowly varying envelope, which is called memory effects. The short memory effects are caused by the bandpass characteristics of the PA input and output matching networks and, sometimes, also the low-pass characteristics of the active device. The long time constant memory effects are typically due to the thermal time constants of the devices and some of the components in the biasing circuit. These memory effects means that the output of the PA no longer only depends on its instantaneous input but also on the previous inputs. To effectively linearize these wideband systems, one must compensate for both static nonlinearities as well as memory effects. There are many behavioral models available in the literature and several excellent review papers were published [7]–[9] which give various comparisons of PA and DPD models. Here we only introduce some of the most representative and widely used models below.

Volterra series

A Volterra series is a combination of linear convolution and nonlinear power series, which can be used to describe the input/output relationship of a general nonlinear, causal and time-invariant system with fading memory. In the discrete time domain, a Volterra series can be written as

$$y(n) = \sum_{p=1}^{\infty} \sum_{i_1=0}^{\infty} \cdots \sum_{i_p=0}^{\infty} h_p(i_1, \dots, i_p) \prod_{j=1}^p x(n - i_j) \quad (9.2)$$

where $x(n)$ and $y(n)$ represents the input and the output, respectively, and $h_p(i_1, \dots, i_p)$ is called the p th order “*Volterra kernel*”. In real applications, the Volterra series is normally truncated to finite nonlinear order P and finite memory length M [10][11].

For baseband digital predistortion, Equation (9.2) must be transformed to a complex baseband representation, which has the form [12],

$$\begin{aligned}
\tilde{y}(n) = & \sum_{i=0}^M \tilde{h}_1(i) \times \tilde{x}(n-i) \\
& + \sum_{i_1=0}^M \sum_{i_2=i_1}^M \sum_{i_3=0}^M \tilde{h}_3(i_1, i_2, i_3) \times \tilde{x}(n-i_1) \tilde{x}(n-i_2) \tilde{x}^*(n-i_3) \\
& + \sum_{i_1=0}^M \sum_{i_2=i_1}^M \sum_{i_3=i_2}^M \sum_{i_4=0}^M \sum_{i_5=i_4}^M \tilde{h}_5(i_1, i_2, i_3, i_4, i_5) \times \\
& \tilde{x}(n-i_1) \tilde{x}(n-i_2) \tilde{x}(n-i_3) \tilde{x}^*(n-i_4) \tilde{x}^*(n-i_5) + \dots
\end{aligned} \tag{9.3}$$

where $\tilde{x}(n)$ and $\tilde{y}(n)$ represents the baseband input and output signal, respectively, and $\tilde{h}_p(i_1, i_2, \dots, i_p)$ is the complex Volterra kernel, $()^*$ represents the conjugate transpose. In the above equation, we have removed the redundant items associated with kernel symmetry, and also the even-order kernels, whose effects can be omitted in band-limited modulation systems [12].

Volterra series can be employed to describe the relationship between the input and output of an amplifier with memory in a very general way. The advantage of Volterra series is that, the output of the Volterra model is linear with respect to its coefficients, as the polynomial model. Under the assumption of stationarity, if we solve for the coefficients with respect to a minimum mean or least square error criterion, we will have a single global minimum. Therefore, it is possible to extract the nonlinear Volterra model in a direct way by using linear system identification algorithms, e.g., least squares. However, direct use of the general Volterra series in digital predistortion is rather impractical because the number of parameters to be estimated increases exponentially with the degree of nonlinearity and memory length of the system. In practical situations, simplified models must be developed.

Memory polynomial

Probably the simplest non-trivial pruned Volterra model is the diagonal Volterra model, called *memory polynomial* [13]. In this model, all off-diagonal terms of the Volterra series are set to zero, i.e., $|i_m - i_n| = 0$ where i_m or i_n represents the delay of the input signal, and the model is written as

$$\tilde{y}(n) = \sum_{k=0}^{\frac{P-1}{2}} \sum_{i=0}^M a_{2k+1,i} |\tilde{x}(n-i)|^{2k} \tilde{x}(n-i). \tag{9.4}$$

This model can be implemented by using a group of filter-banks, and it has been widely used in many DPD systems. This diagonal restriction reduces the model complexity tremendously; however, it sometime has consequences, e.g., decreasing fidelity of the model, because, in some cases, the off-diagonal terms, also called cross terms, may play a very important role in the characteristics of RF PAs.

Generalized memory polynomial

To improve the accuracy of the memory polynomial model, the all-zero off-diagonal restriction condition, $|i_m - i_n| = 0$, can be relaxed to the “near-diagonality” restriction,

i.e., $|i_m - i_n| \leq l$, where l is a small integral [14]. This means that only the coefficients, which are “far away” from the main diagonal in the Volterra model, are removed. This model was further developed in [15], and defined as *Generalized Memory Polynomial*, which is written as,

$$\begin{aligned} \tilde{y}(n) = & \sum_{k=0}^{K_a-1} \sum_{i=0}^{L_a-1} a_{kl} |\tilde{x}(n-i)|^k \tilde{x}(n-i) \\ & + \sum_{k=1}^{K_b} \sum_{i=0}^{L_b-1} \sum_{m=1}^{M_b} b_{kim} |\tilde{x}(n-i-m)|^k \tilde{x}(n-i) \\ & + \sum_{k=1}^{K_c} \sum_{i=0}^{L_c-1} \sum_{m=1}^{M_c} c_{kim} |\tilde{x}(n-i+m)|^k \tilde{x}(n-i) \end{aligned} \quad (9.5)$$

where, K_a, L_a are the index arrays for aligned signal and envelope; K_b, L_b, M_b are the index arrays for signal and lagging envelope; and K_c, L_c, M_c are index arrays for signal and leading envelope, respectively.

Although the number of coefficients is increased in comparison with the memory polynomial model, this solution gives us more flexibility on trading off the model accuracy and the model complexity, by including different cross terms. Measurement results indicated that the model performance is significantly improved with this model [15]. This model can also be implemented by using a group of filter-banks, where the coefficients in the same diagonal line of the weight vector fall in the same FIR filter.

Dynamic deviation reduction (DDR)-based Volterra model

To overcome the complexity of the general Volterra series, another effective model-order reduction method, called *dynamic deviation reduction* (DDR), was first proposed in [16] [17], and further developed and reformatted in the discrete time domain in [18]. This is based on a new Volterra series representation:

$$\begin{aligned} y(n) = & \sum_{p=1}^P h_{p,0}(0, \dots, 0) x^p(n) \\ & + \sum_{p=1}^P \left\{ \sum_{r=1}^p [x^{p-r}(n) \sum_{i_1=1}^M \dots \sum_{i_r=i_{r-1}}^M \right. \\ & \left. \cdot h_{p,r}(0, \dots, 0, i_1, \dots, i_r) \prod_{j=1}^r x(n-i_j) \right\} \end{aligned} \quad (9.6)$$

where $x(n)$ and $y(n)$ are the input and the output, respectively; P is the order of non-linearity and M represents memory length. In this representation, the input elements are re-organized according to the order of dynamics involved in the model, with a variable, r , introduced to represent the order of the dynamics, and $h_{p,r}(0, \dots, 0, i_1, \dots, i_r)$ is the Volterra kernel with p th-order nonlinearity and r th-order dynamics.

Since the effects of dynamics tend to fade with increasing order in many real PAs, the high-order dynamics can be removed by controlling the order of dynamics, i.e., the value

of r , in the model. For instance, the first-order dynamic truncation of the Volterra model in baseband can be written as

$$\begin{aligned} \tilde{y}(n) = & \sum_{k=0}^{\frac{P-1}{2}} \sum_{i=0}^M \tilde{g}_{2k+1,1}(i) |\tilde{x}(n)|^{2k} \tilde{x}(n-i) \\ & + \sum_{k=1}^{\frac{P-1}{2}} \sum_{i=1}^M \tilde{g}_{2k+1,2}(i) |\tilde{x}(n)|^{2(k-1)} \tilde{x}^2(n) \tilde{x}^*(n-i). \end{aligned} \quad (9.7)$$

Compared to the general Volterra series, this dynamic deviation-based model reduction approach gives us an extra degree-of-freedom, through choosing the dynamic order to effectively prune the model, which makes the application of the Volterra model more flexible and more efficient. We can see that only a one-dimensional convolution is involved in the model above after the first-order dynamic reduction. This dramatically simplifies the model structure. If higher-order dynamics affect the distortion significantly, we can increase the value of r to include higher-order dynamics to improve the system performance, while, on the other hand, if only static nonlinearities need to be compensated, we can simply set r to zero, and then the model becomes a memoryless polynomial function. This DDR-based Volterra model has been used in digital predistortion in linearizing various PAs. Experimental results showed that outstanding performance can be achieved by using this model [19].

Vector threshold decomposition

In recent years, to improve efficiency of power amplifiers, new power amplifier architectures have been proposed. For example, in an Envelope-Tracking (ET) system, the drain/collector supply voltage of the RF PA is controlled by an envelope modulator that tracks the input envelope level of the transmission signal. This scheme enables the RFPA to be operated in an efficient region over a wide range of input power levels [20]. However, due to dynamic changes of the supply voltage, the ET PA exploits very distinct characteristics in different power regions, which is very different from what we normally see in the typical linear amplifiers with a fixed supply voltage. For instance, Figure 9.5 shows an example of gain and phase characteristics of a gallium nitride (GaN) ET PA excited by a WCDMA signal, where we can see that large gain reduction occurs as the input signal is changed to the lower values, which leads to gain expansion. While the power is increased to the higher level, the gain is compressed again. The phase remains flat at low and medium power, but it changes rapidly in high power [21]. This nonlinear behavior is very difficult to linearize by employing a single analytic function based DPD. This is because the polynomial or Volterra-like function based model is only suitable for modeling Class A/AB type of PAs where the PA is linear in the small signal region and tends to become nonlinear when the amplitude of the signal increases. Strong nonlinearity in the small signal region in the ET system means that very high nonlinear orders are required when fitting a DPD using a single polynomial-like model. The high-order nonlinear terms lead to very poor extrapolation of the model, which sometimes dramatically degrades the system performance.

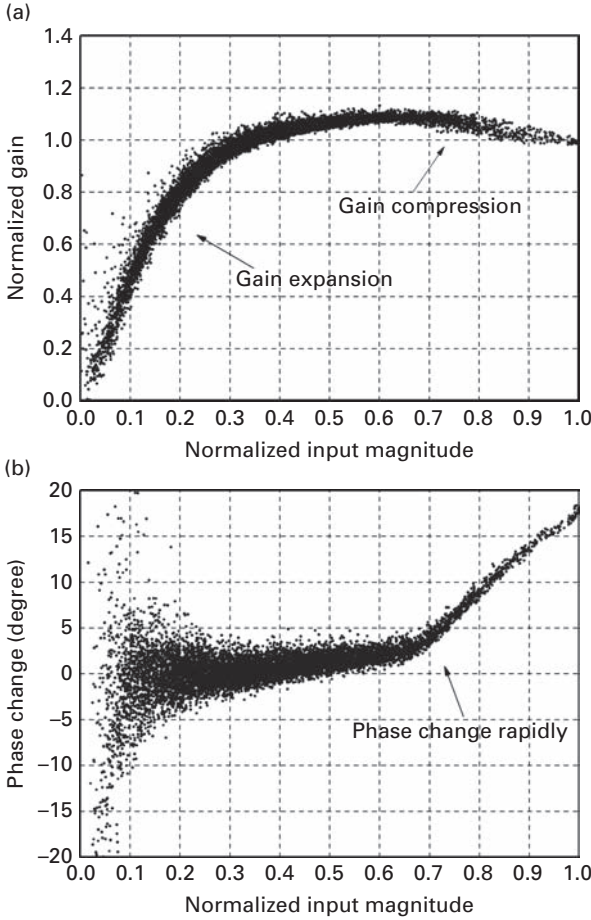


Figure 9.5 Sample characteristics of an ET PA. (a) Gain and (b) Phase.

To cope with this difficulty, a vector threshold decomposition technique was proposed in [21]. In this approach, we first set several thresholds in the input power level according to the characteristics of the PA, and then decompose the input complex envelope signal into several sub-signals by using these thresholds. For instance, the s th sub-signal in the zone s can be obtained from

$$\tilde{x}_s(n) = \begin{cases} 0 & |\tilde{x}(n)| \leq \lambda_{s-1} \\ [|\tilde{x}(n)| - \lambda_{s-1}]e^{j\phi} & \lambda_{s-1} < |\tilde{x}(n)| \leq \lambda_s \\ [\lambda_s - \lambda_{s-1}]e^{j\phi} & |\tilde{x}(n)| > \lambda_s \end{cases} \quad (9.8)$$

where $|\tilde{x}(n)|$ returns the magnitude value of $\tilde{x}(n)$ and ϕ is the phase of $\tilde{x}(n)$. Here we assume $\lambda_0 = 0$ and $1 \leq s \leq S+1$. We then continue to process each sub-signal separately by employing a low-order dynamic model, e.g., deviation reduction-based Volterra series, and finally recombine them together to produce the predistorted output. The total output of the DPD can be obtained from the sum of the outputs of all sub-functions:

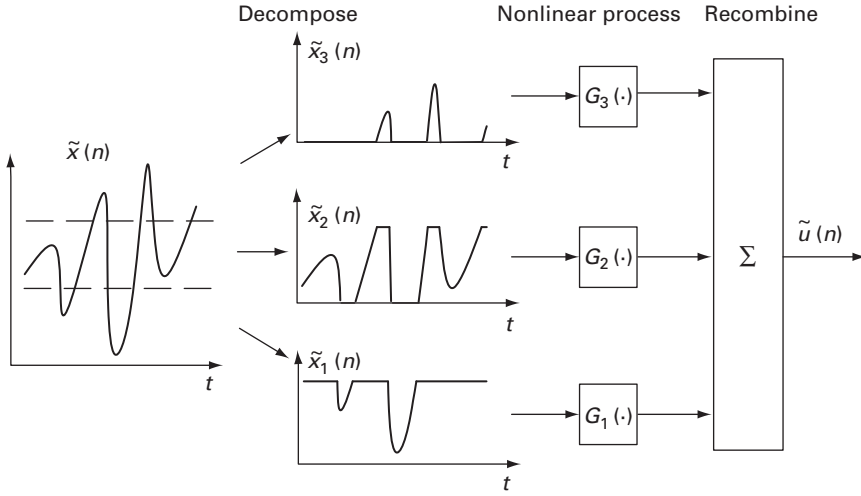


Figure 9.6 Signal decomposition and recombination [21] © 2008 IEEE.

$$\begin{aligned}
 \tilde{u}(n) &= \sum_{s=1}^{S+1} \tilde{u}_s(n) \\
 &= \sum_{s=1}^{S+1} \left\{ \sum_{k=0}^{\frac{P_s-1}{2}} \sum_{i=0}^{M_s} \tilde{g}_{s,2k+1,1}(i) |\tilde{x}_s(n)|^{2k} \tilde{x}_s(n-i) \right. \\
 &\quad \left. + \sum_{k=1}^{\frac{P_s-1}{2}} \sum_{i=1}^{M_s} \tilde{g}_{s,2k+1,2}(i) |\tilde{x}_s(n)|^{2(k-1)} \tilde{x}_s^2(n) \tilde{x}_s^*(n-i) \right\}
 \end{aligned} \tag{9.9}$$

An overview of the waveform decomposition and recombination process is outlined in Figure 9.6. By employing this piecewise type of DPD model, the distinct characteristics of the PA at different signal power levels can be accurately modeled, and thus, the distortion caused by the nonlinear behavior of the system can be effectively compensated.

Unlike using normal piecewise curve fitting, where memory effects are difficult to be embedded into the model since the sub-samples are processed in series in time and by different sub-functions, in this DPD model, memory effects can be easily taken into account because the sub-signals are processed in parallel in time, namely, the samples with time delays (memory terms) can be included in each sub-functions, and then included in the final output. Therefore, memory effects and static nonlinearities, induced by the PA, can be simultaneously compensated by employing this decomposed DPD model. More importantly, despite the fact that the input signal is decomposed into several sub-signals and more than one Volterra function is employed, the output of this model is still linear with respect to its coefficients, i.e., the final output of the DPD is linear in relation to all coefficients. This important feature implies that this nonlinear model can also be extracted by using linear system identification algorithms, e.g., least squares, and

all parameters can be extracted by using a single estimation in one matrix. This significantly simplifies the model extraction process and also minimizes model estimation errors.

9.3. Model extraction structures

There are two main structures currently employed in digital predistortion systems for model extraction: Model Inverse and Model Reference.

Model inverse structure

The model inverse structure is to find an inverse function of the nonlinear behavior of the PA to be used as the predistortion function. In the control theory, this structure is referred to as the self-tuning regulator. This often includes two parameter learning strategies: direct learning and indirect learning.

In direct learning, also referred as p th-order pre-inverse, the input–output transfer function of the power amplifier is first estimated, and the predistortion function is obtained directly by pre-inverting the PA function, as shown in Figure 9.7. This approach is suitable for LUT or curve fitting based memoryless systems, where the inverse functions/characteristics can be easily identified. However, for a system with memory, it is very difficult to find the exact inverse of the transfer function of the PA. Moreover, this normally involves a very complicated procedure, and additional fitting errors may be introduced during the model inversion [1].

Indirect learning, also referred to as p th-order post-inverse, is based on the assumption that, the p th-order pre-inverse of the system is identical to its p th-order post-inverse if linearized within p th-order nonlinearities [19][22][23]. This is reasonable because the nonlinearities beyond p th-order are normally negligible in a real system. This means that we can use two identical models for the predistorter (pre-inverse) and the postdistorter (post-inverse). During the model extraction process, the input and output of the PA are swapped, namely, the output of the PA is used as the input, and the input of the PA as the expected output, to extract the post-inverse model of the PA, and then the extracted

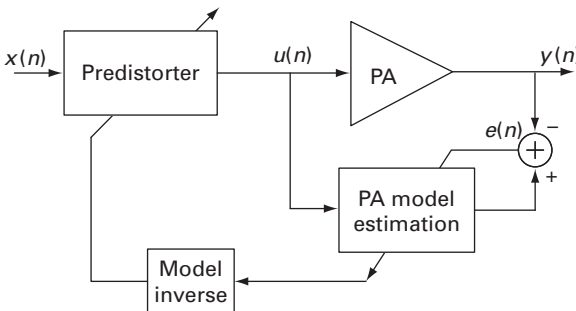


Figure 9.7 Model inverse structure with direct learning.

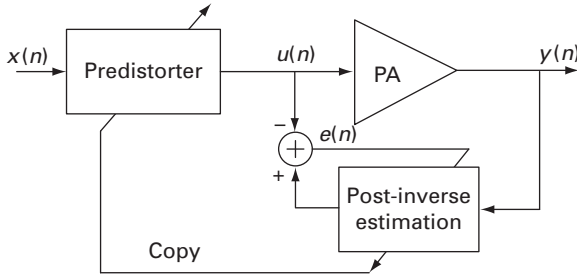


Figure 9.8 Model inverse structure with indirect learning.

parameters are directly copied to the pre-inverse (DPD) block to carry out the predistortion. The model parameters can be extracted by using a linear least square (LS) method from a set of measured input–output data points, or by using adaptive LMS/RLS algorithms on a sample-by-sample basis, if the model is linear in parameters. The basic diagram of an indirect learning system is shown in Figure 9.8. This structure is widely employed in memory DPD systems today.

To enable the model inverse structure to work, we must guarantee that there is one-to-one mapping between the input and the output; otherwise, the inversion of the system may not be possible. For example, if the PA operates near the saturation point, the high peaks of the output signal will be clipped. This clipping operation of the PA causes a non-invertible problem, which means that the DPD parameters might be no longer obtainable.

Model reference structure

Another architecture used in digital predistortion is referred to as model reference controller (MRC) or model reference adaptive system (MRAS). The MRAS is a closed-loop process. The control strategy used in this structure is made by comparing differences in the actual PA output and the reference or target output, which is the original input. The coefficients adaptation process is based on the standard delta rule or steepest descent algorithm to minimize the tracking error. The structure diagram is shown in Figure 9.9. The iteration algorithm based on the LMS iteration method is described as follows,

$$u(n) = \sum_k a_k B_k(n) \quad (9.10)$$

where $u(n)$ is the output of DPD block, $B_k(n)$ is the matrix made of original input $x(n)$, a_k are the coefficients for the predistorter which can be determined by the following function:

$$\varepsilon(n) = r(n) - \sum_k \Delta a_k B_k(n); \quad \text{LMS estimation} \quad (9.11)$$

$$r(n) = \frac{y(n)}{G_0} - x(n); \quad \text{residual error} \quad (9.12)$$

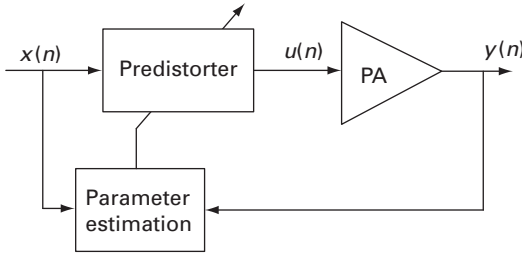


Figure 9.9 Model reference structure.

where G_0 is the nominal gain of PA. Then estimate the Δa_k to minimize the $E[|\varepsilon(n)|^2]$ and update the coefficients as

$$a_k = a_{k-1} - \lambda \Delta a_k; \quad 0 < \lambda < 1. \quad (9.13)$$

In the MRAS, the system inversion is not required because the coefficients update is governed by the first derivative of the error. As a result, the estimation process is more robust and not getting stuck in local minima. However, this approach only works best if the residual distortion in the PA output is small, because the adaptation is based on a first-order approximation in which large errors can cause the coefficient adaptation to diverge. In real applications, this system may be combined with the model inverse structure together to produce a better performance [24].

9.4 Bandwidth and sampling rate requirements

9.4.1 The receiver chain

In order to extract and update the coefficients of the DPD, a small fraction of output signal from the PA must be fed back and transferred to digital baseband, to compare with the original desired signal. The RF power amplifier is inherently nonlinear, and it produces inter-modulation products that cause spectral regrowth in the adjacent and alternate channels. For a modulated signal, third-order products appear within a range of three times the bandwidth of the desired channel. Likewise, fifth-order products appear within a range of five times the bandwidth and seventh-order products within seven times the bandwidth, as shown in Figure 9.10. To capture information regarding the nonlinearity of the PA, the DPD receiver loop must acquire a multiple of the transmit bandwidth equivalent to the order of the inter-modulation products being linearized.

In general, the Nyquist theorem dictates that the sample rate must be at least twice the sampled bandwidth in order to avoid aliasing, and thus high-speed analog-to-digital converters are often employed in the receiver loop. However, since the purpose of the DPD loop is to measure the nonlinear characteristics of the PA, it is not necessary to reconstruct the signal from the digital samples. It is only necessary to capture information for extracting the transfer function of the PA. It has been shown in [28] and [29] that for modeling a nonlinear system excited by a band-limited signal, it is sufficient to sample

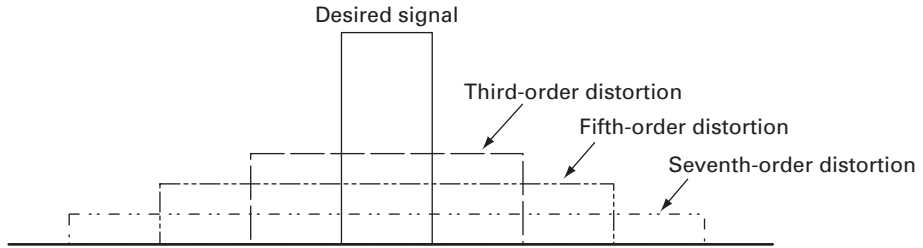


Figure 9.10 Distortion illustration.

the continuous-time input and output signals at twice the maximum input frequency rather than double the output frequency, which is called the “generalized sampling” or “under-sampling” theorem. This under-sampling can be applied for the purpose of identifying the post-inverse function for a Volterra series-based DPD compensation system, as shown in [19]. That is because the one-to-one mapping is valid under the band-limited condition, even if the bandwidth of the input of the post-inverse (the output of the PA) is wider than that of its expected output (the linearly amplified version of the original input of the PA). For a complex baseband I/Q signal, we only need to sample it with a sampling rate equal to its chip rate, which is twice its maximum baseband input frequency.

Sampling with a lower rate substantially eases the requirements on ADCs used in measurement setups [30], especially for future ultra wideband systems, e.g., LTE-Advanced systems where 100 MHz modulation bandwidth is required. For modeling a PA or building a DPD with memory effects in the discrete-time domain, lower-rate sampling also reduces the number of delay taps needed to take into account the same duration of memory effects and, thus, reduces the model complexity. However, as pointed out in [31], reduced-rate sampling may increase the sensitivity to random noise components such as quantization noise and measurement noise, which may degrade the overall system performance. In a real application, we recommend choosing a slightly higher sampling rate than the minimum required. For example, we can choose four times of its chip rate. Nevertheless, a high dynamic range, normally over -65dB , must be guaranteed in the feedback data acquisition process in order to identify an accurate DPD model.

Direct conversion can be used in the feedback loop to transfer the RF signal to baseband. But the trend in current development is to mix the RF signal to an intermediate frequency (IF), and then digitize the IF signal and demodulate it to baseband in the digital domain, to avoid DC offset and distortion caused by analogue demodulators.

9.4.2 The transmit chain

In the transmit chain, the digital predistortion unit is inserted into the signal path before the modulator to preprocess the signal. Since the signal is nonlinearly processed by the DPD, the bandwidth of the predistorted signal after the DPD is much wider than the

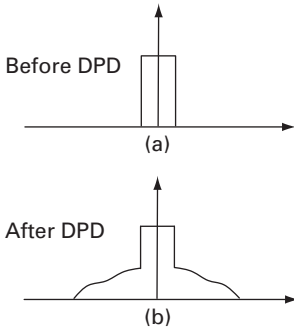


Figure 9.11 Spectrum expansion caused by digital predistortion.

bandwidth of the original input signal. For example, after a fifth-order predistortion, the bandwidth of the input signal will be five times that of the original signal, as shown in [Figure 9.11](#). Unlike the model extraction process where the signal is processed in the digital domain and no reconstruction is needed, and thus low sampling ADCs can be employed, here in the transmission chain, the digital transmission signal must be reconstructed before it is up-converted and transmitted through the analog PA. In order to correctly reconstruct this signal in the analog domain, we must sample it at a rate five times higher compared to that of the input, otherwise aliasing will occur. Therefore, high-speed digital-to-analog converters (DACs) must be employed in the transmit path. If under-sampling is applied during the model extraction process, parameter interpolation must be conducted to match the higher sampling rate in order to produce alias-free output, as described in [19]. A wideband modulation and up-conversion transmission chain is also required to allow the full bandwidth of the predistorted signal to reach the input of the PA, so that the nonlinear distortion can be completely compensated.

9.5 System characterization procedures

9.5.1 Excitation signal selection

In traditional RF tests, single or multiple periodical sinusoid waveforms are used for characterizing devices and circuits. These excitation waveforms are not sufficient for characterizing DPDs for RF PAs since the transfer function of a nonlinear system, e.g., a PA, not only depends on its internal features but also strongly depends on the characteristics of its input signal. Unless the statistical distribution of the sampling data for the DPD model extraction is very close to that of the real signal that the PA transmits, the parameters extracted will not be accurate enough for compensating for the nonlinear distortion induced by the PA. In general nonlinear system identification, to make sure the excitation source is sufficiently rich to excite all important properties of the system, most of the techniques use Gaussian white noise as the excitation signal. However, it is not convenient to use these approaches in a real system, because most commercial simulators

cannot handle a general Gaussian white noise source, and also practical power amplifiers are not excited by Gaussian noise signals. In modern wireless systems, arbitrary waveforms can be used, for example, time-domain W-CDMA or OFDM signals. These waveforms are widely employed in real wireless systems, and they can be easily generated from computer-aided design (CAD) tools or arbitrary waveform generators. These waveforms normally are composed of a wide range of amplitude and phase variations and include wideband frequency components, so that they can be considered as an equivalent band-limited white Gaussian process. And with proper configurations, the statistical distribution of these excitation waveforms can be made very close to that of the real signal that the PA transmits, to ensure the parameters extracted are close to the “true” values required.

9.5.2 Data acquisition

In a real DPD system, as shown in Figure 9.3, a complete feedback loop must be built to take a small fraction of output signal from the PA and transferred to digital baseband, and compare with the original desired signal, to extract the DPD parameters. In a laboratory exercise, time-domain stimulus-response measurement solutions can be used, e.g., the connected solution proposed by Agilent Technologies [32], shown in Figure 9.12. In this test system, the modulated data files are first created at baseband, downloaded to the arbitrary waveform generator, as complex I and Q signals, which are then fed to the IQ-modulator present in the signal generator (ESG). The signal generator produces the test signal to the power amplifier, our device under test (DUT). The output of the DUT is then down-converted and sampled by the Vector Signal Analyzer (VSA). The sampled input and output data are captured and finally used to extract DPD models for the power amplifier. Similar configurations are also used by many other companies and researchers.

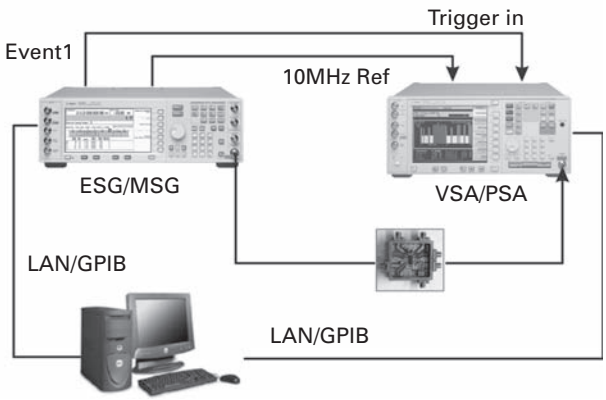


Figure 9.12 Connected solution for data acquisition.

9.5.3 Time alignment

The signal samples obtained from the feedback loop must be time aligned with the original signal before conducting DPD model extraction. Proper and accurate time alignment is critical for a DPD system since misalignment can dramatically degrade system performance. Many time-delay estimation methods have been proposed. A simple approach is to use cross-correlation and interpolation [33]: (i) use cross-correlation between the magnitudes of the input and output samples to find the integer delay between two signals, and adjust the samples to remove the delay; (ii) use sinc interpolation to resample the signals in a higher sampling rate; (iii) use cross-correlation again to estimate the fractional delay and adjust the samples to remove the delay; (iv) resample back to the original sampling rate.

9.5.4 Data normalization

Since superposition cannot be applied to a nonlinear system, generally the power level of the data for model extraction must be normalized to the same power level of the original input, so that the extracted parameters can be directly used in the DPD after model extraction; otherwise a nonlinear scaling must be applied to the parameters. This normalization process strongly depends on the selection of the expected gain. Generally, in a look-up table based predistorter, the maximum gain, which is the linear response of the PA, is chosen as the expected linearized gain. By pre-increasing the amplitudes at the higher power level, gain compression can be compensated. However, because the predistorter can only successfully correct distortion up to the full saturation level of the amplifier (after which point any increase in the input power does not produce an increase in output power), the maximum input allowed for the predistorter can only reach the point where the linear response intersects the saturation limit. Therefore, in this case, the original input and the predistorted input (the output of the DPD) have different peak values, as shown in Figure 9.13(a). This has the consequence that in parameter extraction, the input and the output data measured from the PA must be normalized to different power levels by different scaling factors, in order to correspond to the different peak values in the original and the predistorted inputs. This requires an extra calibration effort and also increases the complexity of the input power control in the final system implementation due to different levels of peak occurring before and after DPD.

However, if we choose the expected gain at the targeted maximum power level, for example, as shown in Figure 9.13 (b), the gain of the overall system k becomes

$$k = \frac{\text{Max}[|\tilde{y}(n)|]}{\text{Max}[|\tilde{x}(n)|]} \quad (9.14)$$

where $\tilde{x}(n)$ and $\tilde{y}(n)$ are the complex envelopes of the input and the output of the PA, respectively. Note that, here we assume that the output power of the PA increases monotonically with the input power. In other words, the peak power only occurs at the maximum input. In this case, both the original input and the predistorted input will reach the same maximum power level. Therefore, all input and output waveforms can be

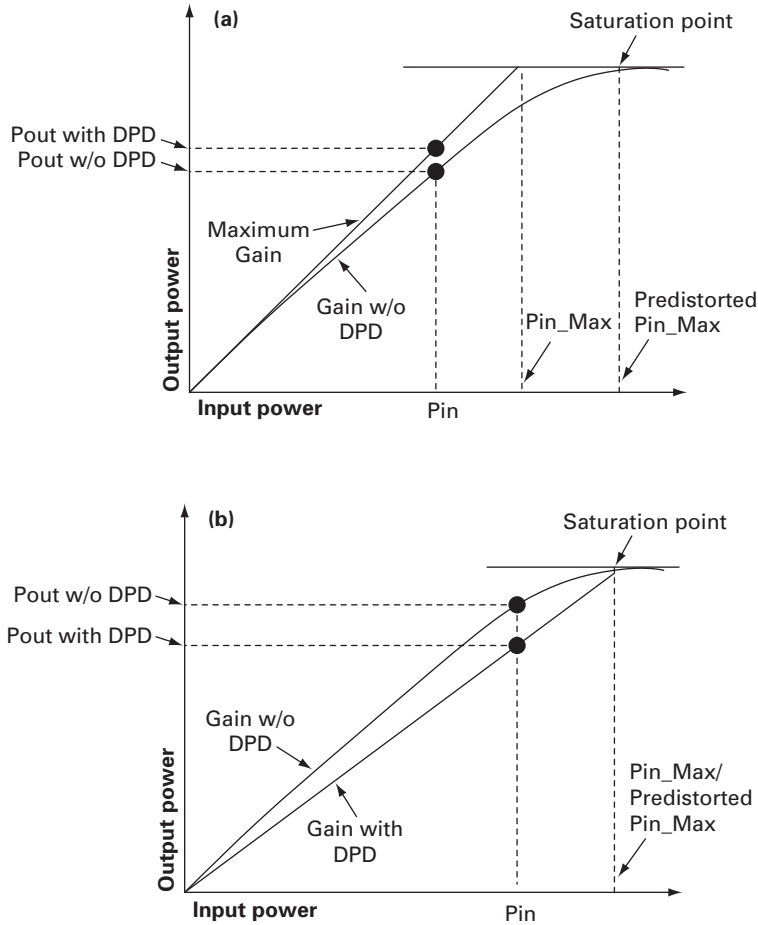


Figure 9.13 Expected gain selection. (a) Linearized for the maximum gain. (b) Linearized for the gain at the targeted maximum power level ([19] © 2008 IEEE).

normalized to unity. This significantly simplifies the model extraction because we can directly normalize the input and the output data measured from the PA by their respective peak values, and then use this data as the output and the input of the DPD function for parameter extraction. Selecting the gain at the maximum power level also makes the final system implementation easier because both the original input and the predistorted input can be normalized by the same scaling factor, which facilitates power control. One of the disadvantages of normalizing at the peak may mean that the average output power is reduced after digital predistortion [19], which must be calibrated and adjusted.

9.5.5 Model extraction algorithms

Most DPD models employed today are derived from polynomials or the Volterra series. Although they are used to represent nonlinear behavior of the DPD, one important

observation is that the output of these models is linear with respect to their coefficients. This leads that it is possible to extract the nonlinear DPD model in a direct way by using linear system identification algorithms, e.g., Least Squares (LS).

There are a number of methods that have been developed to find the coefficients of the nonlinear model using LS. The simplest way to formulate such a problem is to gather a set of input and output data, and form a parameter vector θ containing all of the unknown coefficients, e.g., c_{2k-1} in the polynomial model in Equation (9.1),

$$\theta = \begin{bmatrix} c_1 \\ c_3 \\ \vdots \\ c_P \end{bmatrix} \quad (9.15)$$

and define a matrix $\mathbf{X}$ including all of the product terms appearing in the input of the model for $n = 1, \dots, L$, where L is the total length of the data recorded,

$$\mathbf{X} = \begin{bmatrix} \tilde{x}(1) & |\tilde{x}(1)|^2 \tilde{x}(1) & \cdots & |\tilde{x}(1)|^{P-1} \tilde{x}(1) \\ \tilde{x}(2) & |\tilde{x}(2)|^2 \tilde{x}(2) & \cdots & |\tilde{x}(2)|^{P-1} \tilde{x}(2) \\ \vdots & \vdots & \cdots & \vdots \\ \tilde{x}(L) & |\tilde{x}(L)|^2 \tilde{x}(L) & \cdots & |\tilde{x}(L)|^{P-1} \tilde{x}(L) \end{bmatrix} \quad (9.16)$$

The expected output $\mathbf{y}$ is formed as,

$$\mathbf{y} = \begin{bmatrix} \tilde{y}(1) \\ \tilde{y}(2) \\ \vdots \\ \tilde{y}(L) \end{bmatrix} \quad (9.17)$$

Then, the polynomial model can be written as

$$\mathbf{y} = \mathbf{X}\theta \quad (9.18)$$

Using least squares, a solution for θ can be computed from,

$$\theta = (\mathbf{X}^H \mathbf{X})^{-1} \mathbf{X}^H \mathbf{y} \quad (9.19)$$

where $(\cdot)^H$ represents the Hermitian transpose. In Matlab, the “backslash (\)” operator will solve the equation above.

In many applications, adaptive estimations, e.g., least mean square (LMS) and recursive least squares (RLS), are also often employed. These algorithms are performed on a sample-by-sample basis, which minimizes storage requirements, but may experience slow convergence.

The procedures for DPD system characterization can be summarized as follows, as illustrated in Figure 9.14: (1) excitation signal selection; (2) data acquisition; (3) time alignment; (4) data normalization; (5) model extraction; (6) system implementation.

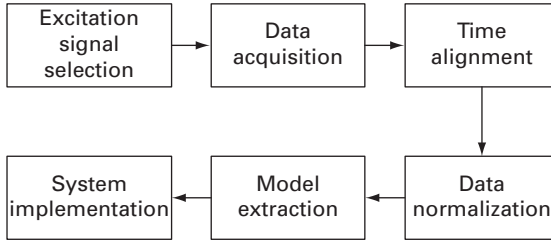


Figure 9.14 DPD system characterization procedures.

9.6 Implementation considerations

After several years development, there are many commercial digital predistortion products available in the market today. They are mostly implemented either in FPGA (field-programmable gate array) or ASIC (application-specific integrated circuit). The implementation of DPD systems mainly includes two parts: (i) the predistortion unit, which predistorts the input signal in the transmit chain; and (ii) the parameter extraction unit, which is used to initialize and update the DPD coefficients. Except for some real-time demanding systems, most RF power amplifiers in wireless transmitters are operated in quasi time-invariant mode, where the characteristics of the PAs only change slowly over the time, e.g., due to temperature, aging, etc. This means that the DPD coefficients are not required to be updated in real-time. The model extraction unit can thus be implemented in a software environment or in general DSP circuits in the system. However, the predistortion unit must be implemented in high-speed digital hardware circuits, e.g., FPGAs or ASICs, because real-time online signal processing is required. As price pressures become ever more demanding, the DPD structure must be implemented in transmitters with very low cost in order to make the system implementation feasible and competitive in the market.

As described earlier, most memory DPD models are analytical function based. These DPD models produce the predistorted signal by multiplying the original input signal with a number of pre-decided coefficients. This avoids one-to-one mapping in the traditional LUT approaches and thus requires much less memory storage. However, direct implementation of these DPD functions requires a large number of complex multipliers and adders to construct the nonlinear and memory terms. For example, 60 multipliers will be needed when $P=5$, $M=3$ in Equation (9.7) of the first-order DDR Volterra model. This number increases significantly when P and M become larger. Because the complex multiplier is one of the most complex and expensive components in digital hardware e.g., FPGA, and it often requires dedicated chip resources to implement [25], the large number of multipliers required dramatically increases system complexity and the final cost of the DPD system.

In order to reduce the use of complex multipliers, an LUT and analytical function combined approach can be adopted [26][27]. From Equation (9.7), we can see that the number of multipliers required mainly results from the generation of the power and the

memory terms, and their multiplications with the coefficients. However, if we replace $\tilde{y}(n)$ with $\tilde{u}(n)$ and re-write Equation (9.7) as

$$\begin{aligned}\tilde{u}(n) &= \tilde{u}_1(n) + \tilde{u}_2(n) \\ &= \sum_{i=0}^M C_{i,1}(n) \tilde{x}(n-i) + \sum_{i=1}^M C_{i,2}(n) \tilde{x}^2(n) \tilde{x}^*(n-i)\end{aligned}\quad (9.20)$$

where

$$C_{i,1}(n) = \sum_{k=0}^{\frac{p-1}{2}} \tilde{g}_{2k+1,1}(i) |\tilde{x}(n)|^{2k} \quad (9.21)$$

$$C_{i,2}(n) = \sum_{k=1}^{\frac{p-1}{2}} \tilde{g}_{2k+1,2}(i) |\tilde{x}(n)|^{2(k-1)} \quad (9.22)$$

we can find that, for a given memory tap i , $C_{i,1}(n)$ and $C_{i,2}(n)$ can be considered as a constant complex gain to $\tilde{x}(n-i)$ or $\tilde{x}^*(n-i)$, respectively, while $C_{i,1}(n)$ or $C_{i,2}(n)$ depends on polynomial power terms of $|\tilde{x}(n)|$ only, since $\tilde{g}_{2k+1,1}(i)$ or $\tilde{g}_{2k+1,2}(i)$ is a constant for each power term after model extraction or in time-invariant operations, when i is fixed.

In a real system, even if the individual samples of the signal, $\tilde{x}(n)$ and $\tilde{x}(n-i)$ are randomly generated, the range of the input signal is normally known in advance. This means that it is possible to calculate the high-order power terms of $|\tilde{x}(n)|$ and thus the values of $C_{i,1}(n)$ and $C_{i,2}(n)$ in advance, and store them in a table and retrieve them later. For example, if the maximum value of the input magnitude $|\tilde{x}(n)|$ is normalized to unity and the LUT size is 1024, the interval value for the LUT, Δ , will be $1/1024$. For $i=2$, $C_{2,1}(n)$ can be obtained by adding all power terms and multiplying with the coefficients for each given interval value, as shown in Figure 9.15. Using this approach, it no longer requires multiplication between input samples to produce the power terms, and all higher-order terms can be built in one table, no matter how many orders there are. These values can be retrieved from the LUTs later and multiplied with memory terms to produce the predistorted output. An implementation of the **first part** of the output, $\tilde{u}_1(n)$, is shown in Figure 9.16.

Unlike most conventional LUT-based DPDs where the gain values are often obtained directly through comparisons between input and output data, which often leads to low accuracy, in this approach, an analytical function-based model, i.e., the truncated Volterra series, is employed to represent the characteristics of the DPD and a least-square algorithm is utilized to extract the coefficients. This allows us to accurately control the nonlinear order and the memory length to be compensated for, and guarantees high accuracy in the model extraction. In summary, this proposed approach takes advantages of the best features of both analytical function-based and LUT-based solutions: (i) high accuracy in model extraction is maintained by using an analytical function-based DPD model; while (ii) complexity and cost in hardware implementation is reduced by the LUT.

$LUT \text{ for } C_{2,1}(n)$	
Address 1	$\tilde{g}_{1,1}(2) + \tilde{g}_{3,1}(2)\Delta^2 + \dots \tilde{g}_{p,1}(2)\Delta^{p-1}$
Address 2	$\tilde{g}_{1,1}(2) + \tilde{g}_{3,1}(2)(2\Delta)^2 + \dots \tilde{g}_{p,1}(2)(2\Delta)^{p-1}$
Address 3	$\tilde{g}_{1,1}(2) + \tilde{g}_{3,1}(2)(3\Delta)^2 + \dots \tilde{g}_{p,1}(2)(3\Delta)^{p-1}$
.	.
.	.
.	.
Address 1023	$\tilde{g}_{1,1}(2) + \tilde{g}_{3,1}(2)(1023\Delta)^2 + \dots \tilde{g}_{p,1}(2)(1023\Delta)^{p-1}$
Address 1024	$\tilde{g}_{1,1}(2) + \tilde{g}_{3,1}(2)(1024\Delta)^2 + \dots \tilde{g}_{p,1}(2)(1024\Delta)^{p-1}$

Figure 9.15 Example of LUT contents generation [26] © 2010 IEEE.

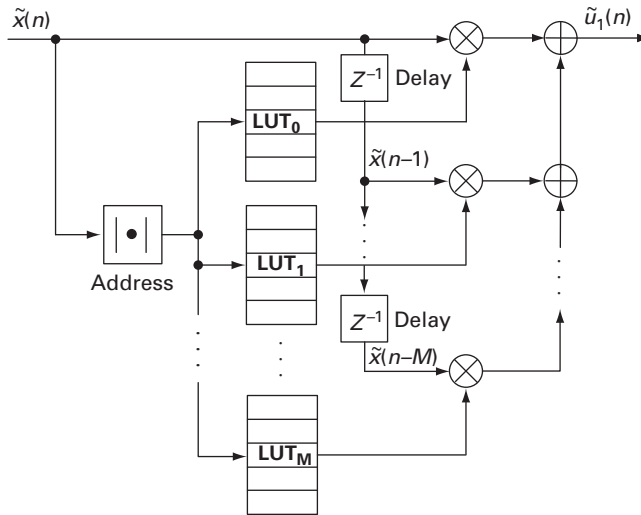


Figure 9.16 Example of LUT-function combined implementation [26] © 2010 IEEE.

Compared to direct analytical function-based DPDs or conventional LUT approaches, the complexity and cost of the implementation shown in Figure 9.16 have been dramatically reduced. However, this structure is not the simplest one because a large number of multipliers are still involved, and this number will increase with the memory length. To further reduce the complexity, time-division multiplexing can be employed as described below.

In Figure 9.16, multipliers are operated in parallel and each operation occupies a whole time slot. However, if we convert the parallel operation to a serial operation, for example, by using a multiplexer, all multiplication operations can be conducted in a serial way with one multiplier. The time sequence of these operations is shown in Figure 9.17.

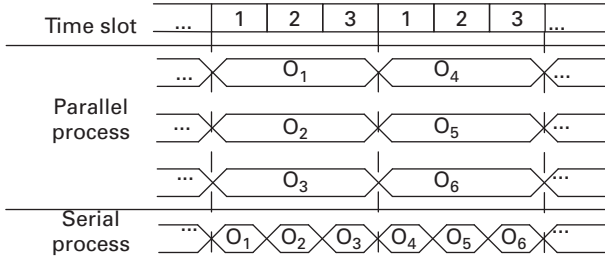


Figure 9.17 Time sequence view of multiplication operations [26] © 2010 IEEE.

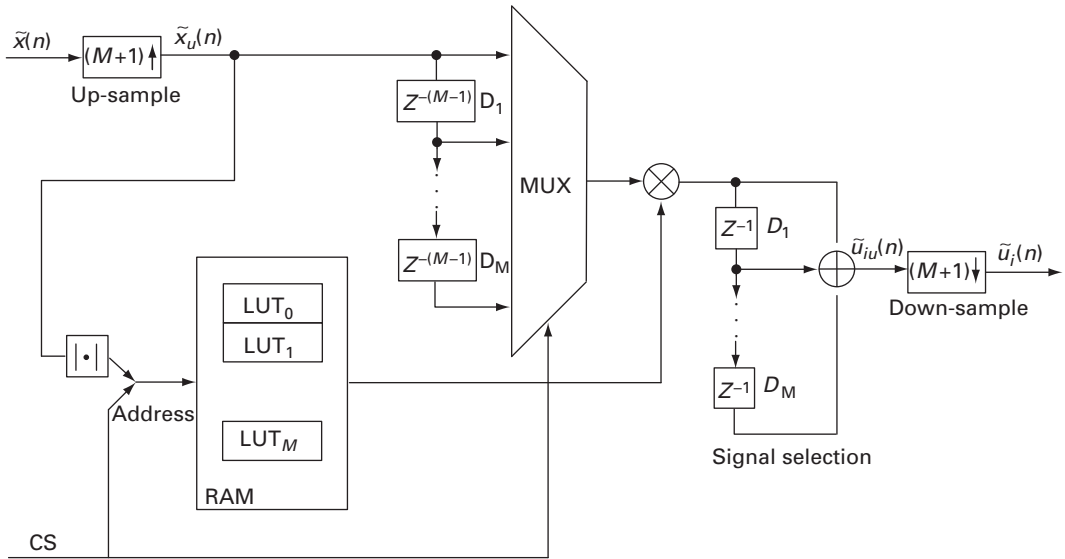


Figure 9.18 DPD implementation using multiplexing [26] © 2010 IEEE.

Transforming from parallel to serial processes, the time duration for each operation becomes shorter, which requires that the original data must be up-sampled and the signal processing needs to be conducted at a higher speed. The block diagram of this implementation of $\tilde{u}_i(n)$ is shown in Figure 9.18, where the original data are up-sampled first and a multiplexer is then used to redistribute the data for conducting the multiplication operations serially. At the end of operations, the output data are recombined and down-sampled back to the original sampling rate to produce the final output. The details of each composite block are given below:

- (1) *Input up-sampling*: this block increases the operation frequency by a factor of $(M+1)$ with data repeating units, i.e., the input data are repeated $(M+1)$ times at a higher sampling rate.

- (2) *Input magnitude calculation*: The CORDIC algorithm [6] is used to calculate the magnitude of input complex I/Q data, and this is used in conjunction with a selection signal CS (chip-selection) as the address for the LUTs.
- (3) *LUT mapping*: Separated LUTs are mapped into different parts of a standalone RAM (random access memory) on board so that indexing the contents is easier; the LUTs parts are indexed with the magnitude of respective input signals and the multiplexed CS signal. This saves the hardware cost for implementing addressing buses. Each LUT contains two parts, one for storing the real part of the coefficients and the other for the imaginary part.
- (4) *Data multiplexing*: This part contains a multiplexer with a selection signal CS and M delay-units, and each delay-unit delays $M + 1$ time cycles, This structure converts parallel multiplications using multiple multipliers to serial operations sharing only one multiplier.
- (5) *Complex multiplication*: The symbol $\otimes$ in Figure 9.18 represents the complex multiplication (CM), which is used to conduct multiplication between the output of MUX and the value retrieved from LUTs.
- (6) *Output accumulation*: This block is used to implement the accumulation (ACC) operation which is used to compensate for the memory effects distortion.
- (7) *Output down-sampling*: Not all the results obtained from the accumulator are useful for producing the final output of the DPD; some values are just the intermediate processing results which can be discarded. The output is only obtained from every $M + 1$ time cycles in the serial data sequences. Thus, a down-sampling block is employed to conduct such “selection.”

The second part of the DPD, $\tilde{u}_2(n)$, can be implemented using a similar structure as in Figure 9.18, with another two complex multipliers to include the $\tilde{x}^2(n)$ term. Combining the two parts, $\tilde{u}_1(n)$ and $\tilde{u}_2(n)$, together with proper delays to compensate processing delays between two branches, we can form the final structure of the DPD, as shown in Figure 9.19. In this structure, only four complex multipliers are required in total, and the number of multipliers used does not increase with the memory length M or the nonlinearity order P . This significantly reduces the cost of the system, especially for the systems with long-term memory effects or high-order nonlinearities.

For comparison, three DPD structures were implemented for the DDR Volterra DPD model in FPGA in [26]: (i) Scheme 1 (S1) is the direct structure, in which the DPD function is directly implemented by using the multipliers and adders; (ii) Scheme 2 (S2) is the structure as shown in Figure 9.16, in which the coefficients and nonlinear power terms are built into LUTs while the memory terms are implemented by using multipliers; (iii) Scheme 3 (S3) is the proposed new structure in which the coefficients and nonlinear power terms are built into LUTs while the memory terms are implemented by using shared multipliers with time-division multiplexing. The hardware utilization for different nonlinear orders and memory lengths for three schemes are shown in Table 9.1, where $\otimes$, $\oplus$, and LUT represents complex multipliers, two-input adders, and look-up-tables, respectively. It can be clearly seen that the cost of Scheme 1 is highest and its complexity

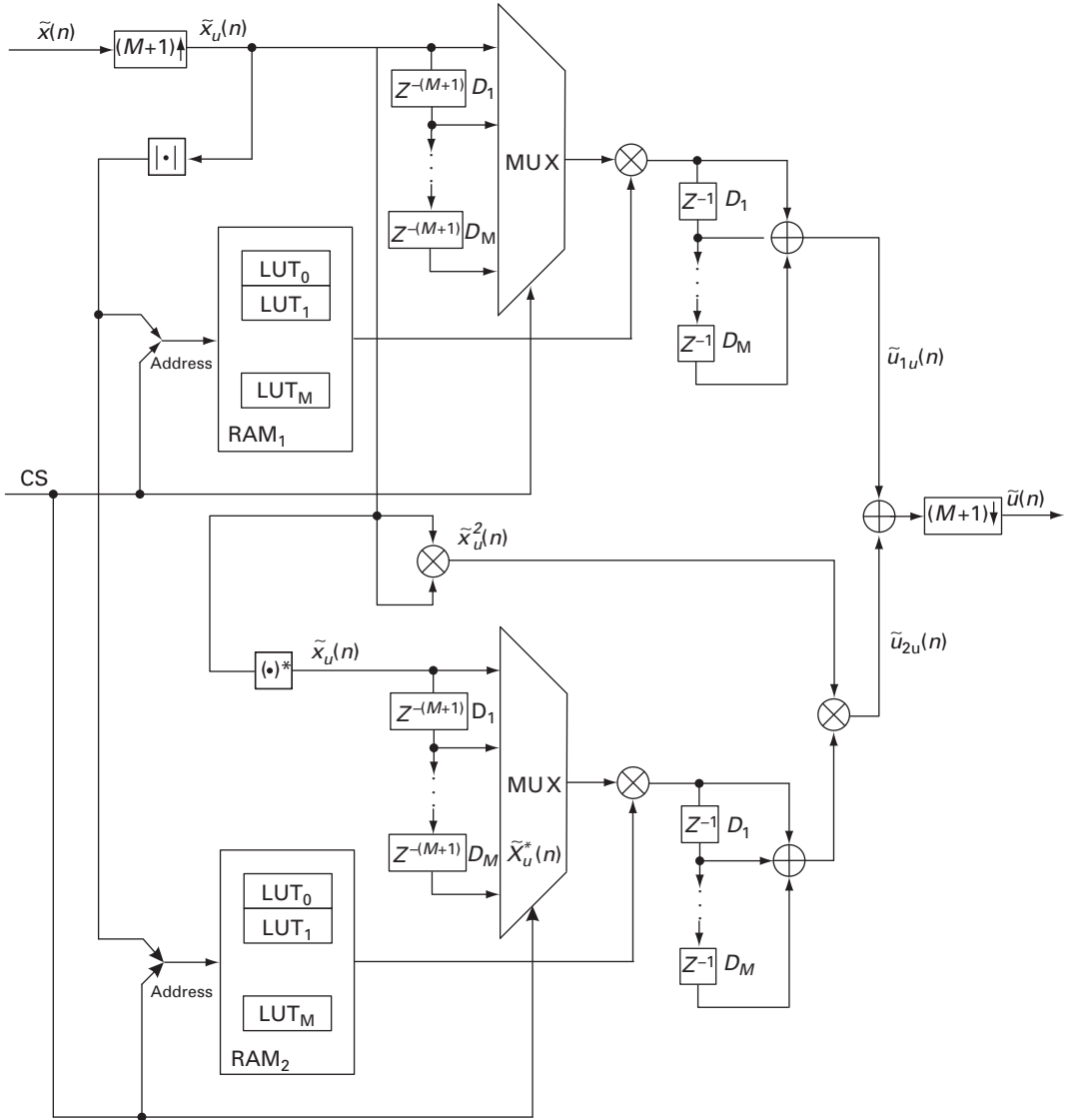


Figure 9.19 Proposed DPD implementation [26] © 2010 IEEE.

significantly increases with the nonlinear order or the memory length. The number of multipliers is reduced in Scheme 2, but it increases with the memory length. Scheme 3 uses much fewer resources and the number of multipliers is the same for all systems with different nonlinear orders and memory lengths, which provides significant benefit when the nonlinear order and the memory length increase.

The resource utilization difference between the three schemes becomes more evident when implementing this DPD on an FPGA board. Table 9.2 shows the Xilinx FPGA resource utilization in implementing the three schemes for the nonlinear order $P=7$ and

Table 9.1 Hardware utilization comparison [26] © 2010 IEEE.

	$P=5, M=2$			$P=7, M=2$			$P=7, M=4$		
	$\otimes$	$\oplus$	LUT	$\otimes$	$\oplus$	LUT	$\otimes$	$\oplus$	LUT
S1	27	12	0	39	17	0	65	31	0
S2	8	4	5	8	4	5	14	8	9
S3	4	4	5	4	4	5	4	8	9

Table 9.2 FPGA resource utilization comparison [26] © 2010 IEEE.

	DSP48	Slice flip flops	Slice	4-input LUTs	RAM 16s
S1	260	8132	5934	9512	17
S2	56	2027	1399	1388	18
S3	16	1282	1220	1418	18

memory length $M=4$. It can be seen that the proposed scheme use much fewer resources on the board, especially the DSP48 slices, which are dedicated high-performance signal processing blocks.

By employing multiplexers and conducting serial operations, the system clock in the proposed scheme must be M times higher than those in the parallel processing, such as in Scheme 1 and 2. This does not create a problem for digital circuits since the clock rate of digital chips is continuously pushing higher, for example, 600 MHz signal processing capability is currently available in Xilinx Virtex-6 [25]. The price of such chips is continuously falling. However, although significant improvements have been made over the last decades, the complex multipliers are still complicated to implement, requiring dedicated DSP slices and occupying a large area of silicon chip. The number of these special designed DSP units, e.g., DSP48, is limited on FPGA boards. Therefore, by employing the LUT and multiplexing combined approach, the implementation structure of the DPD is much simpler and the overall cost will be much lower compared to conventional approaches. The benefit of the proposed approach will become more evident when the FPGA implementation is transformed to stand alone integrated circuits or ASICs for mass production.

9.7 Combination with crest factor reduction

Crest factor (CF) is the ratio of the peak amplitude to the root mean square (RMS) value of the waveform, which is also called peak-to-average ratio (PAR) or peak-to-average power ratio (PAPR). In the discrete time domain, CF is often calculated from the peak power divided by the time-averaged power level, i.e., the expectation of the data sequence, and it is often expressed in decibels (dB). In a wideband wireless system, the transmit signal often has a very high crest factor. For example, the CF of an OFDM-

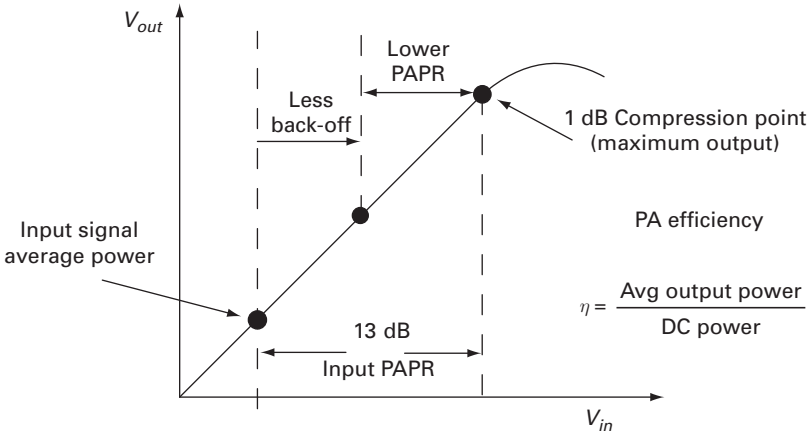


Figure 9.20 Effects of PAPR.

based LTE signal often exceeds 10 dB. This occurs due to a large number of independent subcarriers with random phases added together at the modulator.

Although, with digital predistortion, the RF PA can be driven to the saturation region without introducing severe distortion, this linearity can only be guaranteed under the condition that the peak of the signal does not exceed the saturation point, e.g., P-1dB compression. In other words, DPD only extends the linear range of the PA to its saturation, but the maximum power that the PA can deliver is still limited by the signal peak value. This means that, for non-constant envelope signals with a high PAPR, the PA average output power must be backed off to a very low level in order to avoid hard clipping in the peaks. For example, the input of the PA must be backed off by 13 dB from the maximum input point for a signal with 13 dB PAPR, as shown in Figure 9.20. This consequently results in very low power efficiency.

Crest factor reduction (CFR) techniques are proposed to reduce the peaks of the transmit signal to a satisfactory level before digital predistortion to allow the PA to be operated at higher average power to achieve higher power efficiency. For instance, by decreasing the PAPR from 13 dB to 7 dB, the PA can be operated at 6 dB less back-off while still keeping the peaks below the saturation point. This significantly increases the average power of the PA and thus increases the power efficiency. A number of CFR techniques have been proposed in the past few years [34]. In this section, we briefly describe some of the widely used approaches for wideband multi-carrier signals.

9.7.1 Clipping and filtering

The simplest technique for crest factor reduction is direct amplitude clipping, which limits the peak envelope of the input signal to a predetermined value. It can be expressed as

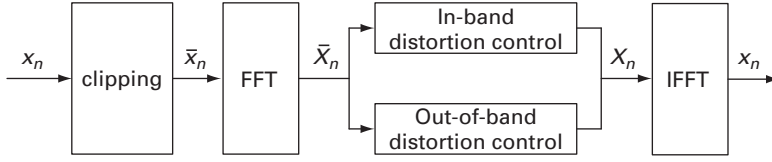


Figure 9.21 Constrained clipping [37] © 2007 IEEE.

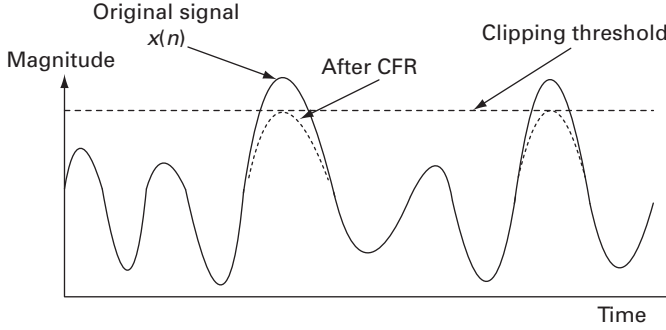


Figure 9.22 Peak windowing.

$$\tilde{x}_d(n) = \begin{cases} \tilde{x}(n) & |\tilde{x}(n)| < R \\ R \cdot e^{j\theta(n)} & |\tilde{x}(n)| \geq R \end{cases} \quad (9.23)$$

where $\tilde{x}(n)$ is the original sample of the time-domain complex envelope signal, $\tilde{x}_d(n)$ is the sample after amplitude clipping, R is the predefined magnitude threshold, $\theta(n)$ is the phase of $\tilde{x}(n)$, and $|\tilde{x}(n)|$ returns the magnitude of $\tilde{x}(n)$. Direct clipping can effectively reduce high peaks in the signal, but it introduces significant in-band and out-of-band distortion. Filtering can be conducted after clipping to reduce out-of-band distortion but may cause some peak re-growth since the signal after filtering may exceed the clipping level again at some points. Generally, clipping and filtering must be repeatedly conducted in several iterations to bring the PAPR to a desired level [35][36]. In-band distortion cannot be reduced by filtering, which may result in a high bit error rate. To resolve this problem, a constrained clipping approach was proposed to limit both in-band and out-of-band distortion in [37], as shown in Figure 9.21. In this approach, both out-of-band spectrum expansion and in-band EVM are limited to a certain level after clipping, so that it can ensure that the processed signal satisfies the specification requirement after the PAPR reduction.

9.7.2 Peak windowing

Another effective CFR technique is the peak windowing approach. Instead of filtering the clipped signal in the frequency domain, this technique multiplies large signal peaks by a non-rectangular window to smooth the sharp signal edges at each clipping point, and thus minimizes the distortion caused by clipping, as shown in Figure 9.22. This can be expressed with a scaling function as below,

$$\tilde{x}_d(n) = c_{pw}(n)\tilde{x}(n) \quad (9.24)$$

$$c_{pw}(n) = 1 - \sum_{i=-\infty}^{\infty} a(i)w(n - n_i) \quad (9.25)$$

where $w(n)$ is a common symmetric window function, $a(n)$ is the scaling coefficient. The criterion to choose the coefficients $a(n)$ is that the resultant envelope of processed signals must not exceed the given threshold. Any window function with good spectral characteristics can be adopted in Equation (9.25), such as Hamming, Gaussian, or Kaiser Window [38]. Peak windowing works in the time domain and uses less logical resource than the frequency domain approaches, such as clipping and filtering.

9.7.3 Partial transmit sequencing

In the partial transmit sequencing (PTS) technique, an input data block comprising N symbols is partitioned into M disjoint sub-blocks. The subcarriers in each sub-block are weighted by a phase factor for that sub-block after IDFT (inverse discrete Fourier transform). The phase factors are selected such that the PAPR of the combined signal is minimized. A block diagram of PTS is shown in Figure 9.23 [39]. In general, the selection of the phase factors is limited to a set with a finite number of elements to reduce the search complexity. The search complexity increases exponentially with the number of sub-blocks. Information about the phase factors selected must be transmitted to the receiver in order to decode the signal accurately.

9.7.4 Selected mapping

In the selected mapping (SLM) technique, the transmitter generates a set of different candidate data blocks, all representing the same information as the original data block. The most favorable data block is then selected for transmission [40][41]. A block diagram of the SLM technique is shown in Figure 9.24. Each data block is multiplied by M different phase sequences, each of length N , before IDFT. The one with the lowest PAPR is selected for transmission. Information about the selected phase sequence should

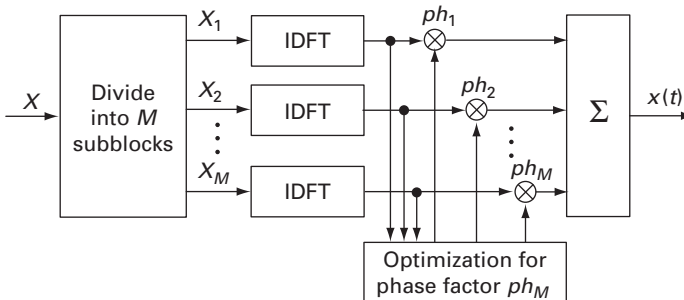


Figure 9.23 Partial transmit sequencing (after [39]).

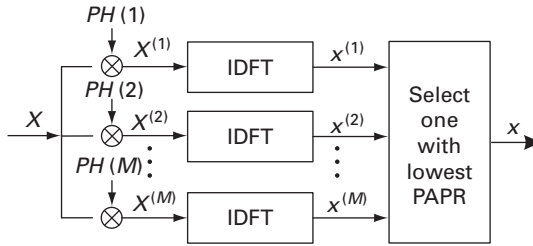


Figure 9.24 Selected mapping (after [40]).

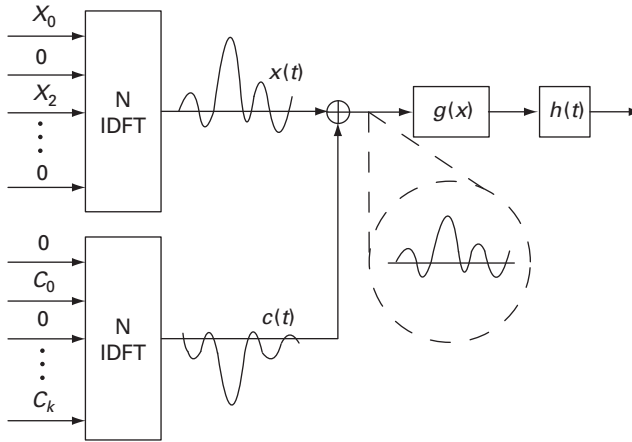


Figure 9.25 Tone reservation.

be transmitted to the receiver as side information. At the receiver, the reverse operation is performed to recover the original data block. This approach is applicable with all types of modulation and any number of subcarriers. The amount of PAPR reduction for SLM depends on the number of phase sequences M and the design of the phase sequences.

9.7.5 Tone reservation

The tone reservation (TR) algorithm was developed by Tellado [42], whereby a small number of subcarriers (tones) are reserved in the frequency domain to create a signal in the time domain which can cancel the high peaks in the information-carrying signals at the transmitter, as shown in Figure 9.25. In this system, the data-bearing vector X , and the reserved tone vector C in the frequency domain lie in disjoint frequency subspaces, namely, they cannot both be non-zero at a given tone, which can be expressed as

$$X_k + C_k = \begin{cases} C_k & k \in L \\ X_k & k \in L^c \end{cases} \quad (9.26)$$

where X_k and C_k is the component signal of the data-bearing vector X and the reserved tone vector C , respectively. The subset of reserved tones can be denoted by $L = [i_0, \dots, i_L]$

with $L \ll N$, where N represents the set of all tones in the multi-carrier symbol; L^c is the complement of L in N and represents information carriers. In the time domain, the addition of these reserved tones c to a data-bearing signal x produces a new composite signal

$$\bar{x}[n] = x[n] + c[n] = IDFT(X_k + C_k) \tag{9.27}$$

Through properly optimizing the reserved vector C and thus its time domain waveform c , the peak of the composite signal can be made smaller than that of the original data-bearing signal. Since the subcarriers are orthogonal and symbol demodulation is performed in the frequency domain on a tone-by-tone basis, the reserved sub-channels can be discarded at the receiver, and only the data-bearing sub-channels are used to determine the transmitted bit stream. This approach can reduce the PAPR of the OFDM signals without introducing any additional distortions to the information data and does not require side information. However, TR suffers a high computational cost due to the difficulties of finding an effective cancellation signal in the time domain from only a small number of reserved tones in the frequency domain.

9.7.6 Tone injection

The idea of tone injection (TI) is to extend the original signal constellation cyclically so that each of the symbol points in the original basic constellation can be mapped into several equivalent points in the expanded constellation. For instance, the extended constellation for QPSK is illustrated in Figure 9.26 [43], where the original constellation is duplicated cyclically onto its surrounding region in the complex plane. The resulting

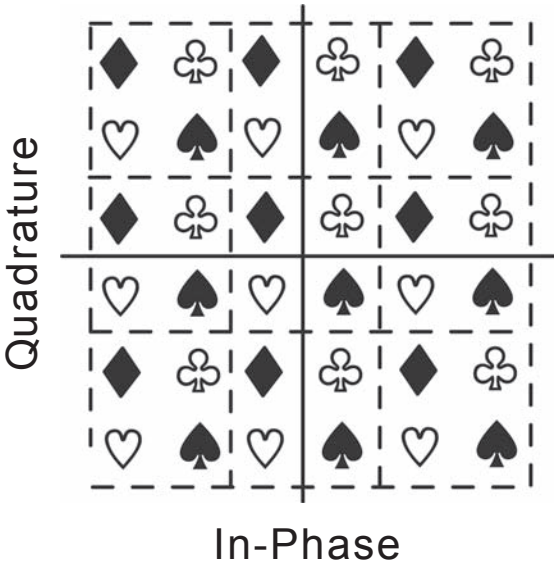


Figure 9.26 The cyclically extended QPSK constellation diagram.

constellation includes nine QPSK sub-constellations in which the equivalent points are spaced by the extension-size D along the real and/or imaginary axes. This extra freedom allows us to reduce the PAPR of the transmit signal in the time domain by modifying the signal in the frequency domain through injecting tones into the symbol constellation. For example, the modified transmit signal after tone injection is

$$\bar{x}[n] = x[n] + c[n] = \text{IDFT}(X_k + C_k) \quad (9.28)$$

where the extension vectors C_k are defined as

$$C_k = p_k D + jq_k D \quad (9.29)$$

Since it is sufficient to use the first set of neighboring sub-constellations, p_k and q_k normally take values from the set $[-1, 0, 1]$, where the horizontal and vertical translations between sub-constellations are D . This approach can effectively reduce the PAPR for OFDM signals and the only addition to the standard receiver is a modulo- D operation after the symbol decision. However, finding the optimal solution for the values of p_k and q_k to obtain the lowest PAPR for $x[n]$ becomes an integer programming problem that is known to be an NP-hard problem [42]. The TI technique may also result in a power increase in the transmit signal due to the injected signal.

9.8 Conclusions and future outlook

In this chapter, we have given a comprehensive overview of digital predistortion techniques, from operational principles, to behavioral modeling and parameter extraction, and from system characterization procedures to hardware implementations. Combination with crest factor reduction was also discussed. It can be seen that, by digitally pre-processing the input signal before amplification, digital predistortion techniques allow analog RF power amplifiers to be operated at a higher drive level for optimal efficiency without introducing severe distortion, which can significantly improve linearity and power efficiency of the system. This also allows one to employ cheaper low-power amplifiers to replace more expensive high-power amplifiers in the transmitters, which reduces development and operation costs of the wireless systems.

After many years development, digital predistortion has become a mature technique, and it has been widely employed in modern wireless communication systems, especially in high-power base-stations in cellular mobile networks. However, with the increasing demand for higher data rates in future wireless systems, the current digital predistortion concept will face severe challenges, due to its inherent high bandwidth requirement in the transmission and feedback chains. For example, 100 MHz instantaneous modulation bandwidth will be required in LTE-Advanced systems [44]. For that case, if we compensate for up to the fifth-order distortion, a 500 MHz bandwidth of predistortion system must be employed, which is very difficult to implement in a real system. Furthermore, current DPDs are normally employed to compensate for nonlinearities induced by PAs operated in a certain frequency band and with a fixed modulation bandwidth. In LTE-Advanced, carrier aggregation techniques will be employed, in which two or more so-

called component carriers (CCs) are coupled together in order to support wider transmission bandwidths. In these systems, spectrum deployment may be either contiguous with adjacent component carriers or non-contiguous with non-adjacent component carriers. Data may be sent either in the same frequency band or different frequency bands. This brings severe challenges for DPD design. For instance, if non-contiguous bandwidths in different frequency bands are used, how can the DPD simultaneously handle the signals in the separate bands? Further research must be conducted and new linearization solutions must be sought in these ultra wideband systems.

References

- [1] P. B. Kennington, *High Linearity RF Amplifier Design*. Norwood, MA: Artech House, 2000.
- [2] K. Mekechuk, W.-J. Kim, S. P. Stapleton, "Linearizing power amplifiers using digital predistortion, EDA tools and test hardware," *High Frequency Electronics*, pp. 18–27, April 2004.
- [3] Y. Nagata, "Linear amplification technique for digital mobile communications," *In Proceedings of the 39th IEEE Vehicular Technology Conference*, pp. 159–164, May 1989.
- [4] M. Faulkner and M. Johansson, "Adaptive linearization using predistortion – experimental results," *IEEE Trans. Veh. Technol.*, vol. **43**, no. 2, pp. 323–332, May 1994.
- [5] J. K. Cavers, "Amplifier linearization using a digital predistorter with fast adaptation and low memory requirements," *IEEE Trans. Veh. Technol.*, vol. **39**, no. 4, pp. 374–382, Nov. 1990.
- [6] Y. H. Hu, "CORDIC-based VLSI architectures for digital signal processing," *IEEE Signal Processing Magazine*, pp. 17–34, July 1992.
- [7] J. C. Pedro, and S. A. Maas, "A comparative overview of microwave and wireless power-amplifier behavioral modeling approaches," *IEEE Trans. Microw. Theory Tech.*, vol. **53**, no. 4, pp. 1150–1163, April 2005.
- [8] F. Ghannouchi, O. Hammi, "Behavioral modeling and predistortion," *IEEE Microwave Magazine*, vol. **10**, pp. 52–64, Dec. 2009.
- [9] A. Tehrani H. Cao, S. Afsardoost, *et al.*, "A comparative analysis of the complexity/accuracy tradeoff in power amplifier behavioral models", *IEEE Trans. Microw. Theory Tech.*, vol. **58**, no. 6, pp. 1510–1520, June 2010.
- [10] M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems*, reprint edn. Malabar, FL: Krieger, 2006.
- [11] V. J. Mathews and G. L. Sicuranza, *Polynomial Signal Processing*, New York: Wiley, 2000.
- [12] S. Benedetto, E. Biglieri, and R. Daffara, "Modeling and performance evaluation of nonlinear satellite links – a Volterra series approach," *IEEE Trans. Aero. Electronic Syst.*, vol. **15**, no. 4, pp. 494–506, April 1979.
- [13] J. Kim and K. Konstantinou, "Digital predistortion of wideband signals based on power amplifier model with memory," *Electron. Lett.*, vol. **37**, no. 23, pp. 1417–1418, Nov. 2001.
- [14] A. Zhu, and T. J. Brazil, "Behavioral modelling of RF power amplifiers based on pruned Volterra series," *IEEE Microw. Wireless Compon. Lett.*, vol. **14**, pp. 563–565, Dec. 2004.
- [15] D. R. Morgan, Z. Ma, J. Kim, M. G. Zierdt, and J. Pastalan, "A generalized memory polynomial model for digital predistortion of RF power amplifiers," *IEEE Trans. Signal Process.*, vol. **54**, no. 10, pp. 3852–3860, Oct. 2006.

- [16] F. Filicori, and G. Vannini, "Mathematical approach to large-signal modeling of electron devices," *Electron. Lett.*, vol. **27**, no. 4, pp. 357–359, 1991.
- [17] D. Mirri *et al.*, "A modified Volterra series approach for nonlinear dynamic systems modeling," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. **49**, no. 8, pp. 1118–1128, Aug. 2002.
- [18] A. Zhu, J. C. Pedro, and T. J. Brazil, "Dynamic deviation reduction-based Volterra behavioral modeling of RF power amplifiers," *IEEE Trans. Microw. Theory Tech.*, vol. **54**, no. 12, pp. 4323–4332, Dec. 2006.
- [19] A. Zhu, P. J. Draxler, J. J. Yan *et al.*, "Open-loop digital predistorter for RF power amplifiers using dynamic deviation reduction-based Volterra series," *IEEE Trans. Microw. Theory Tech.*, vol. **56**, no. 7, pp. 1524–1534, July 2008.
- [20] S. C. Cripps, *RF Power Amplifiers for Wireless Communications*, 2nd edn., Norwood, MA: Artech House, 2006.
- [21] A. Zhu, P. J. Draxler, H. Chin, *et al.*, "Digital predistortion for envelope-tracking power amplifiers using decomposed piecewise Volterra series," *IEEE Trans. Microw. Theory Tech.*, vol. **56**, no. 10, pp. 2237–2247, Oct. 2008.
- [22] C. Eun and E. J. Powers, "A new Volterra predistorter based on the indirect learning architecture," *IEEE Trans. Signal Process.*, vol. **45**, no. 1, pp. 223–227, Jan. 1997.
- [23] L. Ding, G. T. Zhou, D. R. Morgan, *et al.*, "A robust digital baseband predistorter constructed using memory polynomials," *IEEE Trans. Commun.*, vol. **52**, no. 1, pp. 159–165, Jan. 2004.
- [24] R. N. Braithwaite, "An improved Doherty amplifier using cascaded digital predistortion and digital gate voltage enhancement," *IEEE Trans. Microw. Theory Tech.*, vol. **57**, no. 12, pp. 3118–3126, Dec. 2009.
- [25] [online] www.xilinx.com.
- [26] L. Guan and A. Zhu, "Low-cost FPGA implementation of Volterra series-based digital predistorter for RF power amplifiers," *IEEE Trans. Microw. Theory Tech.*, vol. **58**, no. 4, pp. 866–872, April 2010.
- [27] A. Cesari and P. L. Gilabert, "A FPGA based digital predistorter for RF amplifiers with memory effects," in *Proc. 2nd European. Microw. Conf. Integrated Circuits*, Munich Germany, pp. 135–138, Oct. 2007.
- [28] Y. M. Zhu, "Generalized sampling theorem," *IEEE Trans. on Circuits Syst. II, Analog Digit. Signal Process.*, vol. **39**, no. 8, pp. 587–588, Aug. 1992.
- [29] W. Frank, "Sampling requirements for Volterra system identification," *IEEE Signal Process. Lett.*, vol. **3**, no. 9, pp. 266–268, Sep. 1996.
- [30] D. Wisell and P. Handel, "Implementation considerations on the use of Zhu's general sampling theorem for characterization of power amplifiers," in *Proc. IEEE Instrum. Meas. Technol. Conf.*, Warsaw, Poland, pp. 1–4, May 2007.
- [31] J. Tsimbinos and K. V. Lever, "Input Nyquist sampling suffices to identify and compensate nonlinear systems," *IEEE Trans. Signal Process.*, vol. **46**, no. 10, pp. 2833–2837, Oct. 1998.
- [32] "Connected simulation and test solutions using the Advanced Design System" Agilent Technol., Palo Alto, CA, Application Notes 1394, 2000.
- [33] H. Cao, H. Nemati, A. Tehrani, *et al.*, "Linearization of efficiency-optimized dynamic load modulation transmitter architectures," *IEEE Trans. Microw. Theory Tech.*, vol. **58**, no. 4, pp. 873–881, April 2010.
- [34] S. H. Han and J. H. Lee, "An overview of peak-to-average power ratio reduction techniques for multicarrier transmission," *IEEE Wireless Comm.*, vol. **12**, no. 2, pp. 56–65, April 2005.

- [35] X. Li and L. J. Cimini, Jr., "Effect of Clipping and Filtering on the Performance of OFDM," *IEEE Commun. Lett.*, vol. **2**, no. 5, May 1998, pp. 131–33.
- [36] J. Armstrong, "Peak-to-average power reduction for OFDM by repeated clipping and frequency domain filtering," *Elect. Lett.*, vol. **38**, no. 8, Feb. 2002, pp. 246–47.
- [37] C. Zhao, R. J. Baxley, G. T. Zhou, D. Boppana, and J. S. Kenney, "Constrained clipping for crest factor reduction in multiple-user OFDM," in *Proc. IEEE Radio and Wireless Symposium*, pp. 341–344, Jan. 2007.
- [38] O. Vaananen, J. Vankka, and K. Halonen, "Simple algorithm for peak windowing and its application in GSM, EDGE and WCDMA systems," *IEE Communication Proceedings*, vol. **152**, no. 3, pp. 357–362, June 2005.
- [39] S. H. Müller and J. B. Huber, "OFDM with Reduced Peak-to-Average Power Ratio by Optimum Combination of Partial Transmit Sequences," *Elect. Lett.*, vol. **33**, no. 5, Feb. 1997, pp. 368–69.
- [40] S. H. Müller and J. B. Huber, "A Comparison of Peak Power Reduction Schemes for OFDM," *Proc. IEEE GLOBECOM '97*, Phoenix, AZ, Nov. 1997, pp. 1–5.
- [41] R. W. Bäuml, R. F. H. Fisher, and J. B. Huber, "Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping," *Elect. Lett.*, vol. **32**, no. 22, Oct. 1996, pp. 2056–57.
- [42] J. Tellado, Peak to Average Power Reduction for Multicarrier Modulation, Ph.D. dissertation, Stanford University, 2000.
- [43] C. Tuna and D. L. Jones, "Tone injection with aggressive clipping projection for OFDM PAPR reduction," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 3278–3282, March 2010.
- [44] [online] <http://lteworld.org>.

10 Adaptive digital baseband predistortion: design and implementation

Hua Qian and G. Tong Zhou

10.1 Introduction

The radio frequency (RF) power amplifier (PA) is the most power consuming element in a wireless transmission system, and can account for more than 50 percent of the total power consumed by the transmitter [1]. Improving the PA efficiency saves energy and drives down the overall system costs. A study described in [2, p. 13] provides a compelling reason for PA linearization: application of PA linearization technologies can yield annual savings of millions of dollars for a typical network service provider.

Efficient PAs are usually nonlinear. Nonlinearity generates both in-band distortion and out-of-band interference, which manifest in terms of transmitter error vector magnitude (EVM) degradation and spectral regrowth. In wireless communication systems, many signal formats, such as Code Division Multiple Access (CDMA) and Orthogonal Frequency Division Multiplexing (OFDM) transmission, have been introduced to improve spectrum efficiency and data rate. However, these non-constant-envelope signals are not power efficient in the presence of nonlinear PAs as large back-offs are needed for linear transmission. In most commercial wireless communication systems, PAs are still the dominant source of signal quality degradation since the PAs are usually biased in a mildly nonlinear region to gain a reasonable amount of efficiency.

Power amplifier linearization techniques can be sought to improve the linearity of the PA while still achieving a certain degree of power efficiency. Ideally, after linearization, the linear region of the PA is extended close to the saturation point of the PA, allowing a smaller power back-off than the original PA, which in turn improves the PA power efficiency. Power amplifier linearization has been widely studied. Available techniques include feedback, feedforward, and predistortion [2–5].

Feedback is widely used in control systems. In the context of PA linearization, if we can extract the error signal as the difference between the (scaled) input and output to the PA, a negative feedback system can be constructed by feeding the error signal back to the PA input [6, 7]. The gain-bandwidth trade-off and the instability problem, however, limit the feedback linearization performance of RF PAs, especially for wide-band applications.

Feedforward linearization reproduces the error signal and compensates for it at the PA output [8]. In contrast, with feedback linearization, the compensation takes place at the

PA input. In feedforward linearization, the error amplifier, however, also consumes power and thus the overall system power efficiency improvement may not be significant.

In predistortion linearization, a nonlinear block (predistorter, PD), which has inverse characteristic of the PA, is inserted into the data path. The signal first passes through the predistorter. After subsequent distortion introduced by the PA, an approximately linearly amplified version of the original signal can be obtained. Predistortion can be performed in RF, IF, as well as in baseband. The predistorter can be static or adaptive. In this chapter, the focus is on adaptive digital baseband predistortion (DPD) techniques as they offer a good compromise among complexity, cost, and linearization performance.

A few other methods that cannot be characterized as PA linearization methods can also help to improve the PA efficiency. These methods include dynamic biasing, envelope elimination and restoration (EER), and linear amplification with nonlinear components (LINC), etc. [2–5]. These methods all have their own design challenges.

The rest of this chapter is organized as follows: In [Section 10.2](#), the PA nonlinear effects and their system level behavior modeling are discussed in detail. In [Section 10.3](#), predistorter modeling and model parameter estimation are described. Digital predistortion implementation considerations are covered in [Section 10.4](#). [Section 10.5](#) concludes this chapter and suggests future research directions.

10.2

Power amplifier nonlinearity behavior and modeling

The basic function of an RF PA is to linearly amplify the input signal. The efficiency and linearity trade-offs of a few well-known PA classes are summarized in [Table 10.1](#) [1, 8]. All PA designs in this table, except for the class A PA, are nonlinear. Moreover, even a class A PA becomes nonlinear when the input signal approaches the saturation region. In this section, we will study how to measure and quantify the PA nonlinear effects.

10.2.1

Quantifying PA nonlinearity

Amplitude-to-amplitude (AM/AM) and amplitude-to-phase (AM/PM) responses are widely used by RF engineers to represent the PA nonlinear effects. The AM/AM response is the relationship between the input power (or amplitude) and the output

Table 10.1 Performance comparison of various operating class power amplifiers.

Class	Maximum efficiency	Linearity
A	50%	Good
AB	50% – 78.5%	Moderate
B	78.5%	Poor
C	100%	Poor
D	100%	Poor
E	100%	Poor
F	100%	Poor

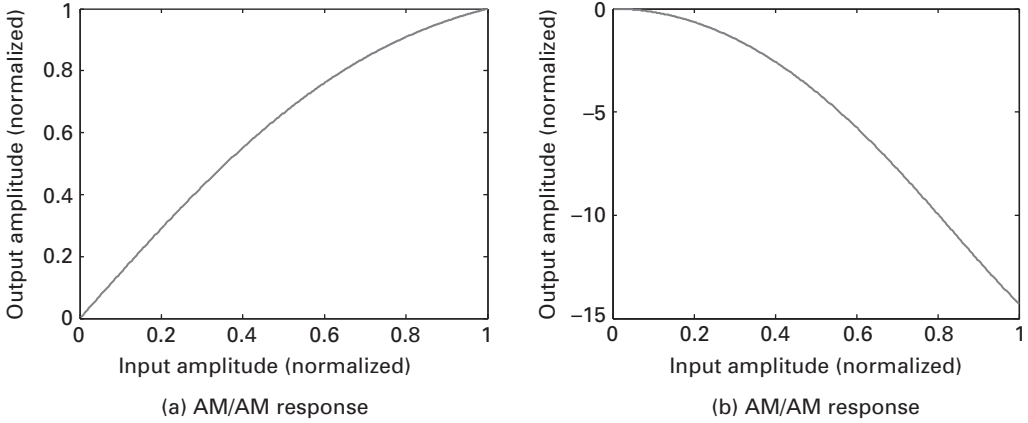


Figure 10.1 AM/AM and AM/PM response of a 2.4GHz PA.

power (or amplitude) [2],¹ and the AM/PM response is used to characterize the dependency of output phase deviation on the input power (or amplitude).

The AM/AM and AM/PM responses can be obtained by sweeping the signal power of a sinusoidal signal on the carrier frequency. The AM/AM response and AM/PM response of a representative 2.4GHz class AB PA used in mobile devices is shown in Figure 10.1.

Similarly, the two-tone test, which generates intermodulation (IMD) components, provides additional information regarding the nonlinearity distortion. With an equal amplitude two-tone signal, the n th-order IMD intercept point, IP_n , is defined as the signal power level where the output power at the fundamental frequency and that at the n th-order IMD product are equal [9]. As a convention, IP_n usually refers to the input, which can also be written as IIP_n .

From a mathematical point of view, we would like to further discuss how the above quantities relate to PA nonlinearities. In theory, a weak memoryless nonlinearity over a fixed interval can be approximated by a power series, i.e.,

$$\tilde{y}(t) = \sum_{l=1}^L \tilde{a}_l \tilde{x}^l(t), \quad (10.1)$$

where $\tilde{x}(t)$ is the input passband signal, $\tilde{y}(t)$ is the output passband signal, and $\{\tilde{a}_l\}_{l=1}^L$ are the model coefficients. In (10.1), the DC (direct current) term is omitted as the DC power is not radiated into the air.

Let us consider a single-tone test by substituting $x(t) = A \sin(\omega_c t)$ into (10.1), where A is the amplitude of the sinusoidal signal, and ω_c is the carrier frequency. Without loss of

¹ In some literature, the output gain function is referred to as the AM/AM response.

generality, the initial phase of the input signal is set to 0. For simplicity, let us consider the $L = 5$ case. Equation (10.1) can be rewritten as:

$$\begin{aligned}
 \tilde{y}(t) &= \sum_{l=1}^5 \tilde{a}_l A^l \sin^l(\omega_c t), \\
 &= \tilde{a}_1 A \sin(\omega_c t) \\
 &\quad + \frac{1}{2} \tilde{a}_2 A^2 - \frac{1}{2} \tilde{a}_2 A^2 \cos(2\omega_c t) \\
 &\quad + \frac{3}{4} \tilde{a}_3 A^3 \sin(\omega_c t) - \frac{1}{4} \tilde{a}_3 A^3 \sin(3\omega_c t) \\
 &\quad + \frac{3}{8} \tilde{a}_4 A^4 - \frac{1}{2} \tilde{a}_4 A^4 \cos(2\omega_c t) + \frac{1}{8} \tilde{a}_4 A^4 \cos(4\omega_c t) \\
 &\quad + \frac{5}{8} \tilde{a}_5 A^5 \sin(\omega_c t) - \frac{5}{16} \tilde{a}_5 A^5 \sin(3\omega_c t) + \frac{1}{16} \tilde{a}_5 A^5 \sin(5\omega_c t). \quad (10.2)
 \end{aligned}$$

In (10.2), in addition to the original carrier frequency, additional frequency components also appear at the DC, and the second-order through fifth-order harmonic frequencies. At the carrier frequency ω_c , the output $\tilde{y}_c(t)$ is

$$\tilde{y}_c(t) = (\tilde{a}_1 + \frac{3}{4} \tilde{a}_3 A^2 + \frac{5}{8} \tilde{a}_5 A^4) A \sin(\omega_c t). \quad (10.3)$$

The output amplitude of $\tilde{y}_c(t)$ as a function of the input amplitude, forms the AM/AM response of the PA, which is nonlinear. From (10.3), we also notice that the above memoryless model does not have AM/PM response since the output phase deviation ($\angle \tilde{y}_c - \angle \tilde{x}$) is a constant 0.

In a two-tone test, the amplitudes of the two input tones are usually assumed to be equal. Furthermore, the input is assumed to be a narrow band signal, that is, $\tilde{x}(t) = A \sin(\omega_1 t) + A \sin(\omega_2 t)$, where $\omega_2 > \omega_c > \omega_1 \gg 0$. Substituting $\tilde{x}(t)$ into (10.1), the PA output can be obtained. The resulting frequency components of the PA output are illustrated in Figure 10.2, for the case with $L = 5$.

In Figure 10.2, we observe:

1. Similar to the single-tone test, we observe that additional frequency components appeared around the DC, and the second-order through fifth-order harmonics at the output of the PA nonlinearity.
2. Only odd-order nonlinearities contribute to the PA output around the carrier frequency.
3. Besides the original input frequencies, additional frequency components are generated around the carrier frequency. The frequency components located at $2\omega_2 - \omega_1$ and $2\omega_1 - \omega_2$ are referred to as the IMD3; the frequency components located at $3\omega_2 - 2\omega_1$ and $3\omega_1 - 2\omega_2$ are referred to as the IMD5.
4. IMD3 components are generated by both the third-order nonlinearity and other higher odd-order nonlinearities (fifth order in the above example).

If we extend the analysis to a modulated signal, out-of-band IMDs manifest as adjacent channel interference, alternate channel interference, etc. Adjacent channel power ratio (ACPR) is widely used to quantify the adjacent channel interference [2, 10].

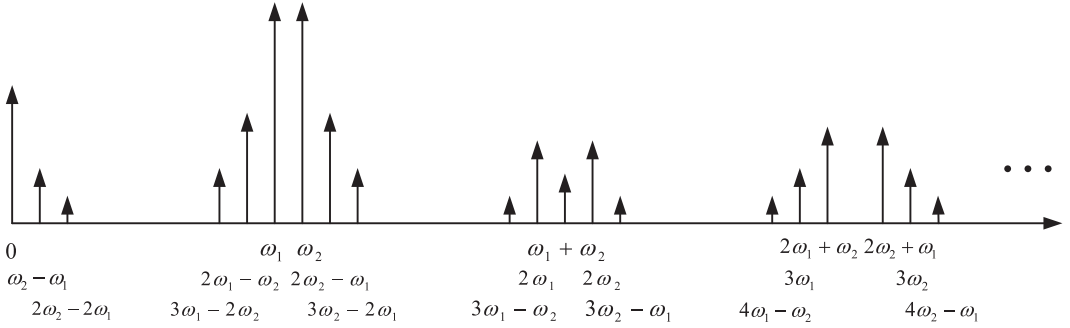


Figure 10.2 Frequency response of the PA output in two-tone test.

As modern PA design becomes more and more sophisticated, AM/AM, AM/PM response or IMD products IIP_n may not be accurate enough to describe the Power amplifier nonlinearity. PA modeling is needed.

10.2.2 System modeling of PA nonlinearity

Physics-based modeling of the PA is often considered by RF engineers and RF design tools. These models, however, are limited in circuit-level modeling accuracy and suffer from lengthy simulation time. Physics-based models are often tied to the internal structure of the PA, which may not be available to the user and often changes with the design.

Power amplifier behavioral models, or PA system-level models are preferred in system-level studies. The behavioral models only require input and output to describe; i.e., they treat the PA as a black box.

For a general nonlinear passband model with memory, the Volterra series can be used [11] [12, Chapter 4]:

$$\tilde{y}(t) = \sum_{l=1}^L \int \cdots \int \tilde{h}_l(\tau_1, \dots, \tau_l) \prod_{i=1}^l \tilde{x}(t - \tau_i) d\tau_1 d\tau_2 \cdots d\tau_l, \quad (10.4)$$

where $\tilde{h}_l(\cdot)$ is the l th-order Volterra kernel. Note that both even and odd l 's are included in (10.4), and all quantities in (10.4) are real-valued. To describe a memoryless nonlinearity, we set $\tilde{h}_l(\tau_1, \dots, \tau_l) = \tilde{a}_l \delta(\tau_1, \dots, \tau_l)$, and obtain the familiar polynomial (or power series) model in (10.1) [12, p. 156], [13, p. 181].

To evaluate the impact of PA nonlinearities, baseband equivalent models may be considered since the important performance metrics in the transmitter, such as the spectral spreading behavior and the error vector magnitude (EVM) can be evaluated in the baseband. The relationship between a passband model and a baseband equivalent model can be expressed as [14]:

$$\tilde{y}(t) = \text{Re}\{y(t)e^{j(\omega_c t)}\}, \quad (10.5)$$

where $\tilde{y}(t)$ is the passband signal and $y(t)$ is its baseband equivalent version.

In the presence of nonlinearity, the relationship between passband and baseband signals becomes complicated. In the above one-tone test and two-tone test, high-order harmonics and DC components are generated due to the nonlinearity, which may not be easily represented by the baseband equivalent model. In most cases, the PA output matching network and the antenna are frequency selective around the carrier frequency. The high-order harmonics and the DC component are suppressed after these frequency selective elements and can be ignored in the baseband representation. In other words, in the nonlinear PA modeling, the baseband model only attempts to model the passband signal behavior around the carrier frequency.

Retaining only the $l = 2k + 1$ terms, the baseband equivalent Volterra model that corresponds to the passband model (10.4) is given by [11, 15]:

$$y(t) = \sum_{k=0}^K \int \cdots \int h_{2k+1}(\cdot) \prod_{i=1}^{k+1} x(t - \tau_i) \prod_{i=k+2}^{2k+1} x^*(t - \tau_i) d\tau_1 d\tau_2 \cdots d\tau_{2k+1}, \quad (10.6)$$

where

$$\begin{aligned} h_{2k+1}(\cdot) &= h_{2k+1}(\tau_1, \dots, \tau_{2k+1}) \\ &= \frac{1}{2^{2k}} \frac{(2k+1)!}{k!(k+1)!} \tilde{h}_{2k+1}(\tau_1, \dots, \tau_{2k+1}) e^{-j\omega_c \left(\sum_{i=1}^{k+1} \tau_i - \sum_{i=k+2}^{2k+1} \tau_i \right)}, \end{aligned} \quad (10.7)$$

and

$$2K+1 = \begin{cases} L, & \text{if } L \text{ is odd,} \\ L-1, & \text{if } L \text{ is even.} \end{cases}$$

If the PA is strictly memoryless, then by substituting $\tilde{h}_l(\tau_1, \dots, \tau_l) = \tilde{a}_l \delta(\tau_1, \dots, \tau_l)$ into (10.6) and (10.7), the corresponding baseband equivalent expression of (10.1) is

$$y(t) = \sum_{k=0}^K a_{2k+1} [x(t)]^{k+1} [x^*(t)]^k, \quad (10.8)$$

where

$$a_{2k+1} = \frac{1}{2^{2k}} \frac{(2k+1)!}{k!(k+1)!} \tilde{a}_{2k+1}. \quad (10.9)$$

Denote by $H_{2k+1}(\cdot)$ the multi-dimensional Fourier transform of $h_{2k+1}(\cdot)$ w.r.t. $\tau_1, \tau_2, \dots, \tau_{2k+1}$. When the PA exhibits short-term memory effects, and if the input signal $x(t)$ is narrowband, we may treat $x(t - \tau_k) \approx x(t)$ [15, 16]. Equation (10.6) can be reduced to (10.8) with

$$\begin{aligned} a_{2k+1} &= H_{2k+1}(\underbrace{0, \dots, 0}_{2k+1}) \\ &= \frac{1}{2^{2k}} \frac{(2k+1)!}{k!(k+1)!} \tilde{H}_{2k+1}(\underbrace{f_c, \dots, f_c}_{k+1}, \underbrace{-f_c, \dots, -f_c}_k), \end{aligned} \quad (10.10)$$

where $\tilde{H}_{2k+1}(\cdot)$ on the right-hand side of (10.10) is the $(2k+1)$ -dimensional Fourier transform of the Volterra kernel $\tilde{h}_{2k+1}(\tau_1, \dots, \tau_{2k+1})$ evaluated at the above frequency tuple.

The model described by (10.8) still has the form of a memoryless model. While the a_{2k+1} in (10.9) is real valued, the a_{2k+1} in (10.10) is generally complex valued. This model is also called a quasi-memoryless model. Equation (10.8) can be rewritten as $y(t) = x(t) \cdot \sum_{k=0}^K a_{2k+1} |x(t)|^{2k}$. The AM/AM response is then $|\sum_{k=0}^K a_{2k+1} |x(t)|^{2k}|$ as a function of $|x(t)|$; the AM/PM response is $\angle(\sum_{k=0}^K a_{2k+1} |x(t)|^{2k})$ as a function of $|x(t)|$.

In References [17–19], some researchers attempt to model the AM/AM and AM/PM responses as two real-valued polynomials, e.g.,

$$|y(t)| = \sum_{k=1}^K \beta_k |x(t)|^k, \quad (10.11)$$

$$\angle y(t) - \angle x(t) = \sum_{l=1}^L \gamma_l |x(t)|^l. \quad (10.12)$$

We prefer to model the PA input and output behaviors as one complex-valued polynomial model, which provides a direct link between the passband nonlinearity and the baseband model, and can extend to the modeling of memory effects.

10.3 Digital predistortion modeling

In Section 10.2, we discussed the PA nonlinearities. Since the ideal predistorter is an inverse of the PA, the PA nonlinearity modeling can guide the predistorter design as well. In this section, we will first apply the knowledge that we learned from the PA to the digital predistorter modeling. We will then discuss different optimization criteria, provide solutions, and also show the predistorter performance from experiments.

10.3.1 Memoryless nonlinear models

The baseband equivalent polynomial model for a memoryless PA can be written as (10.8). It is natural to model the predistorter with the same structure,

$$\begin{aligned} z(t) &= \sum_{k=0}^K a_{2k+1} [x(t)]^{k+1} [x^*(t)]^k \\ &= \sum_{k=0}^K a_{2k+1} |x(t)|^{2k} x(t), \end{aligned} \quad (10.13)$$

where $x(t)$ is the input of the predistorter and $z(t)$ is the output of the predistorter.

For a memoryless PA, the AM/AM and AM/PM responses are sufficient to represent the PA nonlinearity. Many memoryless predistortion models have been derived on the

basis of this observation [17–19]. For example, we assume that the PA AM/AM and AM/PM nonlinearities can be modeled as nonlinear functions of the amplitude of the input signal,

$$|y(t)| = f_1(|x(t)|), \quad (10.14)$$

$$\angle y(t) - \angle x(t) = f_2(|x(t)|). \quad (10.15)$$

The nonlinear functions $f_1(\cdot)$ and $f_2(\cdot)$ can be approximated by real polynomials as in (10.11) and (10.12) respectively. These functions can also be obtained by look-up-tables (LUTs) from the measurements. The predistorter just needs to provide the inverse of the AM/AM response and the opposite phase shift as in the AM/PM response.

$$|z(t)| = f_1^{-1}(|x(t)|), \quad (10.16)$$

$$\angle z(t) - \angle x(t) = -f_2(|z(t)|). \quad (10.17)$$

Please note that (10.17) is not explicit in terms of the input amplitude $|x(t)|$. In practice, (10.16) should be updated before the computation of (10.17). The AM/AM, AM/PM separated memoryless nonlinear model is very attractive in attempts of the predistorter design [20] as it is very simple yet practical. However, there are still some implementation concerns about this model:

1. The measurement may not be accurate in the AM/PM path as the phase is vulnerable to the noise when the signal amplitude is small. Averaging over a large sample size provides little help because the phase wrapping for different samples can be different. Averaging on these samples may introduce irreducible error to the measurement.
2. In real-time adaptation, available data are $z(t)$ and $y(t)$, thus models that are parsimonious in $z(t)$ and $y(t)$ are preferred.
3. It is not easy to extend the AM/AM, AM/PM separated memoryless nonlinear model to the memory case.

10.3.2 Memory nonlinear models

In Section 10.2, it is shown that power amplifiers with short-term memory effects or narrow-band signals lead to memoryless baseband representations. However, as the input signal bandwidth becomes wider, such as in Wideband CDMA (WCDMA), or IEEE 802.11b/g/n signal, the time span of the power amplifier memory becomes comparable with the time scale of the input signal envelope. The memory effects of the power amplifier in the passband can no longer be considered as short term. The baseband equivalent Volterra series [15, 16, 21] is given in (10.6). Volterra series is a general nonlinear system with memory [22]. It can also be applied to the predistorter modeling [23],

$$z(t) = \sum_{k=0}^K \int \cdots \int g_{2k+1}(\cdot) \prod_{i=1}^{k+1} x(t - \tau_i) \prod_{i=k+2}^{2k+1} x^*(t - \tau_i) d\tau_1 d\tau_2 \cdots d\tau_{2k+1}. \quad (10.18)$$

In discrete-time domain, (10.18) becomes

$$z(n) = \sum_k \sum_{l_1} \sum_{l_2} \cdots \sum_{l_{2k+1}} g_{2k+1}(l_1, l_2, \dots, l_{2k+1}) \prod_{i=1}^{k+1} x(n - l_i) \prod_{i=k+2}^{2k+1} x^*(n - l_i). \quad (10.19)$$

From (10.19), it can be seen that the number of coefficients of the Volterra series increases dramatically as the memory length and the nonlinearity order increase. The model can become so complex that model coefficients estimation becomes impractical given constraints in the sample size and processing time. Various special cases of the Volterra model have been proposed and studied for the purpose of predistorter design.

A simple model to incorporate frequency-dependent effects in the nonlinear model is to introduce a linear time-invariant (LTI) block before or after the memoryless nonlinear (NL) block. Both solid-state PAs (SSPAs) and traveling-wave tube PAs (TWTAs) have been modeled by a cascade of linear filters and a memoryless nonlinearity [24–27]. Introducing linear memory effects at the input and/or output can be physically linked to the PA's input and output matching networks.

For example, the LTI block can be placed before the NL block, leading to the so-called Wiener model [24] (see Figure 10.3).

In the Wiener model, the two subsystems are given by

$$z(n) = \sum_{l=0}^{L-1} a_l x(n - l), \quad (10.20)$$

$$y(n) = \sum_{k=0}^{K-1} b_{2k+1} |z(n)|^{2k} z(n), \quad (10.21)$$

where $\{a_l\}$ is the impulse response of the LTI block and $\{b_{2k+1}\}$ are the coefficients of the odd-order polynomials describing the memoryless nonlinearity. Substitution of (10.20) into (10.21) yields

$$y(n) = \sum_{k=0}^{K-1} b_{2k+1} \left| \sum_{l=0}^{L-1} a_l x(n - l) \right|^{2k} \left(\sum_{l=0}^{L-1} a_l x(n - l) \right). \quad (10.22)$$

On the other hand, by placing the NL block before the LTI block, a Hammerstein model can be obtained [25, 26] (see Figure 10.4).

In the Hammerstein model, the two subsystems are given by

$$z(n) = \sum_{k=0}^{K-1} a_{2k+1} |x(n)|^{2k} x(n), \quad (10.23)$$

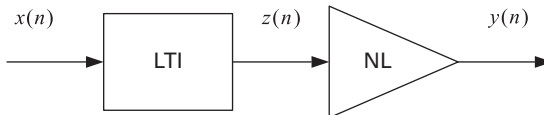


Figure 10.3 The Wiener model.

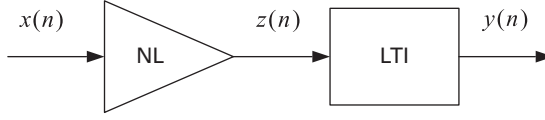


Figure 10.4 The Hammerstein model.

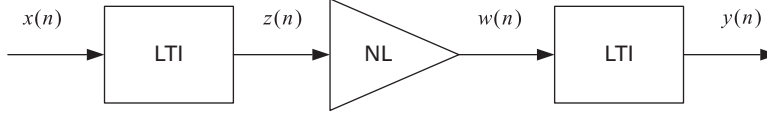


Figure 10.5 The Wiener-Hammerstein model.

$$y(n) = \sum_{l=0}^{L-1} b_l z(n-l), \quad (10.24)$$

where $\{b_{2k+1}\}$ are the coefficients of the polynomial nonlinearity and $\{a_l\}$ are the impulse response values of the LTI block. Substituting (10.23) into (10.24), we obtain

$$y(n) = \sum_{l=0}^{L-1} b_l \sum_{k=0}^{K-1} a_{2k+1} |x(n-l)|^{2k} x(n-l). \quad (10.25)$$

The Wiener-Hammerstein model has three cascaded blocks (see Figure 10.5). The memoryless nonlinearity is saddled between two LTI systems. Such a configuration is commonly used for satellite communication channels, where the power amplifier at the satellite transponder is driven near saturation to exploit the maximum power efficiency for the downlink [27].

The three subsystems in this model are described by

$$z(n) = \sum_{l=0}^{L-1} a_l x(n-l), \quad (10.26)$$

$$w(n) = \sum_{k=0}^{K-1} b_{2k+1} |z(n)|^{2k} z(n), \quad (10.27)$$

$$y(n) = \sum_{m=0}^{M-1} c_m w(n-m), \quad (10.28)$$

where $\{a_l\}$ and $\{c_m\}$ are the impulse responses of the LTI blocks before and after the NL block respectively, and b_{2k+1} are the coefficients of polynomial nonlinearity. The overall input-output response of the system is given by

$$z(n) = \sum_{m=0}^{M-1} c_m \sum_{k=0}^{K-1} b_{2k+1} \left| \sum_{l=0}^{L-1} a_l x(n-l-m) \right|^{2k} \left(\sum_{l=0}^{L-1} a_l x(n-l-m) \right). \quad (10.29)$$

Besides these linear memory effects, there are also some memory effects tightly coupled within the nonlinear circuits [28–31]. For example, thermal memory effects are caused by electro-thermal couplings. The power dissipated in the device in the past can change the junction temperature, which in turn changes PA characteristics. Simple non-interacting linear filter and memoryless nonlinearity models cannot fully describe the phenomena.

In (10.19), if we add the constraint

$$h_{2k+1}(l_1, l_2, \dots, l_{2k+1}) = \begin{cases} a_{2k+1,l}, & \text{if } l_1 = l_2 = \dots = l_{2k+1} = l, \\ 0, & \text{otherwise,} \end{cases} \quad (10.30)$$

i.e., all off-diagonal kernels of the Volterra series are set to zero, we obtain the memory polynomial model [32]:

$$y(n) = \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} a_{2k+1,l} |x(n-l)|^{2k} x(n-l). \quad (10.31)$$

The memory polynomial model can be viewed as a parallel Wiener system (nonlinear tapped-delay line) (see in Figure 10.6) or a parallel Hammerstein system (see in Figure 10.7). The memory polynomial model has proven to be effective for predistortion of actual power amplifiers under typical operating conditions [32, 33]. In (10.31), we observe that the model coefficients are explicitly available. Additionally, although the output is nonlinear w.r.t. the input, it is linear w.r.t. the model coefficients. Classical techniques such as the least squares algorithm can be applied to estimate model coefficients.

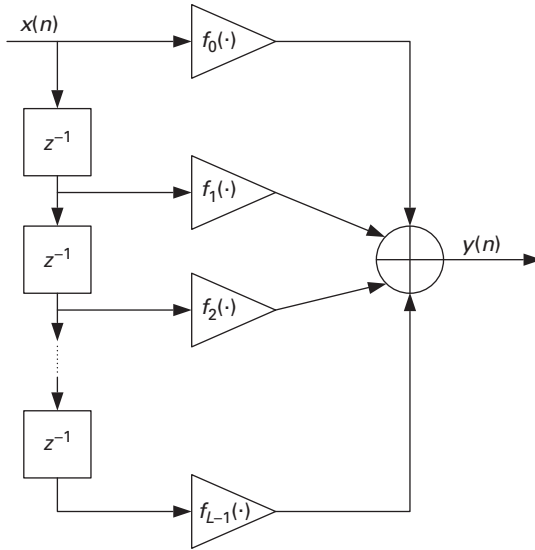


Figure 10.6 The parallel Wiener representation of the memory polynomial model. The parallel LTI part is reduced to the tapped delay lines, and the memoryless nonlinear block is given by $f_l(x) = \sum_{k=0}^{K-1} a_{2k+1,l} |x|^{2k} x$.

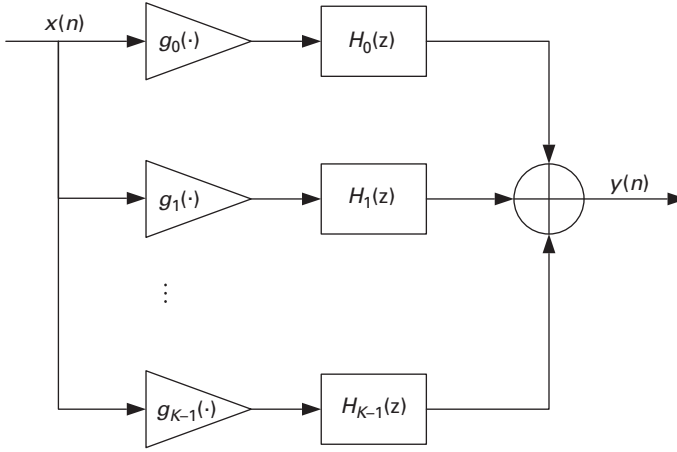


Figure 10.7 The parallel Hammerstein representation of the memory polynomial model. The memoryless nonlinear block is given by $g_k(x) = x(n)|x(n)|^{2k}$, and the LTI block is given by $H_k(z) = \sum_{l=0}^{L-1} a_{2k+1,l}z^{-l}$.

In the memory polynomial model, all product terms involve input samples at the same time instant. The cross terms, or the off-diagonal kernels in the Volterra model, on the other hand, may still help to improve the modeling accuracy. The generalized memory polynomial model can be adapted to include some of the cross terms [34]:

$$\begin{aligned}
 y(n) = & \sum_{k=0}^{K_a-1} \sum_{l=0}^{L_a-1} a_{2k+1,l} |x(n-l)|^{2k} x(n-l) \\
 & + \sum_{k=1}^{K_b} \sum_{l=0}^{L_b-1} \sum_{m=1}^{M_b} b_{2k+1,l,m} |x(n-l-m)|^{2k} x(n-l) \\
 & + \sum_{k=1}^{K_c} \sum_{l=0}^{L_c-1} \sum_{m=1}^{M_c} c_{2k+1,l,m} |x(n-l+m)|^{2k} x(n-l), \quad (10.32)
 \end{aligned}$$

where K_a, L_a are the number of coefficients for aligned signal and envelope (memory polynomial); K_b, L_b, M_b are the number of coefficients for signal and lagging envelope; and K_c, L_c, M_c are the number of coefficients for signal and leading envelope. This model is reported to yield better performance than the memory polynomial model at the expense of more model coefficients [34].

10.3.3 Indirectly learning architecture for system identification

In the [previous section](#), we discussed different models that can be applied to digital predistortion. Various system identification approaches can be applied to obtain the model coefficients. For example, for an LUT-based AM/AM, AM/PM separated predistorter design, we can simply repeat the one-tone test and obtain the PA AM/AM, AM/PM response. Entries of the predistorter AM/AM table can be obtained by inverting the PA AM/AM response; and entries of the predistorter AM/PM table can be obtained by

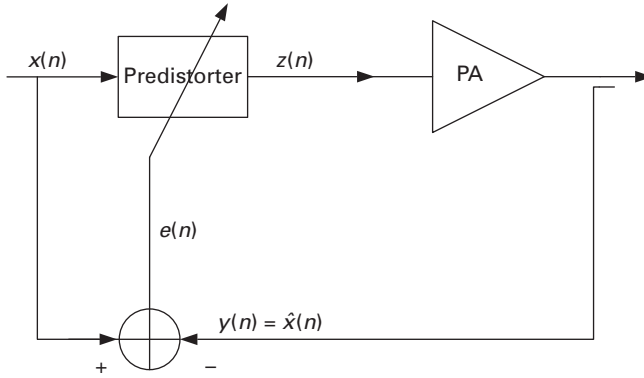


Figure 10.8 A direct learning architecture for estimation of predistorter coefficients. The error signal is produced to compare the input and output of the predistorter + PA system.

applying the opposite phase shift of the AM/PM response, where the amplitude index is obtained after the AM/AM correction.

As predistorter models become more complex, especially when memory structures are introduced, instant samples of the PA input and output are widely used in the estimation of predistorter coefficients. Figure 10.8 shows a basic structure for predistorter system identification.

In this system, samples of $x(n)$ are the input of the predistorter, $z(n)$ are the output of the predistorter, which are also the input to the PA, and $y(n)$ are the (normalized) output of the PA; $x(n)$, $z(n)$ and $y(n)$ are collected in baseband. The predistorter and the PA models are the baseband equivalent versions. An error signal $e(n) = x(n) - y(n)$ is introduced. The overall predistorter + PA system is approximately linear when the error energy in $e(n)$ tends to zero. In fact, the feedback linearization works with similar principles.

This structure, however, is not easy to work with. There is no explicit relationship between the model coefficients of the predistorter and the error signal.

Figure 10.9 shows a second structure to work around this obstacle. We first introduce a model for the PA under test. The error function $e(n)$ is the error between the actual output of the PA and the output of the model $e(n) = y(n) - \hat{y}(n)$. Note that when the PA is estimated perfectly, $\hat{y}(n) = y(n)$. This architecture allows us to identify the model of the PA first, then find its inverse (the predistorter) subsequently [26]. However, obtaining the inverse of a nonlinear system with memory is generally a difficult task.

Figure 10.10 shows the indirect learning architecture [23, 35, 36], which estimates the predistorter directly (and thus learns the PA indirectly). The input to the predistorter estimation block is $y(n)$, the output is $\hat{z}(n)$. The error signal $e(n)$ is given by $e(n) = z(n) - \hat{z}(n)$. By minimizing the error energy in $e(n)$, predistorter parameters can be obtained. Note that when the PA is linearized, $y(n) = x(n)$ and thus $\hat{z}(n) = z(n)$. Then, the estimated parameters of the inverse model are copied to an identical model that is used in the data path to predistort the input of the PA.

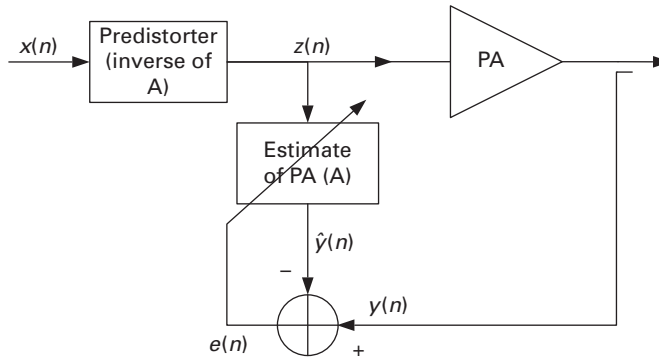


Figure 10.9 A direct learning architecture for estimation of predistorter coefficients. The error signal is produced at the PA output.

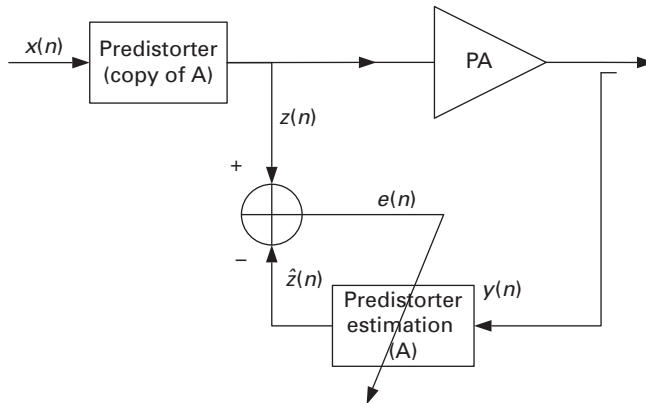


Figure 10.10 The indirect learning architecture estimates the predistorter directly and learns about the PA indirectly. The error signal is produced at the PA input.

10.3.4 Estimation of predistorter coefficients

Our model discussions in the above mostly involve the Volterra series model or its simplifications. These models share the same advantage: model coefficients can be viewed as linear weights of nonlinear signals. This fact greatly simplifies the estimation of model coefficients. Some mature quadratic programming or convex optimization methods such as the least squares, or the least mean square can be applied to obtain the model coefficients.

Let us formulate the optimization problem first. The memory polynomial model is discussed here as an example. The result can be easily extended to the general Volterra series model or other nonlinear model that is linear in the model coefficients. The system setup is given by the indirect learning architecture; $x(n)$ is the baseband input; $z(n)$ is the output from the predistorter, which is also the input of the PA; and $y(n)$ is the normalized

baseband equivalent PA output. The delay along the data path has been carefully calibrated. The predistorter estimation block (A) has the form of a memory polynomial model:

$$\hat{z}(n; \theta) = \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} a_{2k+1,l} |y(n-l)|^{2k} y(n-l), \quad (10.33)$$

where $\theta = \{a_{2k+1,l}\}_{k,l}$. With the knowledge of $z(n)$ and $y(n)$, we would like to estimate model coefficients θ .

The least squares (LS) method obtains the estimate of these parameters as the values which minimize the sum of the squares between the measurements $z(n)$ and the model output, i.e., the predicted values, $\hat{z}(n)$. The objective function is

$$\begin{aligned} & \arg \min_{\theta} \sum_{n=0}^{N-1} |z(n) - \hat{z}(n; \theta)|^2 \\ &= \arg \min_{\theta} \sum_{n=0}^{N-1} |z(n) - \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} a_{2k+1,l} |y(n-l)|^{2k} y(n-l)|^2. \end{aligned} \quad (10.34)$$

In (10.34), we further assume that $y(n) = 0$ at $n = -1, \dots, -(L-1)$ and $N > KL$. Reorganizing the data samples and the coefficients into vectors or matrix, we get

$$z = [z(0), z(1), \dots, z(N-1)]^T, \quad (10.35)$$

$$Y(n) = [\phi_0(n), \phi_1(n), \dots, \phi_{K-1}(n), \dots, \phi_0(n-L+1), \dots, \phi_{K-1}(n-L+1)], \quad (10.36)$$

$$R = \begin{bmatrix} Y(0) \\ Y(1) \\ \vdots \\ Y(N-1) \end{bmatrix}, \quad (10.37)$$

$$a = [a_{1,0}, a_{3,0}, \dots, a_{2K-1,0}, \dots, a_{1,L-1}, a_{3,L-1}, \dots, a_{2K-1,L-1}]^T, \quad (10.38)$$

where

$$\phi_k(n) = |y(n)|^{2k} y(n), \quad (10.39)$$

and $(^T)$ denotes matrix transpose.

The objective function (10.34) can be rewritten as

$$\arg \min_a \|z - Ra\|^2. \quad (10.40)$$

The least squares solution to (10.40) is

$$\hat{a} = (R^H R)^{-1} R^H z, \quad (10.41)$$

where $(^H)$ denotes Hermitian transpose.

The least squares algorithm works on a block of data samples. In real-time applications, sample-by-sample processing can be advantageous and one may consider the recursive least squares (RLS). The RLS algorithm can be derived from the least squares algorithm by using the matrix inversion lemma [37]. The procedure of the RLS adaptation is given by [37]:

1. The recursive least squares algorithm is initialized with

$$P(-1) = \delta^{-1}I, \quad (10.42)$$

$$a(-1) = [0, 0, \dots, 0], \quad (10.43)$$

where $a(n)$ denotes the estimate of model coefficients at the n th iteration, $P(n)$ denotes the inverse of the covariance matrix, $(R^H R)^{-1}$, obtained from the first n samples, $\delta > 0$ is an arbitrary small constant, and I is the identity matrix.

2. For the n th iteration, $P(n)$ and $a(n)$ are given by:

$$P(n) = \left(I - \frac{P(n-1)Y^H(n)Y(n)}{1 + Y(n)P(n-1)Y^H(n)} \right) P(n-1), \quad (10.44)$$

$$a(n) = a(n-1) + \frac{P(n-1)Y^H(n)}{1 + Y(n)P(n-1)Y^H(n)} (z(n) - Y(n)a(n-1)). \quad (10.45)$$

The least squares (recursive least squares) algorithm yields the exact solution to the above optimization problem formulated in a deterministic way. The least mean square (LMS) algorithm, on the other hand, can be considered as an approximate solution to the statistical optimization problem that tries to minimize the mean square value of the estimation error at each time. The objective function to the LMS algorithm is

$$\arg \min_a E(|z - \hat{z}|^2) = \arg \min_a E(|z - Ya|^2), \quad (10.46)$$

where z is a random variable at the PA input, $\hat{z}$ is a random variable at the output of the predistorter estimation block, Y, a are vectors that have the same structure as in (10.36), (10.38).

To obtain the LMS solution of (10.46) [37], we start with an arbitrary initial estimate of a and a positive step size $\mu > 0$. In the context of predistorter coefficients estimation, it is natural to set

$$a(0) = [1, 0, \dots, 0]^T, \quad (10.47)$$

where $a(n)$ denotes the value of estimated coefficients a in the n th iteration.

With the prior knowledge of a coefficients from the n th iteration and new input data $z(n)$ and $Y(n)$, the new error term is given by

$$e(n) = z(n) - Y(n)a(n), \quad (10.48)$$

and the updated a coefficients can be obtained by

$$\begin{aligned}
 a(n+1) &= a(n) + \mu e(n) Y^H(n) \\
 &= a(n) + \mu(z(n) - Y(n)a(n)) Y^H(n).
 \end{aligned} \tag{10.49}$$

As indicated in [37, 38], if the step size $0 < \mu < 2/\lambda_{\max}$, where $\lambda_{\max}$ is the maximum eigenvalue of the autocorrelation matrix of the input vector Y , the mean values of the coefficients converge (for stationary environments) to their optimal values. The LMS algorithm is attractive to use in the digital predistorter implementation because of its simplicity. The algorithm complexity is on the order of $O((KL)^2)$ per sample, where KL is the total number of coefficients in the predistorter model. For comparison, the RLS algorithm has a complexity on the order of $O((KL)^4)$ per sample, and the LS algorithm has a complexity on the order of $O((KL)^6)$ per sample.

On the other hand, the convergence speed of the LMS algorithm becomes a concern for the predistortion application. Different kernels of Volterra series are highly correlated, thereby resulting in a poorly conditioned covariance matrix. We will discuss more on this issue in the [next section](#) when we discuss predistorter implementation issues.

Model coefficients estimation for models such as the Saleh model [39], which does not provide a quadratic relationship between the objective function and the model coefficients, is still possible with other optimization tools. However, the global optimality is not guaranteed and the computational complexity may also be high.

10.3.5 DPD performance and experimental results

Digital predistortion has been a widely studied problem in the literature. In [Table 10.2](#), the performance of a number of digital baseband predistortion linearization algorithms is shown. The performance metric is chosen as the improvement in the ACPR after predistortion.

From [Table 10.2](#), we know that in the early years, when digital signal processing was still not that powerful, the LUT-based digital predistorter gained popularity since the predistorter can be trained offline and the online computation is simple. On the other hand, the performance of the LUT-based predistorter is rather limited. In recent years, the digital signal processing capability has advanced rapidly. Online predistorter training becomes possible. More sophisticated predistorters have been applied to the predistorter design leading to satisfactory results.

Table 10.2 Digital baseband predistortion performances reported in the literature.

Reference	Test PA size	Predistorter type	Test signal (bandwidth)	ACPR improvement
Kusunoki <i>et al.</i> [40]	0.5 W	LUT	CDMA (1.25 MHz)	7 dB
Zavosh <i>et al.</i> [41]	4 W	LUT	modulated (1.25 MHz)	12 dB
Kubo <i>et al.</i> [42]	30 W	LUT	4-carrier WCDMA (20 MHz)	15 dB
Boumaiza <i>et al.</i> [43]	90 W	LUT	CDMA 2000 (3.75 MHz)	10 dB
Jung <i>et al.</i> [44]	300 W	Multiple LUTs	WCDMA (5 MHz)	12 dB
Qian <i>et al.</i> [45]	45 W	Memory polynomial	2-carrier CDMA (2.5 MHz)	15 dB
Morgan <i>et al.</i> [34]	30 W	Generalized memory polynomial	3-carrier UMTS (15 MHz)	20 dB

For predistorter validations, experiments were carried out on an Ericsson PA that is a silicon bipolar transistor based class-AB basestation PA and operates at the cellular band with a center frequency of 881 MHz. The input signal was a 3.84 MHz bandwidth WCDMA signal² with a peak-to-average power ratio (PAPR) of 8.5 dB (after amplitude clipping). The peak envelop power of the PA output was 36.5 dBm. Figure 10.11 shows the performances of the memoryless polynomial predistorter and the memory polynomial predistorter in terms of power spectral density (PSD) of the PA output. The memoryless polynomial predistorter is given by (10.13) with $K = 2$ (fifth order) and the memory polynomial predistorter is given by (10.31) with $K = 2$ and $L = 4$ (fifth order, 5 memory taps). The predistorter coefficients were obtained using the least squares algorithm with 50 000 samples. In Figure 10.11, from top to bottom, the curves are: the PA output PSD without predistortion; the PA output PSD with memoryless polynomial predistortion; and the PA output PSD with memory polynomial predistortion, respectively.

In this experiment, we noticed that the PA output spectrum had significant spectral regrowth since the PA was driven deep into the nonlinear region. Such PA output does not meet the standard spectral mask requirement. The spectral regrowth at the lower sideband

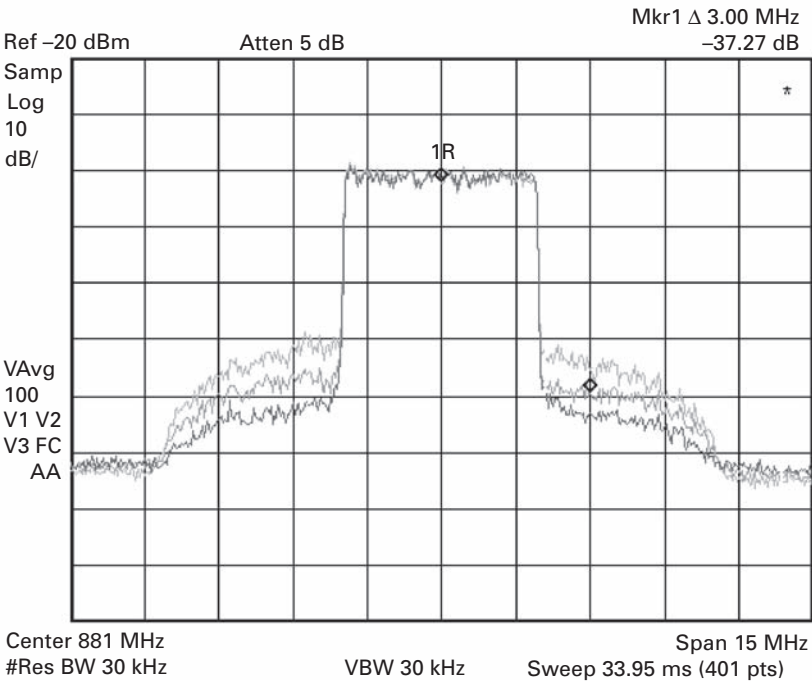


Figure 10.11 Measured PA output PSD for the Ericsson 45 W PA: From top to bottom, PSDs were obtained: without predistortion; with a fifth-order memoryless predistorter; with a fifth-order, 5-tap delay memory polynomial predistorter.

² In this experiment, our focus is on the PA response with a wideband signal. The fact that the WCDMA signal does not operate at 880MHz frequency band does not change the experiment result.

of the main channel and the spectral regrowth at the upper sideband of the main channel were not symmetric. This observation is evidence that the Ericsson 45 W PA exhibits memory effects since a memoryless PA always has symmetric intermodulation products or spectral regrowth. The memoryless polynomial predistorter as well as the memory polynomial predistorter provided some improvement to the adjacent channel power ratio (ACPR). The memory polynomial predistorter suppressed the IMD in the adjacent channel by 12 dB, whereas the memoryless polynomial predistorter only gave 7 dB of IMD reduction. The results showed that the memory polynomial predistorter had a significant advantage over the memoryless polynomial predistorter for the PA with memory effects.

10.4 Predistortion implementation challenges

Predistortion linearization helps to improve the PA efficiency while still maintaining the overall linearity of the system. In this section, we will discuss predistortion implementation considerations and challenges.

10.4.1 Adaptive DPD architecture

The characteristics of PAs may change over time due to the environment temperature changes,³ voltage variations, channel changes, antenna loading changes, and aging [2, 46]. A practical predistorter design should accommodate PA characteristics changes. On the other hand, power amplifier characteristics change very slowly compared to the time needed for the model coefficients adaptation. Digital predistorter designs reported in [34, 40–42, 44, 45] all have adaptive structures.

Figure 10.12 shows an adaptive DPD architecture. In this figure, a general transmitter chain is included. The digital predistorter is inserted in the digital domain right before the digital-to-analog converter (DAC). The predistorted signal is then converted into analog domain and passed through the up conversion blocks before entering the PA (the device under test, DUT). To achieve adaptation, a down-conversion chain is added to the system to convert the passband signal back to baseband. In Figure 10.12, the shaded area is the predistorter estimation block. The whole data path needs to be carefully designed to achieve satisfactory results:

1. Linearity of the receive path. In order to avoid interfering with the normal transmission operation, only a small fraction of the PA output is fed back via the feedback path to facilitate estimation of the predistorter parameters. An attenuator is used to adjust the input power level of the receiver chain. The input power of the receiver needs to be carefully monitored to make sure that the receiver path does not introduce additional nonlinearity.

³ The transistor junction temperature changes caused by the input amplitude variation is signal dependent and is considered as memory effects.

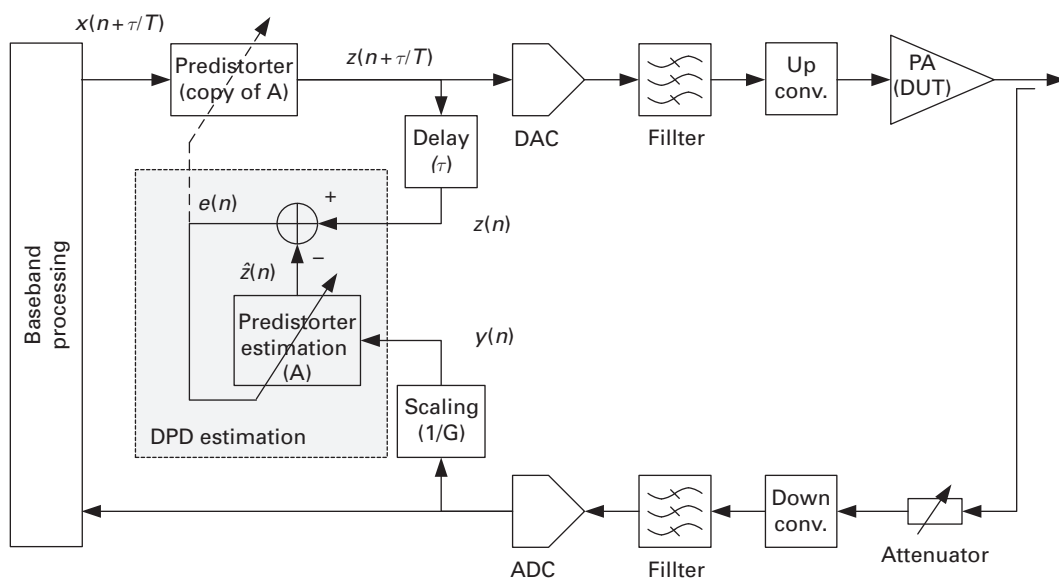


Figure 10.12 An adaptive digital predistortion implementation architecture.

2. Transmit/receive filter selection. Along the transmit/receive data path, there are filters in between different blocks. For example, an anti-aliasing filter after the DAC, a frequency band selection filter after each frequency conversion and the power amplifier matching network. The memory introduced by filters, when coupled with nonlinearity of the PA, increases the predistorter design complexity. The impact of filters can be minimized if the in-band frequency response and the in-band group delay of the filters are flat.
3. Delay compensation. As the signal propagates through the transmitter, over the air and then arrives at the receiver, there is a non-negligible propagation delay between the data transmitted and the data received. An input/output delay matching block, or the synchronization block is introduced to estimate and compensate for the delay mismatch.
4. Scaling of the received signal. At the receiver, a power scaling block is often necessary to scale the PA output power level to the input power level of the predistorter. The analog-to-digital converter (ADC) has limited dynamic range, so the automatic gain control (AGC) is applied to fully utilize the ADC dynamic range. In addition, if the power level of the input and the output to the predistorter estimation block are different, the predistorter estimation algorithm will find the difference and compensate for it in predistorter coefficients. The resulting predistorter has a non-unity gain, which is undesirable from the transmitter power control point of view.

10.4.2 Fixed-point challenges

The predistortion estimation is rather computationally intensive. Compared with floating-point, fixed-point representation and implementation are very attractive in terms of power consumption and hardware complexity. At the same time, the finite

precision associated with the fixed-point implementation needs to be carefully examined to avoid precision loss or even numerical instability.

For example, let us rewrite the memoryless polynomial model (10.13) in the vector form:

$$z = \Phi a, \quad (10.50)$$

where

$$z = [z(0), z(1), \dots, z(N-1)]^T, \quad (10.51)$$

$$\phi_k = [\phi_k(0), \phi_k(1), \dots, \phi_k(N-1)]^T, \quad (10.52)$$

$$\Phi = [\phi_0, \phi_1, \dots, \phi_{K-1}], \quad (10.53)$$

$$a = [a_1, a_3, \dots, a_{2K-1}]^T, \quad (10.54)$$

and

$$\phi_k(n) = |y(n)|^{2k} y(n). \quad (10.55)$$

The least squares solution for (10.50) is given by

$$\hat{a} = (\Phi^H \Phi)^{-1} \Phi^H z. \quad (10.56)$$

The (k, l) th element of the matrix $(\Phi^H \Phi)$ is given by $\phi_k^* \phi_l$. If the sample size N is large, the sample average approaches its mean value. We obtain:

$$\begin{aligned} \phi_k^H \phi_l &= \sum_{n=0}^{N-1} |y(n)|^{2k} y^*(n) |y(n)|^{2l} y(n) \\ &= \sum_{n=0}^{N-1} |y(n)|^{2k+2l+2} \\ &\approx NE[|y|^{2k+2l+2}]. \end{aligned} \quad (10.57)$$

From (10.57), we know that each element of matrix $(\Phi^H \Phi)$ depends on the amplitude of the input signal. Define the condition number of a matrix $\rho = |\lambda_{\max}/\lambda_{\min}|$, where $\lambda_{\max}$ and $\lambda_{\min}$ are largest and smallest eigenvalues of the matrix, respectively. Elements in the matrix $(\Phi^H \Phi)$ are highly correlated and the condition number grows exponentially with the increase of the nonlinear order K . Figure 10.13 shows that the condition number of $E[\Phi^H \Phi]$ increases exponentially with the increase of nonlinear order K . The input signal amplitude is uniformly distributed within $[0, 1]$. Similar results are observed for signals that have other distributions.

In the fixed-point implementation, when the condition number is much greater than 1, the numerical error involved in the matrix inversion can be significant [47]. For a matrix with condition number greater than 10^5 (when $K \geq 4$), it is difficult to invert the matrix in fixed-point (with a reasonable finite word length) accurately.

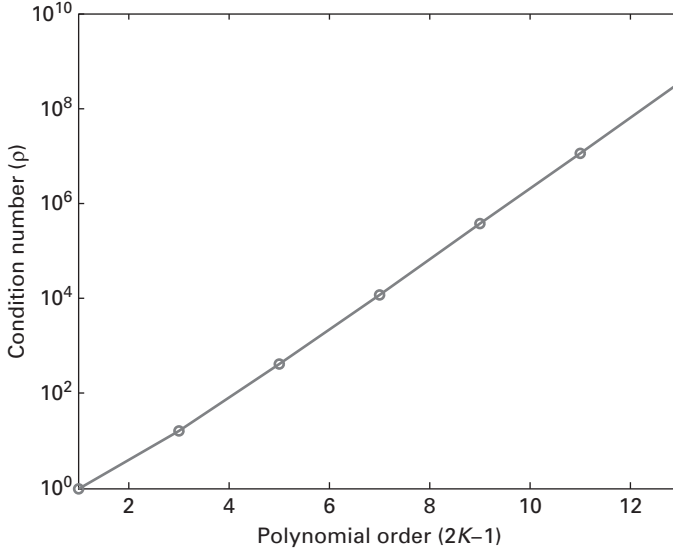


Figure 10.13 Condition number of $(\Phi^H \Phi)$ as a function of polynomial order. The input signal amplitude is uniformly distributed within $[0,1]$.

An orthogonal polynomial is introduced [48, 49] to improve the numerical stability of the model by greatly reducing the condition number of the target matrix. Let us define a new set of basis functions as

$$\begin{aligned}\psi_k(y(n)) &= \sum_{q=0}^k \beta_q \phi_q(y(n)) \\ &= \sum_{q=0}^k \beta_q |y(n)|^{2q} y(n),\end{aligned}\tag{10.58}$$

where coefficients β_q is determined by the orthogonality requirement [48]:

$$E[\psi_k^*(y) \psi_l(y)] = \begin{cases} E[\phi_k^*(y) \phi_k(y)] = E[|y|^{4k+2}], & \text{if } k = l, \\ 0, & \text{otherwise.} \end{cases}\tag{10.59}$$

The new orthogonal polynomial predistorter has the following form

$$z = \Psi \alpha,\tag{10.60}$$

where

$$z = [z(0), z(1), \dots, z(N-1)]^T,\tag{10.61}$$

$$\psi_k = [\psi_k(0), \psi_k(1), \dots, \psi_k(N-1)]^T,\tag{10.62}$$

$$\Psi = [\psi_0, \psi_1, \dots, \psi_{K-1}],\tag{10.63}$$

$$\alpha = [\alpha_1, \alpha_3, \dots, \alpha_{2K-1}]^T. \quad (10.64)$$

The data matrix involved in the predistorter coefficients estimation $E[\Psi^H \Psi]$ becomes a diagonal matrix. The condition number of this matrix is well controlled. The orthogonal polynomial is defined statistically. If the distribution of the input amplitude signal changes, the set of orthogonal polynomial basis functions changes. In [49], a set of orthogonal polynomials is obtained in closed form for the input signal with complex Gaussian distribution:

$$\psi_k(n) = \sum_{q=0}^k (-1)^{m-k} \frac{\sqrt{m+1} m!}{(k+1)! k! (m-k)!} \phi_q(n). \quad (10.65)$$

Similarly, in [48], a set of orthogonal polynomials is obtained in closed form for the input signal whose amplitude is uniformly distributed within $[0,1]$. In reality, the exact distribution of the input signal amplitude may not be available. For different realization, the matrix $(\Psi^H \Psi)$ may not converge to its statistical mean. The exact orthogonality of the matrix is not guaranteed. However, the orthogonal polynomial basis sets provided by [48, 49] can still help to reduce the condition number of the matrix $(\Psi^H \Psi)$ and thus improve the numerical stability. For models of the memory polynomial type, the orthogonal polynomials also help to improve the numerical stability of the model.

The orthogonal polynomials are particularly suitable for fixed-point implementation. The real-time computation of $\phi_k(y)$ can consume a lot of hardware resources involving multiple multiplications and additions. In addition, it may suffer from limited precision. In the actual implementation, $\phi_k(y)$ is realized with a one-dimensional LUT with $|y|$ as its index. The orthogonal polynomial set $\psi_k(y)$, on the other hand, can be realized with the same hardware. The least mean square algorithm is a favorable algorithm in the actual implementation for its low complexity and sample-by-sample processing structure. However, for the predistortion application, different kernels of the Volterra series are highly correlated, thereby resulting in a poorly conditioned covariance matrix. The convergence speed and accuracy of the LMS algorithm are limited in the fixed-point implementation. Orthogonal polynomials can also be applied to reduce the correlation between different input data vectors. The convergence speed and the accuracy of the LMS algorithm can be greatly improved.

The benefit of orthogonal polynomials in LMS algorithm can be demonstrated in the following simulation example. The PA is assumed to be quasi-memoryless following model 8 with $K = 3$ and

$$\begin{aligned} a_1 &= 0.05859375 - j0.80468750, \\ a_3 &= 0.11328125 - j0.00781250, \\ a_5 &= 0.01562500 - j0.01953125, \end{aligned} \quad (10.66)$$

which were extracted from an actual class AB power amplifier. The baseband input signal was an 802.11g WLAN signal with 20MHz bandwidth sampled at 160MHz. The total length of the input signal was 3×10^5 . At the output of the PA, a -25 dBc white noise was added to account for noise and errors in the receive data path and errors in the sampling

process. The LMS algorithm with the conventional polynomials and the LMS algorithm with the orthogonal polynomials were applied to estimate the coefficients of the associated predistorters. To evaluate the performance of predistorters, intermediate predistorter coefficients at each step were recorded. These intermediate results were applied to the DPD + PA system and the PA output y_{est} was obtained for the entire sequence, which was then compared with the linearly amplified PA output y_{lin} . The performance metric, normalized mean square error (NMSE), is defined as

$$\text{NMSE} = 10 \log_{10} \left(\frac{\sum_{n=0}^{N-1} |y_{est}(n) - y_{lin}(n)|^2}{\sum_{n=0}^{N-1} |y_{lin}(n)|^2} \right). \quad (10.67)$$

Smaller NMSE indicates that the predistorter is closer to ideal.

Figure 10.14 shows the NMSE of the above two algorithms at different number of iterations (number of samples processed for predistorter estimation). The NMSE converges to a high level of around -16dB for the LMS algorithm with conventional polynomials, indicating a poor predistorter modeling accuracy. The conventional LMS algorithm does not provide a satisfactory result in DPD parameter estimation. On the other hand, the NMSE for the LMS algorithm with orthogonal polynomials converges to a low level of -43 dB. This result shows the effectiveness of the LMS algorithm with orthogonal polynomials.

Figure 10.15 shows NMSE of RLS and LMS algorithms at different number of iterations. Orthogonal polynomials are applied in both cases. Compared to LMS, RLS shows a clear advantage in the convergence speed and the performance in NMSE. However, the improved performance comes at the cost of a significant increase in the computational complexity. The actual selection of predistorter estimation algorithm depends on target applications.

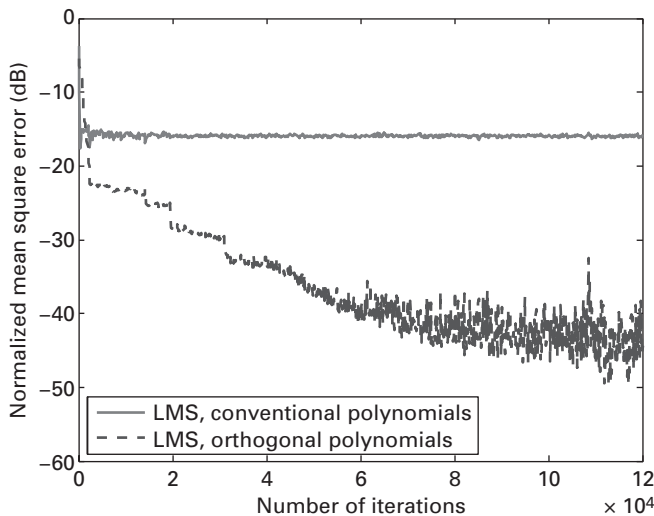


Figure 10.14 Speed of convergence of LMS algorithm with conventional polynomials and LMS algorithm with orthogonal polynomials.

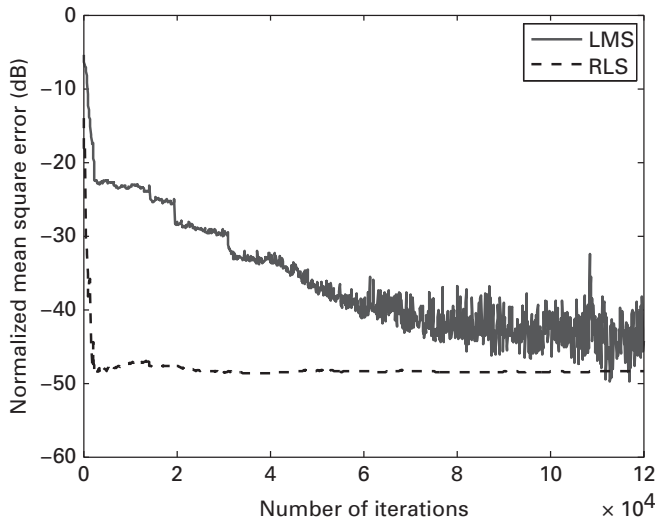


Figure 10.15 Speed of convergence of RLS algorithm with orthogonal polynomials and LMS algorithm with orthogonal polynomials.

10.4.3 Implementation considerations

The purpose of digital predistortion linearization of the PA is to improve the PA efficiency and reduce the overall power consumption of the system. The predistorter and the coefficients estimation, on the other hand, consume power. For example, a floating-point digital signal processor (DSP) can easily consume 0.5W of power. For a basestation PA, 0.5W of power can be considered to be a small fraction to the saved power that is in the order of hundreds of watts. However, for a PA in a mobile device, 0.5W of power may be even higher than the power saved and is no longer negligible. We need to discuss these two scenarios separately.

For the basestation PA, the average transmitted power is in the order of 15 to 300 Watts. The input to the basestation PA may also have wide bandwidth when transmitting modern communication signals such as WCDMA, Worldwide Interoperability for Microwave Access (WiMAX), or Digital Video Broadcasting-Terrestrial (DVB-T) signals. The input signal bandwidth can be even wider when the PA is transmitting a multi-carrier signal. The nonlinear memory effects are significant in the basestation PA case. In this case, the power consumption from the predistorter and coefficients estimation is very small compared with the saved power. The predistortion performance is of primary concern. Sophisticated predistortion algorithms with memory effects can be applied. More expensive DSP or field programmable gate array (FPGA) with floating-point computation capability can be used for better estimation accuracy. The dedicated receive signal path ensures good linearity and noise suppression at the receiver. Reference [50] gave a digital predistortion solution based on DSP implementation.

On the other hand, for a PA in a mobile device, the average transmitted power is in the range of 0.2 to 2 Watts. The PA only costs \$0.2 to \$10. The benefit of applying DPD is marginal. Design trade-offs need to be carefully considered to achieve the overall power

saving and cost reduction. In this case, DSP or FGPA are no longer a good choice for the predistortion implementation. Application specific integrated circuits (ASICs) implementation for the digital predistortion has certain advantages in terms of power consumption and cost. Afsahi *et al.* [51] reported a predistortion implementation for power amplifiers integrated in an SoC (system on a chip) for the wireless local area networks (WLAN) applications.

10.5 Conclusions

In this chapter, nonlinear models are studied to represent the PA as well as the digital predistorter. Nonlinear models with memory structure are introduced to account for the memory effects. The general nonlinear model with memory, the Volterra series, and its simplified versions are of particular interest, since the output is linear w.r.t. the model coefficients in these models and classical optimization tools such as the least squares algorithm can be applied to estimate model coefficients. To make the DPD feasible for real-time implementation, we have presented techniques to further reduce the computational complexity.

With a predistortion block in the baseband, we are able to drive the PA into the nonlinear region without sacrificing the overall system linearity, thereby significantly increasing the power efficiency of the system.

One area that has not been well explored is the joint optimization of the PA design and the DPD design. Conventionally, the PA is designed and evaluated as a standalone unit. Additional power is consumed to make the PA itself much linear. With joint optimization of the PA design and the DPD design, we may obtain more power savings than with conventional PA design.

References

- [1] G. Zhang, S. Khesbak, A. Agarwal, and S. Chin, "Evolution of RFIC handset PAs," *IEEE Microwave Magazine*, pp. 60–69, Feb. 2010.
- [2] P. B. Kenington, *High-Linearity RF Amplifier Design*. Boston, MA: Artech House, 2000.
- [3] S. C. Cripps, *Advanced Techniques in RF Power Amplifier Design*. Norwood, MA: Artech House, 2002.
- [4] L. Sundström, Digital RF Power Amplifier Linearisers. Ph.D. thesis, Lund University, Lund, Sweden, 1995.
- [5] J. L. Dawson and T. H. Lee, *Feedback Linearization of RF Power Amplifiers*. New York: Springer, 2004.
- [6] B. Razavi, *RF Microelectronics*. Upper Saddle River, NY: Prentice Hall, 1997.
- [7] D. Cox, "Linear amplification by sampling techniques: A new application for delta coders," *IEEE Transactions on Communications*, vol. 23, pp. 793–798, Aug. 1975.
- [8] F. H. Raab, P. Asbeck, S. Cripps, *et al.*, "Power amplifiers and transmitters for RF and microwave," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, pp. 814–826, March 2002.

- [9] K. S. Kundert, "Accurate and rapid measurement of IP2 and IP3." Website, 2002. www.designers-guide.org/analysis/intercept-point.pdf.
- [10] K. Chang, I. Bahl, and V. Nair, *RF and Microwave Circuits and Component Design for Wireless Systems*. New York: John Wiley & Sons, 2002.
- [11] S. Benedetto and E. Biglieri, *Principles of Digital Transmission with Wireless Applications*. Dordrecht: Kluwer Academic Publishers, July 1999.
- [12] S. A. Maas, *Nonlinear Microwave Circuits*. Piscataway, NJ: IEEE Press, 1997.
- [13] S. C. Cripps, *RF Power Amplifiers for Wireless Communications*. Norwood, MA: Artech House, 1999.
- [14] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, 3rd edn., 1995.
- [15] G. T. Zhou, H. Qian, L. Ding, and R. Raich, "On the baseband representation of a bandpass nonlinearity," *IEEE Transactions on Signal Processing*, vol. **53**, pp. 2953–2957, Aug. 2005.
- [16] R. Raich and G. T. Zhou, "On the modeling of memory nonlinear effects of power amplifiers for communication applications," in *Proc. the 10th IEEE DSP Workshop*, (Pine Mountain, GA), pp. 7–10, Oct. 2002.
- [17] A. N. D'Andrea, V. Lottici, and R. Reggiannini, "RF power amplifier linearization through amplitude phase predistortion," *IEEE Transactions on Communications*, vol. **44**, pp. 1477–1484, Nov. 1996.
- [18] M. Jin, S. K. Shin, and D. Oh, "A novel predistorter for power amplifier," in *Proc. International Conference on Microwave and Millimeter Wave Technology*, pp. 1129–1133, Aug. 2002.
- [19] T. H. Vo and T. Le-Ngoe, "Baseband predistortion techniques for M-QAM transmission using non-linear power amplifiers," in *Proc. IEEE Vehicular Technology Conference*, vol. **1**, pp. 687–691, Oct. 2003.
- [20] M. Faulkner and M. Johansson, "Adaptive linearization using predistortion – experimental results," *IEEE Transactions on Vehicular Technology*, vol. **43**, pp. 323–332, May 1994.
- [21] S. Benedetto, E. Biglieri, and R. Daffara, "Modeling and performance evaluation of nonlinear satellite links - a Volterra series approach," *IEEE Transactions on Aerospace and Electronic Systems*, vol. **15**, pp. 494–507, July 1979.
- [22] V. Z. Marmarelis, *Nonlinear Dynamic Modeling of Physiological Systems*. Hoboken, NJ: Wiley-IEEE Press, 2004.
- [23] C. Eun and E. J. Powers, "A new Volterra predistorter based on the indirect learning architecture," *IEEE Transactions on Signal Processing*, vol. **45**, pp. 223–227, Jan. 1997.
- [24] C. J. Clark, G. Chrisikos, M. S. Muha, A. A. Moulthrop, and C. P. Silva, "Time-domain envelope measurement technique with application to wideband power amplifier modeling," *IEEE Transactions on Microwave Theory and Techniques*, vol. **46**, pp. 2531–2540, Dec. 1998.
- [25] L. Ding, R. Raich, and G. T. Zhou, "A Hammerstein predistortion linearization design based on the indirect learning architecture," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, vol. **3**, (Orlando, FL), pp. 2689–2692, May 2002.
- [26] H. W. Kang, Y. S. Cho, and D. H. Youn, "On compensating nonlinear distortions of an OFDM system using an efficient adaptive predistorter," *IEEE Transactions on Communications*, vol. **47**, pp. 522–526, April 1999.
- [27] S. Benedetto and E. Biglieri, "Nonlinear equalization of digital satellite channels," *IEEE Journal on Selected Areas in Communications*, vol. **SAC-1**, pp. 57–62, Jan. 1983.
- [28] W. Bosch and G. Gatti, "Measurements and simulation of memory effects in predistortion linearizers," *IEEE Transactions on Microwave Theory and Techniques*, vol. **37**, pp. 1885–1890, Dec. 1989.

- [29] J. Vuolevi, T. Rahkonen, and J. Manninen, "Measurement technique for characterizing memory effects in RF power amplifiers," *IEEE Transactions on Microwave Theory and Techniques*, vol. **49**, pp. 1383–1388, Aug. 2001.
- [30] J. Pedro, N. Carvalho, and P. Lavrador, "Modeling nonlinear behavior of band-pass memory-less and dynamic systems," in *IEEE MTT-S International Microwave Symposium Digest*, (Philadelphia, PA), pp. 2133–2136, June 2003.
- [31] P. Asbeck, H. Kobayashi, M. Iwamoto, *et al.* "Augmented behavioral characterization for modeling the nonlinear response of power amplifiers," in *IEEE MTT-S International Microwave Symposium Digest*, (Seattle, WA), pp. 135–138, June 2002.
- [32] L. Ding, G. T. Zhou, D. R. Morgan, *et al.*, "A robust predistorter constructed using memory polynomials," *IEEE Transactions on Communications*, vol. **52**, pp. 159–165, Jan. 2004.
- [33] J. Kim and K. Konstantinou, "Digital predistortion of wideband signals based on power amplifier model with memory," *Electronic Letters*, vol. **37**, pp. 1417–1418, Nov. 2001.
- [34] D. R. Morgan, Z. Ma, J. Kim, M. G. Zierdt, and P. John, "A generalized memory polynomial model for digital predistortion of RF power amplifiers," *IEEE Transactions on Signal Processing*, vol. **54**, pp. 3852–3860, Oct. 2006.
- [35] C. Eun and E. J. Powers, "A predistorter design for a memory-less nonlinearity preceded by a dynamic linear system," in *Proc. Global Telecommunications Conference*, vol. **1**, (Singapore), pp. 152–156, Nov. 1995.
- [36] S. Chang and E. J. Powers, "A simplified predistorter for compensatoin of nonlinear distortion in OFDM systems," in *Proc. IEEE Global Telecommunications Conference, 2001*, vol. **5**, (San Antonio, TX), pp. 3080–3084, Nov. 2001.
- [37] V. J. Mathews, "Adaptive polynomial filters," *IEEE Signal Processing Magazine*, pp. 10–26, July 1991.
- [38] V. J. Mathews and G. L. Sicuranza, *Polynomial Signal Processing*. New York: John Wiley & Sons, 2000.
- [39] A. M. Saleh, "Frequency-independent and frequency-dependent nonlinear models of TWT amplifiers," *IEEE Transactions on Communications*, vol. **29**, pp. 1715–1720, Nov. 1981.
- [40] S. Kusunoki, K. Yamamoto, T. Hatsugai, *et al.*, "Power-amplifier module with digital adaptive predistortion for cellular phones," *IEEE Transactions on Microwave Theory and Techniques*, vol. **50**, pp. 2979–2986, Dec. 2002.
- [41] F. Zavosh, M. Thomas, T. H. C. Thron, *et al.*, "Digital predistortion techniques for RF power amplifiers with CDMA applications," *Microwave Journal*, pp. 22–30, Oct. 1999.
- [42] T. Kubo, N. Fudaba, H. Ishikawa, *et al.*, "A highly efficient adaptive digital predistortion amplifier for IMT-2000 base stations," in *Proc. IEEE Vehicular Technology Conference*, pp. 2206–2210, April 2003.
- [43] S. Boumaiza and F. M. Ghannouchi, "Realistic power-amplifiers characterization with application to baseband digital predistortion for 3G base stations," *IEEE Transactions on Microwave Theory and Techniques*, vol. **50**, pp. 3016–3021, Dec. 2002.
- [44] W. J. Jung, W. R. Kim, K. M. Kim, and K. B. Lee, "Digital predistorter using multiple lookup tables," *Electronics Letters*, vol. **39**, pp. 1386–1388, Sept. 2003.
- [45] H. Qian, L. Ding, G. T. Zhou, and J. S. Kenney, "Predistortion linearization measurement results for power amplifiers with memory effects," in *Proc. IEEE Topical Workshop on Power Amplifiers for Wireless Communications*, (San Diego, CA), Sept. 2004.
- [46] S. P. Stapleton and J. K. Cavers, "A new technique for adaptation of linearizing predistorters," in *Proc. IEEE Vehicular Technology Conference*, (St. Louis, MO), pp. 753–758, May 1991.

- [47] P. Salmela, A. Happonen, A. Burian, and J. Takala, "Several approaches to fixed-point implementation of matrix inversion," in *Proc. International Symposium on Signals, Circuits and Systems*, vol. **2**, pp. 497–500, July 2005.
- [48] R. Raich, H. Qian, and G. T. Zhou, "Orthogonal polynomials for power amplifier modeling and predistorter design," *IEEE Transactions on Vehicular Technology*, vol. **53**, pp. 1468–1479, Sept. 2004.
- [49] R. Raich and G. T. Zhou, "Orthogonal polynomials for complex Gaussian processes," *IEEE Transactions on Signal Processing*, vol. **52**, pp. 2788–2797, Oct. 2004.
- [50] L. Ding, H. Qian, N. Chen, and G. T. Zhou, "A memory polynomial predistorter implemented using TMS320C67xx," in *Proc. TI Developer Conference 2004*, (Houston, TX), pp. 1–5, Feb. 2004.
- [51] A. Afsahi, A. Behzad, V. Magoon, and L. E. Larson, "Linearized dual-band power amplifiers with integrated baluns in 65 nm CMOS for a 2×2 802.11n MIMO WLAN SoC," *IEEE Journal of Solid-State Circuits*, vol. **45**, pp. 955–966, May 2010.

11 Crest factor reduction techniques

Nuno Borges Carvalho and Wohoon Jang

11.1 Introduction

Signal-to-noise ratio (SNR) is the principal figure of merit for any electronic system. The relation between the signal power and the noise power has been the main challenge for electronic engineers.

In order to study and optimize this type of system, the system design engineers have looked mainly to the noise optimization, that is, to the minimization of all sorts of noise in communication systems. This noise arises mainly from thermal noise [1], nonlinear distortion noise [2], and/or quantization noise in digital systems [3].

Thus the correct identification of the noise contributions is of fundamental importance to the calculation of the noise budget. This is one of the reasons why most of the communication engineers start by worrying about the time domain waveform characteristics of electronic systems. Mainly what they have called the peak of the signal versus its average value, or what is normally known as the peak-to-average ratio (PAR) [4] or crest factor (CF) [5].

Actually the first known reference to this problem, as far as we are aware, was made by Landon in 1936 [1], when he was studying the noise characteristics. A more intensive work [6] was also published by the same author in 1941, related to the distribution of amplitude with noise.

Several other authors have considered that problem since then and in fact in 1964 an instrument called the “PAR Meter” was even proposed by Andersson and Favin [4] for measuring the PAR in loudspeaker tests. What they were searching for was mainly the distortion generated at the loud speakers by the value of the peak-to-average ratio. At that time they proposed a factor of merit called the PAR Rating, that is a relation between the peak value, P , divided by the full-wave average of the received signal, A , and the corresponding values of the undistorted signal, P_0 and A_0 , as expressed in the following.

$$PARRating(R) = 2 \frac{P/A}{P_0/A_0} - 1 \quad (11.1)$$

This rating has a value of unity when the signal is undistorted and it tends to zero if the signal is somehow compressed.

This work has been studied in depth by other authors [7, 8]. The PAR problem continues then to be studied, mainly due to the fact that some signals behave strangely and present some values of peak power to average power that are quite high.

This fact was recognized by the design engineers of radar, sonar and many other applications, which has also driven the mathematician Schroeder [9] to publish a first study on the arrangements of phases angles combinations in periodic signals. In that paper [9] he proposed a phase arrangement specially tailored for multi-sines. This phase arrangement algorithm is called the Schroeder Phase Algorithm [9]. In order to understand the Schroeder algorithm, consider a periodic signal of the form:

$$r(t) = \sum_{k=1}^N \left(\frac{p_k}{2}\right)^{1/2} \cos\left(\frac{2\pi kt}{T} + \theta_k\right). \quad (11.2)$$

The Schroeder phases should be:

$$\phi_n = \phi_1 - 2\pi \sum_{l=1}^{n-1} p_l n = 1, 2, \dots, N \quad (11.3)$$

Schroeder proved that this phase arrangement algorithm is better than an in-phase arrangement of the periodic signal in terms of peak value.

This started the research work on algorithms to minimize the PAR, via a careful choice of the phase arrangements in a periodic signal. In fact in [10] Greenstein *et al.*, proposed an algorithm based on an iterative method to obtain those phases, either at baseband, or at RF, showing that there are differences between the two approaches. This was then further improved by Boyd in [5] where he has defined the Crest Factor, as:

$$CF(u) \equiv \frac{\|u\|_{\infty}}{\|u\|_2} \quad (11.4)$$

where

$$\begin{aligned} \|u\|_{\infty} &\equiv \sup |u(t)| \\ \|u\|_2 &\equiv \left(\int_0^T (u(t)^2) dt\right)^{1/2}. \end{aligned} \quad (11.5)$$

In this work he states that a signal has a CF of one only if the signal is squared. He then proves that different arrangements of the phase between signal harmonics present some good results, such as the Rudin–Shapiro Signs, and the Newman phases.

An important aspect of [5] is that the minimization of the phases is considered in the framework of modeling and for nonparametric measurements of the frequency response of linear systems. So it is a fundamental requirement that a low PAR signal be generated in order not to distort the measured signal, and only capture the linear behaviour of the device under test.

In this line of study, Van Den Bos [11] proposes a new iterative algorithm for CF minimization, comparing the obtained results with the ones of Schroeder, stating that the new method achieves a low value of CF.

In Ouderaa *et al.* [12] there are some comments on the paper from Boyd stating that the correct formula for measuring the CF should be:

$$K_r = \frac{M^+ - M^-}{2 \|u\|_2} \quad (11.6)$$

where M^+ and M^- are the largest positive and the largest negative value of a multi-sine. The same authors then propose in [13] a new type of algorithm for the CF minimization, based on an iterative method that commutes between time and frequency for minimizing the peak value of the time waveform.

The same group later published a large amount of improvements to those methods of CF minimization, mainly and fundamentally in order to extract the underlying linear model for a complex system. In those scenarios the minimization of the CF is fundamental in order to not excite the nonlinear behaviour of those systems. In that respect the selection and design of excitation signals with constant signal-to-noise ratios [14], with minimum crest factors for MIMO systems [15], etc. were studied.

Other scientific groups have also studied this phenomenon, like Evans [2], where specially tailored test signals with reduced CF are designed for the identification of the best linear model in systems presenting nonlinear distortion.

The CF minimization has been a very important line of study since the end of the 1980s and beginning of the 1990s.

In 1992, some authors started to give more importance to the fact that complex digital modulated signals present large values of peak-to-average ratios, and that could degrade the communication path [3, 16]. Since in the presence of high values of PAR, the efficiency of the Power Amplifiers (PAs) were extremely degraded, and the dynamic range of DAC/ADC increase enormously when the value of PAR is high.

This fact separates the study of the PAR in two directions. One is completely devoted to system identification, which continues the previous works of Schouckens and Evans, being mainly in the Instrumentation and Measurement Journals. The other direction is related to communications systems, both wired like x-DSL and wireless systems, mainly those based on multi-carrier modulations, such as OFDM (Orthogonal Frequency Division Multiplexing) or MC-CDMA (Multi-carrier CDMA).

These two areas of study, despite focusing on different scenarios, somehow are complementary, since they both seek the minimization of the time domain waveform peaks.

The rest of this chapter is organized into four sections. Section 11.2 is devoted to multi-carrier communication systems. In Section 11.3, we will focus on CF reduction using clipping plus filtering. In the final section, we will provide a summary and further discussions.

11.2 Multi-carrier communication systems

When the time domain waveforms in the communication signals start to approximate the multi-sine shape usually used in ULS measurement instruments, the problem of CF, or peak-to-average power ratio (PAPR), as it is known in the communication world, starts to attract a greater attention.

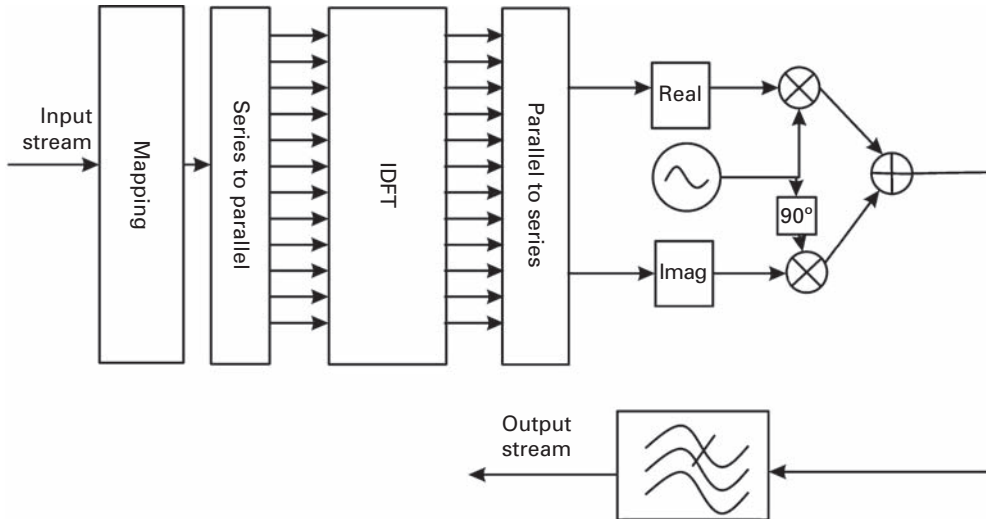


Figure 11.1 OFDM system blocks.

If we look back, the problem in this case is similar to the system identification area, since most of the high values of PAPR arise in multi-carrier communication signals, as in OFDM, or xDSL systems.

An OFDM system can be represented as in [Figure 11.1](#). As can be seen from the figure, the time domain waveform is nothing more than a summation of sines at different frequencies for each time instant, so the problem to address for PAPR reduction is exactly the same as in the system identification case. In fact this problem starts to be first discussed, in relation to communication systems, in [\[16\]](#) where a method for measuring the PAPR of Digital TV was presented and the value of those PAPR values compared for different modulation formats. At that time this problem was considered very important, since it imposed more care in the design of linear PA's and an increase of the DAC/ADC dynamic range.

Related to this dynamic range problem, reference [\[3\]](#) presents some requirements of the A/D converter in QAM receivers, where some guidelines for the selection of the appropriate bit-precision of the A/D converter is given for QAM modulated signals when they present high values of PAPR. At this time the problem also starts to have industrial significance, with some measuring companies presenting some application notes referring to the problem of PAPR [\[17\]](#), and how to measure it.

The increase in PAPR in multi-carrier systems becomes so complex that all the possible schemes for its minimization have become a goal for wireless system design engineers.

The approaches for that minimization span from special coding for PAPR minimization [\[18\]](#), tone reservation [\[19\]](#), tone injection [\[20\]](#), clipping [\[21\]](#), clipping followed by filtering [\[22\]](#), change of the constellation diagram [\[23\]](#) to others based on representation of the signals to transmit [\[24\]](#), like partial transmit sequences, PTS, interleaving and selected mapping SLM, etc.

Recently the study has centered on not on how to minimize the PAPR exclusively, but on how to have a good compromise among the minimum PAPR, bit-rate, and BER optimization [25].

11.2.1 Definition of the relation between peak power and average power

In order to better understand the different approaches followed, let us now first define the diverse definitions on the relation between the peaks and the average power of the time domain waveforms.

Some authors [26] normally used PAPR as the ratio of the power that would result if the envelope were sustained at its peak magnitude to the average power of the overall signal. The PAPR has thus the mathematical form of:

$$PAPR = \frac{\max_{0 \leq t \leq NT} |x(t)|^2}{1/NT \int_0^{NT} |x(t)|^2 dt}. \quad (11.7)$$

A version of the PAPR for sampled signal can also be used and is defined as:

$$PAPR_s = \frac{\max_{0 \leq k \leq NL-1} |x_k|^2}{E[|x(k)|^2]}. \quad (11.8)$$

Despite there being some controversy about the relation of the sampled PAPR to its continuous version, the overall discussion states that we are allowed to choose either one to use. Although one could be different from the other, or at least the minimization of one could not mean the minimization of the other [27, 28], they all converge to a solution where the PAPR will be in fact minimized when compared to the previous results.

Other authors have also studied and proposed new forms of PAPR specially tailored for multi-carrier communication systems. One of those examples is what was called by Wulich [25], the Efficient PAPR, where the measurement of the PAPR is not only regarded as a reduction of the peak power, but as an optimum relation of the PAPR that minimizes the BER.

This way a better value of compromise could be obtained among system complexity, peak reduction, bit-rate optimization, and BER values.

11.2.2 PAPR/CF reduction techniques

After the observation that a high value of PAPR can appear in multi-carrier systems, and thus degrade the SNDR, the scientific community started to search for ways to minimize the PAPR. A very large amount of work has been undertaken on this.

In [29] Seung made a very good overview of the PAPR reduction techniques that were available to that date for multi-carrier transmission. The techniques used span from hardware based schemes to software exclusively based ones. These include Active Constellation Extension [23, 30–33], Tone Injection [20, 34], and Tone Reservation [19, 35–37].

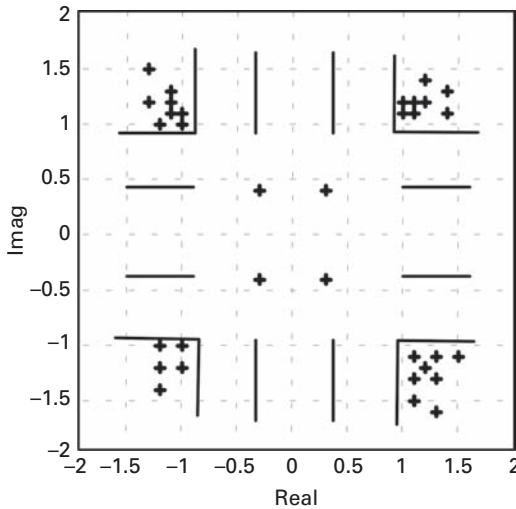


Figure 11.2 ACE scheme.

One obvious solution is Amplitude Clipping and all its combinations with other techniques although it has some major problems as pointed out in [21, 22, 38–43]. More recently also Companding techniques [44–48] was considered.

In addition to these, we also have reduction techniques based on the multiple signal representation such as Partial Transmit Sequences, PTS [21, 49–52], Selective Mapping [49, 53, 54] and Interleaving [55, 56]. This list will not be complete without referring to Coding [18, 57–63] and other more orthodox techniques such as Symbol Shaping [64, 65].

We will now briefly address those different techniques individually.

Active Constellation Extension

In the Active Constellation Extension (ACE) scheme [23, 30–33], the main idea is to change the constellation diagram in order to reduce the PAPR with a low cost impact on the BER. If we look at Figure 11.2, and consider the case of QPSK modulation, the transmitted signal is extended towards the out bounds of the constellation diagram, which can have a slight impact on the BER but it can be proved that has a strong impact on the reduction of PAPR.

What the technique does is nothing more than to transmit a slightly different value of the modulator symbol in order to minimize the PAPR. This changed value should be made at certain boundaries in order that the demodulator could demodulate the signal with low values of BER. Otherwise the BER could be completely degraded due to the reduction of the distance between symbols in the constellation diagram.

Tone Injection

A similar solution to ACE is so-called Tone Injection (TI) [20, 34, 66]. In this scheme the main idea is the creation of several constellation diagrams which can be dynamically chosen in order to reduce the PAPR.

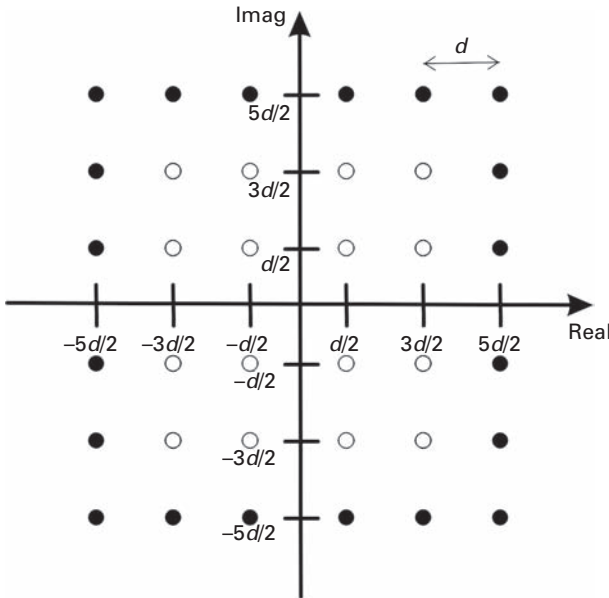


Figure 11.3 TI based on extra constellation diagrams.

In fact the addition of new constellation points in the constellation diagram is the equivalent of injecting new tones. This new constellation point in the augmented constellation could be selected as to maintain the time waveform at a stable PAPR. The idea from an OFDM point of view is nothing more than to add constants C 's to the OFDM symbol, which are carefully selected to reduce the PAPR and not to increase the BER. The effect of the added constant is to increase the constellation magnitude so that each of the points in the original constellation is mapped into the expanded constellation.

The values of C 's are $C_k = p_k D + jq_k D$, p_k and q_k are chosen to minimize the PAPR, and the constant D is known both at the transmitter and the receiver. Figure 11.3 presents this solution for the case of a 16-QAM. In the figure, what we can see is that the black points could also be transmitted, but no new information is added. This means that we can transmit the same digital symbol either using the white points or the black points for the same base information bit, so the modulator could have some redundancy, which is chosen in order to minimize the PAPR.

The main problem is the increase in BER. Nevertheless the augmented capability to reduce PAPR is quite satisfactory.

Tone Reservation

In the case of Tone Reservation [19, 35–37] the underlying idea is to reserve, or say, to select some sub carriers in order that the overall RF signal has a reduced PAPR. In DSL communication systems this is normally done in the low SNR tones, since they will not be very important for the overall signal demodulation. So in this case we will add some information, C , to the un-used tones to reduce the overall PAPR in the time domain scenario.

The unused tones are called the reserved tones and normally do not carry data or they cannot carry data reliably due to their low SNR. It is exactly these tones that are used to send the optimum vector C that was selected to reduce large peak power samples of OFDM symbols. The method is very simple to implement, and the receiver could ignore the symbols carried on the un-used tones, without any complex demodulation process, or extra tail bits.

Amplitude clipping

Amplitude clipping [21, 22, 38–43] is obviously the one that can achieve improved results and is less complex to apply. Nevertheless the clipping increases the occupied bandwidth and simultaneously degrades significantly the in-band and out-of-band distortion, giving rise to the increase of BER, due to its nonlinearity nature.

The technique is based mainly on the following procedure: if the signal is below a certain threshold, then the output is exactly the signal itself. Nevertheless if it passes that threshold then the signal should be clipped as is presented in the next expression.

$$y = \begin{cases} x, & |x| \leq A \\ Ae^{j\phi(x)}, & |x| > A \end{cases} \quad (11.9)$$

where $\phi(x)$ is the phase of x . The main problem of this technique is that somehow we are distorting the signal generating nonlinear distortion both in-band and out-of-band. The in-band distortion cannot be filtered out, and some form of linearizer should be used [67], or other form of reconstruction of the signal prior to the reception block as was done in [68], which somehow has the opposite effect, since one of the reasons to reduce the PAPR is exactly to minimize the impact of the nonlinear PA.

The out-of-band emission, usually called spectral regrowth, can be filtered out but the filtering process will increase again the PAPR. That is why some algorithms are used sequentially with clipping and filtering in order to converge to a minimum value.

This technique can be further associated with other schemes to improve the PAPR overall solution.

Companding techniques

Companding technique [44–48] is a similar technique to clipping, but the signal is not actually clipped, but rather companded or expanded according to its amplitude.

This is probably the oldest technique for reducing PAPR, and actually it was used in the old analog telephone lines, where the voice was companded in order to reduce the need for high dynamic range. Most of the authors have dedicated their time to select the optimum form of the companding function in order to simultaneously reduce the PAPR and improve the BER. Figure 11.4 presents one implementation of these schemes [48].

One possibility for the companding function is the well-known relationship, as expressed in the following:

$$F(x) = \text{sgn}(x) \frac{\ln(1 + u|x|)}{1 + u} \quad -1 \leq x \leq 1. \quad (11.10)$$

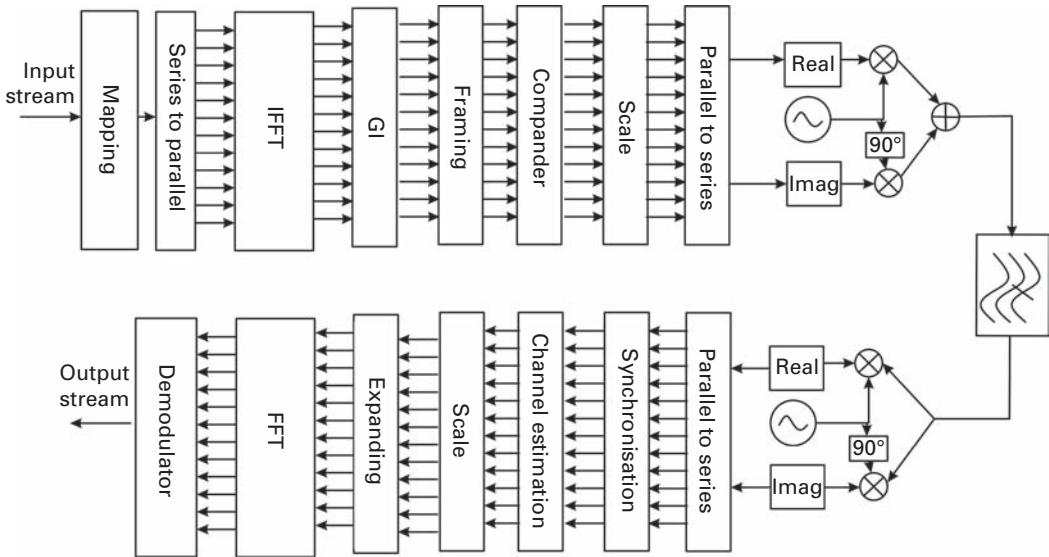


Figure 11.4 Companding implementation.

The drawbacks of this solution are similar to those of the clipping technique, but in this case the nonlinear distortion can be somehow post-distorted at the receiver more efficiently, since the nonlinearity is not as severe as that in the clipping form.

Partial transmit sequences

Other forms of PAPR minimization are related to transmitting only the selected small PAPR waveforms. That can be achieved, for instance, in the Partial Transmit Sequence, PTS [24, 49–52], technique.

In this case the information to be transmitted is first sub-divided into several blocks, [Figure 11.5](#). The blocks are then multiplied with a phase arrangement weighting factor that will allow the reduction of the PAPR. The main objective of the algorithm is to select those B vectors that will minimize the PAPR. The receiving part of the overall system should receive those values in order to demodulate conveniently the information signal. This is one of the drawbacks of this technique since it will need to use an extra carrier to send this information, or an extra overhead, reducing the bit-rate.

Selected mapping

In the select mapping solution [49, 53, 54] the basic idea is to choose from a large amount of possible data blocks and to transmit the one with lowest PAPR. If we look at [Figure 11.6](#), the data information is first replicated into several equal blocks. Those blocks are further multiplied by a B vector, corresponding to different phase sequences.

Then these different blocks are converted to the time domain, by using the IDFT, and the one with lowest PAPR is selected and transmitted. Some side information should also be transmitted to the receiver, in order to demodulate the data sequence.

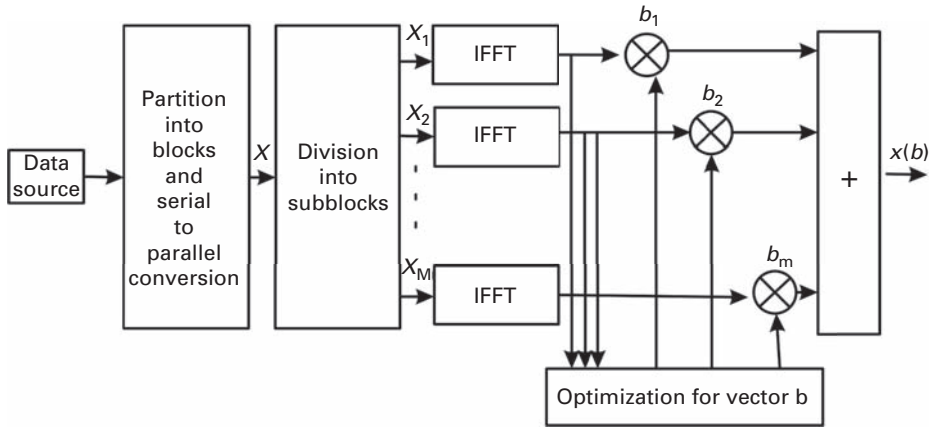


Figure 11.5 PTS technique.

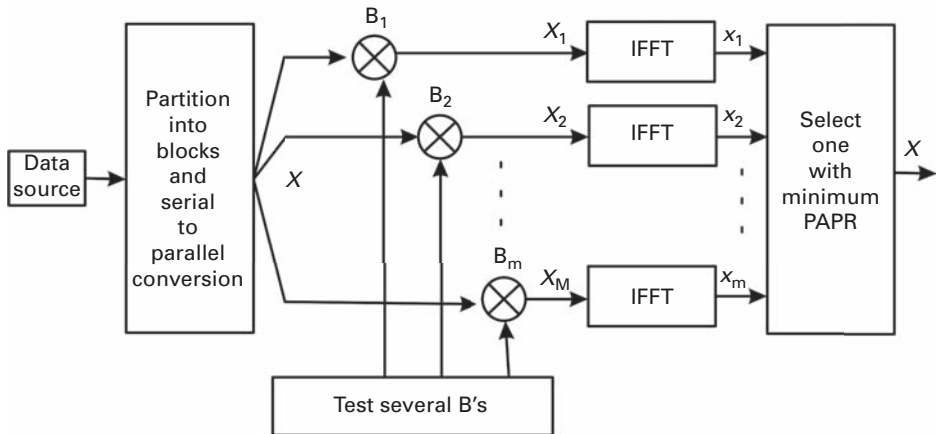


Figure 11.6 SLM technique.

This is one of the drawbacks of this technique since we will lose some bit-rate due to the side information.

Interleaving

In the case of interleaving [55, 56], the approach is similar to the SLM technique, since in this case we should have several interleavers at the transmitter and at the receiver.

The algorithm starts by first replicating the data to be transmitted into several equal blocks, then each one of the interleavers is applied to the data blocks and after the IDFT the one with low PAPR is selected for transmission, Figure 11.7. Again at the receiver we must know which interleaver was selected, so the amount of side information to be sent is on the order of $\log_2(K)$, where K is the number of interleavers used.

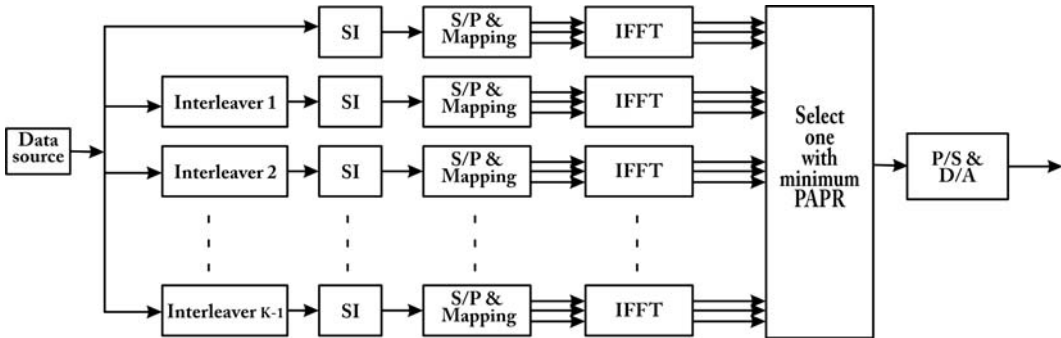


Figure 11.7 Interleaving technique.

Coding

Coding [18, 57–63] is one of the simplest techniques that was used for PAPR reduction. In this case the main idea is to select from a possible set of words to transmit, the one with low PAPR. This fact is achieved by adding to the data words a number of block codes in order to generate a minimum PAPR final word. Consider Figure 11.8, where the PAPR of the 4 word bit is presented. We can see that a maximum of $16W$ PEP (Peak Envelope Power) is achieved.

The main idea of the block coding is to select the sequences that have a higher PAPR and do not consider them for transmission. That can be done, for instance in the case of Figure 11.8, by block coding with encoding the data into 3 bit word and then add an extra fourth bit, so that only the sequences with low PAPR are transmitted.

The main problem with this technique is the large amount of look-up tables needed for coding and decoding of the block coding. Nevertheless this technique could be associated with other coding mechanisms available at the wireless system. For example we can use specific error code sequences to reduce the PAPR, or as some authors present recently we can use only specific Walsh codes in a CDMA system to reduce the PAPR, and do not allow other codes that will degrade the PAPR for that user.

Symbol shaping

More recently some authors have dedicated their work to the study of how the base band filter shaping [64, 65] of the bit pulses affects the PAPR, and how to use that information for minimizing the PAPR.

This proposal is similar to clipping and filtering, since what those authors are proposing is a pre-filter that is able to be designed not for ISI, but for PAPR reduction.

11.3 CF reduction using clipping plus filtering

In order to better understand some of the concepts previously presented, in this section one of the most frequently used techniques for the RF/Wireless system engineers

d_{sec}	d_1	d_2	d_3	d_4	PEP
0	0	0	0	0	$\frac{w}{16.00}$
1	1	0	0	0	7.07
2	0	1	0	0	7.07
3	1	1	0	0	9.45
4	0	0	1	0	7.07
5	1	0	1	0	16.00
6	0	1	1	0	9.45
7	1	1	1	0	7.07
8	0	0	0	1	7.07
9	1	0	0	1	9.45
10	0	1	0	1	16.00
11	1	1	0	1	7.07
12	0	0	1	1	9.45
13	1	0	1	1	7.07
14	0	1	1	1	7.07
15	1	1	1	1	16.00

Figure 11.8 Coding sequence.

will be presented. The technique is called the clipping plus filtering approach, and we will show how to apply it at the baseband waveforms, and at the RF waveforms when in presence of OFDM signals.

The first point that we should account for is that clipping on OFDM signals causes in-band and out-of-band distortions on the frequency domain. The out-of-band distortion can be filtered out but the in-band distortion remains after filtering and it contributes to error vector magnitude (EVM) on the receiver side. The in-band distortion on the constellation diagram can be described as amplitude scaling and scattering around desired constellation points [69]–[70]. Reduction in average amplitude of received symbols around every desired constellation point increases and more scattering occurs as clipping goes intense. Unlike scattering in single carrier systems, it is observed that circle-shaped scattering occurs at each desired constellation point, which means that received symbols are uniformly distributed in angle. Two-dimensional distributions of the received symbols at each desired constellation point are identical and are not Gaussian. As clipping goes deep the distributions come close to Gaussian.

Clipping can be conducted on an envelope or bandpass signal. Other terms for envelope clipping include baseband clipping and soft envelope clipping. The envelope clipping is described in (11.9) and bandpass clipping can be described as

$$y = \begin{cases} x, & |x| \leq A \\ A, & x > A \\ -A, & x < -A \end{cases}. \quad (11.11)$$

In the following subsections both the clipping and filtering techniques are simulated using Matlab and Simulink and then the results are compared.

11.3.1 System models

Model for baseband clipping-and-filtering simulation

A model used for envelope clipping-and-filtering simulation as in Figure 11.9 consists of a transmitter part, a receiver part, and a clipping block in between. At the transmitter side, a random binary signal is mapped into a 16-QAM signal and the QAM signal is changed to a complex envelope signal by zero-padding and IFFT in the OFDM transmitter. There are 203 zeros padded to 53 numbers of constellation vectors in each frame, which results in a frequency bandwidth of around 70 MHz, in order to capture up to fifth-order nonlinear distortion around the baseband. The resulting size of IFFT is 256. The reverse processes are conducted at the receiver side. The clipping-and-filtering block between the transmitter and the receiver makes an output of a clipped-and-filtered complex envelope signal. All blocks in the model are built and simulated in Simulink.

Model for bandpass clipping-and-filtering simulation

A model for bandpass clipping-and-filtering simulation as shown in Figure 11.10 has two parts; a baseband and an RF. The baseband part is the same as in the previous model for envelope clipping-and-filtering simulation. The RF part includes a quadrature modulator, a clipping-and-filtering block, a quadrature demodulator, and a low-pass filter in the order that a signal flows. A complex envelope taken from the output of the OFDM transmitter is up-converted to the center frequency of 80 MHz by the quadrature modulator. The resulting time-domain signal is clipped-and-filtered, down-converted by the quadrature demodulator, and then filtered before the signal reaches the baseband blocks. The

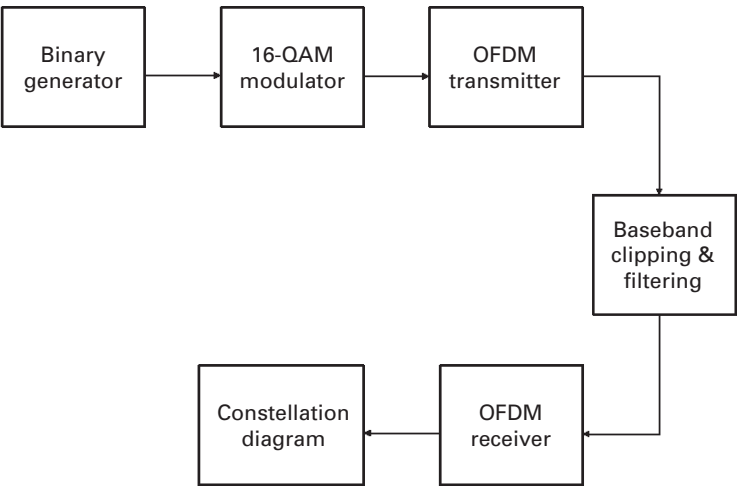


Figure 11.9 A model for simulating envelope-clipping effects.

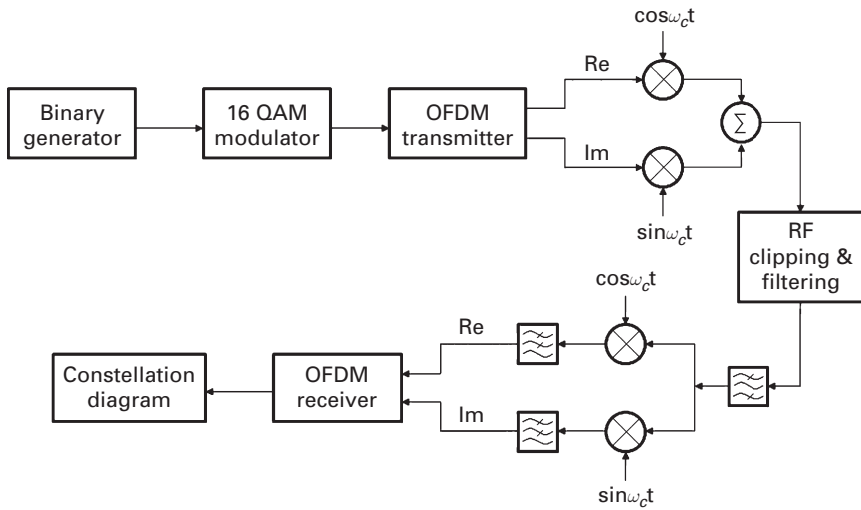


Figure 11.10 A model for simulating bandpass-clipping effects.

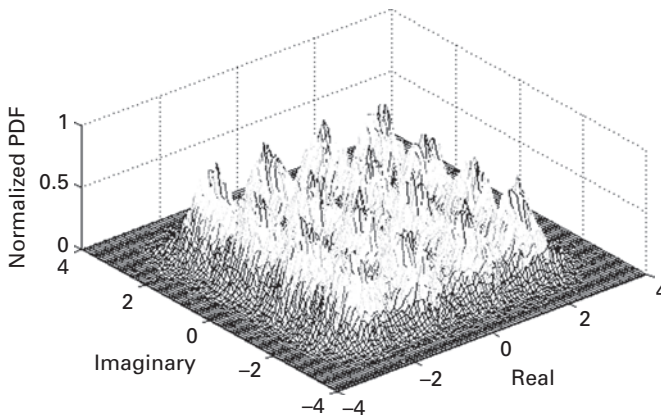


Figure 11.11 Constellation PDF of an envelope-clipped signal.

baseband and RF parts are built and simulated in the Simulink and Matlab environments respectively.

The total frequency bandwidth taken into account is about 760 MHz including the negative and positive frequency bands. This bandwidth is wide enough to capture up to fifth-order nonlinearities.

11.3.2 Simulation results and comparison

In each simulation 250 000 symbols are transmitted and received. [Figures 11.11](#) and [11.12](#) show the resulting constellation diagrams at the receiver side from envelope and

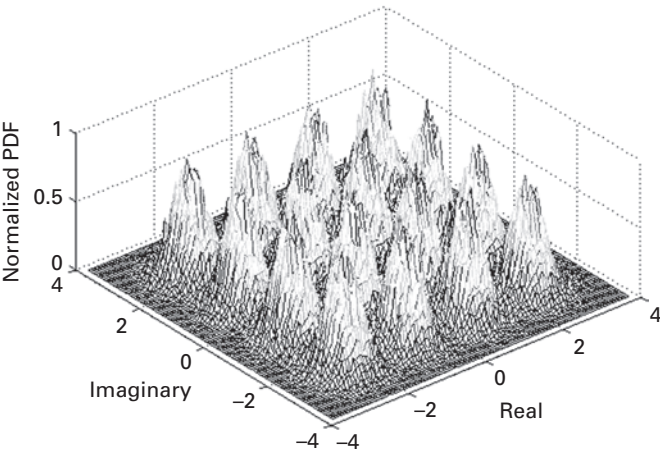


Figure 11.12 Constellation PDF of a bandpass-clipped signal.

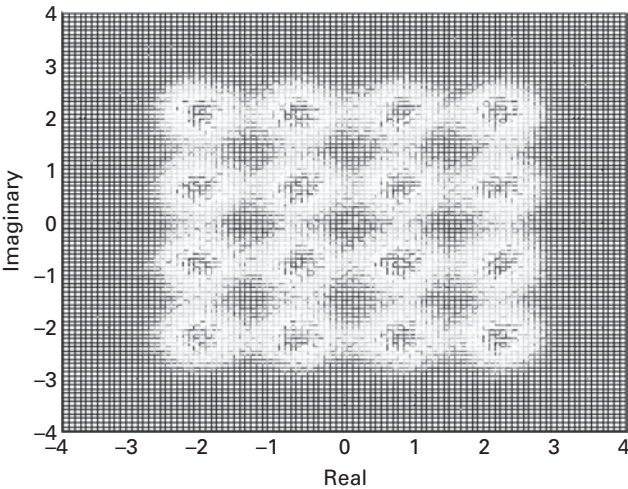


Figure 11.13 Top view of Figure 11.11.

bandpass clipping-and-filtering simulations respectively. Comparing these two figures, symbols in Figure 11.11 are scattered more around the centers of each constellation groups than those in Figure 11.12. This is observed more clearly in Figure 11.13 and Figure 11.14 which are respectively top views of Figure 11.11 and Figure 11.12. It is also observed that the centers of each constellation groups in Figure 11.13 moved more toward the origin than those in Figure 11.14. This means that at the same level of clipping, envelope clipping-and-filtering creates more distortion and BER than bandpass clipping-and-filtering.

In order to observe the difference between the two techniques in terms of in-band and out-of-band distortions, clipped only signals without being filtered are compared.

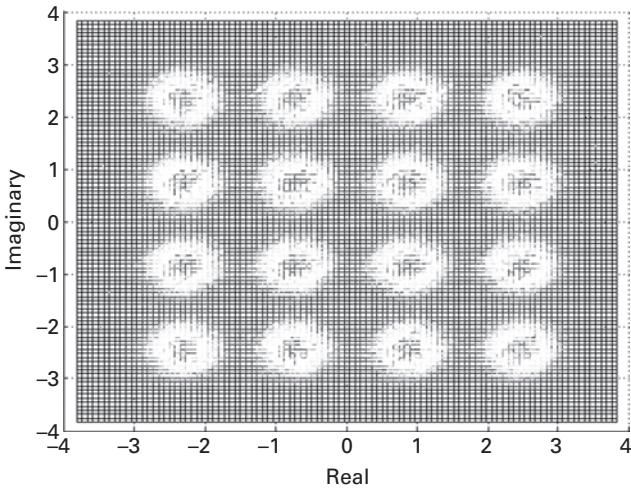


Figure 11.14 Top view of Figure 11.12.

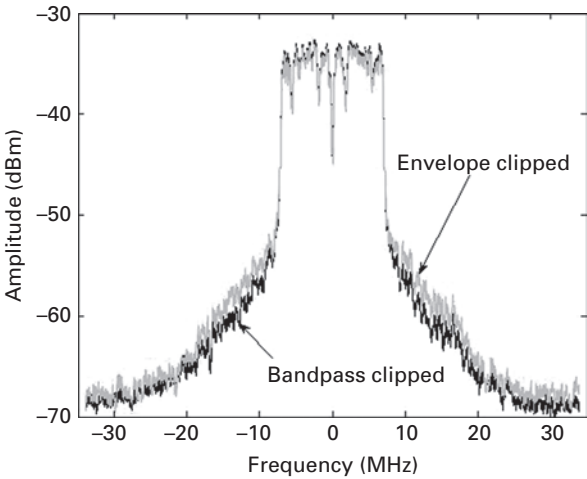


Figure 11.15 Frequency spectra of envelope- and bandpass-clipped signals with the same clipping level.

Frequency spectra of the clipped signals are shown in Figure 11.15. The envelope-clipped spectrum has more in-band suppression than the bandpass-clipped spectrum by about 0.5 dB, which is expected from the constellation diagrams in Figure 11.13 and Figure 11.14.

The spectrum of envelope clipping also has more spectral regrowth (or out-of-band distortion) than that of bandpass clipping. The difference of the spectral regrowths is about 2 dB.

So far, it seems that bandpass clipping-and-filtering has more advantages over the envelope-processed version due to lower EVM and spectral regrowth at the same clipping level. However, it can be observed where the “advantages” come from by

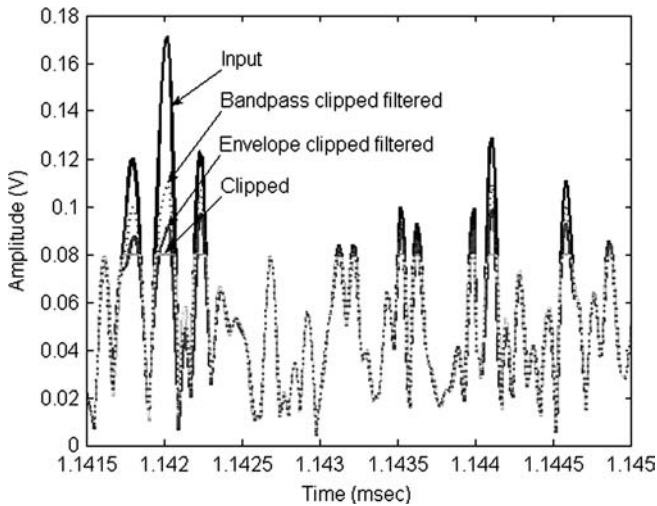


Figure 11.16 Envelope- and bandpass-clipped-and-filtered signals with the same clipping level.

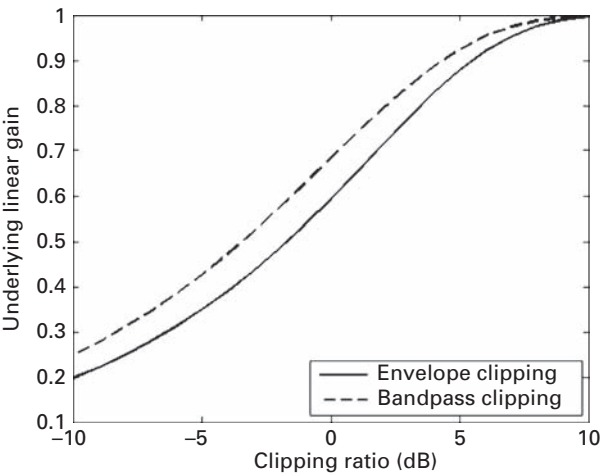


Figure 11.17 Underlying linear gains of envelope and bandpass clipping depending on clipping ratios.

comparing time-domain signals after clipping-and-filtering. As shown in Figure 11.16, the bandpass-clipped-and-filtered signal has higher peak values (or PAPR) than the envelope processed signal does even though both signals are clipped at the same level.

This can be explained in the same concept as underlying linear gain of a memoryless system. Given a clipping level, the underlying linear gain of the bandpass-clipped signal is larger than that of the envelope-clipped signal [71]–[73]. Figure 11.17 shows underlying linear gains of envelope and bandpass clipping depending on clipping ratios. Therefore, a lower EVM of the bandpass clipping results in the cost of a higher PAPR.

11.4 Conclusion

Recently multiband-multi-carrier systems have been attracting more attention as future candidates for communications. One of the foreseeable issues with the systems is the high PAPR problem. In the chapter we have revisited various techniques for CFR to address the problem, and briefly explained how they work and what are the disadvantages. Each technique has different positives and negatives in terms of throughput, EVM (or BER), and system complexity so it is difficult to conclude that a specific one is better than another due to trade-offs in performance and cost factors. Therefore, depending on system requirements it is expected that one technique or a combination of the techniques would be chosen in future systems.

Clipping-and-filtering is one of the most simple and intuitive techniques for CFR. An example of clipping plus filtering simulation is included to show how clipping-and-filtering affects frequency spectra and constellations of symbols. Differences between envelope and bandpass clipping-and-filtering are also shown.

Future schemes for CF reduction are mainly concentrated on reduction techniques focusing not on the value of the peak but combining with linearizers in order to optimize the values of BER in transmission chains. Moreover the combination of these techniques with software defined radio approaches is expected to improve the overall transmission quality in future wireless systems.

References

- [1] Landon, V. D. A study of the characteristics of noise, *Proceedings of the IRE* volume **24**, issue 11, Nov. 1936 page(s):1514–1521.
- [2] Evans, D. C., Rees, D., Jones, D. L. Design of test signals for identification of linear systems with nonlinear distortions, *Instrumentation and Measurement, IEEE Transactions on*, volume **41**, issue 6, Dec. 1992 page(s):768–774.
- [3] Guan, Z.-Y., Hulyalkar, S. N. Bit-precision requirements on the A/D converter in a QAM receiver, *Consumer Electronics, IEEE Transactions on*, volume **39**, issue 3, Aug. 1993 page(s): 692–695.
- [4] Anderson, T., Favin, D. Loudspeaker testing by PAR meter method, *Audio, IEEE Transactions on*, volume **12**, issue 5, part 1, Sep 1964 page(s):94–100.
- [5] Boyd, S. Multitone signals with low crest factor, *Circuits and Systems, IEEE Transactions on*, volume **33**, issue 10, Oct. 1986 page(s):1018–1022.
- [6] Landon, V. D. The distribution of amplitude with time in fluctuation noise, *Proceedings of the IRE* volume **29**, issue 2, Feb. 1941 page(s):50–55.
- [7] Campbell, L., Jr. The PAR meter: characteristics of a new voiceband rating system, *Communications, IEEE Transactions on [legacy, pre-1988]*, volume **18**, issue 2, Apr. 1970 page(s):147–153.
- [8] Fennick, J. The PAR meter: applications in telecommunications systems, *Communications, IEEE Transactions on [legacy, pre-1988]*, volume **18**, issue 1, Feb. 1970 page(s):68–73.
- [9] Schroeder, M. Synthesis of low-peak-factor signals and binary sequences with low autocorrelation (Corresp.), *Information Theory, IEEE Transactions on*, volume **16**, issue 1, Jan. 1970 page(s):85–89.

- [10] Greenstein, L., Fitzgerald, P. Phasing multitone signals to minimize peak Factors, *Communications, IEEE Transactions on [legacy, pre-1988]*, volume **29**, issue 7, July 1981 page(s):1072–1074.
- [11] Van Den Bos, A. A new method for synthesis of low-peak-factor signals, *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing]*, *IEEE Transactions on*, volume **35**, issue 1, Jan. 1987 page(s):120–122.
- [12] Van der Ouderaa, E., Schoukens, J., Renneboog, J. Comments on “Multitone Signals with Low Crest Factor,” *Circuits and Systems, IEEE Transactions on*, volume **34**, issue 9, Sep. 1987 page(s):1125–1127.
- [13] Van der Ouderaa, E., Schoukens, J., Renneboog, J.; *Instrumentation and Measurement, IEEE Transactions on*, volume **37**, issue 1, March 1988 Page(s):145–147.
- [14] Guillaume, P., Pintelon, R., Schoukens, J. On the use of signals with a constant signal-to-noise ratio in the frequency domain, *Instrumentation and Measurement, IEEE Transactions on*, volume **39**, issue 6, Dec. 1990 page(s):835–842.
- [15] Guillaume, P., Schoukens, J., Pintelon, R., Kollar, I., Crest-factor minimization using non-linear Chebyshev approximation methods, *Instrumentation and Measurement, IEEE Transactions on*, volume **40**, issue 6, Dec. 1991 page(s):982–989.
- [16] Rhodes, C. W. Measuring peak and average power of digitally modulated advanced television systems, *Broadcasting, IEEE Transactions on*, volume **38**, issue 4, Dec. 1992 page(s): 197–201.
- [17] Meyer, C. J. Measuring the peak-to-average power of digitally modulated signals, Boonton Electronics Application Note AN-50 Apr, 1993.
- [18] Jones, A. E., Wilkinson, T. A., Barton, S. K. Block coding scheme for reduction of peak to mean envelope power ratio of multicarrier transmission schemes, *Electronics Letters* volume **30**, issue 25, 8 Dec. 1994 page(s):2098–2099.
- [19] Krongold, B. S., Jones, D. L. An active-set approach for OFDM PAR reduction via tone reservation, *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, volume **52**, issue 2, Feb. 2004 page(s):495–509.
- [20] Tellado-Mourelo, J., Peak to average power reduction for multicarrier modulation, Ph.D. dissertation, Stanford University, Sept. 1999.
- [21] Wulich, D., Goldfeld, L. Reduction of peak factor in orthogonal multicarrier modulation by amplitude limiting and coding, *Communications, IEEE Transactions on*, volume **47**, issue 1, Jan. 1999 page(s):18–21.
- [22] Armstrong, J. Peak-to-average power reduction for OFDM by repeated clipping and frequency domain filtering, *Electronics Letters* volume **38**, issue 5, 28 Feb. 2002 page(s): 246–247.
- [23] Krongold, B. S., Jones, D. L. PAR reduction in OFDM via active constellation extension, *Broadcasting, IEEE Transactions on*, volume **49**, issue 3, Sept. 2003 page(s): 258–268.
- [24] Muller, S. H., Huber, J. B. OFDM with reduced peak-to-average power ratio by optimum combination of partial transmit sequences, *Electronics Letters* volume **33**, issue 5, 27 Feb. 1997 page(s):368–369.
- [25] Wulich, D. Definition of efficient PAPR in OFDM, *Communications Letters, IEEE*, volume **9**, issue 9, Sep. 2005 page(s):832–834.
- [26] Gimlin, D. R., Patisaul, C. R. On minimizing the peak-to-average power ratio for the sum of N sinusoids, *Communications, IEEE Transactions on*, volume **41**, issue 4, April 1993 page(s):631–635.

- [27] Wulich, D. Comments on the peak factor of sampled and continuous signals, *Communications Letters, IEEE*, volume 4, issue 7, July 2000 page(s):213–214.
- [28] Minn, H., Tellambura, C., Bhargava, V.K. On the peak factors of sampled and continuous signals, *Communications Letters, IEEE*, volume 5, issue 4, April 2001 page(s):129–131.
- [29] S. H. Han, J. H. Lee An overview of peak-to-average power ratio reduction techniques for multicarrier transmission, *Wireless Communications, IEEE [see also IEEE Personal Communications]*, volume 12, issue 2, April 2005 page(s):56–65.
- [30] Kwok, H., Jones, D. PAR reduction via constellation shaping, *Information Theory, 2000. Proceedings. IEEE International Symposium on 25–30 June 2000* page(s):166.
- [31] Kou, Y., Lu, W. S., Antoniou, A. New peak-to-average power-ratio reduction algorithms for multicarrier communications, *Circuits and Systems I: Regular Papers, IEEE Transactions on*, volume 51, issue 9, Sept. 2004 page(s):1790–1800.
- [32] Yang, Z., Fang, H., Pan, C. ACE with frame interleaving scheme to reduce peak-to-average power ratio in OFDM systems, *Broadcasting, IEEE Transactions on*, volume 51, issue 4, Dec. 2005 page(s):571–575.
- [33] Sezginer, S., Sari, H. OFDM peak power reduction with simple amplitude predistortion, *Communications Letters, IEEE*, volume 10, issue 2, Feb. 2006 page(s):65–67.
- [34] Hwang, C.-S. Peak power reduction method for multicarrier transmission, *Electronics Letters* volume 37, issue 17, 16 Aug. 2001 page(s):1075–1077.
- [35] Schmidt, H., Kammeyer, K.-D. Reducing the peak to average power ratio of multicarrier signals by adaptive subcarrier selection, *ICUPC '98. IEEE 1998 International Conference on Universal Personal Communications* volume 2, 5–9 Oct. 1998 page(s):933–937.
- [36] Gatherer, A., Polley, M. O., Redfern, A. J. PAR reduction for DMT systems with unloaded subchannels, *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002* volume 2, 3–6 Nov. 2002 page(s):1847–1851.
- [37] Sumasu, A., Ue, T., Uesugi, M., Kato, O., Homma, K. A method to reduce the peak power with signal space expansion (ESPAR) for OFDM system, *Vehicular Technology Conference Proceedings, VTC 2000-Spring Tokyo*, volume 1, 15–18 May 2000 page(s):405–409.
- [38] Ochiai, H., Imai, H. Performance of the deliberate clipping with adaptive symbol selection for strictly band-limited OFDM systems, *Selected Areas in Communications, IEEE Journal on* Volume 18, Issue 11, Nov. 2000 Page(s):2270–2277.
- [39] Ren, G., Zhang, H., Chang, Y. A complementary clipping transform technique for the reduction of peak-to-average power ratio of OFDM system, *Consumer Electronics, IEEE Transactions on*, volume 49, issue 4, Nov. 2003 page(s):922–926.
- [40] Tarokh, V., Jafarkhani, H. On the computation and reduction of the peak-to-average power ratio in multicarrier communications, *Communications, IEEE Transactions on*, volume 48, issue 1, Jan. 2000 page(s):37–44.
- [41] Vaananen, O., Vankka, J., Viero, T., Halonen, K., Reducing the crest factor of a CDMA downlink signal by adding unused channelization codes, *Communications Letters, IEEE*, volume 6, issue 10, Oct. 2002 page(s):443–445.
- [42] Chow, J. S., Bingham, J. A. C., Flowers, M. S. Mitigating clipping noise in multi-carrier systems, *ICC 97 Montreal, "Towards the Knowledge Millennium," 1997 IEEE International Conference on Communications*, volume 2, 8–12 June 1997 page(s):715–719.
- [43] Raich, R., Qian, H., Zhou, G. T. Optimization of SNDR for amplitude-limited nonlinearities, *Communications, IEEE Transactions on*, volume 53, issue 11, Nov. 2005 page(s):1964–1972.

- [44] Huang, X., Lu, J., Zheng, J., Chuang, J., Gu, J. Reduction of peak-to-average power ratio of OFDM signals with companding transform, *Electronics Letters* volume **37**, issue 8, 12 Apr 2001 page(s):506–507.
- [45] Vaananen, O., Vankka, J., Halonen, K. Reducing the peak to average ratio of multicarrier GSM and edge signals, *The 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 2002*, volume **1**, 15–18 Sept. 2002 Page(s):115–119 vol.1
- [46] Jiang, T., Zhu, G. Nonlinear companding transform for reducing peak-to-average power ratio of OFDM signals, *Broadcasting, IEEE Transactions on*, volume **50**, issue 3, Sept. 2004 page(s): 342–346.
- [47] Huang, X., Lu, J., Zheng, J., Letaief, K. B., Gu, J. Companding transform for reduction in peak-to-average power ratio of OFDM signals, *Wireless Communications, IEEE Transactions on*, volume **3**, issue 6, Nov. 2004 page(s):2030–2039.
- [48] Pratt, T. G., Jones, N., Smee, L., Torrey, M. OFDM link performance with companding for PAPR reduction in the presence of non-linear amplification, *Broadcasting, IEEE Transactions on*, volume **52**, issue 2, June 2006 page(s):261–267
- [49] Jayalath, A. D. S., Tellambura, C. SLM and PTS peak-power reduction of OFDM signals without side information, *Wireless Communications, IEEE Transactions on*, volume **4**, issue 5, Sept. 2005 page(s):2006–2013.
- [50] Peak-to-average ratio reduction algorithm for MCM signal using the selective partial transmit sequence Eun Jung Chang; Ho Yeol Kwon; Intelligent Signal Processing and *ISPACS 2004. Proceedings of 2004 International Symposium on Intelligent Signal Processing and Communication Systems* 18–19 Nov. 2004 page(s):723–727
- [51] Nikookar, H., Lidsheim, K. S. Random phase updating algorithm for OFDM transmission with low PAPR, *Broadcasting, IEEE Transactions on*, volume **48**, issue 2, June 2002 page(s):123–128.
- [52] Ju, S. M., Leung, S. H. Clipping on COFDM with phase on demand, *Communications Letters, IEEE*, volume **7**, issue 2, Feb. 2003 page(s):49–51.
- [53] Friese, M. Multicarrier modulation with low peak-to-average power ratio, *Electronics Letters* volume **32**, issue 8, 11 April 1996 page(s):713–714
- [54] Bauml, R. W., Fischer, R. F. H., Huber, J. B. Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping, *Electronics Letters* volume **32**, issue 22, 24 Oct. 1996 page(s):2056–2057.
- [55] Van Eetvelt, P., Wade, G., Tomlinson, M. Peak to average power reduction for OFDM schemes by selective scrambling, *Electronics Letters* volume **32**, issue 21, 10 Oct. 1996 page(s):1963–1964.
- [56] Choi, K., Kang, K., Kim, S. Peak power reduction scheme based on subcarrier scrambling for MC-CDMA systems, *Communications, IEE Proceedings*, volume **151**, issue 1, Feb 2004 page(s):39–43.
- [57] Shepherd, S. J., Van Eetvelt, P. W. J., Wyatt-Millington, C. W., Barton, S. K. Simple coding scheme to reduce peak factor in QPSK multicarrier modulation, *Electronics Letters* volume **31**, issue 14, 6 July 1995 page(s):1131–1132.
- [58] Gadkari, S., Rose, K. Transmission energy allocation with low peak-to-average ratio, *Communications Letters, IEEE*, volume **1**, issue 6, Nov. 1997 page(s):166–168.
- [59] Henkel, W. Wagner, B. Another application for trellis shaping: PAR reduction for DMT (OFDM), *Communications, IEEE Transactions on*, volume **48**, issue 9, Sept. 2000 page(s): 1471–1476.

- [60] Lau, V.K.N. On the analysis of peak-to-average ratio (PAR) for IS95 and CDMA2000 systems, *Vehicular Technology, IEEE Transactions on*, volume **49**, issue 6, Nov. 2000 page(s): 2174–2188
- [61] Lau, V.K.N. Average of peak-to-average ratio (PAR) of IS95 and CDMA2000 systems-single carrier, *Communications Letters, IEEE*, volume **5**, issue 4, April 2001 page(s):160–162
- [62] Ginige, T., Rajatheva, N., Ahmed, K. M. Dynamic spreading code selection method for PAPR reduction in OFDM-CDMA systems with 4-QAM modulation, *Communications Letters, IEEE*, volume **5**, issue 10, Oct. 2001 page(s):408–410
- [63] Hathi, N. Darwazeh, I., O'Reilly, J. Peak-to-average power ratio performance comparison of different spreading code allocation strategies for MC-CDMA and MC-DS-CDMA, *Electronics Letters* volume **38**, issue 20, 26 Sept. 2002 page(s):1219–1220
- [64] Rha, P. S., Hsu, S. Peak-to-average ratio (PAR) reduction by pulse shaping using a new family of generalized raised cosine filters, *Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th* volume **1**, 6–9 Oct. 2003 page(s):706–710.
- [65] Baas, N. J., Taylor, D. P. Pulse shaping for wireless communication over time- or frequency-selective channels, *Communications, IEEE Transactions on*, volume **52**, issue 9, Sept. 2004 page(s):1477–1479.
- [66] Schurgers, C., Srivastava, M. B., A systematic approach to the peak-to-average power ratio in OFDM, *SPIE's 47th Annual Meeting*, San Diego, CA, pp. 454–464, July 29–August 3, 2001.
- [67] Kennington, P. B., *High-Linearity RF Amplifier Design*, Norwood MA: Artech House, 2000.
- [68] Saeedi, H., Sharif, M., Marvasti, F. Clipping noise cancellation in OFDM systems using oversampled signal reconstruction, *Communications Letters, IEEE*, volume **6**, issue 2, Feb. 2002 page(s):73–75.
- [69] Brajal, A., Chouly, A., Compensation of nonlinear distortions for orthogonal multicarrier schemes using predistortion, in *Proc. IEEE GLOBECOM '94*, San Francisco, CA, Nov. 1994, pp. 1909–1914.
- [70] Santella, G., Mazzenga, F., A hybrid analytical-simulation procedure for performance evaluation in M-QAM-OFDM schemes in presence of nonlinear distortions, *Vehicular Technology, IEEE Transactions on*, volume **47**, issue 1, Feb. 1998 page(s) 142–151.
- [71] Rowe, H., Memoryless nonlinearities with Gaussian inputs: Elementary results, *Bell Systems Technical Journal*, volume **61**, issue 7, Sept. 1982 page(s) 1519–1525.
- [72] Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, 3rd edn. New York: McGraw-Hill, 1991.
- [73] Dardari, D., Tralli, V., Vaccari, A., A theoretical characterization of nonlinear distortion effects in OFDM systems, *Communications, IEEE Transactions on*, volume **48**, issue 10, Oct. 2000 page(s) 1755–1764.

Part III

DUC, DDC, ADC, DAC, and NCO

12 Digital down-conversion and up-conversion: theory and implementation

Yue Zhang, Taiwen Tang, Kok-Keong Loo, Ben Allen, and Dayou Li

Digital conversion is a fundamental part of many digital radio systems, which include up-conversion of the discrete baseband signal stream into a high-resolution radio signal at the transmitter, and down-conversion of a high-resolution radio signal back into a baseband signal at the receiver. In this chapter, we cover the basics of digital conversion (analog to digital and digital to analog), the functionality of the digital up-converter (DUC) and digital down-converter (DDC) in relation to conversion between intermediate frequency (IF) and baseband with the emphasis on the implementation of the DDC and DUC for standard wireless communication systems.

This chapter is organized into six sections. The first section gives an introduction to the digital transceiver and related basic processing blocks. [Section 12.2](#) will present the multi-rate, multi-stage and filter-banks design for a digital front-end. [Section 12.2.1](#) will describe I/Q (In-phase/Quadrature) modulation/demodulation and NCO (numerical control oscillator) design for the DDC and DUC. [Section 12.2.2](#) will present sample rate conversion design in the DDC. The multi-rate, multi-stage filtering and filter bank implementation is covered in this section. [Section 12.2.3](#) will talk about sample rate conversion in the DUC.

[Section 12.3](#) will cover the processing blocks for mixing and combining for multi-channels and multi-carriers design. [Section 12.3.1](#) will present the process for mixing and combining for multi-channels. [Section 12.3.2](#) will present the multimode radio DUC design in current industry standards.

[Section 12.4](#) will describe the hardware implementation for a DDC in wireless communication systems. A reference design for IEEE 802.11n will be presented in this section. [Section 12.4.2](#) will present the process block for DC (Direct Current) offset cancellation. The system design for a DDS (Direct Digital Synthesizer) is in [Section 12.4.3](#). [Section 12.4.4](#) is devoted to the design and implementation for multi-rate decimation filtering. Finally, the performance of DDC design will be described in [Section 12.4.5](#).

[Section 12.5](#) presents the hardware implementation for a DUC in wireless communication systems. [Section 12.5.2](#) is devoted to the process block for multi-rate interpolation filtering design and implementation. [Section 12.5.3](#) will present the performance of DUC design.

Finally, [Section 12.6](#) will give a summary and some further discussions for this chapter.

12.1 Introduction

An overview of the digital radio system is essential in order to introduce the importance of the DDC and DUC.

[Figure 12.1](#) illustrates an example digital radio transceiver presented in a super-heterodyne architecture. This is a dual-stage conversion architecture, in which, at the first state, a radio frequency (RF) signal is down-converted to an intermediate frequency (IF) and then, in the second stage, it is converted from IF to baseband signal. As shown in the figure, the hardware parts in the RF, IF and baseband sections are controllable and reconfigurable by software. This kind of radio transceiver architecture is known as Software Defined Radio (SDR).

In the presented SDR architecture, the RF front-end (also known as analog front-end) is the only analog section, which include mixer, low-noise amplifier (LNA), power amplifier (PA), RF combiner, bandpass filter (anti-aliasing), and antenna. The RF front-end is responsible for conversion between RF and IF, and transmitting/receiving the RF signal. Some advanced RF front-ends allow a certain degree of controllability through software, e.g. frequency tuning.

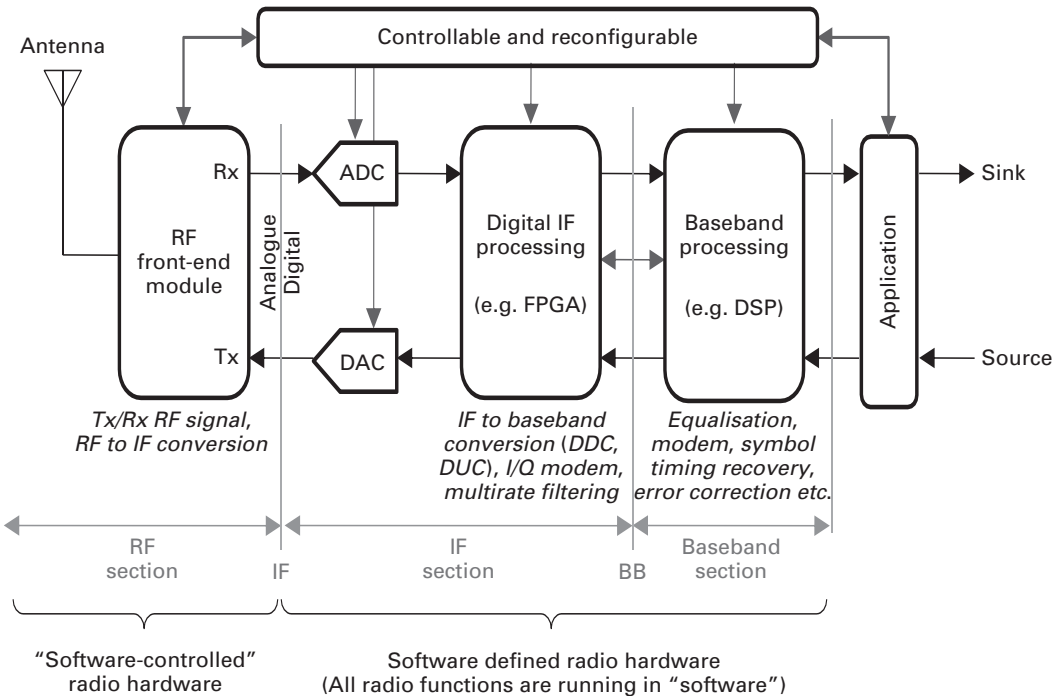


Figure 12.1 Software defined radio block diagram.

The other parts of the SDR architecture are all digital processing components. In the IF section, the sampling and separation of IF carriers, and the up/down frequency conversion to baseband is performed by the digital IF processing normally implemented in FPGA. Take the receiving path as an example; the IF signal is digitized by the analog-to-digital converter (ADC) and converted to a baseband signal by the digital IF processing. The digital IF processing performs DDC operations such as digital synthesizing, digital NCO, digital mixer, I/Q demodulation, multi-rate decimation filtering. Similarly, it performs DUC with the operations in reverse order of DDC. Digital systems such as field programmable gate arrays (FPGAs) are normally used to perform the IF to baseband conversion because they are able to handle the tight real-time constraint due to high-speed sampling and digital conversion. Lastly, in the back-end of the SDR architecture, the baseband process is mainly performing digital communications functions such as the symbol timing recovery, equalization, modulation, channel coding, etc. Those functions are usually computed by digital signal processors (DSPs) with slightly relaxed real-time constraint.

The SDR architecture tends to utilize general-purpose digital systems. Utilizing FPGAs and DSPs in the SDR would allow generation of the transmitted signal and tuning/detection of the received radio signal at the receiver, carried out digitally by software, as opposed to the conventional method of an analog signal going through individual hardware components each performing a specific function. Software defined radio systems offer great extension of programmability, reconfigurability, and definability. The goal of SDR is to realize a radio system that is flexible, versatile, and multi-standard. It is hence future proof for update of new radio functions.

Following the above introduction, this chapter intends to look further into the implementation of digital parts of the SDR, mainly the DDC and DUC with reference to standard wireless communication systems. Prior to that, some practical background related to DDC/DUC are covered in the following subsections, which are not meant to be extensive.

12.1.1 Sampling of radio signals

Radio signals need to be sampled correctly for processing in the DDC/DUC. This subsection discusses practical techniques used in the sampling of radio signals.

12.1.1.1 Bandpass sampling

The Nyquist sampling theorem is a well-covered topic in DSP which states that the sampling frequency, F_s , must be greater than or equal to twice the highest frequency component of a signal, f_{max} (i.e. $F_s \geq 2f_{max}$). Sampling at less than the rate specified by the Nyquist sampling rate leads to folding over or an “aliasing” effect where the original signal cannot be accurately converted back to analog if it is required. This method of converting an analog signal into the digital domain is known as lowpass sampling. The term “lowpass” is associated to this sampling method because all the frequency components from 0 Hz (DC) to F_s are taken into consideration [1].

Lowpass sampling may not be suitable in some applications, such as communication systems. This is because the bandwidth of a narrowband signal, B , is often very small

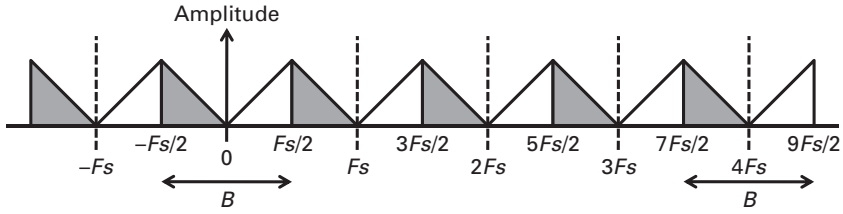


Figure 12.2 Nyquist frequency spectral repetition: a.

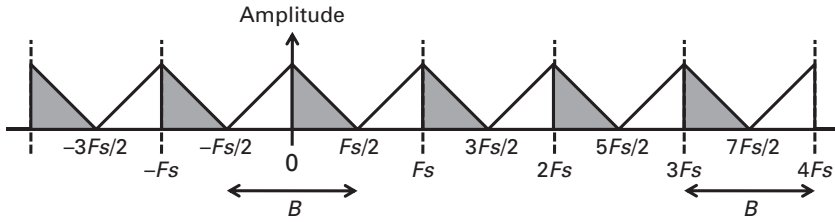


Figure 12.3 Nyquist frequency spectral repetition: b.

compared to the lower and upper band-edge frequencies (f_L and f_H) of the signal, and it is not practical to use lowpass sampling at the rate of $2f_H$ especially, in the case where digital processing hardware cannot handle the data processing at such a high sampling rate.

It is possible to sample the high-frequency narrowband signals at a much reduced sampling rate but to still avoid aliasing [1]. This method is known as bandpass sampling (a.k.a. undersampling which sometimes is called sub-sampling) where the technique takes advantage of the frequency spectral repetition caused by the sampling process as illustrated in Figure 12.2. It can be seen that the narrowband signal, B , repeats itself every multiple of the sampling frequency, F_s .

Using sub-sampling, a narrowband signal can be sampled at a theoretical minimum rate of $2B$ without aliasing. However, the signal band-edge frequencies, f_L and f_H , are required to be integer multiples of the signal bandwidth, B ; for example $n = f_H/B$, where n is an integer number. When this condition is satisfied, the signal band is said to be integer positioned. That said, in the case of a narrowband signal where its center frequency falls exactly on $4F_s$, its upper band-edge frequency, $f_H = 9F_s/2$, satisfies an odd integer positioned at $n = 9$ (see Figure 12.2), thus the signal can be sampled at the rate of $2B$. This shifts the signal from a significantly high frequency i.e. RF/IF to the baseband frequency. As a result, the sampled signal band between DC and $F_s/2$ is in the same frequency spectral order compared to the original signal band between $4F_s$ and $9F_s/2$. With this, the signal band is not aliased and hence the baseband information can be recovered.

In the case of a narrowband signal where its center frequency falls on $7F_s/2$ with its upper band-edge frequency, $f_H = 4F_s$, it satisfies an even integer positioned at $n = 8$ (see Figure 12.3).

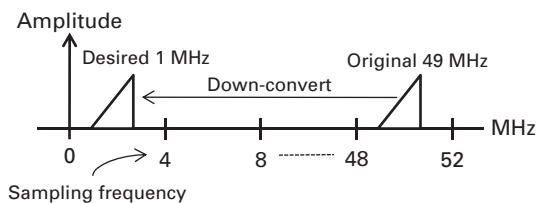


Figure 12.4 Frequency conversion for subsampling.

Therefore, sampling the narrowband signal with the rate of $2B$, produces a sampled signal band between DC and $F_s/2$ in reverse frequency spectral compared to the original signal band between $7F_s/2$ and $4F_s$. Although the sampled signal band is not aliased, the baseband information would have to be recovered correctly through a frequency reversal technique, that is by inverting the quadrature component of the in-phase (I) and quadrature (Q) signals during the sampling process. Note that if a signal band is not integer positioned, the upper band-edge frequency can be extended such that the effective band becomes integer positioned.

In essence, bandpass sampling can also be used as a frequency down-converter. Figure 12.4 shows how the bandpass sampling converts a signal at 49 MHz IF to a desired 1 MHz IF. As seen, the original signal has an offset of 1 MHz from 48 MHz that is a multiple of the sampling frequency (4 MHz). Therefore, applying the principle discussed above, by sampling the signal at 4 MHz, the original signal shifts to DC with an offset of 1 MHz, and achieves the translation to the desired frequency. The same action is carried out by multi-stage multi-rate filtering in DDC.

Based on the discussion so far, it is feasible to use a relatively low-speed ADC to sample the signal at very high frequencies e.g. IF and still recover the baseband information. The bandpass sampling approach enables an all digital implementation of down-conversion at high IF frequency i.e. 40 MHz–250 MHz. In practice, there are a number of additional factors that need to be evaluated, such as the ADC requirements which affect the outcome and overall feasibility to a great extent. Careful analysis of the requirements e.g. dynamic range, dithering, clock jitter, aperture uncertainty, and distortions (SNR, SINAD, SFDR), is required before deciding on an ADC for a particular sub-sampling application. Design issues for a particular application can be found in many ADC device datasheets.

12.1.1.2 Oversampling

Oversampling is the process of sampling a signal with a sampling frequency significantly higher than the Nyquist frequency, usually an integral multiple of the Nyquist frequency. Oversampling the sub-sampled signal would have an advantage of spreading quantization noise energy over a much wider frequency range. This technique is commonly used in the DSP industry to reduce the noise level in the signal band and to overcome some limitations with ADC e.g. extending ADC resolution, reducing noise floor with possible noise shaping so that a lower resolution can be used. Normally, four times oversampling can improve the SNR by 6dB and one more extra bit for the bit width of an ADC.

12.1.1.3 Anti-aliasing filtering

An anti-aliasing filter is commonly used at the input of a digital processing system and during digital IF processing. Analog signals to be digitized may contain frequency components in addition to the folding frequency such as high-frequency interference. Thus, an anti-aliasing filter is used to restrict the signal bandwidth in order to satisfy a sampling requirement, such that all frequency components above the folding frequency are removed. One advantage is the wider separation between the signal and image spectra while meeting the sampling requirement. For signals not centered at DC, the anti-aliasing filter is often a bandpass filter. In that, the filter is normally used for channel selection at RF and/or IF. A lowpass anti-aliasing filter is normally used in digital IF processing especially in digital mixer, harmonic rejection, spectrum masking etc. A general rule of thumb in the design of the anti-aliasing filter is to achieve good performance in terms of passband ripple, phase linearity, group delay and steep transition between passband and stopband, with a minimum of insertion loss.

12.1.2 Frequency conversion

The baseband signal is essentially delivered over an allocated and dedicated RF band which is normally in the region of MHz and GHz. With super-heterodyne radio architecture, it gives more flexibility to digital systems in handling frequency conversion between the baseband and IF. In the digital domain, the frequency conversion performs down-conversion from IF to baseband, and up-conversion from baseband to IF.

The operation of frequency down-conversion includes:

- selecting a target narrowband signal and sampling into the digital domain;
- isolating it out from the wideband source by digital channel filtering;
- converting the selected narrowband signal down in frequency, usually from an IF to baseband;
- reducing the data rate to some integer multiple of the baseband information rate.

The operations of frequency up-conversion include:

- converting a narrowband signal source up in frequency, usually from baseband to IF;
- combining several narrowband signals for up-conversion;
- increasing the data rate to a digital IF rate.

In order to carry out the above operations, a few important signal processing blocks are required, such as a direct digital synthesizer (DDS), mixer, digital filter and multi-rate filter. These are briefly introduced in the following subsections.

12.1.2.1 Direct digital synthesizer

A direct digital synthesizer (DDS), sometime called a numerical control oscillator (NCO), is a simple function that generates orthogonal sinusoidal waves i.e. sine and cosine, which are used for the DDC and DUC, and also digital modulators. A DDS uses

an addressing scheme with a look-up table to create a sinusoid waveform with a desired frequency. In an actual implementation, the look-up table stores numerical representation of sine and cosine waveforms [2]. Depending on the number of bits used in the DDS implementation, the table determines the resolution of the waveform, for example, a 14-bit look-up table will give 16 384 phase increments in the look-up table. A phase accumulator is used to generate the look-up table address for each generated waveform sample. This is done by using a desired frequency clock as a reference and a clock to sample a suitable phase argument that is mapped by the look-up table to the desired output waveform. Further detail in the implementation is discussed in [Section 12.4.3.1](#).

12.1.2.2 Digital mixer

A digital mixer is simply an arithmetic function that multiplies incoming digital source signals, namely I and Q, with carrier waveforms namely sine and cosine generated by an NCO. In a DUC, the digital mixer is used in conjunction with a quadrature modulator to generate a single-sideband IF transmit signal at a desired frequency. Similarly, in a DDC, a single-sideband IF receive signal is fed into the quadrature demodulator in order to differentiate it into I and Q signals.

12.1.2.3 Digital filter

We will see that a finite impulse response (FIR) filter is widely used in the DUC and DDC implementation in contrast to an infinite impulse response (IIR) filter, an FIR filter can achieve an exact linear phase response with unconditional stability in its implementation. This is important to ensure the representation of the output signal after the frequency conversion processing (up/down conversion) which may involve several stages of filtering. It is for this reason that different classes of FIR are used in order to achieve a stable implementation of the DUC and DDC, e.g. Cascaded Integrator-Comb (CIC), Polyphase interpolator and decimator. These classes of FIR filter are often discussed together with multi-stage, multi-rate, and filter-banks which are the topic of the following section

12.2 Multi-rate, multi-stage, and filter-banks in DDC and DUC

A super-heterodyne receiver is a well-known receiver architecture, where an RF signal is converted to an IF frequency, then it is sampled by the ADC and digitally processed by the DDC. At the DDC, I/Q components are extracted and demodulated to the baseband. Due to the sample rate difference at the IF and at the baseband, multi-rate filters are required to implement the sample rate conversion.

Similarly at the transmitter side, the baseband signal is then up-sampled to the IF sampling frequency. The spectrum needs to be translated to the corresponding IF channel by I/Q modulation. After this, the IF signal is DAC-converted and up-converted to the RF frequency.

12.2.1 I/Q demodulation in DDC (I/Q mod. in DUC) and NCO

Wireless signals are usually represented by complex numbers. Thus they require two real components to symbolize this, namely the I (In-phase) component and Q (Quadrature) component. Such signals require special treatment as discussed.

In-phase/Quadrature demodulation in DDC is illustrated in [Figure 12.5](#).

If the conversion rate is an integer, the sample rate conversion module in DDC includes:

- (i) removal of the image signal generated by the digital down-conversion;
- (ii) sample rate conversion by anti-aliasing filtering and down-sampling.

Note that filtering in (i) and (ii) can be jointly designed by using multi-stage or single-stage filtering which provides an equivalent impulse response to perform the image removal and anti-aliasing filtering. The single-stage implementation is illustrated in [Figure 12.6](#). The multi-stage implementation that decouples filtering and down-sampling into multiple stages will be discussed in the [next section](#).

If the conversion rate is a fractional number, the DDC module includes the following functions:

- (i) removal of the image signal generated by the digital down-conversion;
- (ii) sample rate conversion by up-sampling, interpolation/anti-aliasing filtering, and down-sampling.

Note that the filtering operation in (i) and (ii) cannot be designed jointly in this case. We illustrate the single-stage implementation of each filtering operation in [Figure 12.7](#).

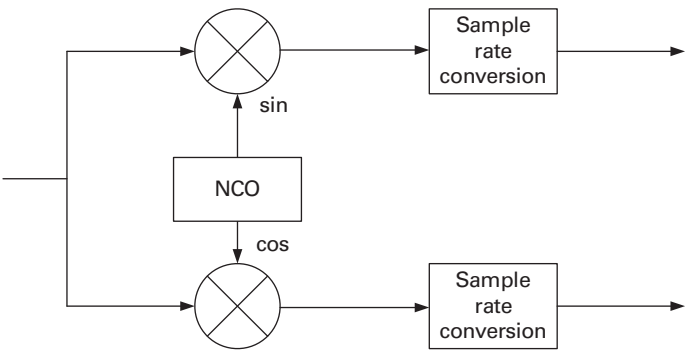


Figure 12.5 The block diagram of a digital I/Q down-converter.



Figure 12.6 Block diagram of the sample rate conversion module in DDC when the conversion rate is an integer.

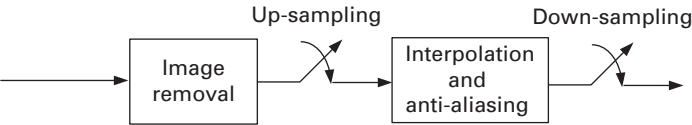


Figure 12.7 Block diagram of the sample rate conversion module in DDC when the conversion rate is a fractional number.

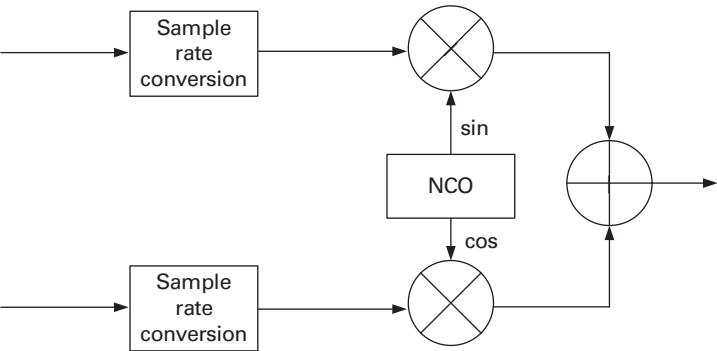


Figure 12.8 The block diagram of a digital I/Q up converter.

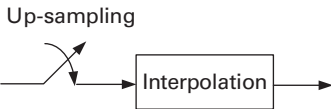


Figure 12.9 Block diagram of the sample rate conversion module in DUC when the conversion rate is an integer.

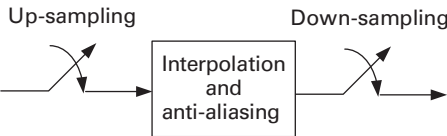


Figure 12.10 Block diagram of the sample rate conversion module in DUC when the conversion rate is a fractional number.

In-phase/Quadrature modulation in DUC is illustrated in [Figure 12.8](#).

The sample rate conversion module in DUC fulfills the up-sampling and interpolation filtering operations when the conversion rate is an integer. We illustrate this operation in [Figure 12.9](#).

If the conversion rate is a fractional number, it can be implemented by up-sampling, interpolation/anti-aliasing filtering, and down-sampling. We illustrate the single-stage implementation of each filtering operation in [Figure 12.10](#).

The NCO structure is shown in Figure 12.11. It consists of a register, an accumulator, and a SIN and COS look-up table (LUT).

12.2.2 Sample rate conversion in DDC

Sample rate conversion in the DDC is discussed in this section. First, multi-stage filtering is explained as polyphase filter bank implementation.

12.2.2.1 Multi-rate and multi-stage filtering

Sample rate conversion from the IF sampling frequency to the baseband sampling frequency requires multi-rate filters. The brute-force method is to design the FIR anti-aliasing filters to meet the sample rate conversion requirements. These filters usually have long impulse responses and are costly to implement on ASIC or FPGA.

Efficient implementation of the multi-rate filter usually decouples the down-sampling and filtering procedure into multiple stages. The following example is provided to motivate the idea of multi-stage filtering.

The brute-force system requires down-sampling of the signal by a factor of four. It is assumed that the desired spectrum lies within $[-\alpha, \alpha]$ of the spectrum in Figure 12.13 (1 stands for the normalized sampling frequency and $\alpha \leq 0.25$). The brute-force design of the anti-aliasing filter requires the transition bandwidth to be $0.25 - \alpha$. However, the two-stage filtering approach requires two filters with transition bandwidths $0.5 - \alpha$ and $0.5 - 2\alpha$. Figure 12.14 illustrates the two-stage filtering approach.

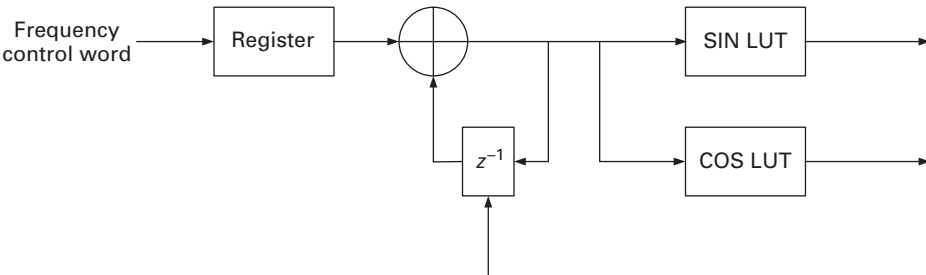


Figure 12.11 The block diagram of numerical controlled oscillator.

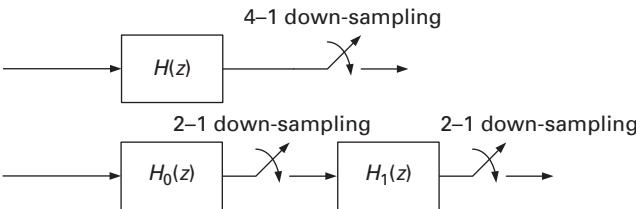


Figure 12.12 The brute-force design versus the two-stage design.

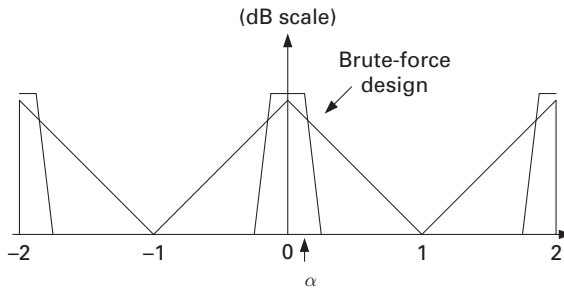


Figure 12.13 The brute-force design of anti-aliasing filtering.

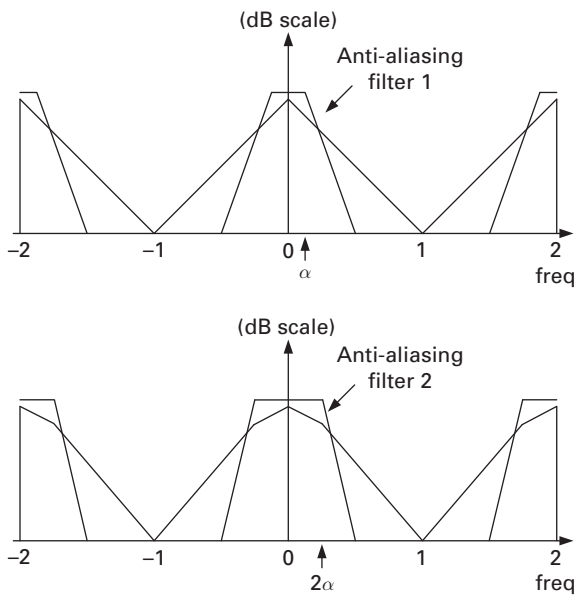


Figure 12.14 The two-stage design of anti-aliasing filtering.

The number of operations per input sample (Ops/Input) of the brute-force design and the two-stage filtering approach are estimated as [3]

$$N \cong K(\alpha) \cdot \frac{Atten(dB)}{22},$$

where $Atten(dB)$ stands for the stopband attenuation in the dB scale. We illustrate the work-load factor, $K(\alpha)$, for the two design approaches in Table 12.1. The reduction in terms of number of operations per input sample is illustrated in Figure 12.15.

The essential idea of multi-stage filtering lies in the fact that the spectrum of the multi-stage filters multiply together, thus in the stopband, the attenuation effects are superimposed by the multi-stage filters. The cost reduction has been seen in the previous example.

Table 12.1 Workload factor of the two design approaches.

	Work-load factor per input sample
Brute-force $K(\alpha)$	$\frac{2}{0.25 - \alpha}$
Two stages $K(\alpha)$	$\frac{2}{0.5 - \alpha} + \frac{1}{0.5 - 2\alpha}$

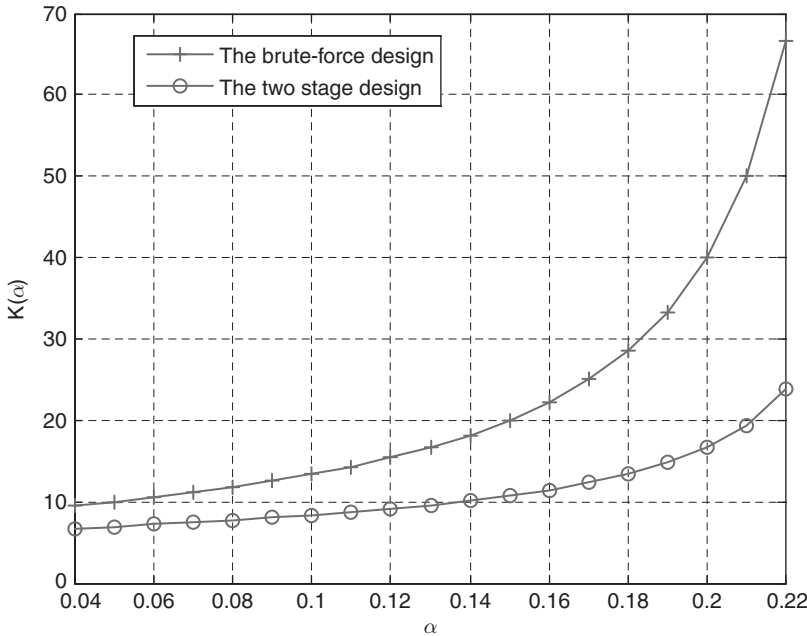


Figure 12.15 Ops/Input comparison of the brute-force design and the two-stage design.

12.2.2.2 Filter bank implementation

Polyphase filter implementation is explained in this section. Note that an anti-aliasing filter is required before down-sampling. In the following example, 1-5 down-sampling is assumed. The anti-aliasing filtering operates according to the following

$$y(n) = \sum_{l=0}^{L-1} h(l)x(n - l), \tag{12.1}$$

where $h(l)$ denotes the l th tap of the time domain channel and L denotes the number of filter taps. The z -transformed channel response $H(z)$ can be written as

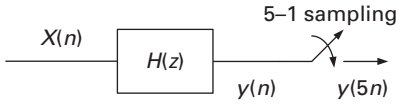


Figure 12.16 5-1 down-sampling and the anti-aliasing filter.

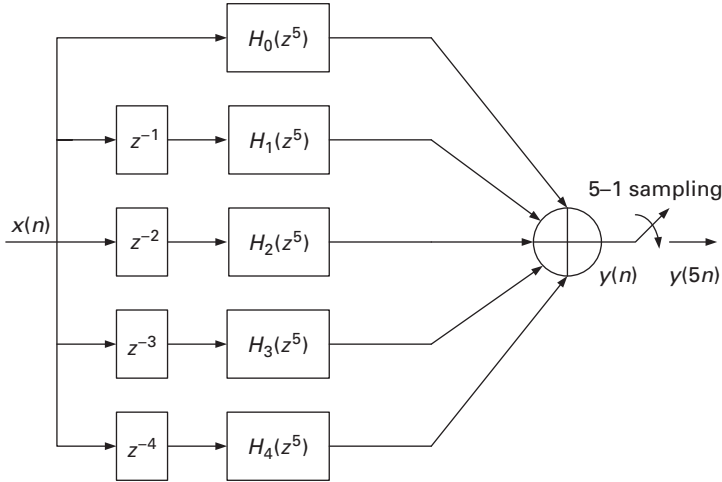


Figure 12.17 5-path representation of the anti-aliasing filter.

$$\begin{aligned}
 H(z) &= \sum_{l=0}^{L-1} h(l)z^{-l} \\
 &= \sum_{k=0}^{\lceil L/5 \rceil - 1} \sum_{r=0}^4 h(r + 5k)z^{-(r+5k)} \\
 &= \sum_{r=0}^4 z^{-r} \sum_{k=0}^{\lceil L/5 \rceil - 1} h(r + 5k)z^{-5k} = \sum_{r=0}^4 z^{-r} H_r(z^5),
 \end{aligned} \tag{12.2}$$

where $\lceil \cdot \rceil$ stands for the ceiling operation for integers. If L is not divisible by 5, zeros are inserted for the channel response $h(l)$, $l = L + 1, \dots, 5\lceil L/5 \rceil - 1$.

Figure 12.16 illustrates the idea of decimation polyphase filters. Note that we can interchange the operation of down-sampling and filtering in each branch based on Figure 12.17. Thus the filtering operations can be done at a lower sampling rate as shown in Figure 12.18 and Figure 12.19.

12.2.3 Sample rate conversion in DUC

Sample rate conversion in the DUC can also be more efficiently implemented using multi-stage filters and polyphase filter banks. The idea of multi-stage filtering for up-

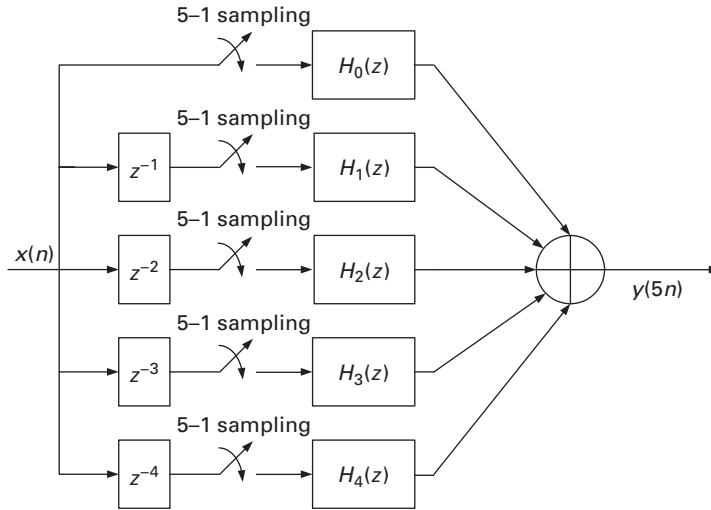


Figure 12.18 5-path polyphase representation of the anti-aliasing filter.

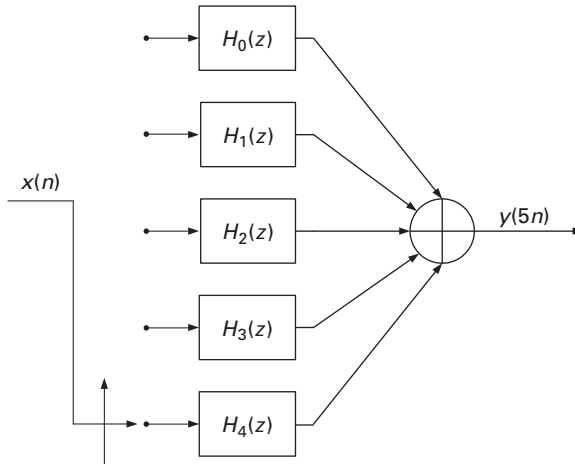


Figure 12.19 5-path polyphase representation with an input commutator.

sampling is motivated in [Chapter 8](#) of [3]. Using multi-stage interpolation filters brings significant cost reduction compared with the brute-force one-stage design.

The main idea of polyphase filtering is illustrated in [figures 12.20](#) and [12.21](#). The derivation is the same as the polyphase filtering for the DDC. Note that we can interchange the up-sampling and filtering for each branch of the polyphase filter. Thus filtering operations can be conducted at a lower sampling rate as shown in [Figure 12.22](#) and [Figure 12.23](#).

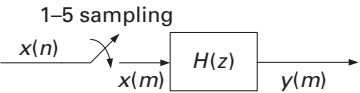


Figure 12.20 1-5 up-sampling and the interpolation filter.

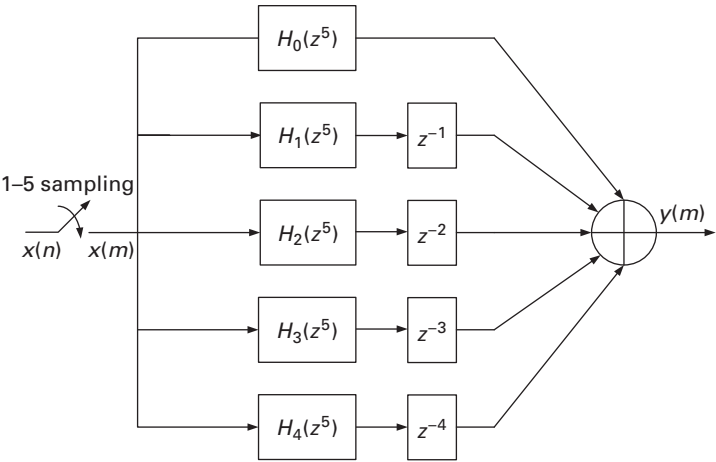


Figure 12.21 5-path representation of the interpolation filter.

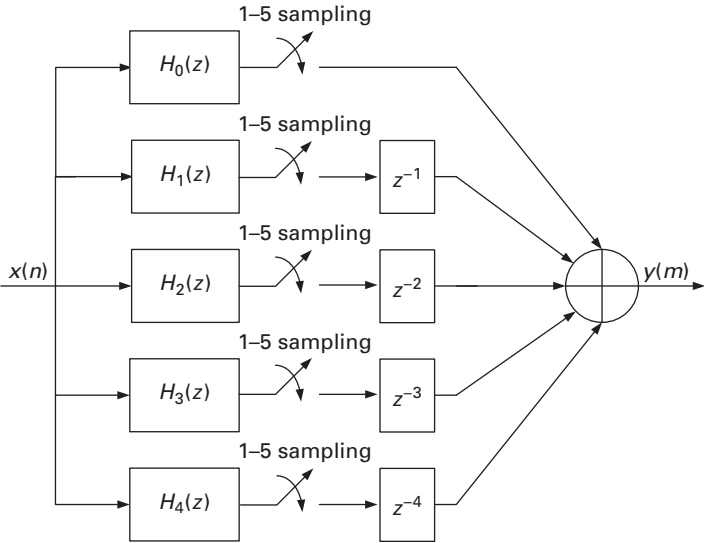


Figure 12.22 5-path polyphase representation of the interpolation filter.

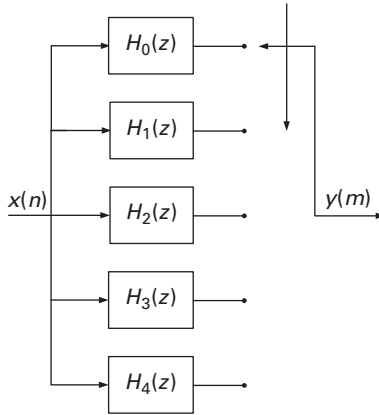


Figure 12.23 5-path polyphase representation with an output commutator.

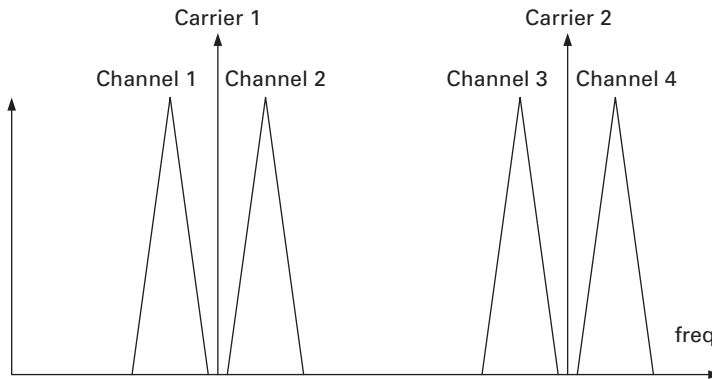


Figure 12.24 Multi-channel and multi-carrier signals of one standard in the frequency domain.

12.3 Mixing and combining for multi-channels and multi-carriers

In this section, we discuss mixing and combining for multi-channels and multi-carriers at the DUC. The DDC operations are similar to those for the DUC and related schemes can directly be generalized for the DDC.

First, it is assumed that each device only consumes one channel of a specific standard. When there are multiple RF carrier frequencies, multiple RF chains are required for every RF carrier. A channel and a carrier have to be selected at the device. The frequency domain signals are illustrated in [Figure 12.24](#) for a two-carrier and four-channel scenario. Though there are different RF carriers, only a single common IF is used. The DUC is designed to select a channel of a specific carrier for mixing and combining. The DUC architecture for this single channel scenario is elaborated in [Section 12.3.1](#).

We assume that each device can operate under multiple modes. Each mode defines a combination of the selected channels of one or multiple standards. For simplicity, the selected channels of each mode are not located across the bands of different RF carriers. The spectrum allocation/frequency plan for the Global System for Mobile Communications (GSM) standard, the Wideband Code Division Multiple Access (WCDMA), and the Long Term Evolution (LTE) standard is illustrated in Figure 12.25. For example, if each mode is restricted to occupy less than or equal to 10 MHz bandwidth, a mode can thus consist of twelve adjacent GSM channels and one 5MHz WCDMA or LTE channel. Similarly, if each mode is restricted to occupy less than or equal to 20 MHz bandwidth, a mode can be defined to take two 10 MHz LTE channels or one 20 MHz LTE channel. The multimode radio DUC is elaborated in Section 12.3.2.

12.3.1 Mixing and combining for multi-channels

Mixing and combining for multi-channel signals is shown in Figure 12.26 for the case of two channels, where F_s is the sampling rate.

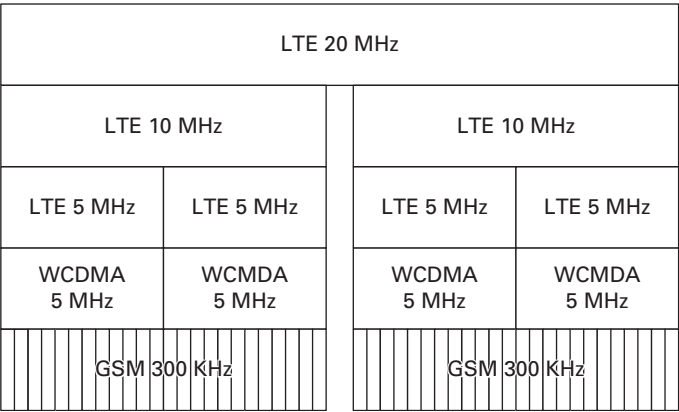


Figure 12.25 Possible spectrum allocation of different standards (referring to Figure 9 of [7]).

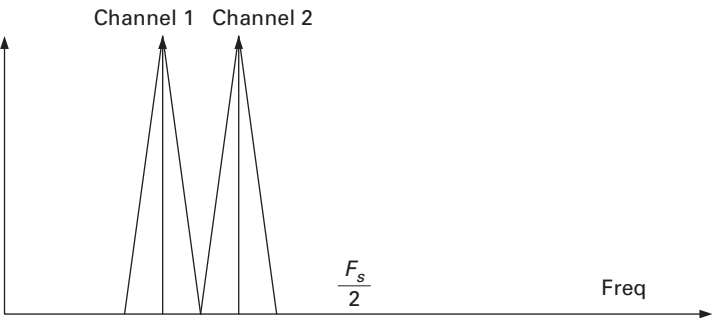


Figure 12.26 Multi-channel signals in the frequency domain.

In the above example, there are two devices which use one channel separately. Mixing and combining for multi-channel signals need to translate the signal to its desired channels in the frequency domain at each DUC device by a programmable multi-channel NCO which is shown in Figure 12.27.

The channel selection is done by configuring a frequency control word to the NCO. The SIN/COS look-up tables need to be designed to include the values corresponding to the phase sequences for the programmable multi-channel NCO.

12.3.2 Multimode radio DUC

We first illustrate a simple spectrum allocation of two standards in Figure 12.28. Standard 1 and standard 2 are both assumed to have two channels. The i th channel of the j th standard is denoted by $C_{i,j}$.

A multimode radio DUC for this spectrum allocation is thus illustrated in Figure 12.29. Thus, by duplicating the hardware for each channel and each standard, a multimode radio DUC can be implemented. The list of available modes is given in Table 12.2.

Note that each NCO in Figure 12.29 has different SIN and COS look-up tables to customize itself to the particular channel. This implementation that consists of parallel blocks, however, is not optimized in hardware because for each standard, the sample rate conversion block should be the same for the I/Q component of each channel. An improved implementation uses serial buses to multiplex the I/Q components of each channel.

Using serial buses, we can time-division multiplex the I and Q signals of each channel. The multiplexing can be represented by the sequence ... I1/Q1 I2/Q2 I1/Q1 I2/Q2 ...,

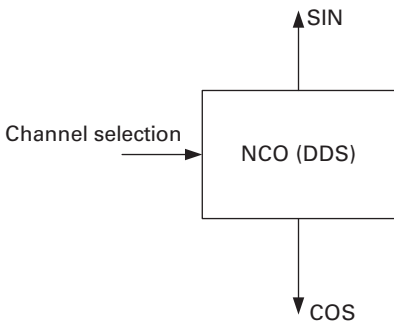


Figure 12.27 A programmable multi-channel NCO.

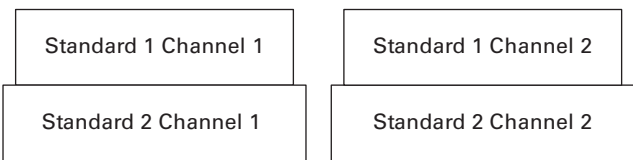


Figure 12.28 A spectrum allocation for two standards.

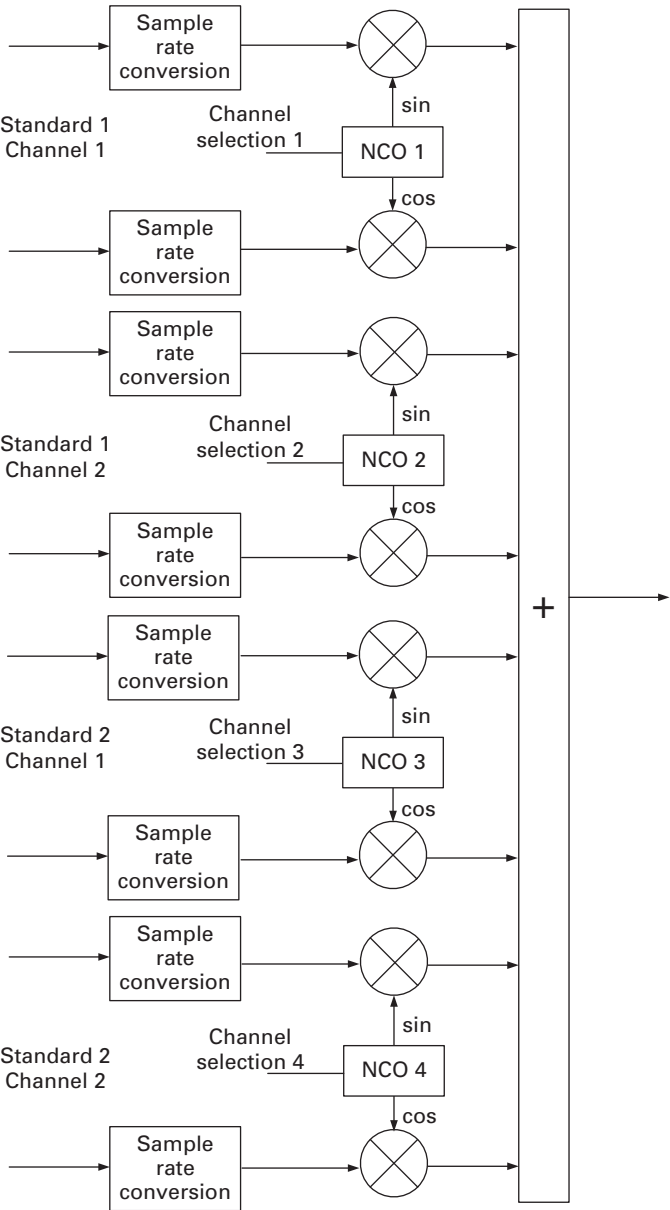


Figure 12.29 Multimode DUC by sample rate conversion to the same IF sampling rate for standard 1 and standard 2 (each standard has two channels).

where $I1/Q1$ represent the I/Q component of channel 1 of one standard and $I2/Q2$ represents the I and Q component of channel 2 of the same standard. The improved implementation is shown in Figure 12.30. The hardware savings come from the reuse of the multipliers for the sample conversion filters.

Table 12.2 List of modes for a two-standard example.

Mode number	Channel combination
Mode 1	C_{11}
Mode 2	C_{21}
Mode 3	C_{12}
Mode 4	C_{22}
Mode 5	C_{11}, C_{21}
Mode 6	C_{12}, C_{22}
Mode 7	C_{11}, C_{22}
Mode 8	C_{12}, C_{21}

In [Figure 12.31](#), the reference design from Altera for the GSM/WCDMA/ LTE multimode radio DUC is illustrated (the spectrum allocation is illustrated in [Figure 3.3](#) of [7]). This multimode DUC supports up to twelve adjacent GSM channels and one 5 MHz WCDMA/LTE channel or two 5 MHz WCDMA/LTE channels or one 10 MHz LTE channel. This architecture uses a Farrow filter ([Chapter 7](#) of [3]) to perform a fractional rate interpolation. It also uses a CIC filter for the GSM chain. The baseband data rate for each channel of GSM is 270.833 ksp/s (kilo samples per second). The mixing and combining happens at the sample rate of 61.44 MHz and the final output of the DUC is at the sample rate of 245.76 MHz.

12.4 Hardware implementation for digital down-conversion

The digital down-converter (DDC) translates one or more intermediate IF channels from a set of specified center frequencies into DC (centre at 0 Hz). The DDC is comprised of a direct digital synthesizer (DDS), matched filters, decimation filters, and other digital circuits to remove adjacent channels and maximize the received signal-to-noise ratio (SNR). The general design of the DDC is shown in [Figure 12.32](#).

12.4.1 Introduction

As shown in [Figure 12.32](#), the desired channel is translated to baseband using a DDS comprising the two multipliers. The sample rate of the signal is then adjusted to match the baseband sample rate of the system. This is performed using a multi-stage multi-rate filter consisting of the CIC filter, the two polyphase decimators $G(z)$ and $H(z)$. In order to provide maximum flexibility in the datapath, each of the filters may be optionally inserted or excluded from the signal processing chain. In the case of multimode systems with different channels, sample rates, and decimation rates (R , $D1$ and $D2$), the multipliers in the mixer, various components in the CIC filter, and the CFIR and PFIR may be time division multiplexed to service both the inphase (I) and quadrature (Q) signals according to [Chapter 2](#) of [8].

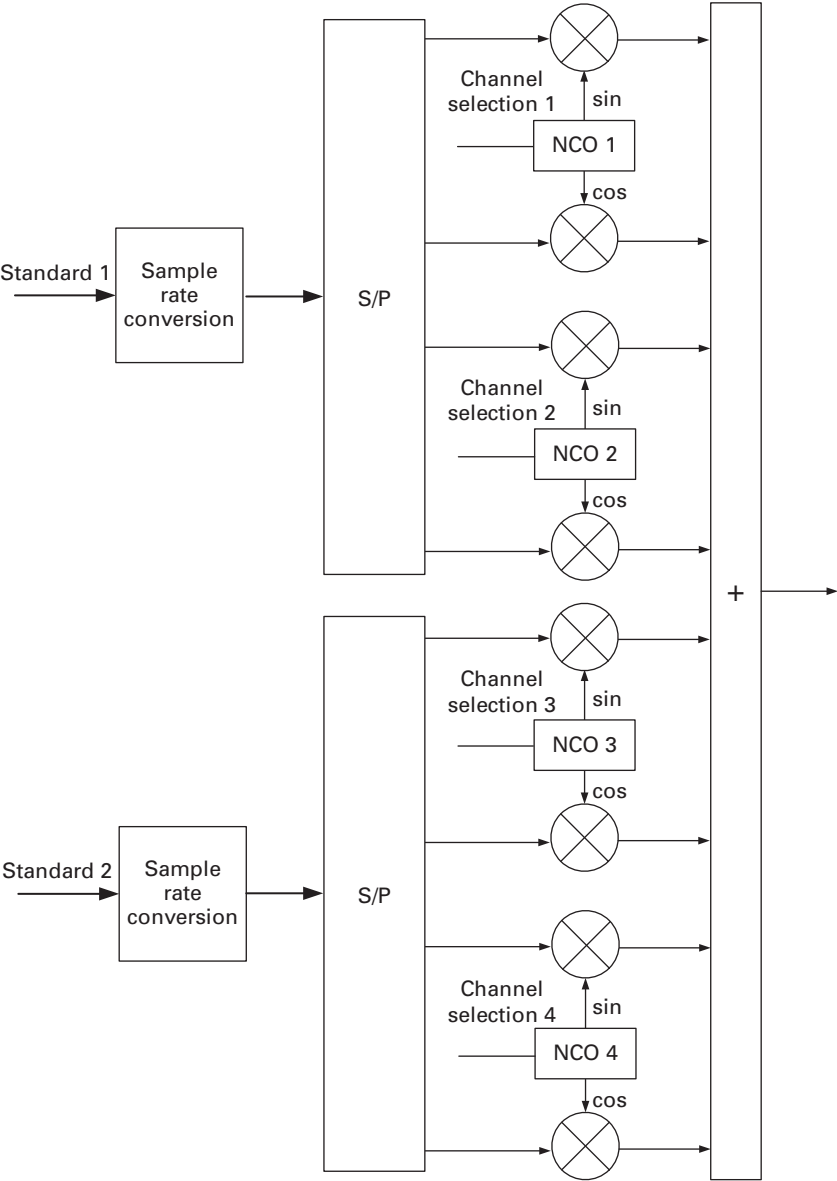


Figure 12.30 Multimode DUC using serial buses for two standards.

A coarse gain adjustment may be applied at the output of the CIC filter to compensate for the bandwidth reduction performed by the CIC decimator in Figure 12.32. The signal gain can be boosted by 42 dB (7 bits). If the signal power across the input bandwidth is flat, as is the case in most frequency division multiplexed (FDM) systems, then one would want to boost the signal power at the output of the CIC filter by a factor of $\sqrt{R}$, where R is the decimation rate of the CIC filter. For narrowband systems such as GSM, the bandwidth of

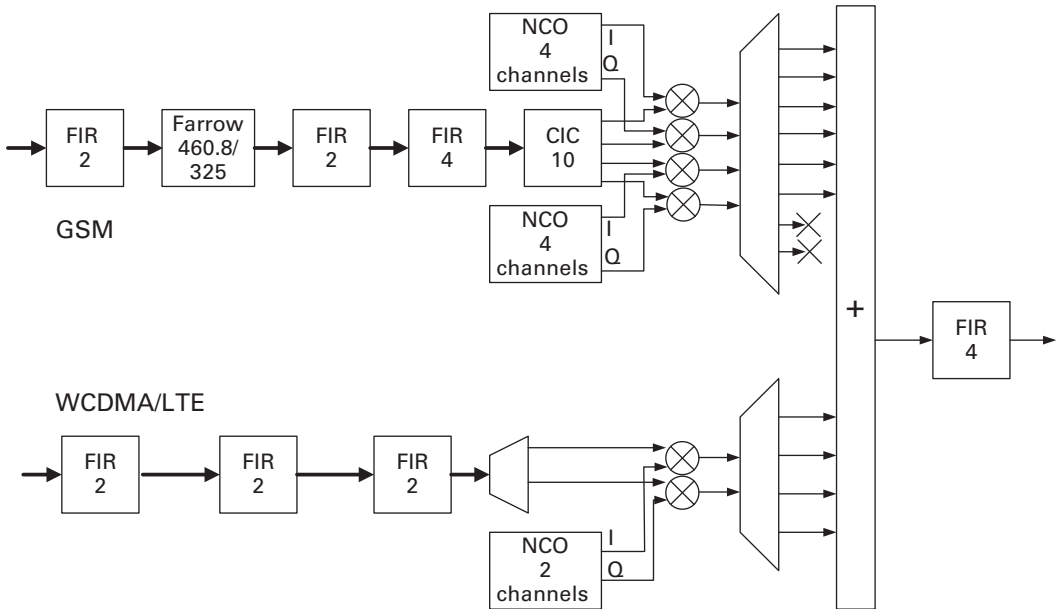


Figure 12.31 Possible multimode system architecture with mixing and combining at 61.44 MHz (referring to Figure 18 of [7]).

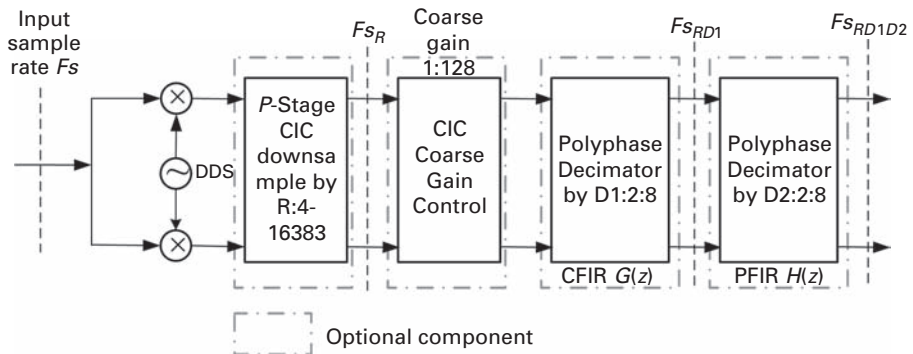


Figure 12.32 Digital down-converter general design.

the signal is significantly less than the sampled bandwidth and large-sample rate changes must be undertaken to efficiently process the signal. The sample rate conversion normally would be hundreds or even thousands. For wideband systems such as DVB-T/H, IEEE802.11 a/g/n, WiMax, and LTE, the sample rate conversion would be much smaller than for narrowband systems. However, the decimation FIR filter design with efficient spurious rejection and small pass band ripple would be a challenge. The critical design parameters for the DDC are total rate change, IF sample rate, number of carriers, and passband bandwidth. The total rate-change is the most critical parameter in DDC design.

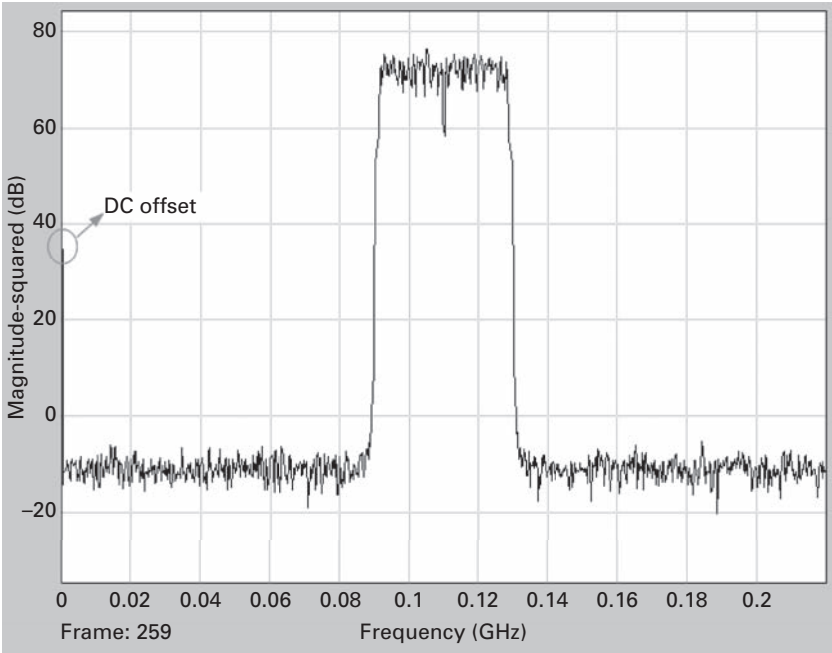


Figure 12.33 Spectrum of the input test signal after ADC.

For narrowband systems with large rate conversion, CIC filters are regarded as efficient decimation filters. For wideband systems with small rate conversion, the decimation stages can be achieved effectively with a cascade of FIR filters. The IF sample-rate determines the number of operations performed in a single sample period. It is normally an integer multiple above the passband width for the wideband system [9], [10].

For the reference design in this section, without losing generality, the IEEE802.11n system is selected [11]. The IEEE802.11n is a wideband wireless system and the main challenge of a wideband DDC is getting enough processing resource to filter the signal. All the processing is performed at a fairly high rate. Therefore, the filters will cost intensive resource in an FPGA. In order to achieve the efficient DDC design in an FPGA, the filters are best implemented as a multi-rate FIR. And the multi-rate FIR will be split into different stages with reduced taps.

The target device for implementation is Xilinx Virtex 5. The testing signal for the DDC reference design is sampled at 440MSPS and quantized to 14 bits. The spectral profile after ADC is shown in Figure 12.33.

According to Figure 12.33, the center frequency of the input signal is at 110 MHz and the bandwidth is 40 MHz. The spectrum is determined by 2048 point FFT and flat-top window. From Figure 12.33, a Direct Current (DC) offset clearly appears in the spectrum. The IF input signal is converted into baseband IQ data through a DDC process. The complex IQ output of the DDC is generated at a sampling rate of 44MSPS. A summary of requirements for the IEEE802.11n receiver DDC design is listed in Table 12.3. The DDC

Table 12.3 DDC parameters for IEEE802.11n 40 MHz mode.

Parameter	Value	
Passband bandwidth	40 MHz	
Total rate change	440:176:44	
Number of carriers	128	
Value of carrier number Kmin	0	
Value of carrier number Kmax	127	
IF sample rate	440 MSPS (Mega Samples Per Second)	
IF center frequency	110 MHz	
DDC output rate	44 MSPS	
Baseband sample rate	40 MSPS	
Input signal quantization	14-bit Real	
Output signal quantization	14-bit I and Q	
Base sample duration T	25 ns	
Spacing between carriers Kmin and Kmax	40 MHz	
Allowed guard interval Δ/T_u	1/4	1/8
Duration of symbol part T_u	$128 \times T$	
	$3.2 \mu\text{s}$	
Duration of symbol interval Δ	$32 \times T$	$16 \times T$
	800 ns	400 ns
Symbol duration	$4 \mu\text{s}$	$3.6 \mu\text{s}$
$T_s = \Delta + T_u$		

input is real directly coming from the ADC. However, in the reference design a DC offset cancellation module has been added between the ADC and DDC.

12.4.2 DC offset cancellation

Direct current offset refers to a situation where the entire signal is shifted up or down slightly from the center line. The DC offset is caused by the IF gain control circuits and real ADC. Direct current offset causes leakage in the spectrum mask after the DDS stage as shown in [Figure 12.34](#).

In order to remove this leakage, the multi-rate FIR will cost intensive resource to implement the coefficient taps. It may also cause other artifacts depending on what is being done with the signal. Direct current offset can be reduced in real-time by a high-pass filter. Ideally, the input IF signal level for the real ADC is 0 dBm in order to get the full dynamic range of ADC. The DC offset cancellation diagram is shown in [Figure 12.35](#).

12.4.2.1 DC offset cancellation design

The detailed circuit design of DC offset cancellation is shown in [Figure 12.36](#). The DC estimation is based on a low-pass filter. In order to save the hardware multiplier in the FPGA, the low-pass filter is composed of a digital integrator and a comb filter also referred to as a moving average (MA) filter.

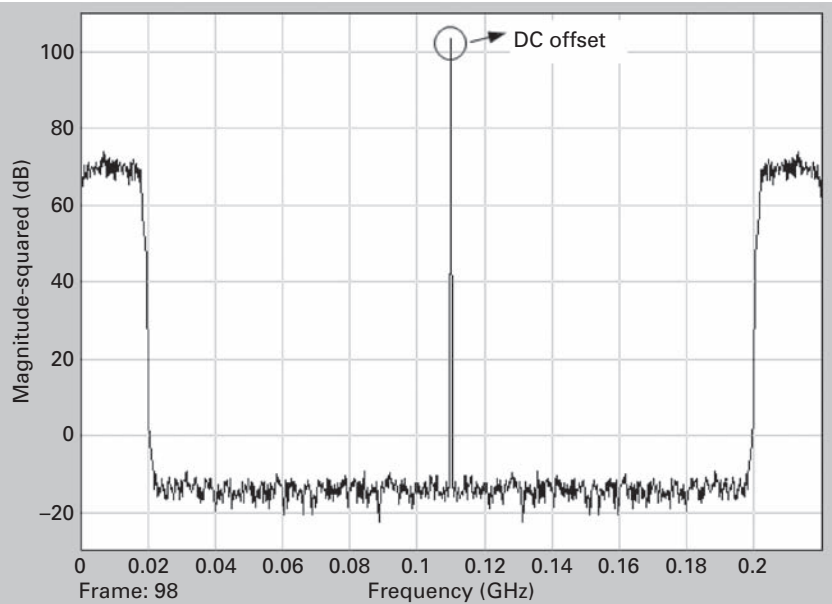


Figure 12.34 Spectrum of the input test signal after the DDS stage without DC canceller.

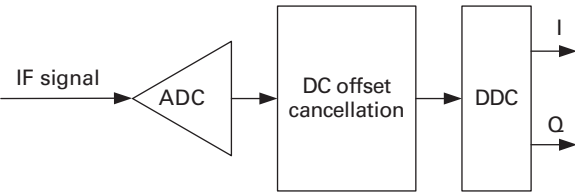


Figure 12.35 DC offset cancellation diagram.

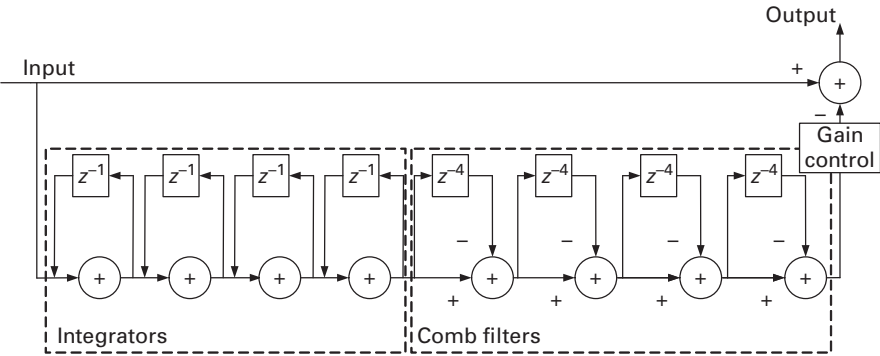


Figure 12.36 DC offset cancellation design.

According to Figure 12.36, the signal's DC offset is removed at the output by subtracting the signal's moving average from the output signal. The MA consists of integrators and comb filters [12]. The 4-point MA in Figure 12.36 has a transfer function defined by

$$H(z) = \frac{(1 - z^{-4})^4}{(1 - z^{-1})^4}. \quad (12.3)$$

Equation 12.3 is merely a recursive running sum comprising a digital integrator followed by a 4-length comb filter. The comb section consists of four comb stages with a differential delay of four samples per stage. This is a one pole and one zero system. And the magnitude response at the output of the circuits is:

$$|H(f)| = \left| \frac{\sin 4\pi f}{\sin \pi f} \right|^4. \quad (12.4)$$

Considering $\sin x \approx x$ for small x , Eq. (12.4) is

$$|H(f)| \approx \left| 4 \frac{\sin 4\pi f}{4\pi f} \right|^4. \quad (12.5)$$

From Eq. 12.5, the output spectrum has nulls at multiples of $f = 1/4$. Here, f is the normalized frequency with F_s . The passband attenuation is a function of the number of stages. Increasing the number of stages can improve the imaging/alias rejection. The gain, G , at the output of the final comb section is

$$G = 4^4 = 256. \quad (12.6)$$

The gain control block will scale back the gain at the final stage before subtraction. The spectrum of the MA input is shown in Figure 12.37.

And the spectrum of the MA output is shown in Figure 12.38.

The spectrum of the output of the DC offset cancellation is shown in Figure 12.39.

Comparing Figure 12.37, Figure 12.38, and Figure 12.39, there is a 96.3dB DC cancellation rate. The further cancellation is done in baseband processing.

12.4.2.2 DC offset cancellation implementation

The top-level screen shot for the System Generator implementation of the DC cancellation is shown in Figure 12.40. The implementation is based on the Xilinx Virtex 5 device. From Figure 12.40, there are two process chains after the IF signals is digitalized by 14bit ADC at 440 MHz. The upper chain is to delay the original signal to maintain the timing balance between the original signal and the subtraction signal for the DC cancellation. The lower chain is the process of DC offset estimation for the subtraction. After the subtraction, there is a gain control block to maintain the subtracted signal level. The implementation of DC offset estimation is shown in Figure 12.41.

There are four accumulators and four integrators for the DC offset estimation. The resource required by the DC offset cancellation for the IEEE802.11n system when targeting the Virtex 5 FPGA family is summarized in Table 12.4.

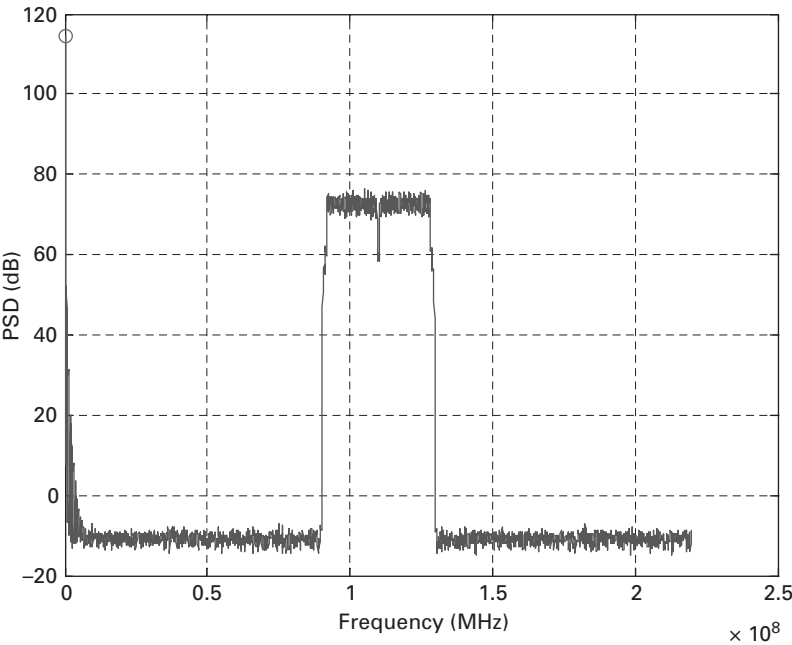


Figure 12.37 Spectrum of signal at the MA input.

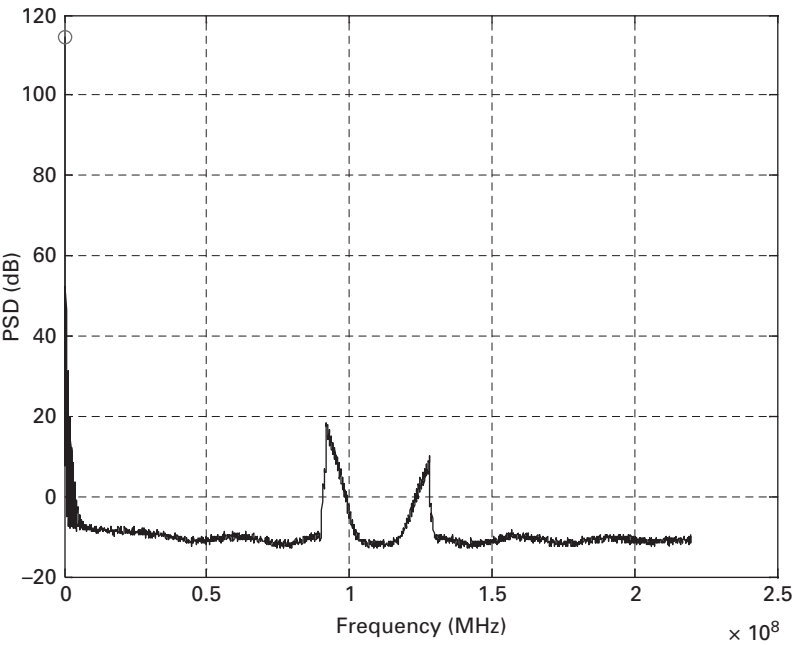


Figure 12.38 Spectrum of signal at the MA output.

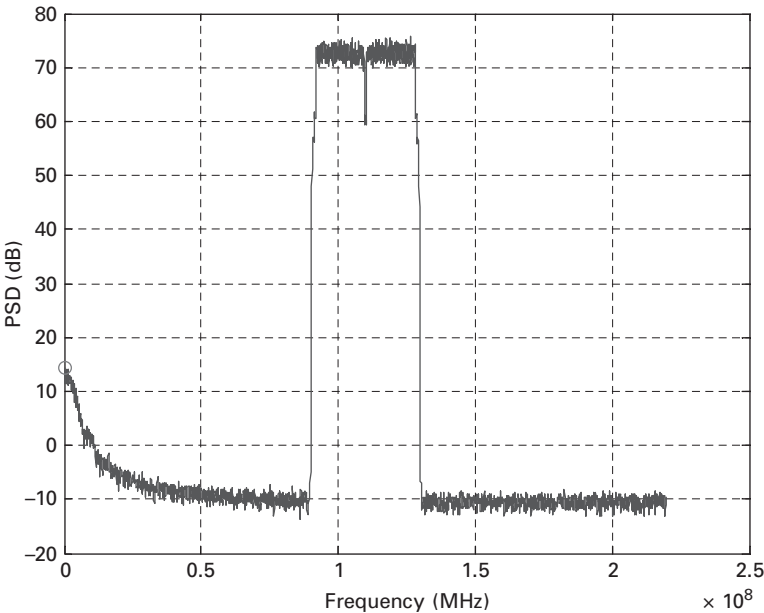


Figure 12.39 Spectrum of signal at the DC offset cancellation output.

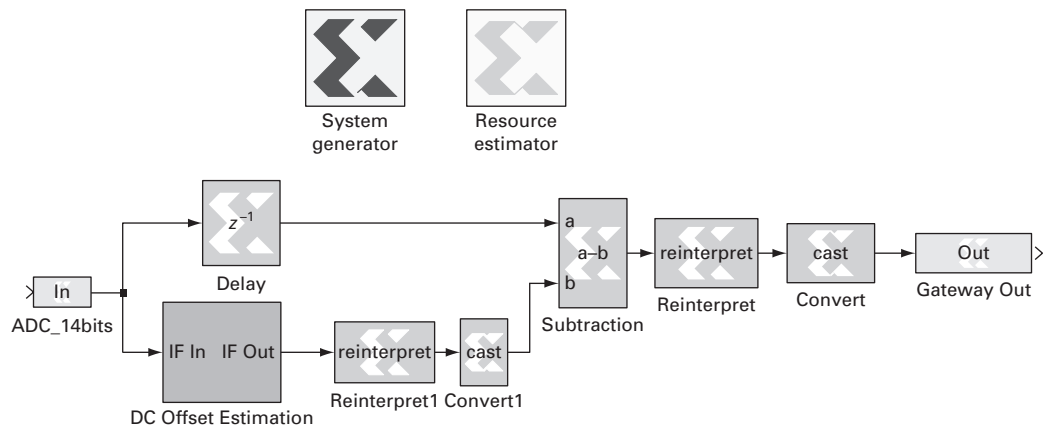


Figure 12.40 DC offset cancellation System Generator implementation.

12.4.3 Direct digital synthesizer

A direct digital synthesizer (DDS) is an important component in many digital communication systems. A DDS is a method of producing an analog waveform by generating a time-varying signal in digital form. Since operations within a DDS device are mainly digital, it can offer fast switching between output frequencies, fine frequency resolution, and operation over a broad spectrum of frequencies.

Table 12.4 Resource utilization for DC offset cancellation.

Parameter	Value
DSP48Es	0
36 kbit Block RAMs	0
Slices	639
Flip-flops	583
Logic cells	406
Fmax (MHz)	440

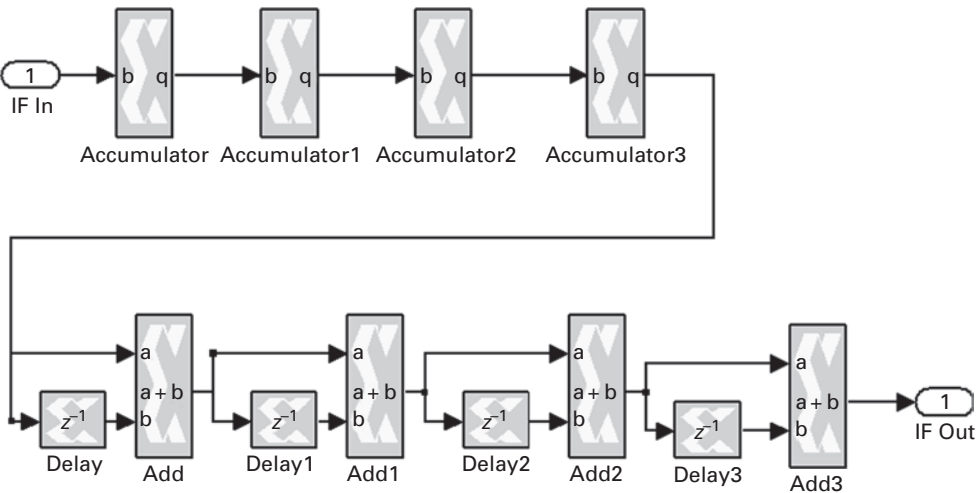


Figure 12.41 DC offset estimation System Generator implementation.

12.4.3.1 Direct digital synthesizer design

A common method of DDS design with a complex or real-valued sinusoid employs a look-up table scheme. The look-up table stores samples of a sinusoid and a digital integrator is used to generate a suitable phase argument that is mapped by the look-up table to the desired output waveform. The actual design of a DDS requires the look-up table to determine the phase output signal at any point in time. Figure 12.42 shows the building blocks for DDS based waveform generation [13].

The system diagram of the DDS presented in Figure 12.42 operates as follows. The phase accumulator computes a phase slope that is mapped to a sinusoid by the look-up table. The phase register is a slicer which accepts the high precision phase and generates a lower precision representation of the angle. This value is presented to the address port of the look-up table that performs the mapping from phase-space to time. The DDS uses an addressing scheme with an appropriate look-up table to form samples of an arbitrary frequency sinusoid.

The DDS is realized using the embedded multipliers available in Virtex 5 FPGAs, for example. The DDS employs an error feed-forward circuit, which produces a high-quality local oscillator with up to 115 dB of spurious free dynamic range (SFDR). The SFDR of the

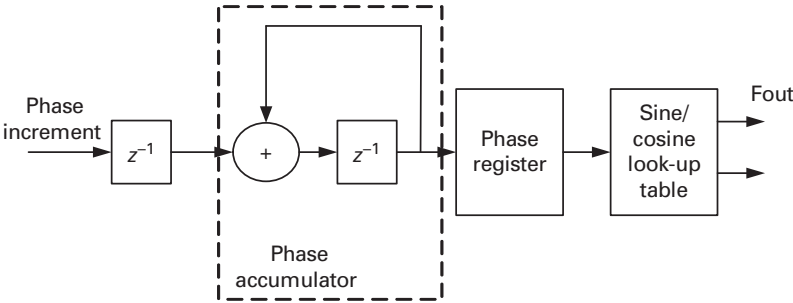


Figure 12.42 DDS system diagram.

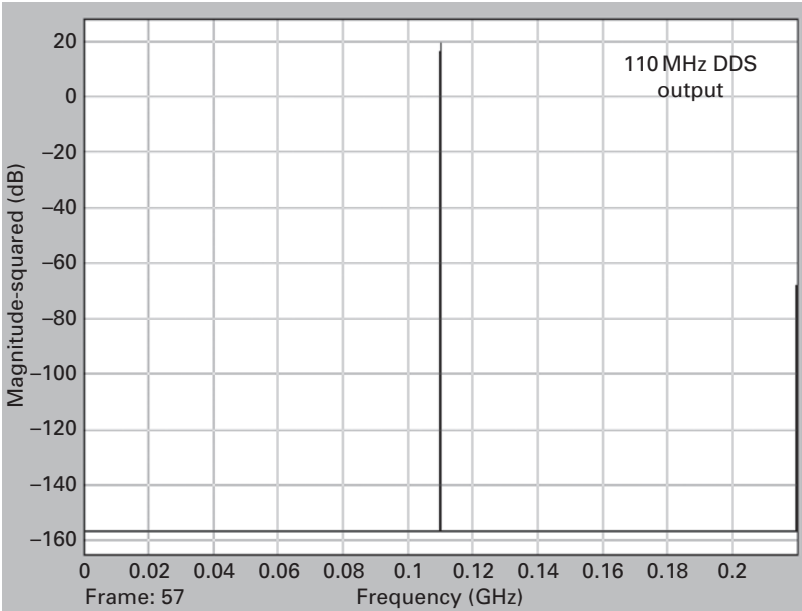


Figure 12.43 DDS output spectrum.

DDS is a selectable parameter with three choices, 115, 80, or 60 dB, to provide the option for reduced utilization of FPGA resources. The implementation requires a single block memory and multipliers. The phase accumulator is fixed at 32 bits. Figure 12.43 provides the spectrum plot of the DDS with F_s 91.4 MHz, SFDR 80 dB, and center 112 MHz. Figure 12.43 is single tone test. The DDS is tuned to a single output frequency.

12.4.3.2 Direct digital synthesizer implementation

The top-level screen shot for the System Generator implementation of the DDS is shown in Figure 12.44 and is based on the Xilinx Virtex 5 device.

In order to save the FPGA resource, the fixed phase increment is considered in the implementation. The Input IF carrier is 110MHz and sample rate is 440MHz, therefore,

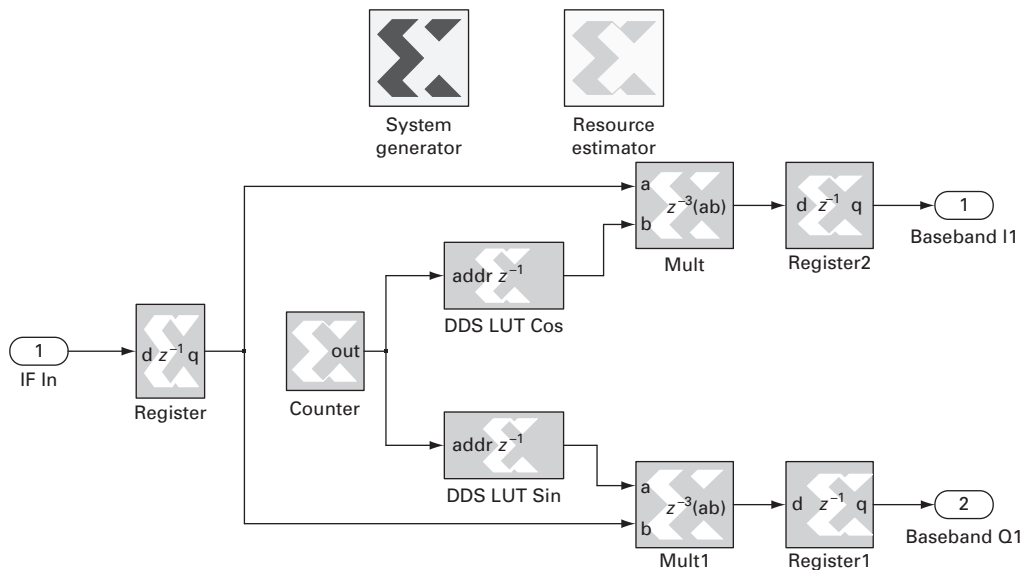


Figure 12.44 DDS mixer System Generator implementation.

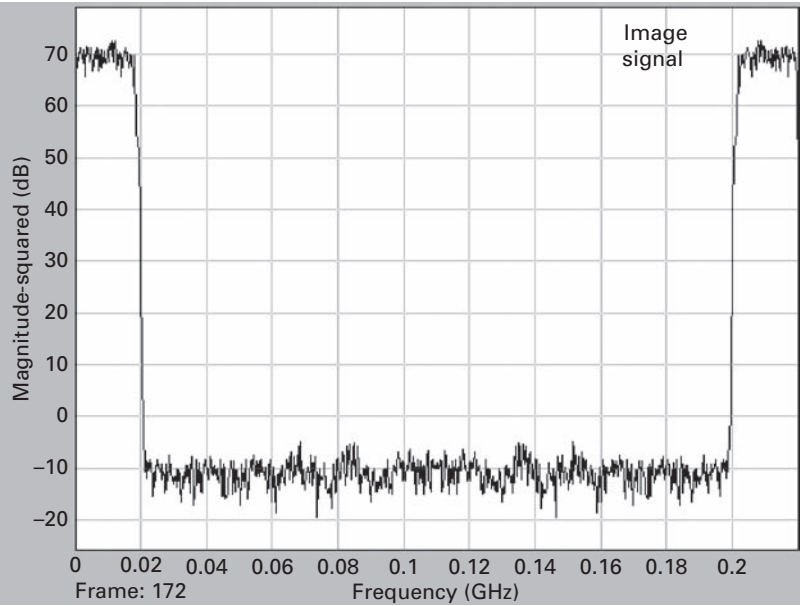


Figure 12.45 Signal spectrum after DDS mixer.

the fixed phase increment for the DDS is $\pi/2$ and the simplified DDS can be implemented by only two block ROMs (DDS LUT Cos and DDS LUT Sin) and one counter. The digital IF data is multiplied by Sine ROM and Cosine ROM to get the baseband IQ data. From Figure 12.45, there is an image signal after the DDS mixer and an anti-alias filter is applied to remove it. The detailed design is presented in the next section. The resource

Table 12.5 Resource utilization for DC offset cancellation and DDS mixer.

Parameter	Value
DSP48Es	0
36 kbit Block RAMs	1
Slices	810
Flip-flops	747
Logic cells	490
Fmax(MHz)	440

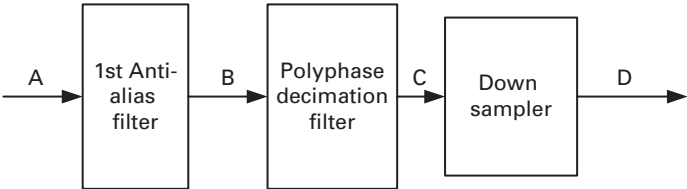


Figure 12.46 System diagram of multi-rate filtering.

required by the DC offset cancellation and DDS mixer for the IEEE802.11n system using a Virtex 5 FPGA is summarized in Table 12.5. From Table 12.5, the resource usage is fairly low for the implementation and no DSP slice is required in the implementation.

12.4.4 Multi-rate decimation filter design and implementation

In this section, a reference design for multi-rate filtering is presented. The decimation stages can be achieved effectively with a cascade of FIR filters. In this design, the multi-rate FIR is split into two stages. The first step is anti-alias filtering and the second step is polyphase decimation filtering [14]. The system diagram is presented in Figure 12.46.

The sample rate at stage A and B is F_s . The sample rate at stage C is $2*F_s/5$ and $2*F_s/20$ at stage D. The detailed design is presented as follows.

12.4.4.1 Anti-alias filtering

The image signal after the DDS is filtered out in this stage. In order to save the taps of the filter, the roll-off rate of the filter is as large as possible. The filter specification is given in Table 12.6. The frequency response is shown in Figure 12.47, which shows a good match to the specification.

The signal spectrum after filtering at stage B is shown in Figure 12.48. From Figure 12.48, the image is clearly filtered out and the level is below quantization noise level.

12.4.4.2 Decimation filtering

In this stage, the signal is up-sampled by two and down-sampled by five. In order to save the taps of the filter, there are two sets of phases to implement the resampling filter. The filter specification is in Table 12.7.

Table 12.6 First anti-alias filter specification.

Parameter	Value
Passband ripple	<0.02 dB
Passband frequency	19.8 MHz
Stopband frequency	187 MHz
Stop attenuation	90 dB
Number of taps	6

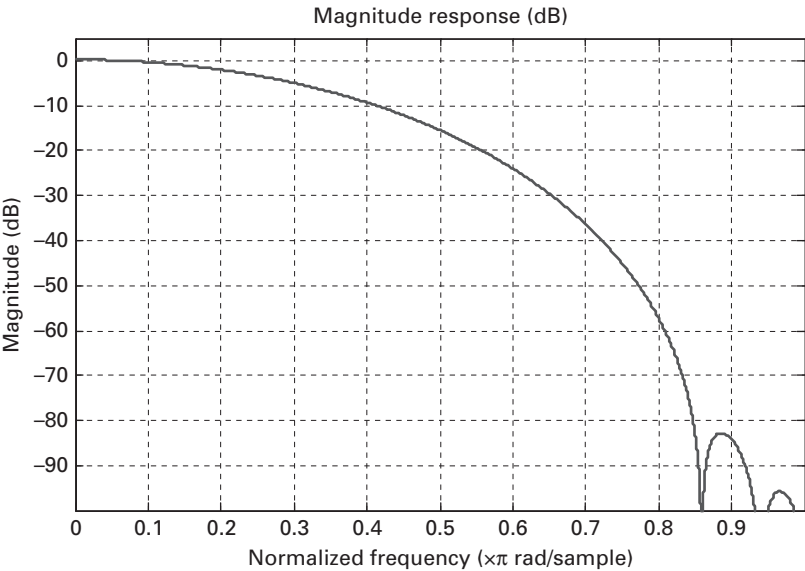


Figure 12.47 Frequency response of the first anti-alias filter.

There are two steps in the filtering process including interpolation and decimation. Firstly, the input data is interpolated by polyphase filtering. Secondly, the filtered data is decimated by the decimation selector. The system diagram is shown in [Figure 12.49](#).

Each phase FIR has three taps. The decimation selector is composed of FIFOs, ROMs, counters, and multipliers. The frequency response of the filter is shown in [Figure 12.50](#). The input sample rate of the filter is 440 MSPS, and the output sample rate is 176 MSPS. The coefficients are quantized to 16 bits. From [Figure 12.50](#), it can be seen that the stopband attenuation is below 85 dB and the passband ripple is less than 0.012 dB. After the filtering stage, a further down sampler is applied to achieve the baseband rate at 44 MSPS. The signal spectrum after the second polyphase filtering is presented in [Figure 12.51](#). According to this figure, the image replica signal and DC are rejected below 75 dBc. A simple decimator is used for the final decimation to obtain the baseband data which is sent to the RAM and DSP for further baseband processing [3]. The baseband spectrum is presented in [Figure 12.52](#), which shows that the clear baseband spectrum is achieved by the polyphase decimation filter.

Table 12.7 Polyphase filter specification.

Parameter	Value
Passband ripple	<0.012 dB
Passband frequency	39.6 MHz
Stopband frequency	374 MHz
Stop attenuation	85 dB
Number of taps	6

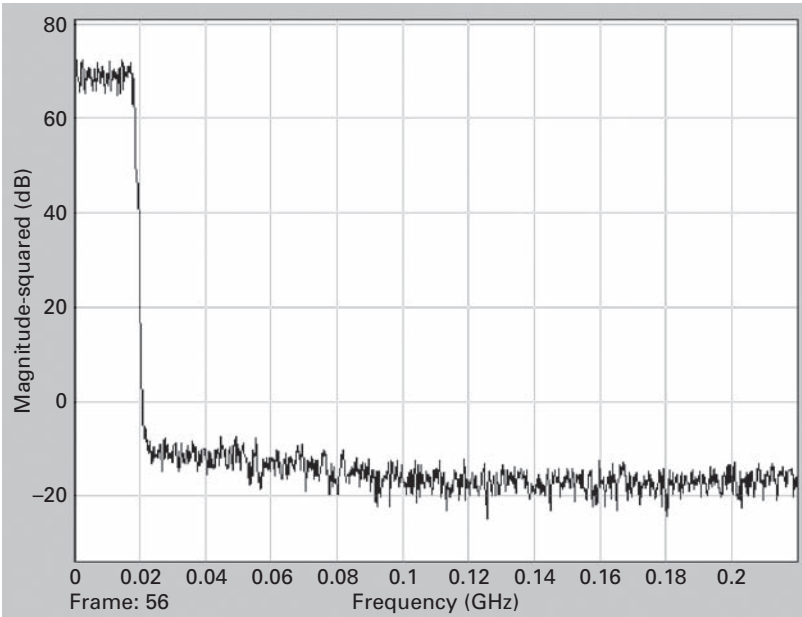


Figure 12.48 Signal spectrum after the first anti-alias filtering.

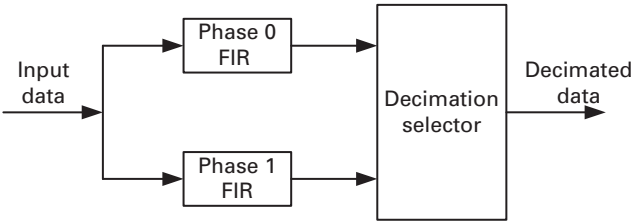


Figure 12.49 System diagram of polyphase decimation filtering.

12. 4.4.3 Multi-rate filter implementation

The top-level screen shot for the System Generator implementation of the polyphase filter is shown in [Figure 12.53](#) and the whole multi-rate filtering System Generator implementation is shown in [Figure 12.54](#).

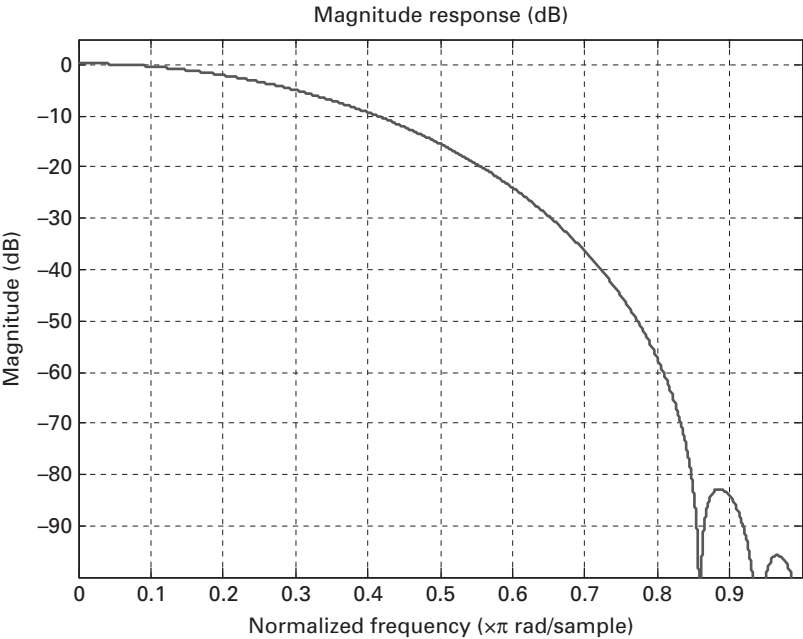


Figure 12.50 Frequency response of the second polyphase filter.

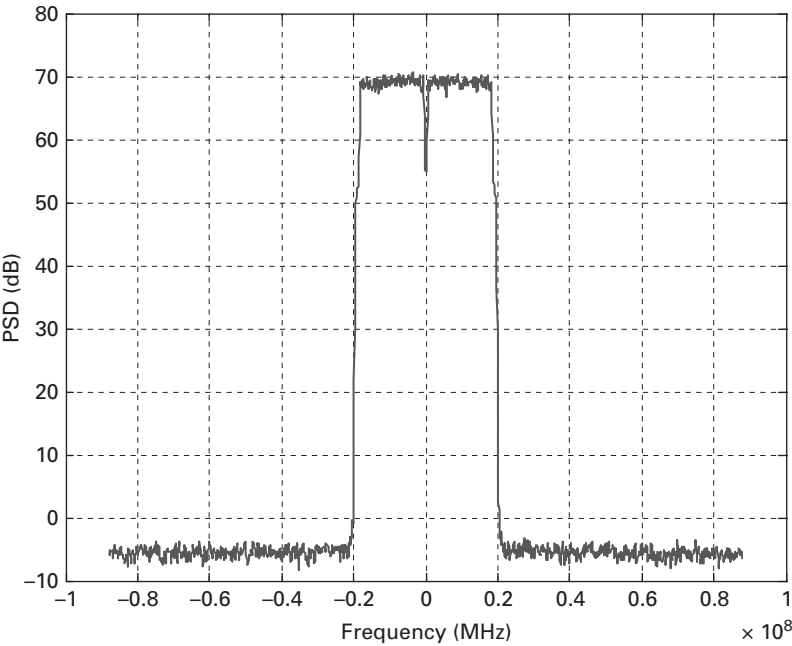


Figure 12.51 Signal spectrum after the second polyphase filter.

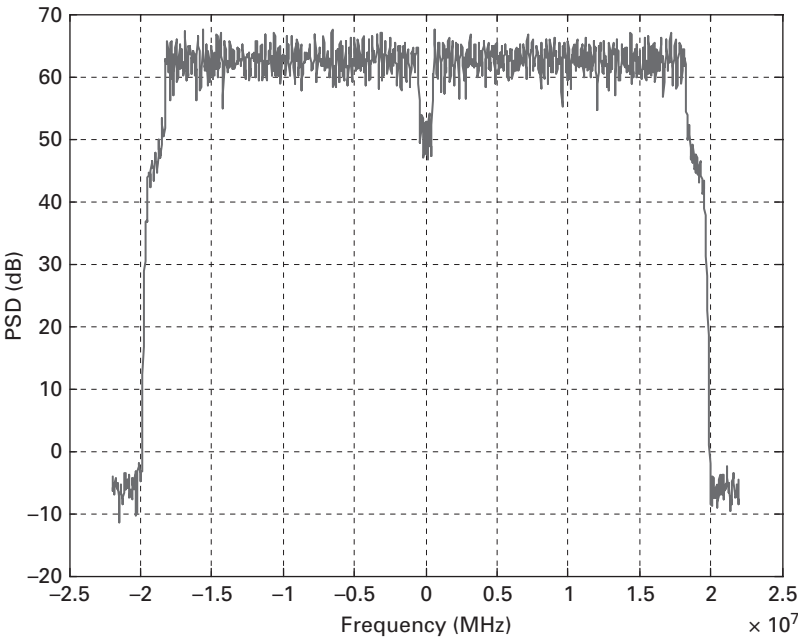


Figure 12.52 Signal spectrum of the baseband data.

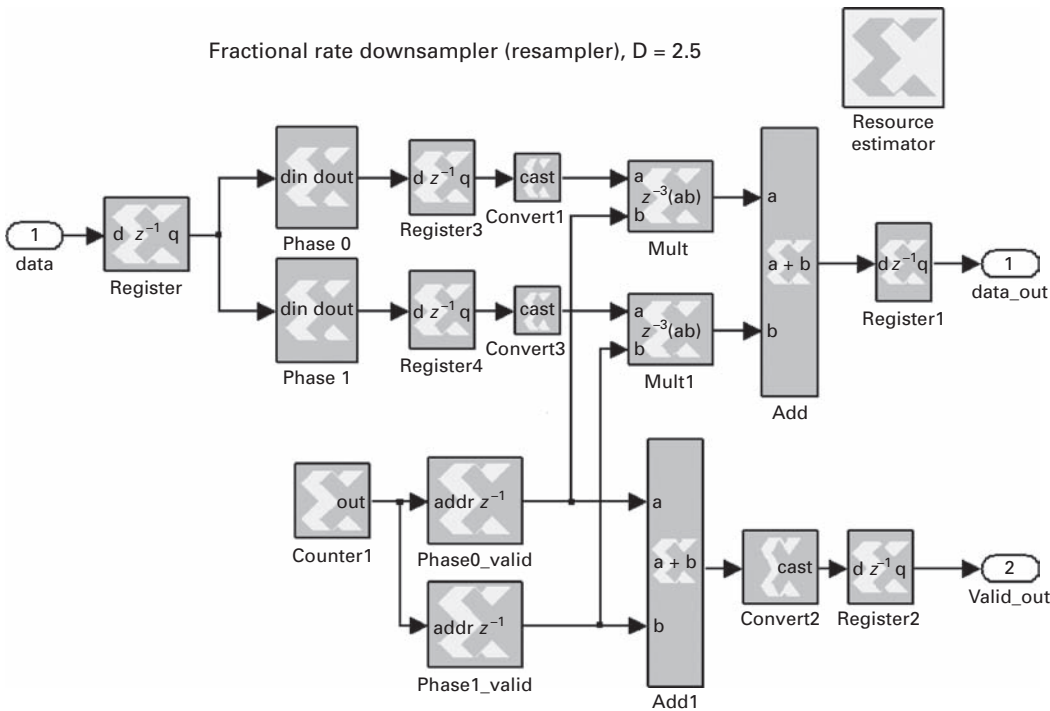


Figure 12.53 Polyphase filter System Generator implementation.

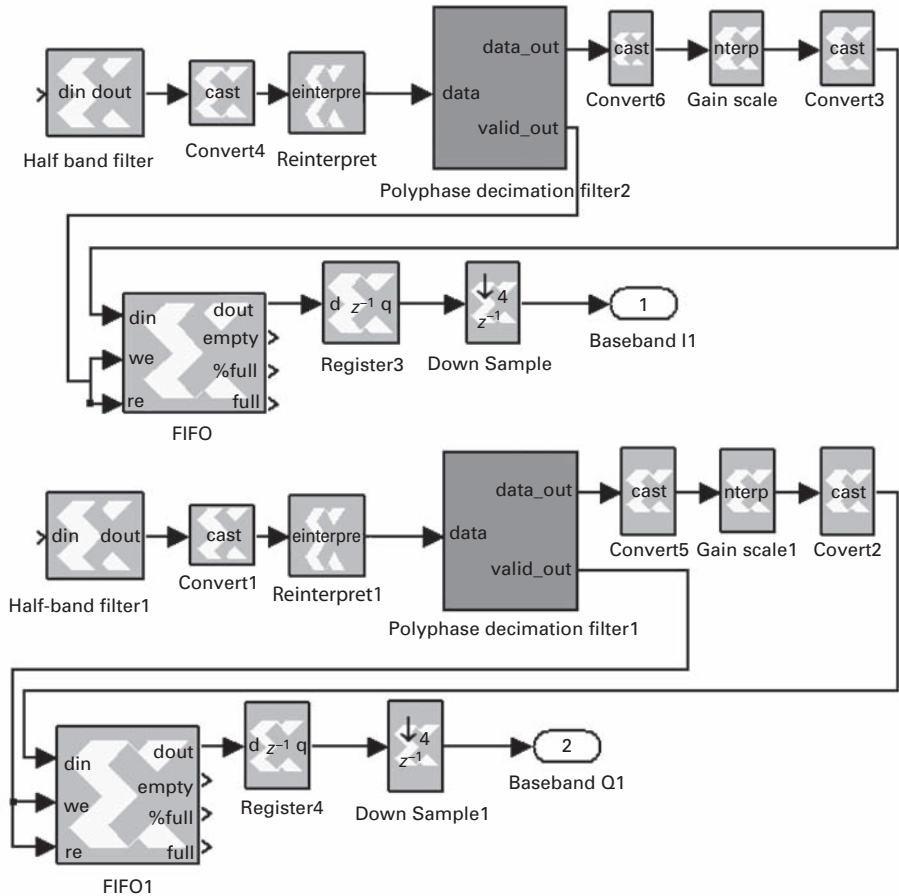


Figure 12.54 Multi-rate filter System Generator implementation.

From Figure 12.53, the polyphase filtering module consists of two FIR filters (block “Phase0” and “Phase1”), two multipliers (block “Mult” and “Mult1”), two ROMs (“Phase0_valid” and “Phase1_valid”), one counter, two combinars (“add” and “add1”), and some registers. Firstly, I or Q data is filtered simultaneously by two FIR filters with two sets of coefficients. The filters are working at 440 MSPS. There are two ROMs with different phase coefficients to generate the valid phase for the selection of the filtered I/Q data. Finally, the two filtered data with different phases are added together.

Figure 12.54 shows the block diagram for the IQ data filtering. Take the upper chain for an example, I-data is firstly filtered by a half-band filter (block “Half-band filter”) to get the clear baseband data without the image band. The half-band filter is generated by the Xilinx FIR core generator with DSP48s. The half-band filter can support multi-channel and multimode systems. The input data width is 14-bit and the coefficients data width is 16 bits. After the half-band filter, the I data is filtered by the polyphase decimation filter module (block “polyphase decimation filter2”). The following FIFO block is to remove the

Table 12.8 Resource utilization for DDC.

Parameter	Value
DSP48Es	24
36 kbit block RAMs	8
Slices	1350
Flip-flops	2141
Fmax (MHz)	440

invalid data in the filtered I data. After this stage, I-data is decimated by 2.5 and running at 176 MHz. Finally, the last decimator is to get the baseband sampling rate at 44 MHz.

The resource required by the DDC for the IEEE802.11n system when targeting the Virtex 5 FPGA family is summarized in Table 12.8. In this design, 24 DSP slices is required. However, considering the high-speed data-rate and the Virtex 5 device resource, this design is still achievable for most of the Virtex 5 devices.

12.4.5 DDC performance

This section discusses the performance of the DDC design and implementation in FPGA. The EVM (Error Vector Magnitude) for different ADC bit-widths is analyzed. The input to the data test is 802.11n composite signals with total noise power in a 40 MHz bandwidth set at the level of 85 dB below the desired signals. The EVM is defined as the difference between the reference waveform and the measured waveform [15]. The EVM can be expressed as:

$$EVM(\%) = \frac{RMS(error\_vector)}{RMS(reference\_signal)} \times 100\% \tag{12.7}$$

In order to achieve the best value for EVM, the frequency synchronization, timing synchronization, channel equalization, pilot timing tracking, and pilot phase/amplitude tracking are applied into the baseband processing. The constellation of 64QAM for the 14bit ADC after DDC is shown in Figure 12.55. The EVM is 0.59%. This EVM result meets the requirement for the residual EVM for the digital system which normally is 0.5%. When the ADC is changed into 8bits, the spectrum profile and constellation of test data is shown Figure 12.56. From the spectrum profile, the SNR is reduced by about 38 dBc because of the increased quantization noise by 8bit ADC. The EVM is 1.9%. Table 12.9 describes the EVM vs. different ADC bit widths.

12.5 Hardware implementation for digital up-conversion

The digital up-converter (DUC) is the reverse process of the DDC. The DUC consists of a direct digital synthesizer (DDS), matched filters, interpolation filters and other digital circuits to remove adjacent channels and maximize the transmitted SNR. The general design of a DUC is shown in Figure 12.57.

Table 12.9 EVM vs. ADC bit widths.

ADC bit widths	EVM
14	0.59%
13	0.59%
12	0.60%
11	0.62%
10	0.71%
9	1.11%
8	1.90%

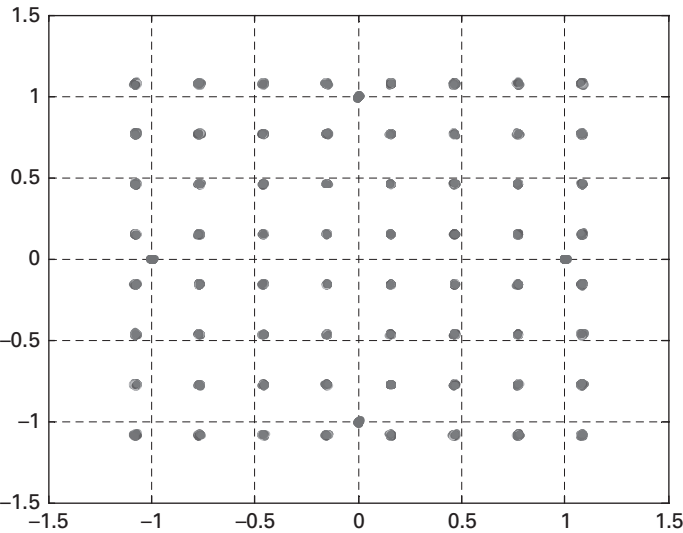


Figure 12.55 The constellation of 64 QAM test data with 14bit ADC after DDC.

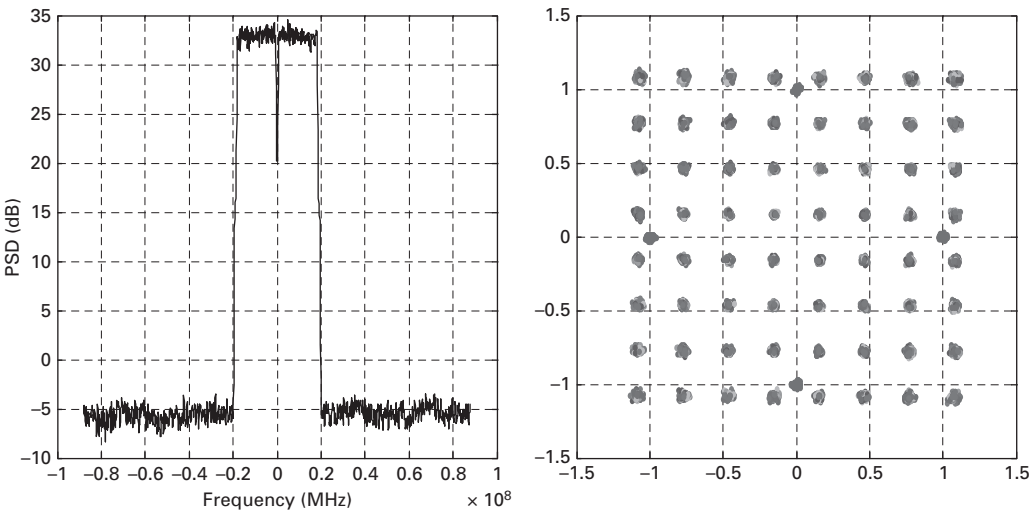


Figure 12.56 The spectrum profile and constellation of 64 QAM test data with 8bit ADC after DDC.

12.5.1 Introduction

Figure 12.57 shows the block diagram of the DUC. The DUC is a digital circuit which implements the conversion of a complex digital baseband signal to a real passband signal. The input complex baseband signal is sampled at a relatively low sampling rate, typically the digital modulation symbol rate. The baseband signal is filtered and converted to a higher sampling rate before being modulated onto a DDS carrier frequency. The DUC performs pulse shaping and modulation of an intermediate carrier frequency appropriate for driving a final analog up-converter. From the beginning of the interpolation filter, the complex input signals are passed through three stages of filtering, each of which performs a sampling-rate change and associated low-pass interpolation filtering.

As shown in Figure 12.57, the three filtering stages are as follows. Firstly, the pulse shaping FIR filter CFIR $G(z)$ provides a sampling rate increased by step 2 from 1 to 8 and performs transmitter Nyquist pulse shaping. Second, the compensation FIR filter PFIR $H(z)$ provides a sampling rate increase of 2 and is used to compensate for the passband distortion of the third stage CIC filter. The FIR filters, $G(z)$ and $H(z)$, are implemented using efficient Multiplier-Accumulator (MAC) blocks. The multipliers are realized using the embedded multipliers available in the Virtex 5. Thirdly, the P -stage CIC interpolation filter implements sampling-rate increased from 4 to 1448. The CIC filter design parameters permit the synthesis of a circuit with a fixed sampling-rate increase, or one whose sampling-rate increase is programmable in real-time. In the fourth step, the CIC coarse gain control block is to maintain the power distortion caused by CIC filtering. Finally, the complex data stream from the filtering stages is up converted to an intermediate frequency band by a mixing operation with a local oscillator generated by DDS in the digital domain. The I and Q mixer outputs are combined to form the final DUC passband result at 110 MHz. The CIC filters are multi-rate filters used for realizing large sample rate changes in digital systems. However, CIC filters do not use multipliers and their structure uses only addition, subtraction and registers. The CIC filter supports interpolation factors from 4 to 1448 [4].

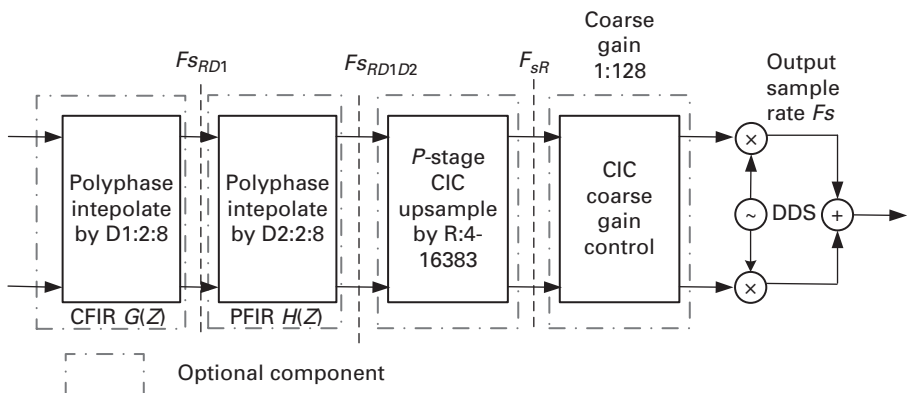


Figure 12.57 Digital up-converter general design.

Table 12.10 DUC parameters for IEEE802.11n 40 MHz mode.

Parameter	Value
Passband bandwidth	40 MHz
Total rate change	88:440
IF sample rate	440 MSPS
IF center frequency	110 MHz
DUC output rate	440 MSPS
Input signal quantization	14-bit signed I and Q
Output signal quantization	14-bit Real

Table 12.11 Anti-alias interpolation filter specification.

Parameter	Value
Passband ripple	<0.3 dB
Passband frequency	19.8 MHz
Stopband frequency	70.4 MHz
Stop attenuation	80 dB
Number of taps	26

A summary of requirements for the IEEE802.11n DUC design is listed in [Table 12.10](#).

12.5.2 Multi-rate interpolation filter design and implementation

In this section, the reference design for multi-rate interpolation filtering is presented. The interpolation stage can be achieved effectively with an FIR filter.

12.5.2.1 Anti-alias interpolation filter

The baseband signal after two times interpolation is filtered out in this stage. In order to save the taps of the filter, the roll off rate of the filter is as large as possible. The filter specification is in [Table 12.11](#). The frequency response is shown in [Figure 12.58](#). The signal spectrum after filtering is shown in [Figure 12.59](#), which shows the interpolation alias is clearly filtered out and the level is below the quantization noise level.

12.5.2.2 Multi-rate interpolation filtering implementation

The top-level screen shot for the System Generator implementation of the interpolation filter is shown in [Figure 12.60](#). From [Figure 12.61](#), the baseband 14-bit I/Q data is firstly up-sampled by 5. The interpolation filter (block “LPF FIR1” and “LPF FIR2”) is to remove the harmonics caused by the up-sampler. The interpolation filter is made by the Xilinx FIR core generator. The FIR has 26 taps and is working at 440 MHz. The interpolated and filtered baseband data is up-converted into an IF carrier frequency of 110 MHz by the DDS. The detailed design of the DDS is similar to the design given in [Section 12.4.3](#). The system generator design of the DDS mixer is in [Figure 12.61](#). From

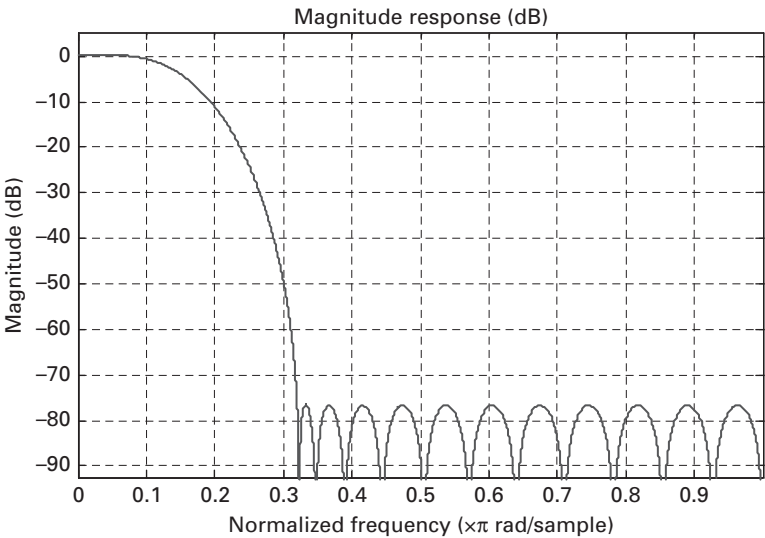


Figure 12.58 Frequency response of the anti-alias interpolation filter.

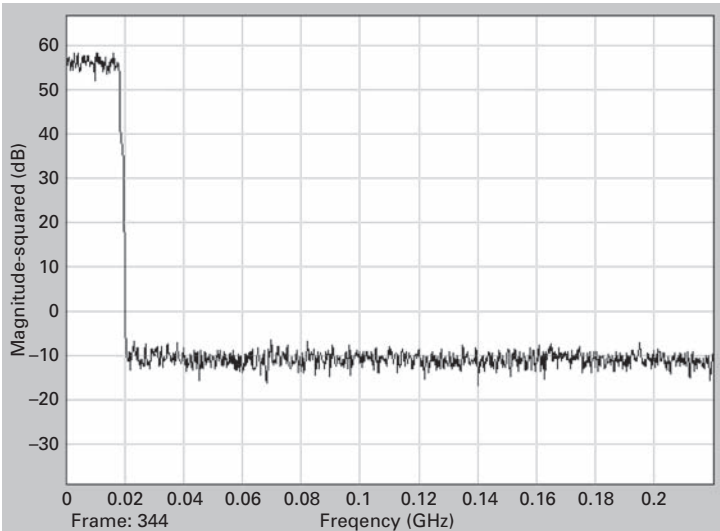


Figure 12.59 Signal spectrum after the anti-alias interpolation filtering.

Figure 12.61, the interpolated I/Q data is mixed with Sine and Cosine saved in the two ROMs (block “DDS LUT Sin” and “DDS LUT Cos”). The two ROMs are driven by a counter running at 440 MHz. As further illustration, the IF signal spectrum after the DUC is shown in Figure 12.62.

The resource required by the DUC for the IEEE802.11n system when targeting the Virtex 5 FPGA family is summarized in Table 12.12.

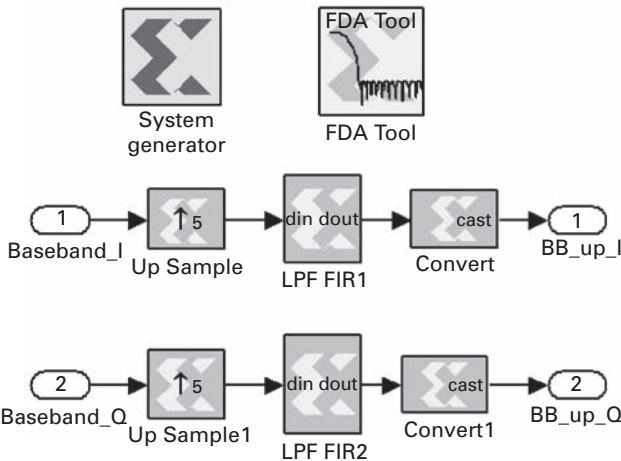


Figure 12.60 Interpolation filter System Generator implementation.

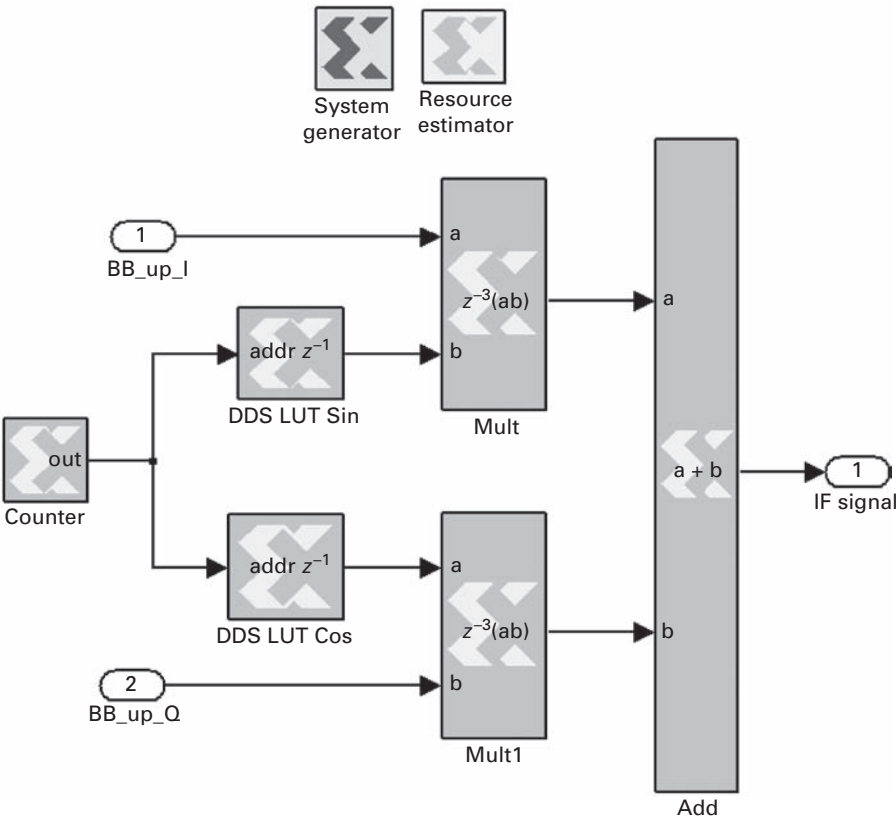


Figure 12.61 DDS Up-Converter System Generator implementation.

Table 12.12 Resource utilization for DUC.

Parameter	Value
DSP48Es	26
36 kbit Block RAMs	1
Slices	831
Flip-flops	1463
Fmax (MHz)	440

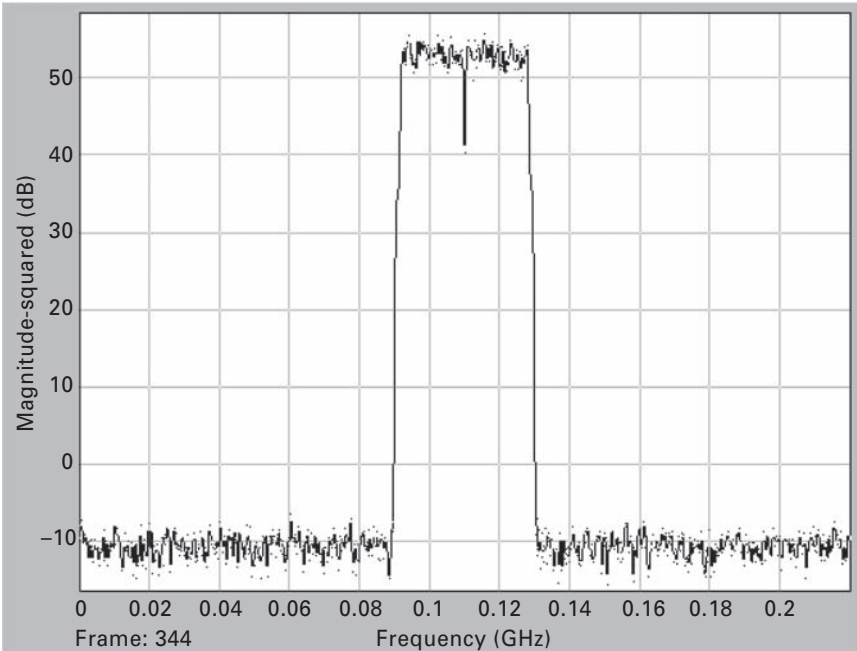


Figure 12.62 Spectrum of IF signal after DUC.

12.5.3 DUC performance

This section discusses the performance of the DUC design and implementation in FPGA. The evaluation result in terms of EVM for different DAC bit widths is considered. The input to the data test is an 802.11n composite signal with total noise power in a 40 MHz bandwidth set at the level of 85 dB below the desired signals. In order to achieve the best for EVM, the frequency synchronization, timing synchronization, channel equalization, pilot timing tracking, pilot phase/amplitude tracking are applied to the baseband processing.

The 64QAM constellation for the 14-bit signed DAC with 14-bit signed I/Q is shown in [Figure 12.63](#). The EVM is 0.65%. This EVM result meets the requirement for the residual EVM for the digital system which is 0.5%. When the bit-width of I/Q data is changed

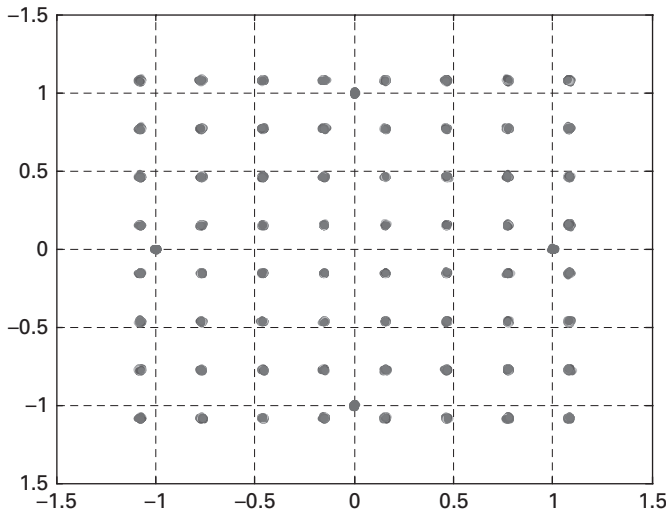


Figure 12.63 The constellation of 64 QAM test data with 14bit signed DAC with 14bit signed I/Q data.

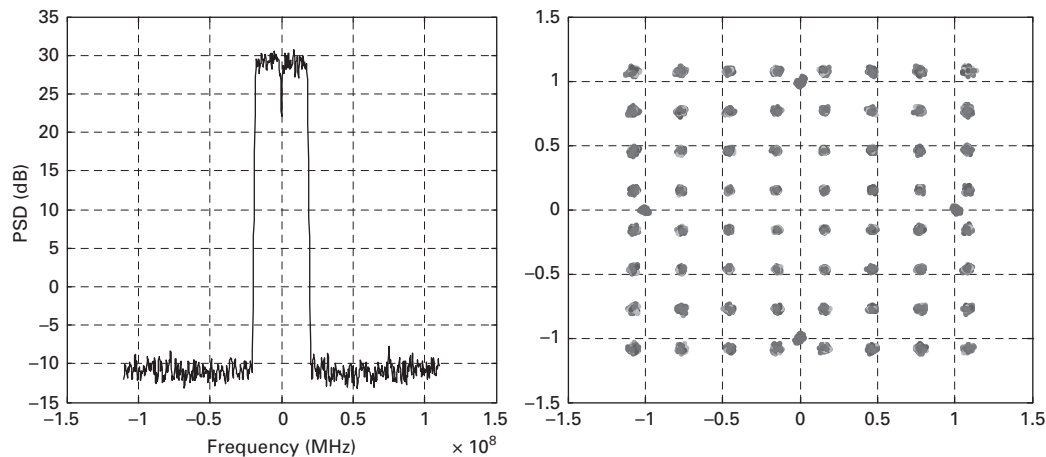


Figure 12.64 The spectrum profile and constellation of 64 QAM test data with 8bit ADC after DDC.

10-bits, the spectrum profile and constellation of test data is shown [Figure 12.64](#). From the spectrum profile, the SNR is reduced by about 38 dBc because of the increased quantization noise. The EVM is 1.63%. [Table 12.13](#) gives the EVM vs. different I/Q data bit-widths.

12.6 Summary

This chapter has described the DDC/DUC design and implementation in wireless communications systems including digital synthesizing, digital NCO, digital mixer, I/Q demodulation, multi-rate decimation filtering, polyphase filtering, multi-rate

Table 12.13 EVM vs. I/Q data bit widths.

I/Q data bit widths	EVM
14	0.65%
13	0.66%
12	0.76%
11	0.98%
10	1.63%

interpolation filtering, and DDS. The theory of bandpass sampling, over-sampling, anti-aliasing filtering, and frequency conversion has also been covered.

Furthermore, the design details of multi-rate, multi-stage, and filter-banks in the DDC and DUC are demonstrated in Section 12.2 of this chapter. The super-heterodyne receiver design is used to describe the multi-rate, multi-stage and filter-banks design in the DDC. Direct up-converter transmitter design is used for the DUC. The details of I/Q demodulation/modulation, NCO, sample rate conversion, multi-rate and multi-stage filtering have been explored in Section 12.2. Mixing and combining for multi-channels and multi-carriers for the DUC has been considered in Section 12.3. The design in this section can support multi-standard radios when there are multiple possible input baseband signals (based on different standards) operating at different sampling-rates.

Finally, the reference design example for an 802.11n wireless system in Sections 12.4 and 12.5 has illustrated creative techniques in the design and implementation of efficient and cost-effective solutions to the meet the demands of constant cost reduction requirements for the DDC/DUC. It has demonstrated the details of DDC/DUC design using a Xilinx FPGA. The multi-rate polyphase filtering for decimation and interpolation has also been described in the chapter. Furthermore, the EVM performance of the DDC with different ADC bit widths and DUC with different I/Q bit widths has been analyzed.

References

[1] E. C. Ifeachor and B. W. Jervis, *Digital Signal Processing: A Practical Approach* (Prentice Hall, 2001)

[2] Xilinx Logicore DDS v5.0. Available at www.xilinx.com/support/documentation/ip_documentation/dds.pdf (2010)

[3] J. F. Harris, *Multirate Signal Processing for Communication Systems* (Prentice Hall, 2008).

[4] H. Tarn, K. Neilson, R. Uribe, and D. Hawke, *Designing Efficient Wireless Digital Up and Down Converters Leveraging CORE Generator and System Generator*, XILINX Application Note: Virtex-5, Spartan-DSP FPGAs (Xilinx Inc, 2007).

[5] P.P. Vaidyanathan, Multirate Digital Filters, Filter Banks, Polyphase Networks, and Applications: A Tutorial, *Proceedings of the IEEE*, vol. **79**, no. 1, (1990) pp. 56–93

- [6] A. V. Oppenheim, R. W. Schaffer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd edition (Prentice Hall, 1999).
- [7] Altera Corporation, *Simplifying Simultaneous Multimode RRH Design*, white paper (2009).
- [8] F.-L. Luo (ed.) *Mobile Multimedia Broadcasting Standards: Technology and Practice* (Springer, 2008).
- [9] Y. Zhang, J. Cosmas, K. K. Loo, and M. Bard, Design and Implementation of Digital Echo Cancellation On-Channel Repeater in DVB-T/H Networks, *57th Annual IEEE Broadcast Symposium, Washington DC, USA*, (2007).
- [10] Y. Zhang, J. Cosmas, M. Bard, and Y. H. Song, Diversity Gain for DVB-H by Using Transmitter/Receiver Cyclic Delay Diversity, *IEEE Transactions on Broadcasting*, vol. **52**, no. 4 (2006) pp 464–474.
- [11] IEEE Standards *IEEE 802.11n-2009 – Amendment 5: Enhancements for Higher Throughput* (IEEE-SA, 2009).
- [12] G.-T. Gil, Nondata-Aided I/Q Mismatch and DC Offset Compensation for Direct-Conversion Receivers, *IEEE Transactions on Signal Processing*, vol. **56**, no. 7 (2007) pp. 2662–2668.
- [13] L. Cordesses, Direct Digital Synthesis: A Tool for Periodic Wave Generation (Part 1), *IEEE Signal Processing Magazine, DSP Tips & Tricks*, vol. **21**, no. 4 (2004) pp. 50–54.
- [14] *IEEE Programs for Digital Signal Processing*, IEEE Press (John Wiley & Sons, 1979).
- [15] M. D. McKinley, K. A. Rewley, M. Myslinski, *et al.*, EVM Calculation for Broadband Modulated Signals, *64th ARFTG Conf. Dig.*, Orlando, FL, (2004) pp. 45–52.

13 A/D and D/A data conversion for wireless communications transceivers

Jerzy Dąbrowski

Contemporary wireless communications have expanded to provide a multitude of services ranging from mobile telephony and satellite navigation to Internet access and image or video transfer. These functions have been made possible by the digital signal processing (DSP) techniques implemented in dedicated hardware. In addition to the primary functions also various digital techniques must be employed to ensure integrity and reliability of communications systems, increasing thereby the demands for the system throughput. To enable ever higher data rates both wider channel bandwidths and advanced DSP techniques such as Orthogonal Frequency-Division Multiplexing (OFDM) and multi-bit quadrature amplitude modulation (QAM) are used. These solutions place more demands on the signal-to-noise ratio (SNR) and peak-to-average power ratio (PAPR) of the signals, and in effect, on the dynamic range (DR) and linearity of the analog front-end of a transceiver.

The requirements on the interface between the digital and analog front-end part are just as important. In fact, the quality of the A/D and D/A data conversion largely decides the performance of the received and transmitted signal. The demands placed on the data converters in wireless transceivers are exacerbated in terms of multimode operation, and are pushed to an extreme for the software defined radio (SDR) where the converters should be placed close to antenna [1].

Even though such a “true SDR” has not been attained yet, hampered by limitations in speed, dynamic range, linearity, and power consumption, RF sampling receiver front-ends based on discrete-time signal processing have been demonstrated as an intermediate step in this direction [2,3,4]. In this case the traditional RF mixers and analog channel filters are replaced by switched-capacitor (SC) decimation filters followed by baseband- or IF-operating A/D converters (ADCs). Also flexible transmitter architectures based on two-point polar modulation [5,6] or poly-phase power amplification [7] with limited need for RF filtering have been proposed. A further progress in RF technologies, especially in programmable RF filters is believed to make the SDR challenge a reality in the near future.

As of today, the new generations, 3G and 4G, of communications systems also place challenging demands on the RF front-ends and consequently on the employed A/D and D/A converters due to speed, resolution, and linearity. In this chapter we discuss A/D and D/A conversion in terms of design requirements typical of contemporary wireless transceivers.

The rest of this chapter is organized into eight sections. [Section 13.1](#) is devoted to A/D conversion in RF receivers. Based on the fundamental specifications like sensitivity, selectivity, and blocking profile we estimate the dynamic range and the equivalent number of bits (ENOB) for an ADC in a receiver. Also the corresponding linearity constraints are put in perspective. In the [second section](#), we discuss A/D conversion in terms of the RF image problem typical of heterodyne and low-IF receiver architectures. Specifically, IF applications of ADCs are considered. In this context in [Section 13.3](#) we also discuss digital down-conversion (DDC) and, consequently, on the transmitter side in [Section 13.6](#), digital up-conversion (DUC) is addressed. Implementation examples of state-of-the-art ADCs based on CMOS technology are presented in [Section 13.4](#). The following three sections are devoted to D/A conversion. In [Section 13.5](#) D/A converters (DACs) are primarily discussed in terms of fundamental transmitter specifications: the error-vector-magnitude (EVM) and the adjacent-channel-power-rejection (ACPR). Based on this the ENOB-estimate and linearity constraints imposed on the transmitter DAC are found. Implementation examples of state-of-the-art DACs using CMOS process are given in [Section 13.7](#), where in addition to baseband and IF D/A applications, we also consider the emerging RF DACs. The [final section](#) provides a summary and further discussions.

13.1 A/D conversion in RF receiver

The design of an adequate A/D converter (ADC) appears a particular challenge as the received signal is exposed to interference and noise that degrade the signal-to-noise ratio (SNR) and thereby the receiver bit error rate (BER). The receiver SNR is affected externally in several ways. Strong in-band interference gives rise to increased noise floor due to the local oscillator phase noise and the reciprocal mixing effect. On the other hand, the limited linearity of the receiver front-end evokes intermodulation products which can fall in the desired signal band and effectively act as extra noise. In quadrature receiver architectures the possible I/Q mismatch degrades image (or mirror) rejection and thereby elevates the noise floor as well. Additionally, in full-duplex (FDD) transceivers not only the external interference but also the transmitter leakage must be taken into account.

To alleviate the interference effects, filters are extensively employed in receivers. Clearly, for the performance of a receiver, effective filtering is essential. However, in modern integrated receivers where the chip area overhead and power consumption are the key factors, relatively simple analog filters are preferred wherever possible. As a consequence the in-band interferers are only partly attenuated and they reach the ADC input to be finally removed in the digital domain. This approach provides more flexibility to a receiver and facilitates multi-standard operation when the filter bandwidths can be controlled. On the other hand, more stringent requirements are imposed on the ADC in terms of dynamic range and linearity.

These requirements are usually exacerbated for A/D converters in wideband base-station receivers, where many channels are processed in parallel. In effect the ADC is exposed to their unfiltered power profile (blocking profile) while channelization is carried

out by digital filters. Consequently, also much larger ADC bandwidth is required but the power consumption limit, as compared to mobile receiver applications, can be relaxed.

13.1.1 ADC dynamic range

Consider a typical terminal receiver architecture, zero-IF or low-IF, where an ADC is placed after a channel filter and variable gain amplifier as shown in Figure 13.1. The channel filter only provides a limited attenuation (rejection) to the in-band interferers (blockers). Consequently, the ADC has to maintain SNR required by the demodulator, when the maximum blocker and a minimum wanted signal power S are present at the ADC input simultaneously. The required ADC dynamic range follows

$$DR = (P_B - R_B)_{max} - (S + G_P - SNR_{min}) + \Delta_{DR} \quad (13.1)$$

where P_B is the blocker power at the receiver input, R_B is the corresponding rejection mainly provided by the channel filter, and G_P is the possible processing gain typical of spectrum spreading systems. Here, $(S + G_P - SNR_{min})$ stands for the maximum allowed noise and distortion level, and it is also a receiver input-referred quantity. A reserve Δ_{DR} is used for possible variations in receiver gain and signal magnitude, such as peak-to-average ratio (PAPR). For a zero-IF receiver this reserve should also cover the maximum DC offset.

To understand the impact of G_P we observe that SNR_{min} is the value required after the spectrum dispreading so before this process $(SNR_{min} - G_P)$ is sufficient as also illustrated in Figure 13.2.

The required DR largely depends on the blocking profile and receiver selectivity that is reflected here by the maximum blocking power $(P_B - R_B)$. The channel filter is responsible for the blocker attenuation (rejection) while the ultimate channel filtering can be completed after ADC in the digital domain. On the other hand the channel filter should band-limit the signal to prevent aliasing in A/D conversion.

For further illustrations, let us consider the UMTS system with $S = -114$ dBm over 4 MHz bandwidth and $G_P = 25$ dB, $P_B = -15$ dBm at 25 MHz offset, $SNR_{min} = 8$ dB, and $\Delta_{DR} = 9$ dB. From (13.1) we find $DR = 91 - R_B$. In zero-IF architecture the signal occupies a band between zero and 2 MHz and the m th order channel filter with cut-off frequency placed at 3 MHz (i.e. at the edge of the adjacent channel spaced by 5 MHz) provides blocker rejection $R_B \approx 20m \cdot \log(25/3) \approx 18m$ dB. To manage the blocker with a simple low-pass filter, e.g. second order, a dynamic range $DR \approx 55$ dB would be sufficient. With a continuous-time $\Sigma\Delta$ A/D converter, additional anti-alias filtering can

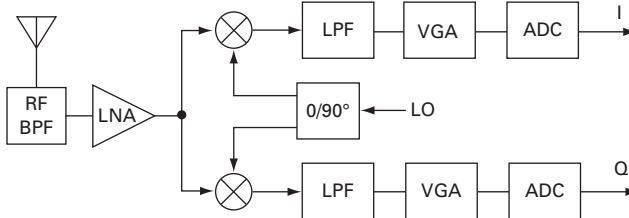


Figure 13.1 Standard architecture of zero-IF or low-IF receiver.

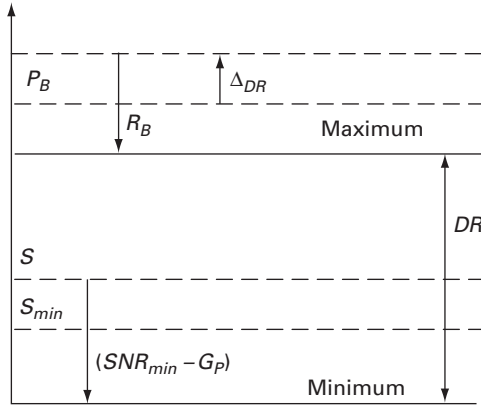


Figure 13.2 Signal and noise levels in ADC.

be achieved so as to support a large enough oversampling ratio. Note that this kind of ADC offers inherent anti-alias filtering and it is becoming an attractive solution in CMOS technology [8,9].

As seen, there is a trade-off between the channel filter order and DR of ADC. The ultimate channel selection can be completed in the digital domain, but to minimize power consumption that is a common objective for mobile sets, the chain composed of the analog channel filter, ADC, and the following digital filter should be carefully optimized. Adoption of a receiver to multi-standard operation appears an additional constraint in this case.

On another extreme, when the desired signal attains its maximum level, say approaching P_B , it should be attenuated by AGC as it is not subject to the filter attenuation. To keep the signal level safely within the ADC voltage range, gain reduction equal to R_B should be provided in this case. In many applications, automatic gain control (AGC) is distributed between the IF (or baseband) stage and low-noise amplifier (LNA) to prevent overloading of the down-converter. Also by placing AGC in LNA less gain mismatch between the receiver I/Q branches can be expected.

13.1.2 Equivalent number of bits

To cover all kinds of noise, i.e. the quantization, thermal, and clock jitter-induced noise in ADC a performance measure referred to as the equivalent number of bits (ENOB) can be employed. The following formula holds

$$ENOB = \frac{SNR_{ADC} - 1.76}{6.02} \quad (13.2)$$

When SNR_{ADC} in (13.2) is replaced by signal-to-noise-and-distortion ratio ($SNDR_{ADC}$) then also ADC distortions are covered by ENOB. The power and noise levels as shown in Figure 13.2, can be related to the ADC voltage range. If the ADC input impedance is Z_{ADC} then (assuming a real value for simplicity)

$$10 \log \frac{V_{Range}^2/8}{Z_{ADC} \times 1 \text{ mW}} - G_{FE} = P_B + \Delta_{DR} - R_B \quad (13.3)$$

where G_{FE} stands for the power gain of the receiver analog front-end. The inherent ADC voltage noise (incl. distortion) can be found from

$$SNDR_{ADC} = 10 \log \frac{V_{Range}^2/8}{V_{n+d,ADC}^2} \quad (13.4)$$

and when combined with the analog front-end noise and distortions contributed by in-band blockers, it decides the receiver SNR. Then according to conditions defined by (13.1) the minimum SNR requirement at the ADC output can be expressed as

$$SNR_{min} - G_P = (S + G_{FE}) - 10 \log \frac{V_{n+d,FE}^2 + V_{n+d,ADC}^2}{Z_{ADC} \times 1 \text{ mW}} \quad (13.5)$$

where $V_{n+d,FE}^2$ is the front-end noise including distortions falling in the signal band. Both $V_{n+d,FE}^2$ and $V_{n+d,ADC}^2$ are calculated over the signal bandwidth. Assuming $V_{n+d,ADC}^2 = \alpha V_{n+d,FE}^2$ with $\alpha < 1$, or alternatively

$$SNDR_{ADC} = SNDR_{FE} - 10 \log \alpha \quad (13.6)$$

we find the ADC contribution to the receiver SNR_{min} (given by (13.5)) as $10 \log(1 + \alpha)$. Also the receiver noise figure would be degraded in a similar way. Clearly, a relatively small α value would be preferred in this case. Observe that for a lower blocking power, AGC would increase the front-end gain and at the same time also the front-end noise would be increased. In effect, the relative contribution of ADC noise would be even more limited.

From (13.1)–(13.6) the following estimate can be derived

$$ENOB = \frac{DR - 1.76 + 10 \log(1 + 1/\alpha)}{6.02} \quad (13.7)$$

which is the required ENOB measured under reference conditions for the stand-alone ADC. Here, the total effect of receiver noise and distortions is reflected by DR and $1/\alpha$. As seen, when reducing the contribution by ADC more demands must be placed on the converter resolution.

As another example, consider the UMTS zero-IF receiver with the dynamic range $DR \approx 55$ dB as discussed in the first example. Assume the front-end SNDR is reduced by the inherent ADC noise and distortion by $\Delta = 0.5$ dB. Then using $\Delta = 10 \log(1 + \alpha)$ we have $\alpha = 0.122$ and from (13.7) we find $ENOB = 10.5$. Assuming also $V_{Range} = 1$ V and $Z_{ADC} = 500 \Omega$ from (13.3) the front-end gain adequate to the blocking condition is $G_{FE} = 36$ dB. When the blocker power drops by ΔP_B the automatic gain control mechanism (AGC) can increase the receiver gain by the same amount.

The ENOB estimate of this example can be compared with ADC design results reported recently. From [11], [10], [9] for UMTS we find 10.6, 10.5, and 11.2/12 bits, respectively.

13.1.3 Linearity of ADC

Most of the harmful nonlinear effects in a receiver are evoked by strong in-band interferers in radio-frequency (RF) stages. Note that blockers for various standards can easily surpass the signal level by 80 to 100 dB. Hence, the intermodulation and harmonic distortions that originate in the desired signal are negligible compared to those evoked by blockers. The corresponding nonlinear products tend to fall in the receive channel and degrade SNR so the design requirements imposed on linearity of the RF front-end used to be strict [12]. The second-order nonlinear effects can be largely mitigated by using differential circuits but for zero-IF or low-IF receivers usually a calibration is also required.

When the signal is large also the distortions originating in the signal are relatively large. However, they are kept in the same proportion to the signal as they are to the blockers for a low signal level.

This can be illustrated for IM2 distortion as follows (Figure 13.3). According to a blocker P_B the baseband IM2 power (input-referred) can be estimated from

$$P_{IM2,B} = 2(P_B - 3) - IIP_2 \quad (13.8)$$

where IIP_2 is the second input-referred intercept point of the receiver. To maintain SNR of a weak signal S we can assume $P_{IM2,B} < N_{floor}$ (receiver noise floor). The blocker is modeled here by two closely spaced tones, each of a power $(P_B - 3)$ dBm.

The nonlinear distortions in a receiver are mainly attributed to the RF stages. The nonlinearities of the IF and baseband stages take less effect or, ideally, no effect since the interferers are attenuated by the channel filter. An ADC can also contribute to the receiver distortions. This contribution is reflected by the converter dynamic performance measures such as IIP2, IIP3 or SFDR (spurious free dynamic range) rather than by INL (integral nonlinearity) and the static input-output characteristic.

For illustration consider the ADC second-order nonlinearity expressed by IIP2. In this case the IM2 power referred to the receiver input, can be calculated from

$$P_{IM2} \cong 2(P_B - 3) + 10 \log \left(10^{\frac{-IIP_{2,FE}}{10}} + 10^{\frac{-IIP_{2,ADC} + G_{FE} - R_B}{10}} \right) \quad (13.9)$$

where the logarithmic term represents the receiver total IIP2 with components by the analog front-end and A/D converter. As seen the ADC contribution is subject to the

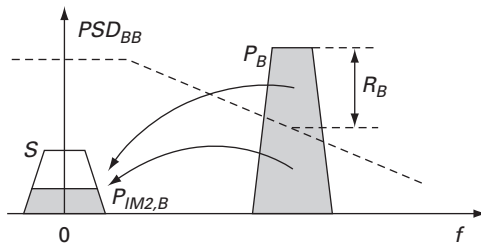


Figure 13.3 Evolution of IM2 distortion in zero-IF receiver under strong blocker.

effective front-end gain ($G_{FE} - R_B$) so it can be reduced by a large enough blocker rejection R_B . However, it is not the case when a simple analog channel filter is used and the blocker is only partly suppressed as discussed in the UMTS example above so the effect of ADC cannot be neglected.

In a similar way the IM3 distortion and the respective ADC contribution can be addressed as well.

As compared to the analog front-end one distinctive feature of the ADC is the differential nonlinearity (DNL) which quantifies discrepancies between subsequent quantization levels in the converter. The impact of DNL on the converter SNR assuming it is the dominant effect [13 Chapter 3.3], can be estimated from

$$SNR_{DNL} = 6.02n - 20 \log(1 + \varepsilon) \quad (13.10)$$

where ε is the average DNL and n is the ADC number of bits. For example, a converter with $\varepsilon = 1.5$ LSB (least significant bit) has SNR degraded by 6.2 dB from its ideal value and in this way the ENOB is reduced by approximately 1 bit.

Altogether for a single channel receiver the nonlinear distortions in ADC can be incorporated in noise and quantified jointly by SNDR as considered in (13.4). The other ADC specification, spurious free dynamic range (SFDR), seems to be less meaningful in this context since the spurs which fall in the signal band are covered by SNDR. If the distortions combined with noise are limited according to (13.5) the receiver performance is maintained.

However, for a multi-channel receiver such as in cellular base-stations, the SFDR appears a critical specification dictated by the blocking profile. In fact, the problem is to maintain SNR in weak channels that are exposed to spurs evoked by blockers. For this reason, for cellular standards like EDGE or UMTS a base-station ADC with a SFDR of 85 to 90 dB is required [14]. To improve SFDR creating thereby a possibly uniform ADC noise floor the analog dithering [13 Chapter 3.3] or digital correction techniques can be used [15 Chapter 6], [16 Chapter 3.6], [17 Chapter 9]. Nevertheless, for wideband spread-spectrum or OFDM systems where the accompanying harmonic or intermodulation distortions are more like noise, these approaches are less meaningful.

13.2 ADC for heterodyne and low-IF receiver

The heterodyne receivers are subject to the image problem. Insufficient suppression of an image results in increased receiver noise floor that degrades sensitivity. An image is located at $2f_{IF}$ distance from the signal, so for larger IF values more image rejection can be achieved. In this case, however, the receiver selectivity suffers as it is more difficult to filter out the close-in interferers (adjacent channels) at higher frequencies.

The trade-off between selectivity and sensitivity is usually mitigated by using a double-IF architecture. Still the heterodyne receivers require bulky image reject filters, which are unsuited for integration. One viable option in this case is image-reject down-conversion by two I/Q mixers as shown in Figure 13.4.

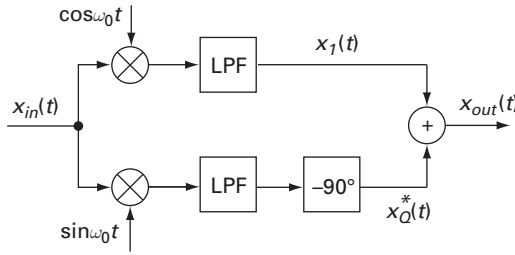


Figure 13.4 Image reject down-converter in Hartley architecture.

If the desired modulated signal and its image (denoted by subscripts S and Im , respectively) are expressed by

$$x_{in}(t) = A_S(t) \cdot \cos(\omega_S t + \varphi_S(t)) + A_{Im}(t) \cdot \cos(\omega_{Im} t + \varphi_{Im}(t)) \quad (13.11)$$

then the I and Q components after a low-pass filtering appear as

$$\begin{aligned} x_{in}(t) \cos \omega_0 t &\xrightarrow{LPF} x_I(t) = A_S(t) \cdot \cos(\omega_{IF} t - \varphi_S(t)) + A_{Im}(t) \cdot \cos(\omega_{IF} t + \varphi_{Im}(t)) \\ x_{in}(t) \sin \omega_0 t &\xrightarrow{LPF} x_Q(t) = A_S(t) \cdot \sin(\omega_{IF} t - \varphi_S(t)) - A_{Im}(t) \cdot \sin(\omega_{IF} t + \varphi_{Im}(t)) \end{aligned} \quad (13.12)$$

where $\omega_0 = (\omega_S + \omega_{Im})/2$ is the local oscillator (LO) frequency and $\omega_{IF} = |\omega_S - \omega_{Im}|/2$. To cancel the image one of the I and Q components needs a phase shift by 90 degrees. Applying the phase shift to x_Q , from (13.12) we find

$$x_Q^*(t) = -A_S(t) \cdot \cos(\omega_{IF} t - \varphi_S(t)) + A_{Im}(t) \cdot \cos(\omega_{IF} t + \varphi_{Im}(t)). \quad (13.13)$$

Ideally, the image at IF is removed by subtraction $x_I - x_Q^*$, but the quality of this process heavily depends on the I/Q match and accuracy of the phase shift. One viable phase shift technique is based on the Hilbert transform implemented by four I/Q mixers architecture. The accuracy of this technique depends again on the I/Q match so for the best performance a digital rather than analog implementation would be preferred.

In Figure 13.5 a low-IF receiver architecture with a quadrature down-converter and a four-mixer phase shifter is shown. The IF would typically be half the channel spacing (bandwidth) so an image is the adjacent channel that cannot be suppressed before down-conversion to IF. The I and Q components (contaminated by an image as in (13.12)) are first low-passed and next converted to the digital domain. Bandpass ADCs would be preferred in this case to limit the quantization in-band noise.

Here, the phase shifter works also as a digital down-converter (DDC) to baseband so at the same time the signal is demodulated and it appears at the output of the I and Q branch (upper and lower) as

$$\begin{aligned} a_I(kT) &= A_S(kT) \cdot \cos \varphi_S(kT) \\ a_Q(kT) &= A_S(kT) \cdot \sin \varphi_S(kT) \end{aligned} \quad (13.14)$$

where kT is a discrete time according to the ADC sampling rate $1/T$.

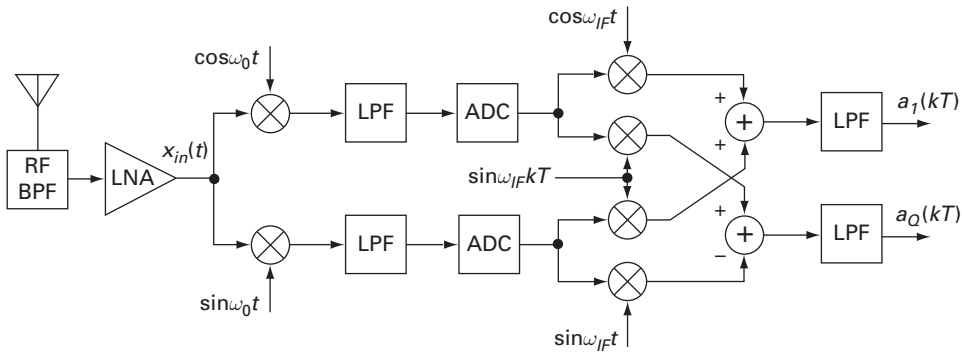


Figure 13.5 Low-IF receiver with DDC to baseband.

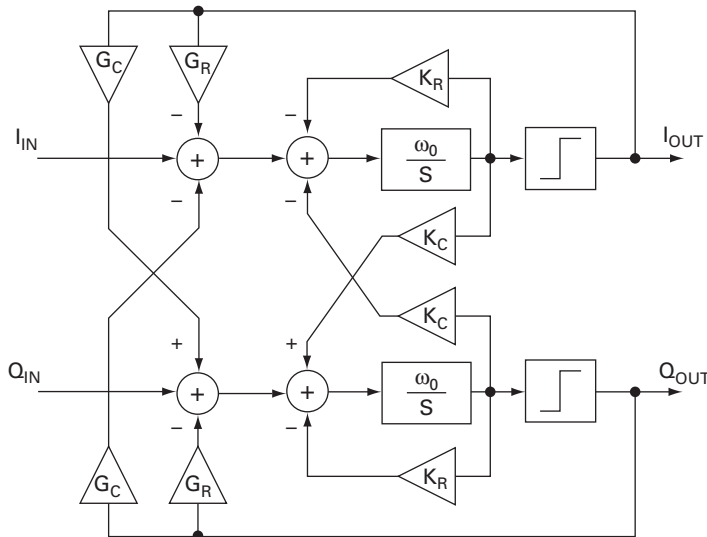


Figure 13.6 Model of quadrature $\Sigma\Delta$ ADC with complex bandpass filter, adapted from [18], © 2002 IEEE.

The advantage of this architecture is that it is well suited for integration while the shortcomings typical of a homodyne (zero-IF) receiver, i.e. DC offset and $1/f$ noise are mitigated. Bulky image rejection filters are avoided and a small IQ mismatch can be tolerated unless the adjacent channel power is large. For this reason the low-IF architecture has been well adopted e.g. to GSM standard where the difference between power of the adjacent and the received channel is limited to 9 dB.

Rather than using two independent ADCs one quadrature ADC can be used to save power. If the ADC is implemented as a continuous-time $\Sigma\Delta$ modulator the requirements on analog low-pass filters (as in Figure 13.6) are relaxed. By implementing bandpass loop filters in the $\Sigma\Delta$ modulator the advantage in power efficiency is taken according to the signal bandwidth to IF ratio [18].

Unlike the low-IF, in a standard heterodyne architecture one mixer per IF stage is sufficient, but image rejection filters are indispensable. In this case one ADC is used and

for down-conversion (DDC) to baseband two mixers operating in quadrature are sufficient. For an arbitrary IF signal defined by $x_{IF}(t) = A_S(t) \cdot \cos(\omega_{IF}t + \phi_S(t))$ the DDC process can be described as follows

$$\begin{aligned} x_{IF}(kT) \cos \omega_{IF}kT &\xrightarrow{LPF} x_I(kT) = A_S(kT) \cdot \cos \phi_S(kT) \\ x_{IF}(kT) \sin \omega_{IF}kT &\xrightarrow{LPF} x_Q(kT) = -A_S(kT) \cdot \sin \phi_S(kT). \end{aligned} \quad (13.15)$$

Low-pass filtering is also used here to suppress the unwanted components and high-frequency noise.

13.3 Digital down-conversion

With increased speed of ADC it is also possible to complete the down-conversion from IF to baseband in the digital domain. The digital down-conversion (DDC) can be carried out with high precision that is hardly achieved in analog circuits. Specifically, the harmful process variations and I/Q mismatch effects are avoided in this way. Moreover, the functions defined in software provide a receiver with the flexibility needed for multimode operation. In this case, the advantages justify some extra power consumption overhead.

13.3.1 Multiplying technique

The conventional digital down-conversion is performed by multiplying the digitized IF signal by I/Q carrier as illustrated in [Figure 13.5](#). The respective cosine and sine values can be stored in a cyclic memory and applied to the DDC at a rate equal to IF. For n -bit resolution the memory size can be estimated as $\sim n 2^n$ bits [19] that results in a significant chip area overhead for large n values. A few techniques have been proposed to overcome this drawback.

Firstly, we observe that the discrete time angle $2\pi f_{IF}kT$ depends on f_{IF} to the sampling rate $(1/T)$ ratio. If the sampling rate is related to IF by

$$\frac{1}{T} = 4f_{IF} \quad (13.16)$$

then the angle becomes $k\pi/2$ so the I/Q carriers only take on values of $\{1, 0, -1\}$. The resolution and memory size problem are overcome in this way and also the four multipliers (mixers) can be replaced by multiplexers. However, to obey the Nyquist criterion, the f_{IF} should be chosen larger than half the signal bandwidth ($BW/2$).

If an oversampling ADC is used and the sampling frequency is constrained by

$$\frac{1}{T} = 4nf_{IF} \quad (13.17)$$

with integer n , then the angle of the discrete time carrier appears as $k\pi/2n$. In this case, decimation by n of the signal is required to maintain the ternary-valued carrier $\{1, 0, -1\}$ and thereby the savings in hardware.

In a more flexible solution the sine and cosine are generated by a digital n -bit IIR filter rather than by a memory. In this case the classical oscillator theory applies necessitating the complex closed-loop gain of the filter to be one with 2π phase shift. The generator scales in frequency by an external clock and the conditions (13.16) and (13.17) can be neglected.

13.3.2 CORDIC technique

An alternative approach to save the memory in DDC is based on the coordinate rotation digital computer algorithm (CORDIC) that was primarily invented for converting between Cartesian and polar coordinates. The iterative CORDIC algorithm can rotate the discrete time IF signal by an angle of $2\pi f_{IF}kT$, performing thereby down- or up-conversion [19]. The required memory size is only $\sim n^2$ bits and the multipliers can be eliminated. However, a relatively large pipeline circuitry must be used to implement iterations required by CORDIC. The corresponding chip area overhead is equivalent to three multipliers, so the most savings are achieved on the memory size.

13.3.3 Subsampling technique

If the image is suppressed in the analog front-end the digitized signal can be down-converted using subsampling (down-sampling) technique. For this purpose, first, the discrete-time IF signal $x_{IF}(kT) = A_0(t) \cdot \cos(2\pi f_{IF}kT + \varphi_0(kT))$ is put in quadrature by interleaving as shown in Figure 13.7. This technique is also referred to as uniform I/Q sampling. In this case f_{IF} should be an odd multiple of the quarter sampling rate, i.e. $f_{IF} = (2m+1)/4T$. This also means the ADC sampling rate can be less than $2f_{IF}$ that is a special case of the bandpass sampling technique [20], which is useful to save power.

For this frequency plan, first, the IF signal is down-converted to $1/4T$ while the angle $2\pi f_{IF}kT$ is effectively equal to $k\pi/2$, so for the odd and even values of k we achieve

$$\begin{aligned} x_{IF}^{odd}(kT) &= (-1)^{\frac{k+1}{2}} A_0(kT) \cdot \sin \varphi_0(kT) \\ x_{IF}^{even}(kT) &= (-1)^{\frac{k}{2}} A_0(kT) \cdot \cos \varphi_0(kT). \end{aligned} \quad (13.18)$$

From (13.18) we see that the I and Q components are separated and their subsequent values are multiplied by ± 1 , respectively. Due to interleaving the data rate in either branch is $1/(2T)$ so the discrete time can be redefined as $2nT$.

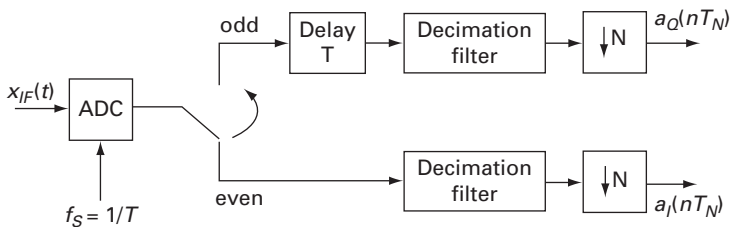


Figure 13.7 Down-conversion by interleaving and sub-sampling.

Next, by down-sampling with a factor N , (the sampling period $2NT = T_N$) the IF signal (13.18) appears as

$$\begin{aligned} x_{IF}^{odd}(nT_N) &= (-1)^{nN} A_0(nT_N) \cdot \sin \phi_0(nT_N) \\ x_{IF}^{even}(nT_N) &= (-1)^{nN} A_0(nT_N) \cdot \cos \phi_0(nT_N) \end{aligned} \quad (13.19)$$

where N should be chosen as an even number to cancel the multiplication by ± 1 for consistent demodulation.

Moreover, N should be large enough to ensure a data rate $1/T_N$ suitable for further baseband processing. On the other hand, the Nyquist criterion must be obeyed to avoid a possible spectrum folding so the upper bound for N applies

$$N < \frac{1}{4TB} \quad (13.20)$$

where B is the guarded signal bandwidth.

At the same time, the signal spectrum is down-converted to baseband with multiple replicas spaced by $1/T_N$ as shown in Figure 13.8. Some fragments of this spectrum at $1/4T$, $3/4T$, $5/4T$, ... can be attributed to sampling in ADC whereas the remaining originate with the down-sampling (decimation). In fact, this spectrum also undergoes shaping by the decimation filter as illustrated in Figure 13.8 by dotted line. According to multiple notches this filter prevents interferers and noise from the image bands, located at m/T_N , from folding to baseband.

Image rejection in a heterodyne receiver can also be achieved by simple manipulation on the digital I/Q data. Finally, down-conversion to baseband can be completed by subsampling with decimation filters as already discussed. In this case also the condition $1/T = 4f_{IF}$ applies and like for (13.12) we find

$$\begin{aligned} x_I(k) &= A_S(k) \cdot \cos(k\pi/2 - \varphi_S(k)) + A_{Im}(k) \cdot \cos(k\pi/2 + \varphi_{Im}(k)) \\ x_Q(k) &= A_S(k) \cdot \sin(k\pi/2 - \varphi_S(k)) - A_{Im}(k) \cdot \sin(k\pi/2 + \varphi_{Im}(k)) \end{aligned} \quad (13.21)$$

where T has been omitted to simplify the notation while the index S denotes the wanted signal and Im the image. Using a delay by T (equivalent to $\pi/2$ at f_{IF}) the I component follows

$$\begin{aligned} x_I(k-1) &= A_S(k-1) \cdot \sin(k\pi/2 - \varphi_S(k-1)) + A_{Im}(k-1) \cdot \sin(k\pi/2 + \varphi_{Im}(k-1)) \\ &\cong A_S(k) \cdot \sin(k\pi/2 - \varphi_S(k)) + A_{Im}(k) \cdot \sin(k\pi/2 + \varphi_{Im}(k)). \end{aligned} \quad (13.22)$$

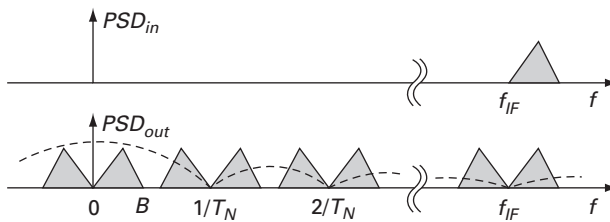


Figure 13.8 Power spectral density (PSD) of signal at ADC input and after down-conversion by sub-sampling for single-sideband modulated signal.

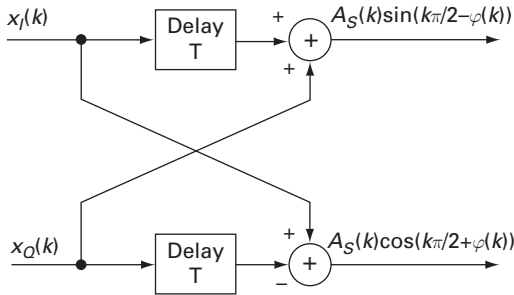


Figure 13.9 Discrete-time image reject filter.

Hence, it is easy to see that by simple combination of $x_I(k-1)$ and $x_Q(k)$ the image is removed from the Q component and in a similar way also the I component can be achieved as illustrated in Figure 13.9. This circuit can also be considered a complex discrete time filter with a transfer function

$$H(z) = z^{-1} + j \quad (13.23)$$

which has a notch at $f = 1/4T$, responsible for image rejection. Observe that the approximation used in (13.22) holds provided the signal (and image) bandwidth is much less than $1/T$ so in practice the image rejection is limited. Improvement can be achieved by increasing the notch bandwidth. For this purpose a cascade of a few filter stages shown in Figure 13.9 can be used [21]. Still, this technique is not suited for a low-IF receiver for which f_{IF} is less than the signal bandwidth.

A disadvantage of the down-conversion techniques using the sampling frequency of $4f_{IF}$ or $4f_{IF}/(2m+1)$ occurs due to harmonic distortions. Observe that during A/D conversion the third-order harmonic at $3f_{IF}$ appears as an image, which folds back to f_{IF} . Obviously, the SNDR suffers in this way, but this effect is usually less meaningful compared to IM2 and IM3 intermodulation according to in-band blockers.

13.4 Implementation of ADCs in communications receivers

Among various types of A/D converters mainly the pipeline and $\Sigma\Delta$ architectures have been implemented in contemporary communications receivers (transceivers). The advantage of a pipeline ADC is that the data rate is independent of the number of the used stages so a high speed and high resolution can be achieved at the same time. A basic architecture of this converter is shown in Figure 13.10, where each stage resolves a certain number of bits and generates a residue that is digitized by the next stages. In effect, the hardware overhead is approximately proportional to the converter resolution. Moreover, the accuracy requirements for the later stages are relaxed compared to the input stages that allows the ADC power consumption to be reduced. Digital correction techniques can be used to cancel the effects of analog component nonlinearities in this converter [15]. As the signal needs to propagate through all stages until the final

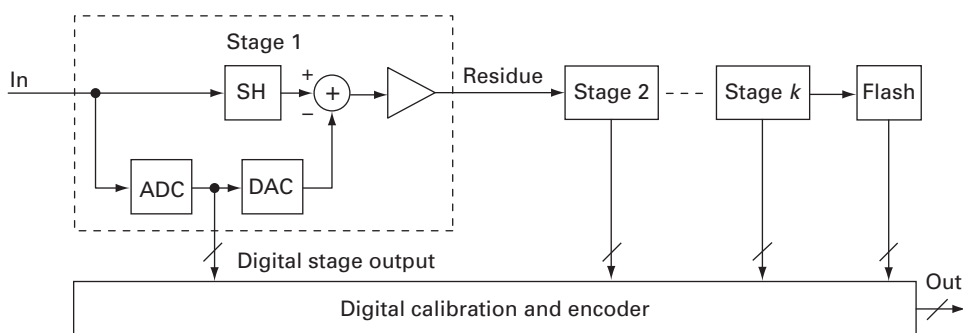


Figure 13.10 Basic architecture of pipeline ADC.

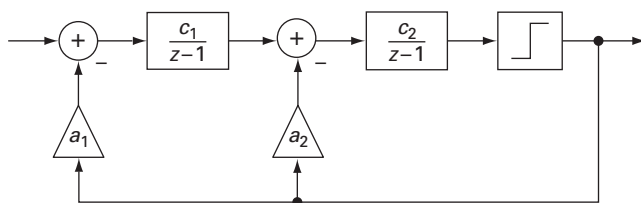


Figure 13.11 Basic architecture of second-order low-pass $\Sigma\Delta$ ADC with one-bit quantizer.

conversion is completed a conversion latency occurs, which must be tolerated during the system start-up.

Driven by the need for giga-sample per second data rates that is especially useful in massive data transfer, time-interleaving ADCs can be also introduced [22], [17 Chapter 10]. The challenge with interleaving is to maintain accurate phase alignment of sampling clock edges and also to compensate for manufacturing variations that occur between ADCs. In practice, these problems are solved by precise synchronization and digital correction techniques that are not addressed in this chapter.

The $\Sigma\Delta$ A/D converters for communications have emerged from the audio applications where they have been appreciated for their excellent noise performance and high linearity [23]. A basic architecture of the second-order low-pass $\Sigma\Delta$ ADC with one-bit quantizer is depicted in Figure 13.11. Using the negative feedback the $\Sigma\Delta$ converters push the quantization noise out of the signal band that is commonly known as noise shaping. To accomplish this task the oversampling technique must be used, i.e. $1/T \gg B/2$. Recently, ever faster semiconductor devices have enabled migration of the $\Sigma\Delta$ conversion technique from the audio field to radio communications applications, especially in CMOS technology. Low-pass or bandpass $\Sigma\Delta$ A/D converters have been used depending on the receiver architecture to support the common requirement for a low noise figure and limited power consumption at the same time. However, satisfying the oversampling condition for standards with signal bandwidths, which demand clock frequencies in GHz range (such as WiMAX) still appears a design challenge in the CMOS technology. The reasons are parasitics which deteriorate the performance of devices at GHz frequencies, and especially the limitations in power, gain-bandwidth product, and linearity of the analog blocks.

The $\Sigma\Delta$ ADCs can be built as continuous-time (CT) or discrete-time (DT) systems, the latter also called SC for switched-capacitor technique commonly used for this purpose. The main advantage of the CT $\Sigma\Delta$ architecture is its inherent anti-alias filtering function. Also the CT integrators are better suited to handle wideband signals than their SC counterparts which are largely limited in frequency by settling times. On the other hand, unlike DT $\Sigma\Delta$ the CT $\Sigma\Delta$ converters are prone to instability evoked by a loop delay and are very sensitive to clock jitter boosting the converter noise floor. Moreover, the transfer functions of the CT analog blocks are subject to process variations and mismatch, whereas for SC circuits they rely on robust capacitance ratios.

Following the need for multi-standard (multimode) mobile terminals, also dubbed “software defined radio” (SDR), flexible communications A/D converters have been developed, recently. Especially, the $\Sigma\Delta$ architectures have proven flexible in terms of programmable bandwidth and resolution while saving power.

Examples of state-of-the-art communications ADCs based on CMOS technology are discussed below.

13.4.1 Multi-standard $\Sigma\Delta$ ADCs

A continuous-time third-order $\Sigma\Delta$ A/D converter reported in [9] achieves performance adequate for WLAN, DVB, UMTS, and BT standards with very low power consumption. The integrators and DACs used in this circuit are programmed according to the signal bandwidth varying between 500 kHz for BT and 10 MHz for WLAN in a zero-IF reception scenario. This flexibility allows the lowest power consumption to be obtained for a given bandwidth. More importantly, the gain-bandwidth product requirement for the integrator in the input stage is much relaxed compared to the following stages, while its linearity and noise performance must be superior. The noise contribution of the following stages is less meaningful according to the gain provided by the first stage. Single-bit rather than multi-bit quantization is used for best linearity while avoiding costly dynamic element matching circuits (DEM) typical of multi-bit converters. A threshold-configurable comparator is used in the quantizer to compensate for the loop delay, which could badly affect the stability of the converter. To reduce the converter sensitivity to clock jitter an SC DAC is used in the first stage. The circuit is integrated using a digital 90 nm CMOS process with a 1.2 V supply and it occupies an area of 0.4 mm². In Table 13.1 the main performance

Table 13.1 Performance measures of CT $\Sigma\Delta$ ADC, adapted from [9], © 2010 IEEE.

Mode	BT	UMTS2	UMTS3	DVB	WLAN
BW [MHz]	0.5	1.92	1.92	4	10
f_s [MHz]	90	245.8	245.8	512	640
DR [dB]	84.2	74.6	79.6	76.4	69.6
$SNDR_{max}$ [dB]	77	69	74	70	65
$SFDR$ [dB]	81	82	80	86	72
$IM3$ [dB]	> 75	> 75	> 75	> 75	70
Dif. input range [V _{pp}]	1.6	1.7	2.1	1.9	1.9
P_{DC} [mW]	5	4.34	6.44	5.5	6.8

Table 13.2 Performance measures of DT $\Sigma\Delta$ ADC, adapted from [11], © 2010 IEEE.

Mode	GSM	BT	UMTS
Architecture	first stage	first stage	first and second stage, multi-rate
BW [MHz]	0.1	0.5	1.92
f_s [MHz]	50	90	80/320
DR [dB]	85	77	66
$SNDR_{max}$ [dB]	77	76	65.5
$SFDR$ [dB]	83	83	79
$IM3$ [dB]	76	76	76
Dif. input range [V_{pp}]	0.8	0.8	0.8
P_{DC} [mW]	3.43	3.7	6.83

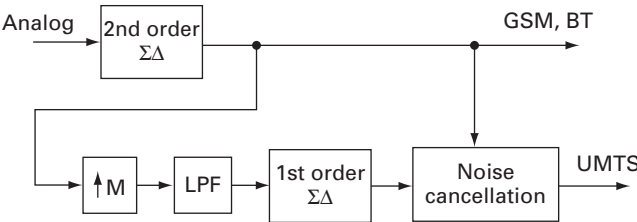


Figure 13.12 Architecture of multi-standard multi-rate DT $\Sigma\Delta$ ADC, adapted from [11], © 2010 IEEE.

measures are summarized. Observe that the maximum DR and SNDR is achieved in the Bluetooth mode (BT) for which also the largest oversampling ratio is used.

A similar performance has been achieved for a discrete-time $\Sigma\Delta$ ADC reported in [11]. To avoid stability problems typical of higher-order $\Sigma\Delta$ modulators a MESH architecture has been used [23], composed of two cascaded lowpass modulators, second- and first-order. The advantage of this structure is in cancellation of the quantization noise of the first stage based on extra digital filtering. The cancellation requires a perfect match between the transfer functions of the modulator and the digital filter. This is feasible for DT modulators due to robustness of the SC circuits used in this case (the transfer functions rely on capacitor ratios). Additionally, the presented DT converter adopts multi-rate operation so that the first modulator operates at reduced sampling rate compared to the following stage. This allows the power consumption of the first stage and finally of the whole circuit to be saved. The resulting loss in resolution is recovered by increasing the clock speed in the following stage only with a little power overhead. To support the multi-rate operation, an interpolator (up-sampler) is placed between the stages as also illustrated in Figure 13.12.

If the first modulator in the cascade operates at a sampling rate f_s and the other at $4f_s$ then ideally, the converter resolution is the same as the resolution (ENOB) of a cascade with both modulators operating at $2f_s$, while power savings of up to 35% can be achieved. The circuit is implemented in 1.2 V digital 90 nm CMOS technology with a core area of only 0.076 mm². The performance of the multi-rate DT $\Sigma\Delta$ converter is summarized in Table 13.2. Observe that in GSM or BT operation mode the first stage is sufficient

(second-order modulator) while the other ADC blocks are off. In UMTS mode both modulators are necessary and the multi-rate operation is applied. Interestingly, in BT and UMTS mode the maximum SNDR is close to the corresponding DR that means the modulators can operate only slightly backed off from the full-scale input before overloading occurs.

The achieved performance is comparable to the CT $\Sigma\Delta$ ADC (Table 13.1) for UMTS and is better for BT operation. However, the wideband standards (DVB, WLAN) are not addressed by this design, certainly due to settling time limitation of the SC integrators and the resulting inter-symbol interference problem (ISI).

13.4.2 Pipeline ADC

A pipeline A/D converter suited for embedding in a wideband cellular base-station receiver has been reported in [14]. As this converter should process many channels in parallel the corresponding requirements for linearity and DR are extremely high. For a given standard the DR can be estimated from (13.1) assuming $R_B=0$ and the P_B value adequate to the blocking profile, usually for a limited number of channels to make the DR feasible. In this case, the converter SFDR is of utmost importance to maintain SNR in the weakest channels, and this requirement is also dictated by the blocking profile.

The reported converter consists of eleven stages with the largest resolution of 4 bits in the first stage and 2.5-bit resolution in the following two stages. This choice is based on the fact that in a pipeline ADC each stage amplifies a residue signal, so the requirements in terms of noise and linearity decrease from stage to stage according to the preceding amplification. Consequently, it is advantageous to implement a large gain, and hence a large resolution in the first stages. Range-scaling is implemented in the first stage to maximize the input signal swing and a charge-reset is used to eliminate the ISI-induced distortion. Power savings in the opamps prevent a high open-loop gain deteriorating thereby linearity of the pipeline stages. However, these nonlinearities are effectively cancelled by the digital background calibration technique to ensure the ADC performance, 14-bit resolution with 73 dB SNR and 90 dB SFDR at 100 MS/s sampling rate. The SNDR performance is maintained within 3 dB up to the Nyquist frequency. The ADC is implemented in 90 nm CMOS technology with 1.2 V supply voltage. The performance summary is given in Table 13.3.

Table 13.3 Performance measures of pipeline ADC, adapted from [14], © 2009 IEEE.

Resolution [bit]	14
Sampling rate[MS/s]	100
DNL/INL [LSB]	0.9/1.3
SNR [dB]	73 ($f_{in} = 21$ MHz), 70.7 ($f_{in} = 64.5$ MHz)
$SNDR$ [dB]	73 ($f_{in} = 21$ MHz), 69.3 ($f_{in} = 64.5$ MHz)
$SFDR$ [dB]	90 ($f_{in} = 21$ MHz), 78.7 ($f_{in} = 64.5$ MHz)
Dif. input range [V_{pp}]	1.6
P_{DC} [mW]	250
	200 (excluding on-chip ref. buffers)

13.5 D/A conversion in RF transmitter

On the transmitter side the signal characteristics are less problematic than on the receiver side. A signal to be transmitted is usually appropriately conditioned in the digital domain and after D/A conversion it achieves enough power to maintain SNR in terms of additive noise from the following stages. Still, some effects in the analog front-end like LO phase noise, I/Q mismatch, and power amplifier nonlinearity can largely deteriorate the signal performance and evoke spurious emission. The corresponding transmitter performance measures, the error vector magnitude (EVM) and the adjacent channel power ratio (ACPR) are also subject to the DAC performance.

In Figure 13.13 the standard architecture of I/Q up-converter typical of a communications transmitter is shown. It can be used as IF or RF stage, the latter equivalent to the one-step transmitter architecture (homodyne). The digital baseband signal is transformed to the analog domain and reconstructed by a low-pass filter. The combined I/Q output is available after up-conversion in the quadrature mixer. Clearly, the DAC intermodulation distortion and noise enlarged by DNL errors are passed by I and Q branches to the output and contribute to the ACPR and EVM, accordingly.

13.5.1 EVM and ENOB

The EVM that is usually illustrated by spread of constellation points on the I/Q plane, can also be defined by the (S/N) ratio at the transmitter output. Using the linear scale we have

$$EVM^2 = \frac{1}{(S/N)_{Tx}}. \quad (13.24)$$

The main factors contributing to $(S/N)_{Tx}$ are: the LO phase noise, LO leakage (in homodyne transmitter), I/Q mismatch, and DAC noise. The contribution due to the analog front-end noise figure is usually less meaningful and can be neglected in this case [24]. Assuming the noise sources are uncorrelated, (13.24) can be rewritten as

$$EVM^2 = \sum_k EVM_k^2 \quad (13.25)$$

where for the k th noise component $EVM_k^2 = (S/N)_k^{-1}$. In a design perspective (13.25) defines an upper bound on the EVM components when the transmitter target EVM value is specified. Having estimated the EVM that falls to DAC the corresponding ENOB can be calculated like for an ADC in (13.2)

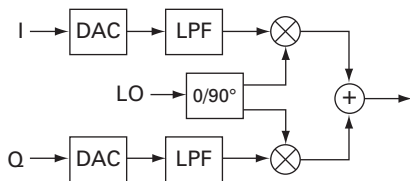


Figure 13.13 Standard architecture of I/Q up-converter (modulator).

$$ENOB = \frac{SNR_{DAC} - 1.76}{6.02}. \quad (13.26)$$

Importantly, SNR_{DAC} should be taken with a reserve for signal PAR variation and a necessary design margin, also to cover DNL errors and clock jitter. Then denoting the reserve by Δ_{DAC} and also including the oversampling effect we find

$$SNR_{DAC} = -20 \log EVM_{DAC} + \Delta_{DAC} - 10 \log OSR. \quad (13.27)$$

In this case the OSR not only enables reducing the number of bits but also can relax the requirements for the reconstruction filter.

For illustration purposes, let us consider a homodyne transmitter design with the target EVM of 8% ($SNR_{Tx} = 22$ dB) and PAR of 3.85 dB such as in a CDMA system. In this case DAC operates at baseband with signal bandwidth of 1.25 MHz while the available sampling clock has a jitter of 16 ps (rms). Assume the LO phase noise integrated over the channel bandwidth according to [24, Chapter 5.3]

$$EVM_{PN}^2 = \int_{-B/2}^{B/2} 10^{PN(\Delta f)/10} d(\Delta f)$$

results in $EVM_{PN} = 2\%$ (PN expressed in dBc/Hz). For the maximum I/Q mismatch of 0.5 dB and 2 deg the image suppression would be $IMG_S = -25$ dB [24 p. 178]. Then using

$$EVM_{IMG_S}^2 = 10^{IMG_S/10}$$

we obtain $EVM_{IMG_S} = 5.6\%$. Further assume the LO leakage (due to limited isolation on chip) in low-gain transmitter mode, results in $EVM_{LO,leak} = 4\%$ [25].

Then from (13.25) with $EVM = 8\%$ as the upper bound the remaining EVM headroom would be

$$\sqrt{EVM^2 - EVM_{PN}^2 - EVM_{IMG_S}^2 - EVM_{LO,leak}^2} = 3.5\%.$$

Putting safely $EVM_{DAC} = 1\%$ and using a reserve $\Delta_{DAC} = 12$ dB the $ENOB$ can be estimated from (13.26) as

$$ENOB = 8.35 - 1.66 \log OSR.$$

Here, the choice of Δ_{DAC} is mainly justified by the signal PAR and the effect of DNL elevating the noise level in DAC. Assuming the average DNL of 1 LSB from (13.10) we find the SNR decreases by 4.24 dB.

Compared to this the jitter induced noise is less meaningful since DAC operates at baseband with a signal of $f_{mx} = 1.25$ MHz. The corresponding SNR according to [13 Chapter 4.1] is

$$SNR_j = -20 \log(2\pi f_{mx} t_j) = 78 \text{ dB}$$

whereas SNR due to quantization noise for DAC using up to 11 bits is less than 68 dB.

The joint contribution of PAR and DNL amounts to 8.1 dB, and with 3.9 dB design margin the reserve $\Delta_{DAC} = 12$ dB is met. In fact, this margin is equivalent to 0.65 bit.

Clearly, by oversampling the required number of bits can be largely reduced. In this case, the converter with $OSR = 4$ achieves $ENOB = 7.35$ bits and on every doubling of OSR, an extra 0.5 bit can be saved.

13.5.2 Linearity of DAC

While the converter DNL mainly results in the elevated noise floor and claims thereby more bits to maintain the required EVM, the linearity of the DAC in a communications transmitter is considered in terms of the intermodulation and harmonic distortions. It is the third-order intermodulation effect (IM3) that is mainly responsible for the so-called spectral regrowth in a transmitter, i.e. a spectral leakage to the adjacent channels, quantified by the ACPR specification (adjacent channel power rejection). This effect is difficult to cancel by analog filters for limited selectivity even at baseband while at RF the cancellation is infeasible. Observe that at baseband also the second-order intermodulation and harmonic distortions evoke the spectral regrowth. For example, the third-order harmonics originating in tones between frequencies $B/3$ and $2B/3$ fall perfectly in the adjacent channel. This leakage is amplified by the following stages and adds to the front-end IM3 products deteriorating the transmitter ACPR performance. Additionally, for lower OSR values some spurious DAC components can fold back into the baseband before they are attenuated by the following filter. Thus, the overall linearity of a DAC operating at baseband is a vital design issue, which should be considered individually with respect to the modulated signal spectrum, transmission mask and ACPR specification.

For a transmitter with output signal power P_{Tx} and a leakage to the adjacent channel $P_{Leak,Adj}$ the ACPR can be estimated from [24 Chapter 5.4]

$$ACPR_{Adj} \cong \left(P_{Leak,Adj} + 10 \log \frac{\Delta B_{ACP}}{B} \right) - P_{Tx} + C_0. \quad (13.28)$$

As seen the leakage power is only measured over a fragment ΔB_{ACP} of the channel bandwidth B and $C_0 \approx 0.85(PAR - 3)$ is an empirical correction factor. For simplicity the leakage roll-off and the position of ΔB_{ACP} within the bandwidth are omitted. The adjacent leakage power is attributed to nonlinearity both of the DAC and the transmitter front-end, where the power amplifier (PA) is the most critical block, especially for envelope-varying modulation formats such as QAM.

Here, we also assume that the signal is shaped by a digital filter (such as raised-root cosine) to fit the channel bandwidth appropriately.

For a DAC operating at baseband various nonlinear effects can raise the leakage power and deteriorate ACPR as discussed above. In contrast, for a DAC operating at IF or RF the second-order intermodulation products and harmonics are away from the transmitted channel and can be suppressed by the following analog filter while possible aliasing effects are usually not meaningful which we discuss in the [next section](#). In this case, the leakage to the adjacent channel is mainly attributed to IM3 distortion in the transmitter

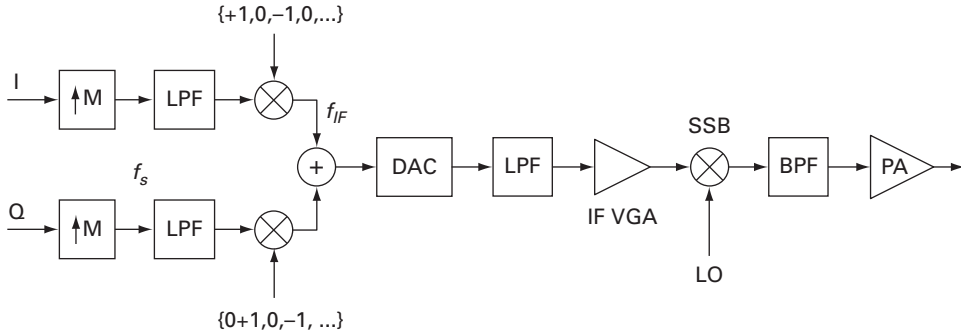


Figure 13.14 Typical transmitter architecture using digital up-conversion.

front-end and the DAC. A typical architecture using an IF DAC is shown in Figure 13.14. After D/A conversion and low-pass filtering the analog IF signal is up-converted to RF by a single-sideband mixer and applied to a power amplifier (PA).

Assuming a simple two-tone model for the transmitter output signal the adjacent channel leakage can be estimated as IM3 power using the basic formula

$$P_{Leak,Adj} \cong 3(P_{Tx} - 3) - 2OIP_3 \quad (13.29)$$

where the output third intercept point of the transmitter can be expressed as

$$OIP_3 \cong -10 \log \left(10^{\frac{-OIP_{3,DAC} - G_{FE}}{10}} + 10^{\frac{-OIP_{3,FE}}{10}} \right). \quad (13.30)$$

Clearly, $OIP_{3,DAC}$ and $OIP_{3,FE}$ are the output IP3 points of the DAC and front-end, respectively, and G_{FE} stands for the transmitter front-end gain. Based on this estimate (13.29) can be rewritten as

$$P_{Leak,Adj} \cong 3(P_{DAC} - 3) + 20 \log \left(10^{\frac{-OIP_{3,DAC}}{10}} + 10^{\frac{-OIP_{3,FE}}{10}} \right) + G_{FE} \quad (13.31)$$

where P_{DAC} is the signal power at the DAC output. From (13.31) we infer that the IP3 performance of the DAC and the front-end are equally important. In practice, however, DACs prove to be more linear than transmitter front-ends and especially more than PAs. In effect, the transmitter IP3 and ACPR performance are mainly decided by the PA while DAC linearity is kept sufficiently high.

The DAC linearity is often measured as IM3 applying a two-tone test, but in the context of the previous discussion the following useful relation holds for a modulated signal too

$$OIP_{3,DAC} \cong (P_{DAC} - 3) + \frac{|IM_{3,DAC}|}{2}. \quad (13.32)$$

For further illustrations, let us consider a transmitter with 1 W output power (30 dBm), $PAPR = 3.85$ dB and $ACPR_{Adj} = -48$ dB measured for $(\Delta B_{ACP}/B) = 1/20$. Assume the DAC output power is 1 mW (0 dBm), then from (13.28) we find the allowed leakage $P_{Leak,Adj} = -5.7$ dBm and from (13.29) the transmitter minimum $OIP_3 = 43.4$ dBm.

Consequently, we have to assume $OIP_{3,FE} > OIP_3$. Choosing the minimum $OIP_{3,FE} = 45$ dBm, from (13.30) we find $OIP_{3,DAC} = 18.5$ dBm. For the specified DAC power this is equivalent to $|IM_{3,DAC}| = 43$ dB that is rather a mild condition since IM3 performance of 60–70 dB is not unusual. Compared to this the corresponding $|IM_{3,FE}| = 36$ dB. Finally, note that the transmitter compression should also support the maximum power level, i.e. the 1 dB compression point $P_{1dB} > P_{Tx} + PAR = 33.85$ dB should be achieved with some reserve. Usually, P_{1dB} is a few dB up to ten dB less than OIP_3 . As a consequence the minimum OIP_3 larger than calculated from (13.29) might be required putting thereby more stringent constraints on the transmitter blocks.

13.5.3 Digital images in DAC

Using oversampling helps to lower the DAC noise floor and limit the number of bits as discussed previously. Due to oversampling also the digital image spectra can be moved to higher frequencies and relax requirements for the analog reconstruction filter. For this purpose the interpolation technique is applied, i.e. the Nyquist-rate digital signal is re-sampled at higher frequency and zeros are inserted between the original samples. Next, a digital low-pass filter (interpolation filter) is used to remove the unwanted spectral images as shown in Figure 13.15. Clearly the oversampling puts far greater demands on the DAC, but it allows the baseband signal processing at the minimum possible sampling rate.

Observe that the image next to the signal corresponds to the transmitter adjacent channel, so unless filtered out appropriately it is like a leakage evoked by nonlinearities badly affecting ACPR. Thus, to be efficient the interpolation filtering may occur in two steps using a combination of up-sampling and FIR or comb filtering in the [first section](#), and comb filtering in the following section. The remaining images centered at integer multiples of the sampling frequency are suppressed by the analog reconstruction filter.

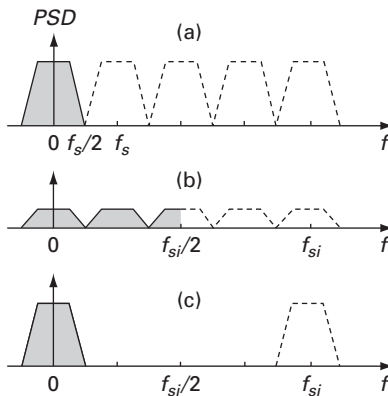


Figure 13.15 Signal spectra in interpolation process: (a) Nyquist-rate sampled signal, (b) re-sampled signal at $f_{si} = 4f_s$, (c) digitally low-pass filtered signal after amplification.

In fact, the analog filtering is also supported by $\text{sinc}(f/f_s)$ shaping, inherent to the zero-order hold (ZOH) waveform of DAC. To improve attenuation achieved in this way the first- or second-order hold waveforms can be implemented as well, relaxing further the reconstruction filter requirements. Digital or mixed-signal techniques can be used for this purpose [25,26,27].

13.6 Digital up-conversion

Like DDC in receivers digital up-conversion (DUC) has also been widely adopted in communications transmitters. As more functions are implemented in the digital domain also more immunity to process variations and possible mismatch is attained by a transceiver. Like in a receiver the programmable functions provide flexibility facilitating multimode or multi-channel operation of a communications transmitter. Specifically, in a multi-channel transmitter suitable DUCs allocate the transmission channels in the selected frequency slots. The channels can be converted into the analog domain by a common DAC. Compared to a multi-channel receiver where ADC linearity is dictated by the blocking profile, here the average power in each channel would be the same, so less stringent demands on DAC linearity would be imposed. On the other hand, as it is a multi-carrier signal the PAPR of the signal tends to increase. In effect, greater demands must be placed on the DAC dynamic range (resolution) in a multi-channel transmitter.

Typically, the DUC is carried out in I/Q mode and is preceded by interpolation as shown in Figure 13.14. To avoid costly multiplications with discrete sine and cosine values it is common to set the symbol rate four times f_{IF} , and use simple multiplexers rather than multi-bit multipliers, like in the DDC as discussed in Section 13.3. Interestingly, this task can be further simplified by acknowledging that the I/Q discrete carriers $\sin(k\pi/2)$ and $\cos(k\pi/2)$ can be shifted in phase by $\pi/4$. In effect, double-valued rather than triple-valued carrier sequences are achieved, which in practice take the form of $\{+1, +1, -1, -1, \dots\}$ and $\{-1, +1, +1, -1, \dots\}$.

As mentioned previously, choosing the sampling rate $4f_{IF}$ creates an image due to the third-order harmonic distortion (HD3) in a receiver ADC. Also for a transmitter DAC a similar image occurs that folds back in the transmit channel. Observe that the HD3 power is typically 50–60 dB below the signal and additionally it is shaped by the spectral $\text{sinc}(f/f_s)$ envelope. For a ZOH DAC this extra attenuation amounts to $20\log \text{sinc}(3/4) = -10.45$ dB and for a first-order hold DAC it is -20.9 dB. In effect, the DAC HD3 distortion is usually not a serious drawback with this frequency plan.

13.7 Implementation of DACs in communications transmitters

Due to the demands for high resolution and bandwidth current-steering DACs are commonly employed in communications transmitters [28,29,13]. Compared to other popular DAC architectures, they avoid amplifiers which limit the speed of analog systems. In a current-steering DAC, the component current sources are properly switched

so that the combined current is proportional to the digital input code as shown in Figure 13.16. The converter is usually arranged as a segmented structure where less significant bits switch binary weighted current sources, while the most significant bits (MSB) are thermometer-coded and switch unary current sources. Thermometer-coded DACs are appreciated for their advantages with respect to DNL, glitch energy, monotonicity, and linearity, while the binary-coded DACs have advantages with respect to compactness and simplicity. Additionally, the binary-coded DACs prevail in speed since they do not use extra decoding.

However, in many applications only thermometer coding is used with equally sized current sources within a segment. With this approach less mismatch and better performance can be achieved. By segmentation the number of current sources is greatly reduced and the chip area is used more efficiently. Observe for example that for a 12-bit DAC with plain thermometer encoding 4096 unary current cells are needed. Compared to this, grouping of bits, for example 7 bits in one segment and 5 bits in the other, only results in 128 and 32 current cells, respectively [30].

On the other hand, the current-steering DAC performance is limited by inter-symbol interference (ISI), imperfect timing synchronization, and clock jitter. Also the device mismatch and finite output impedance of current sources deteriorate DAC linearity. Specifically the transfer function from the digital input word to the output voltage exhibits third-order nonlinearity [31]. For DAC with N unary current cells, each with output impedance Z_{Out} , the resulting SFDR can be estimated from

$$SFDR = 20 \log \frac{4Z_{Out}}{NZ_L} \quad (13.33)$$

where Z_L is the common load impedance. In many DAC applications (13.33) appears as a fundamental performance constraint. Observe that Z_{Out} goes down with frequency due to device parasitic capacitances.

If a $\Sigma\Delta$ modulator is used in front of the DAC the number of current cells can be reduced even more without loss in ENOB. A rough comparison in terms of the quantization noise performance is as follows. For the n -bit Nyquist DAC with oversampling provided by interpolation the S/N ratio can be calculated with

$$(S/N)_{Nyq} = \frac{3}{2} 2^{2n} OSR. \quad (13.34)$$

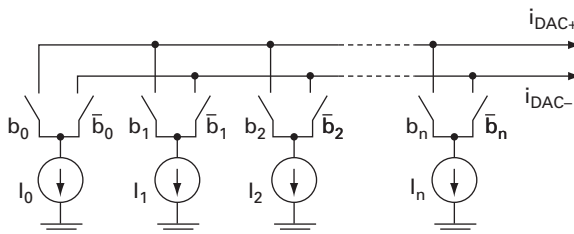


Figure 13.16 Basic architecture of current steering DAC.

Next, consider a digital $\Sigma\Delta$ modulator reducing the number of bits from n to m , followed by a DAC. In this case we have [23]

$$(S/N)_{\Sigma\Delta} = \frac{3}{2} 2^{2m} \frac{2l+1}{\pi^{2l}} OSR^{2l+1} \quad (13.35)$$

where l stands for the modulator order. The relevant condition is $(S/N)_{\Sigma\Delta} \geq (S/N)_{Nyq}$ and hence the required number of the quantization bits is

$$m \geq n - \frac{10 \log(2l+1) - 10l + 20/\log OSR}{6.02}. \quad (13.36)$$

Using for example a third-order modulator, $l=3$ with $OSR=16$, for $n=12$ we find $m \geq 3.6$, thus the minimum number of thermometer-encoded current cells should be $2^{3.6} > 12$. This number can be compared to 4096 or $128+32$ cells for 12-bit DAC discussed previously.

Clearly, with less analog complexity the design becomes simpler and more accurate. The in-band quantization noise is low owing to inherent $\Sigma\Delta$ shaping while for a large enough OSR the out-of-band noise can be suppressed by the reconstruction filter. These advantages come at the expense of extra digital circuit complexity that reflects a trade-off typical of contemporary CMOS circuits.

The $\Sigma\Delta$ modulation technique can be easily applied to baseband or IF operated DACs, whereas for RF DACs – a technique that emerged recently, more sophisticated approaches must be used to cope with the $\Sigma\Delta$ out-of-band noise and spurious emission [32,33,34,26,27].

In fact, the RF DAC is a concept where a D/A-converter and RF-mixer are combined in a single building block as shown in Figure 13.17 [32]. Basically, this structure is only composed of transistor switches and matched current sources avoiding thereby nonlinearities typical of continuous analog circuits, like the Gilbert mixer. Such an RF DAC can be employed both in a heterodyne and homodyne architecture. In either case the transmitter analog front-end is largely reduced whereas the digital baseband and/or IF part can be reconfigured to enable multi-standard, multimode operation.

The RF D/A conversion is also employed in polar transmitters using PLL architecture as shown in Figure 13.18. In this approach the digital baseband data occurs as two streams representing symbol phases and magnitudes rather than Cartesian I/Q coordinates. The phase modulation is realized by fractional- N frequency synthesis in the PLL and for this purpose the phase signal must be converted to frequency using differentiation. The amplitude modulation, synchronized with the phase modulation, is carried out using a DAC which through a power-efficient DC-DC converter provides supply for the PA, also referred to as “envelope restoration.” A special case is constant envelope systems such as GMSK where the amplitude modulation is omitted. Since the amplitude of the phase-modulated signal remains constant, it can be amplified using a very efficient PA, which largely saves power consumption of the transmitter. This is also considered the primary advantage of the polar transmitter architecture.

The phase signal is low-pass filtered by the PLL, and to enable wideband modulation the PLL loop bandwidth should be increased. This, however, boosts the phase noise and

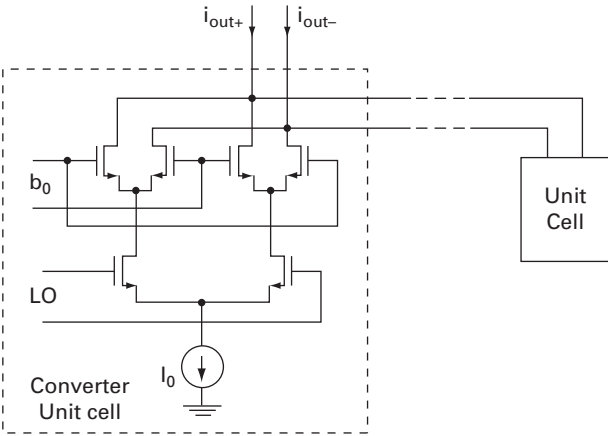


Figure 13.17 Architecture of typical RF DAC.

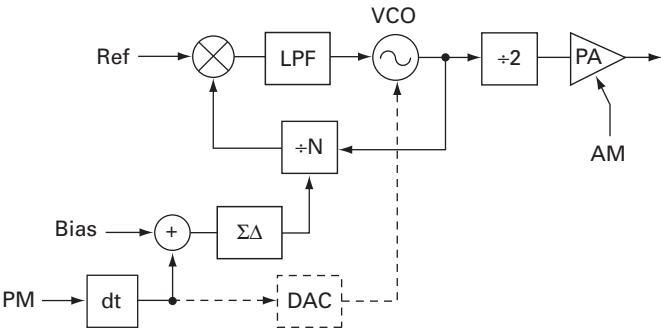


Figure 13.18 Polar-loop transmitter architecture (dotted line part – two-point PLL modulation), adapted from [6], © 2007 IEEE.

thereby degrades the EVM performance. One way to overcome this drawback is to use signal pre-emphasis while keeping the PLL bandwidth relatively low [35]. In another approach, two-point PLL modulation is used, ideally resulting in all-pass transfer function [5]. Obviously, the inevitable component mismatch is a limitation in this case, albeit possible to alleviate by “all-digital” PLL as proposed in [36]. Another shortcoming of the PLL phase modulation is generation of fractional spurs that usually can be minimized by dithering technique and using a higher-order $\Sigma\Delta$ modulator [6].

The D/A conversion in a polar transmitter is subject to similar constraints as in the traditional I/Q topology. A special case is the phase signal conversion where the $\Sigma\Delta$ modulator controls the fractional frequency divider. Observe that the out-of-band $\Sigma\Delta$ noise is low-pass filtered by the PLL, additionally reinforced by ZOH behavior of the involved circuits. Otherwise, as long as the respective DACs are not tightly coupled with the RF circuitry, they function as baseband DACs being separated from RF blocks (PA or VCO) by necessary reconstruction filters. These DACs are also challenged by linearity,

digital image, noise, and spurious emission requirements. Hence, oversampling, noise shaping, and first-order-hold techniques are employed to meet the specifications.

In the following sub-sections we present state-of-the-art DAC implementations based on standard CMOS technology. Although other technologies such as SiGe BiCMOS offer higher speeds, the submicron CMOS is becoming increasingly competitive also in most challenging applications [37].

13.7.1 Baseband and IF DACs

A 2-channel I/Q DAC suited for baseband multimode applications is presented in [30]. Implemented in 90 nm CMOS the converter makes use of two 12-bit current-steering blocks, each with thermometer coded 7- and 5-bit segment. The DAC current cells employ degenerated current switches to minimize energy of glitches improving thereby linearity. With 0.2% of three-sigma current cell mismatch the DAC meets SFDR specification of 74 dB up to the Nyquist frequency at 160 MS/s for signal BW < 1 MHz. Under the same conditions SNR > 56 dB and SNDR > 55 dB are achieved. Matching techniques are carefully applied to attain this performance. The 2-channel DAC occupies 0.13 mm² area and only consumes a power of 10.6 mW.

In the following two baseband/IF DACs suited for UWB or multimode applications are presented.

A 12-bit 2.9 GS/s current-steering DAC implemented in 65 nm CMOS is presented in [38]. The converter is composed of two 6-bit segments, where the MSB segment is thermometer coded and the LSB segment binary coded. For high-frequency operation the design departs from calibration and dynamic element matching (DEM) techniques which prove useful rather at lower frequencies. Instead, proper layout and proper transistor sizing is used to benefit from circuit simplicity and ensure low parasitic capacitances. For this purpose, the basic current cell is significantly modified and also fast latch drivers with a high crossing point are used. To maintain high output impedance the equivalent switching capacitance is kept at minimum. Additionally, cascode switches with small current sources are placed on top of the current switches. These extra switches are always on, increasing thereby the DAC output impedance. In effect, the converter performance at high frequency is largely improved compared to standard current-cell solutions. In measurements the SFDR of 60 dB and IM3 of -70 dB are achieved in bandwidth of 550 MHz. In 1 GHz bandwidth IM3 is -60 dB. The chip power consumption is 188 mW.

Another high-speed current-steering DAC implemented in 130 nm CMOS is reported in [39]. This 6-bit converter is aimed at ultra-wideband systems where high speed and low resolution are required. It is based on pseudo-segmented architecture with unary current cells which are connected in a binary way. By using this approach the advantages of the binary and unary topology are combined. Namely, the nonlinear switching effects typical of a DAC with binary weighted cells are mitigated, while a speed higher than for a thermometer-coded converter is achieved. The switch drivers are merged with the current cells and are synchronized to minimize timing differences and thereby improve DAC linearity. Also careful device sizing and layout techniques are employed to reduce parasitic effects. In measurement an SFDR of more than 36 dB is obtained over the

complete Nyquist interval for sampling frequencies up to 3 GHz. The SFDR is increased to 47 dB for $\text{OSR} = 50$. In the two-tone test, IM3 better than -39 dB is achieved. The DAC area is 0.2 mm^2 while the power consumption is only 29 mW, making it suitable for mobile terminals.

13.7.2 RF DACs

A two-step transmitter with RF DAC reported in [33] is shown in Figure 13.19. The I/Q baseband signal is up-converted using an interpolation filter and next applied to a second-order 3-bit $\Sigma\Delta$ modulator. The transmitter is implemented in a variant of Weaver architecture [40 Chapter 5] that is useful to achieve single-sideband RF output and attenuate input noise. The DUC in the IF stage is realized with four I/Q multipliers, which also provide 90 deg phase shift between I and Q branches. The following stage makes use of a quadrature RF DAC with an embedded bandpass LC reconstruction filter to remove spurs and the out-of-band quantization noise. The current-steering RF DAC is composed of seven unit cells corresponding to the 3-bit IF signal.

The IF frequency is chosen as $f_s/4$ to simplify implementation of the DUC. For RF of 5.25 GHz with 2.625 GHz sampling clock the LO frequency of approx. 4.6 GHz is used. The high-Q LC filter can be tuned from 4.8 GHz to 5.6 GHz. The circuit designed in 130 nm CMOS demonstrates a data rate of 1.2 GS/s using 64-QAM OFDM modulation in a bandwidth of 200 MHz. For output power of -8 dBm the measured SNDR is 49 dB while SFDR is better than 59 dBc without calibration. The RF DAC core power consumption is 10 mW while the complete chip including LO and clock buffers consumes 187 mW. The chip area is 0.72 mm^2 .

Another state-of-the-art transmitter implementation using RF DAC is shown in Figure 13.20 [27]. Here, a baseband modulated I/Q signal is digitally interpolated to

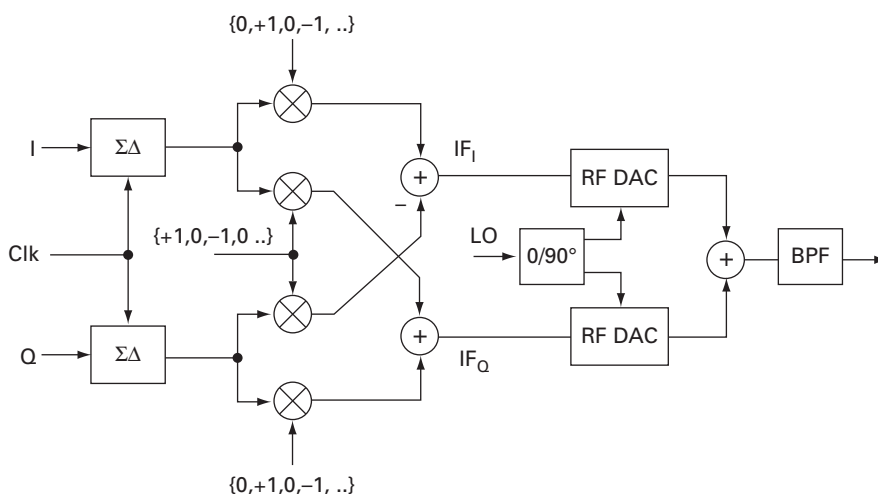


Figure 13.19 Quadrature digital-IF $\Sigma\Delta$ RF modulator, adapted from [33], © 2007 IEEE.

Table 13.4 Performance measures of transmitter with digital IF to RF DAC, adapted from [27], © 2008 IEEE.

RF [MHz]	1062.5/1032
Sampling rate [MHz]	250/125
Bandwidth [MHz]	15
SNR [dB]	72
$SFDR$ [dB]	75
$IM3$ [dBc]	-64.7
Idle channel noise [dBm/Hz]	-146
P_{DC} [mW]	122

The circuit is integrated using a 0.25 μm CMOS process with a 2.5 V supply and it occupies an area of 0.23 mm². The chip power consumption is 122 mW while the RF DAC core only consumes 12.8 mW. The performance is demonstrated for sampling rate $f_S = 250$ MHz, $f_{IF} = 62.5$ MHz, and $f_{RF} = 1062.5$ MHz. In Table 13.4 the main performance measures are summarized.

Both [27] and [33] only demonstrate transmitter RF DAC application for one dedicated band. The design presented in [27] seems to be more flexible in terms of multiband operation. This is due to the embedded FIR filter that could be reprogrammed by clock frequency. However, using more filter taps to improve out-of-band rejection is problematic, unless another smart technique is incorporated, since also nonlinear distortions in the RF DAC go up. The other approach by [33] seems to be more rigid in this regard. Still some tunability is demonstrated and by using additional LC tanks the transmitter could also serve as a multiband terminal, certainly at the expense of an extra chip area overhead.

Interestingly, in both designs the power consumption of the RF DAC core is only a small fraction of the total chip power. Hence, we infer there is headroom left for optimization where the current mismatch can be reduced by scaling up the DAC current cells and hence improvement in linearity can be envisioned. In this case, however, also a limitation due to increasing parasitic capacitances applies.

13.8 Summary

In this chapter state-of-the-art A/D and D/A converters suited for contemporary and future wireless communications systems have been reviewed. These converters come along with emerging 4G systems like WiMAX, LTE, or UWB, which aim at data rates in excess of 100 Mb/s. Unlike base-stations that can extensively exploit parallelism to support massive data rate transfers mobile terminals have to save power supplied usually by small batteries. The power issue is a fundamental constraint exacerbating the design requirements for mobile terminal circuits and specifically for the data converters being challenged by ever-increasing data rates and channel bandwidths specified for 4G systems. Also adoption of the existing standards by 4G systems is reinforcing the status of multi-standard and multimode transceivers. This in turn, calls for circuit flexibility and optimization such as reconfigurable $\Sigma\Delta$ converters, also saving power [41,42].

In this chapter we have focused on implementations in CMOS, which continues to be the leading edge IC technology. Before looking into the converter examples we discussed the design requirements imposed by receiver and transmitter specifications. Estimates for ADC and DAC performance measures like dynamic range, ENOB, or IP3 have been derived and illustrated by practical examples. We discussed converter applications in various receiver and transmitter architectures. In the context of IF operating ADC and DAC also the digital down- and up-conversion techniques were considered as a way to reduce I/Q mismatch and improve the overall performance and robustness of RF transceivers.

On the transmitter side we have emphasized RF DAC implementations as promising solutions for a future radio; D/A conversion for polar transmitter architectures was also discussed as an alternative approach. On the receiver side, however, placing an A/D converter close to the antenna happens to be an even more challenging task. In this case an extreme demand on the ADC dynamic range should be mitigated using dynamic spectrum allocation techniques [43] supported by programmable RF filters. This approach is believed to make the “true SDR” challenge a reality.

References

- [1] J. Mitola, The Software Radio Architecture. *IEEE Communications Magazine*, **33**:5 (1995), 26–38.
- [2] K. Muhammad, D. Leipold, B. Staszewski, *et al.*; A Discrete-time Bluetooth Receiver in a 0.13 μm Digital CMOS Process, *Int. Solid-state Circuit Conf. Dig. Tech. Papers*, (2004), 268–9, 527.
- [3] D. Jakonis, K. Folkesson, J. Dabrowski, *et al.*; A 2.4-GHz RF Sampling Receiver Front-End in 0.18- μm CMOS, *Journal of Solid-State Circuits*, **40**:6 (2005), 1265–77.
- [4] A. Abidi, The Path to the Software-Defined Radio Receiver, *Journal of Solid-State Circuits*, **42**:5 (2007), 954–66.
- [5] C. Durdodt, M. Friedrich, C. Grewing, *et al.*; A Low-IF Rx Two-Point DS-Modulation Tx CMOS Single-Chip Bluetooth Solution, *Transactions on Microwave Theory and Techniques*, **49**:9 (2001), 1531–37.
- [6] J. Groe, Polar Transmitters for Wireless Communications, *IEEE Communications Magazine*, **45**:9 (2007), 58–63.
- [7] R. Shrestha, E. Klumperink, Mensink E., *et al.*; A Polyphase Multipath Technique for Software-Defined Radio Transmitters, *Journal of Solid-State Circuits*, **41**:12 (2006), 2681–92.
- [8] L. Breems and J. H. Huijsing, *Continuous-Time Sigma-Delta Modulation for A/D Conversion in Radio Receivers*, (Dordrecht: Kluwer Academic Publisher, 2001).
- [9] P. Crombez, G. Van der Plas, M. Steyaert, *et al.*; A Single-Bit 500 kHz–10 MHz Multimode Power-Performance Scalable 83-to-67 dB DR CT $\Delta\Sigma$ for SDR in 90 nm Digital CMOS, *Journal of Solid-State Circuits*, **45**:6 (2010), 1159–71.
- [10] T. Christen, T. Burger, and Q. Huang, A 0.13 μm CMOS EDGE/UMTS/WLAN Tri-Mode $\Delta\Sigma$ ADC with -92 dB THD, *Int. Solid-State Circuits Conf., Dig. Tech. Papers*, 2007, 240–1, 599.
- [11] L. Bos, G. Vandersteen, P. Rombouts, *et al.*; Multirate Cascaded Discrete-Time Low-Pass $\Delta\Sigma$ Modulator for GSM/Bluetooth/UMTS, *Journal of Solid-State Circuits*, **45**:6 (2010), 1198–208.

- [12] M. Brandolini, P. Rossi, D. Manstretta, *et al.*; Toward Multistandard Mobile Terminals – Fully Integrated Receivers Requirements and Architectures, *Transactions on Microwave Theory and Techniques*, **53**: 3 (2005), 1026–38.
- [13] P. Kenington, *RF and Baseband Techniques for Software Defined Radio*, (Norwood, MA: Artech House, 2005).
- [14] H. Van de Vel, B. Buter, H. van der Ploeg, *et al.* A 1.2-V 250-mW 14-b 100-MS/s Digitally Calibrated Pipeline ADC in 90-nm CMOS, *Journal of Solid-State Circuits*, **44**:4 (2009), 1047–56.
- [15] B. Murmann and B. E. Boser, *Digitally Assisted Pipeline ADCs: Theory and Implementation*, (Hingham, MA: Kluwer Academic Publishers, 2004).
- [16] K-L. Lin, A. Kemna, and B. Hosticka, *Modular Low-Power, High-Speed CMOS Analog-to-Digital Converter of Embedded Systems*, (Dordrecht: Kluwer Academic Publishers, 2003).
- [17] M. Gustavsson, J. Wikner, and N. Tan, *CMOS Data Converters for Communications*, (New Jersey: Kluwer Academic Publishers, 2000).
- [18] P. Vancorenland, P. Coppejans, W. De Cock *et al.*; A Quadrature Direct Digital Downconverter, *Proc. IEEE Custom Integrated Circuits Conf.*, (2002), 235–8.
- [19] M. Lohning, T. Hentschel, and G. P. Fettweis. Digital Down Conversion in Software Radio Terminals, *Proc. 10 European Signal Processing Conference*, **3** (2000), 1517–20.
- [20] R. G. Vaughan, N. L. Scott, and D. R. White, The Theory of Bandpass Sampling, *Transactions on Signal Processing*, **39**:9 (1991), 1973–84.
- [21] S. Levantino, C. Samori, M. Banu, *et al.*; A CMOS GSM IF-Sampling Circuit with Reduced In-Channel Aliasing, *Journal of Solid-State Circuits*, **38**:6 (2003), 895–904.
- [22] K. Poulton, R. Neff, B. Setterberg, *et al.*; A 20 GS/s 8b ADC with 1 MB Memory in 0.18 μ m CMOS, *Int. Solid-State Circuits Conf.*, **1** (2003), 318–19, 496.
- [23] R. Schreier, G. Temes, *Delta-Sigma Data Converters*, (IEEE-Press, Wiley, 2005).
- [24] Q. Gu, *RF System Design of Transceivers for Wireless Communications*, (New York: Springer, 2005).
- [25] V. Leung, L. Larson, and P. Gudem, Improved Digital-IF Transmitter Architecture for Highly Integrated W-CDMA Mobile Terminals, *Transactions on Vehicular Technology*, **54**:1 (2005), 20–32.
- [26] Y. Zhou and J. Yuan, A 10-bit Wide-Band CMOS Direct Digital RF Amplitude Modulator, *Journal of Solid-State Circuits*, **38**:7 (2003), 1182–8.
- [27] S. Taleie, T. Copani, B. Bakaloglu, *et al.* A Linear $\Sigma\Delta$ Digital IF to RF DAC Transmitter with Embedded Mixer, *Transactions on Microwave Theory and Techniques*, **56**:5 (2008), 1059–68.
- [28] C. Shi, M. Ismail, *Data Converters for Wireless Standards*, (Secaucus, NJ: Kluwer Academic Publishers, 2001).
- [29] F. Maloberti, *Data Converters*, (Dordrecht, Springer, 2007).
- [30] D. Seo and G. McAllister, A Low-Spurious Low-Power 12-bit 160-MS/s DAC in 90-nm CMOS for Baseband Wireless Transmitter, *Journal of Solid-State Circuits*, **42**:3 (2007), 486–95.
- [31] A. Bosch, M. Steyaert, and W. Sansen, SFDR-Bandwidth Limitations for High Speed High Resolution Current Steering CMOS D/A Converters, *Proc. 6th IEEE Int. Conf. Electronics, Circuits and Systems*, **3** (1999), 1193–6.
- [32] S. Luschas, R. Schreier, and L. Hae-Seung, Radio Frequency Digital-to-Analog Converter, *Journal of Solid-State Circuits*, **39**:9 (2004), 1462–7.
- [33] A. Jerng and C. Sodini, A Wideband $\Delta\Sigma$ Digital-RF Modulator for High Data Rate Transmitters, *Journal of Solid-State Circuits*, **42**:8 (2007), 1710–22.

-
- [34] P. Eloranta, P. Seppinen, S. Kallioinen, *et al.*; A Multimode Transmitter in 0.13 μm CMOS Using Direct-Digital RF Modulator, *Journal of Solid-State Circuits*, **42**:12 (2007), 2774–84.
 - [35] S. Lee, S. Fang, D. Allstot, *et al.*; A Quad-Band GSM-GPRS Transmitter With Digital Auto-Calibration, *Journal of Solid-State Circuits*, **39**:12 (2004), 2200–14.
 - [36] R. Staszewski, J. Wallberg, S. Rezek, *et al.*; All-Digital PLL and Transmitter for Mobile Phones, *Journal of Solid-State Circuits*, **40**:12 (2005), 2469–82.
 - [37] B. Razavi, A 60-GHz CMOS Receiver Front-End, *Journal of Solid-State Circuit*, **41**:1 (2006), 17–23.
 - [38] C-H. Lin, F. van der Goes, J. Westra, *et al.*; A 12 bit 2.9 GS/s DAC With IM3 < 60 dBc Beyond 1 GHz in 65 nm CMOS, *Journal of Solid-State Circuits*, **44**:12 (2009), 3285–93.
 - [39] X. Wu, P. Palmers, and M. Steyaert, A 130 nm CMOS 6-bit Full Nyquist 3 GS/s DAC, *Journal of Solid-State Circuits*, **43**:11 (2008), 2396–2403.
 - [40] B. Razavi, *RF Microelectronics*, (New Jersey: Prentice Hall, 1998).
 - [41] Y. Wei, H. Tang, and A. Doboli, Systematic Methodology for Designing Reconfigurable Delta Sigma Modulator Topologies for Multimode Communication Systems, *Transactions on Computer Aided Design*, **26**:3 (2007), 480–96.
 - [42] K. Yi, J. Craninckx, and G. Gielen, A Design Approach for Power-Optimized Fully Reconfigurable $\Sigma\Delta$ A/D Converter for 4G Radios, *Transactions on Circuit Systems II*, **55**:3 (2008), 229–33.
 - [43] I. F. Akyildiz, W. Y. Lee, M. C. Vuran, *et al.*; A Survey on Spectrum Management in Cognitive Radio Networks, *IEEE Communications Magazine*, **46**:4 (2008), 40–8

14 Advanced quadrature sigma-delta modulator designs for A/D interface

Mikko Valkama, Jaakko Marttila, and Markus Allén

14.1 Sigma-delta basics

The principle of oversampling the analog-to-digital converter (ADC) with negative feedback has been invented decades ago and is still being developed further by scientists all over the world. Today's state-of-the-art converters have come a long way from the first ADCs employing the $\Sigma\Delta$ principle in the 1960s. There is quite a broad selection of $\Sigma\Delta$ oriented publications in the literature since the 1960s and early development phases have also been documented in a comprehensive manner, e.g., in [2], [4] and [52]. On the basis of [2], and following the outline of [35], this chapter aims to present the theory and technology of the advanced quadrature sigma-delta modulator designs for A/D interface. The chapter is organized into six sections. In the rest of the [first section](#), we outline the basics of sigma-delta modulation. [Section 14.2](#) is devoted to extending the discussion on some further modulator concepts and selected advanced quadrature structures will be presented in [Section 14.3](#). Related implementation nonidealities are discussed in [Section 14.4](#). [Section 14.5](#) gives some simulation examples on the advanced structures introduced in [Section 14.3](#), taking also circuit nonidealities into account. [Section 14.6](#) will present related conclusions.

The origin of modern $\Sigma\Delta$ modulation is in delta modulation and differential pulse-code modulation (PCM). Delta modulation was invented in ITT laboratories in France in 1946, as was also the classical version of the PCM. Differential PCM system was patented in 1950 by Bell Telephone Labs.

In delta modulation, the term delta refers to transmitting only the changes between consecutive samples. The delta modulator employs a 1-bit ADC and the feedback signal is fed through a 1-bit digital-to-analog converter (DAC), whereafter it is subtracted from the input after an integrator. In this kind of system output “1” means that input signal amplitude is raising and “0” indicates negative slope. Differential PCM uses the same kind of structure as delta modulation, the only difference being that the single-bit ADC and DAC have been replaced with their multi-bit flash counterparts.

Since the output of the modulator depends only on the direction of change in the input signal, the clipping effect can be avoided with a big enough step size and sampling rate. For that reason delta and differential PCM modulators usually demand sampling rates as high as

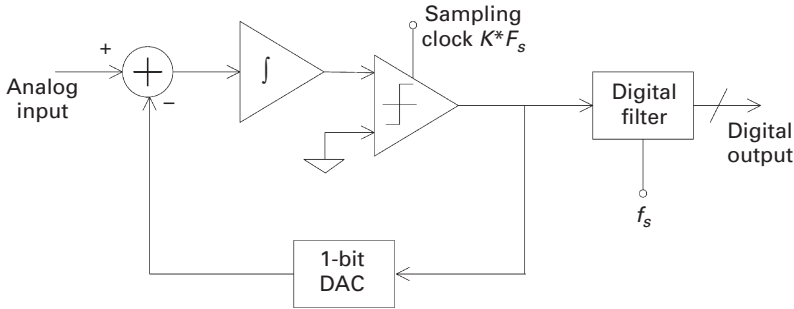


Figure 14.1 Block diagram of $\Sigma\Delta$ converter (first-order structure with 1-bit quantization). Used with permission from [2].

20 times the bandwidth of the signal of interest. In the modern $\Sigma\Delta$ modulator, the integrator block does not operate on the feedback loop, but is transferred to the forward branch and thus operates on the error signal. A block diagram of $\Sigma\Delta$ modulator is shown in Figure 14.1. The publications [19] and [61] can be mentioned as examples of the earlier work on the topic.

In this kind of structure the integrator shapes quantization noise away from the band of interest by processing the error of the ADC and the DAC compared to the input. In this part the oversampling becomes significant, because in the traditional Nyquist rate converter there is no extra frequency band where the noise could be pushed. With oversampling used in $\Sigma\Delta$ modulators there is always more band available than that of the desired signal. This extra band can be used for the noise spectrum shaping, and thus the noise can be filtered out in digital domain after the conversion. In this way, it is possible to cope even with simple 1-bit quantization, if there is enough oversampling. The exact sampling speed requirements are of course application specific. After the filtering, the output signal can be decimated to lower rate or even down to the Nyquist rate, depending on the forthcoming signal processing stages.

If oversampling were the only way to reduce quantization noise, the signal should be oversampled with a factor of 2^{2N} to obtain N -bit increase in resolution [2]. In the $\Sigma\Delta$ principle, the noise shaping properties fortunately ease the situation. The error signal, which is the difference of the original input and the output of the feedback DAC, is fed to the forward integrator. In the integrator output, the noise is pushed towards the high end of the Nyquist band, when considering the lowpass type of $\Sigma\Delta$ modulator. The noise spectra of a traditional Nyquist ADC, an oversampling traditional ADC and an oversampling $\Sigma\Delta$ ADC are compared in Figure 14.2.

Often the transfer function of the $\Sigma\Delta$ modulator is expressed in two parts. The noise shaping is described in the noise transfer function (NTF), usually having one or more zeros on the band of interest. The behavior of the desired signal itself depends on the signal transfer function (STF), usually having unity gain. With these definitions the z -domain output of the converter can be described as

$$V[z] = STF[z]U[z] + NTF[z]E[z], \quad (14.1)$$

where $U[z]$, $E[z]$, $V[z]$, $STF[z]$, and $NTF[z]$ refer to the input, the quantization error, the output, the signal transfer function, and the noise transfer function, respectively [4].

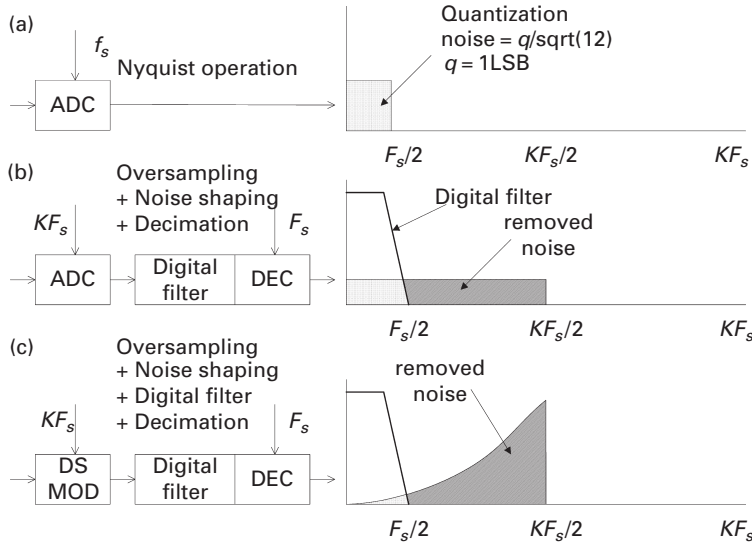


Figure 14.2 Noise spectra of (a) a Nyquist ADC, (b) an oversampling traditional ADC, and (c) an oversampling $\Sigma\Delta$ ADC. Used with permission from [2].

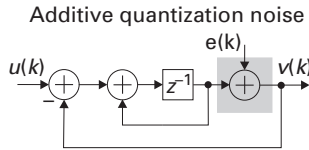


Figure 14.3 Discrete-time linearized model of the first-order lowpass $\Sigma\Delta$ modulator.

14.1.1 Lowpass modulator

A discrete-time block diagram for a first-order feedback lowpass $\Sigma\Delta$ modulator is presented in Figure 14.3 to clarify the main principle. The quantizer is replaced with an additive noise source and the DAC is assumed ideal and thus replaced by a unity gain. For analysis purposes, the quantization error inside the modulator loop is assumed to be additive white Gaussian noise (AWGN) and to be uncorrelated with the input [52]. The assumption is not exactly accurate, as Gray found in [14] that the error actually is correlated with the input, but the uncorrelatedness offers a valid enough base for the analysis to be adopted in use.

Based on the model, the transfer function can be derived giving [4]

$$V[z] = z^{-1}U[z] + (1 - z^{-1})E[z]. \quad (14.2)$$

It can be seen, from (14.2), that for the discussed structure $STF[z] = z^{-1}$ and $NTF[z] = 1 - z^{-1}$, are the multipliers of the signal and noise components ($U[z]$ and $E[z]$), respectively. Thus, the desired signal is only delayed on the way from input to output and some filtered noise is added to the signal. The NTF amplitude response is

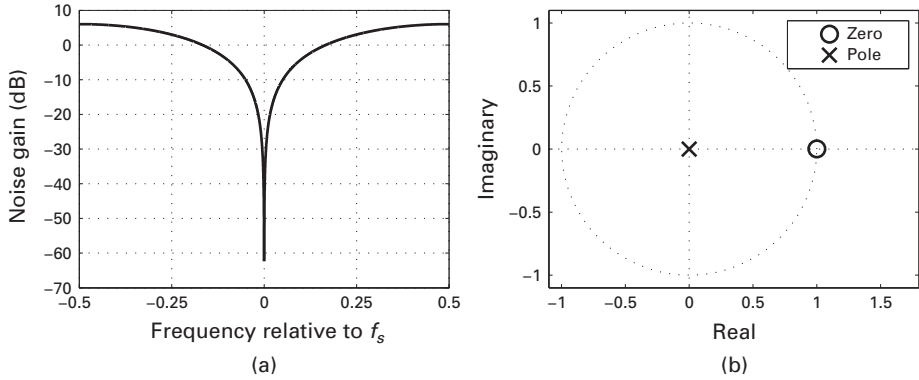


Figure 14.4 Amplitude response (a) and zero-pole plot (b) of the NTF for a first-order lowpass $\Sigma\Delta$ modulator.

plotted in Figure 14.4(a) and the related positions of the modulator zero and pole are given in the zero-pole plot in Figure 14.4(b).

From the amplitude response, it is clear that the noise signal has attenuation level of -60 dB on the zero-frequency. The zero-gain level, which would correspond to non-shaped noise, is roughly at frequencies ± 0.166 relative to f_s . Thus, the band of 0.33 relative to f_s has only attenuated noise present. The zero-pole plot confirms that the modulator zero lies at DC on the unit circle and the pole is in origin.

For a more detailed mathematical perspective, interested readers are advised to consult [15]. Therein, a precise walk through the effects of quantization in single loop, two-stage, and second-order $\Sigma\Delta$ converters is given. Reference [15] also expands the discussion to the effects of dithering, higher-order modulation, leaky integration, and multi-bit ADCs, in respective order.

In [4], the formulas for inband noise power and signal-to-noise ratio (SNR) in case of single-bit quantizer and single-bit DAC assuming high enough oversampling ratio (OSR) to shape noise effectively are given. Reproducing the findings, the output inband noise power of first-order lowpass $\Sigma\Delta$ modulator becomes

$$\sigma_{ey}^2 = \sigma_e^2 \frac{\pi^2}{3} \left(\frac{2f_B}{f_s} \right)^3, \quad (14.3)$$

where σ_e^2 corresponds to the whole band quantization noise power, f_B and f_s are signal bandwidth and the sampling frequency, respectively. This leads to the inband signal-to-noise ratio of

$$SNR = 10 \log_{10}(\sigma_x^2) - 10 \log_{10}(\sigma_e^2) - 10 \log_{10} \left(\frac{\pi^2}{3} \right) + 30 \log_{10} \left(\frac{f_s}{2f_B} \right) \quad (14.4)$$

in decibels, where σ_x^2 is the desired signal power and $f_s/2f_B$ is the oversampling ratio.

The effect of the NTF is confirmed by simulation with a single sine-wave input in Figure 14.5. The STF is assumed frequency-flat with unity response. In simulations, the quantization error is modeled as AWGN, as was the assumption when deriving the linear model for the modulator. Variance of the noise is equivalent to that of the theoretical

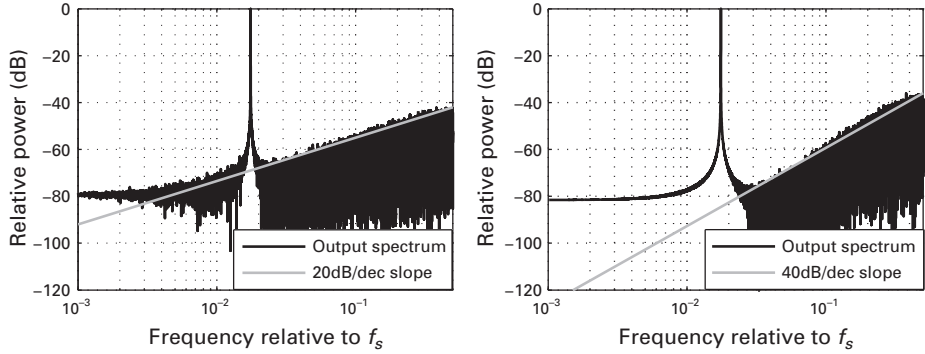


Figure 14.5 Simulated output spectrum for the first- (left) and the second-order (right) lowpass $\Sigma\Delta$ modulators with a sine-wave input and an AWGN modeled quantization error.

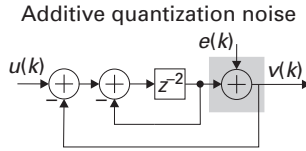


Figure 14.6 Discrete-time linearized model of a first-order BP $\Sigma\Delta$ modulator.

quantization error produced by a 1-bit quantization. Thus, it gives full-band signal-to-noise ratio of $6.02 \times 1 + 1.76 = 7.76$ dB [26]. The plots clearly show the shaped noise spectrum. Highlighted 20 dB and 40 dB per decade slopes are consistent with the theoretical NTFs for first and second order [52]. Noise floor is limited to -80 dB due to simulation precision and FFT properties.

14.1.2 Real bandpass modulator

In a lowpass type $\Sigma\Delta$ modulator the quantization noise is pushed away from the low frequencies towards the high end of the Nyquist band. Correspondingly, a bandpass (BP) converter shapes the noise towards the Nyquist frequency and the DC, and thus away from the band of interest in the middle of the Nyquist band [2]. This can be particularly useful in case of a low IF or an IF receiver [12].

The best-known method to design a bandpass $\Sigma\Delta$ modulator is to transform a lowpass prototype to a non-zero center frequency. This can be done, e.g., with $z^{-1} \rightarrow -z^{-2}$ transformation [47], [51]. In other words, an additional delay is added to the integrator inside the modulator loop and the sign of the feedback is inversed. The resulting bandpass structure is presented in Figure 14.6. With the described method the center frequency of the noise notch becomes $f_s/4$.

If the noise shaping characteristics are desired to be such that the notch frequency of the bandpass modulator differs from a quarter of the sampling frequency, a more sophisticated transformation method has to be used. A tunable bandpass modulator

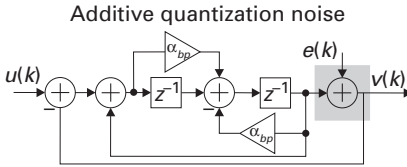


Figure 14.7 Discrete-time linearized model of a first-order tunable BP $\Sigma\Delta$ modulator.

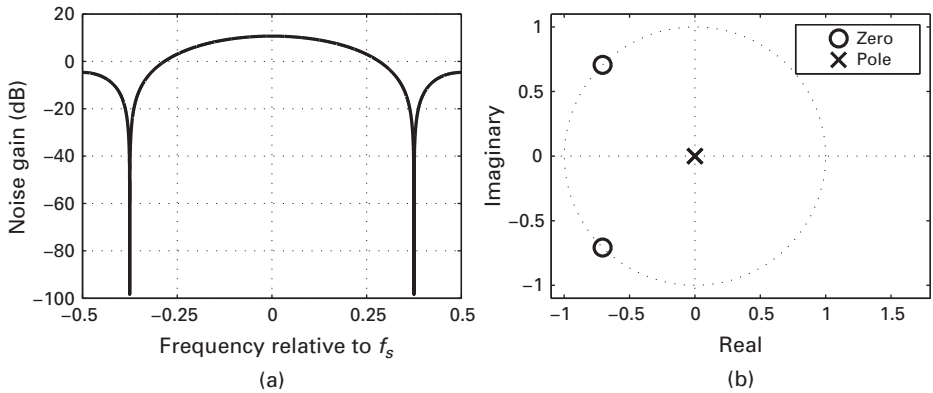


Figure 14.8 Amplitude response (a) and zero-pole plot (b) of the NTF for a second-order real BP $\Sigma\Delta$ modulator.

based on a switched-capacitor resonator was presented in [10]. The basic structure of the modulator is similar to the one presented in Figure 14.6, but an additional summation and the tuning parameter α_{bp} are included in order to make an arbitrary notch frequency possible. This is shown in Figure 14.7.

Bandpass transformation doubles the modulator order [47], which can be seen also from the zero-pole plot in Figure 14.8(b) as a doubled number of zeros and poles. The corresponding NTF response is given in Figure 14.8(a). The two zeros can be straightforwardly seen as two frequency-symmetric NTF notches in the amplitude response. The zeros – and the notches – are placed at a frequency ± 0.35 relative to the sampling frequency. With double the order, the amount of hardware to be implemented is of course increased. However, in practice the extra complexity compared to the lowpass receiver can be avoided by simplifying the overall receiver structure when utilizing a bandpass modulator.

Doubling the modulator order also means that, when considering an asymmetric complex bandpass signal to be converted, only half of the modulator zeros are on the desired band. This affects the expected SNR values on a given modulator order when compared to the lowpass case. For a lowpass modulator, increasing the oversampling ratio by an octave gives $6L + 3\text{dB}$ better SNR, L being the modulator order [4]. In the case of a bandpass $\Sigma\Delta$ modulator the corresponding value is $3L + 3\text{dB}$, from which it can be confirmed that a double-order modulator is indeed needed in order to get similar enhancement in SNR [25], [51].

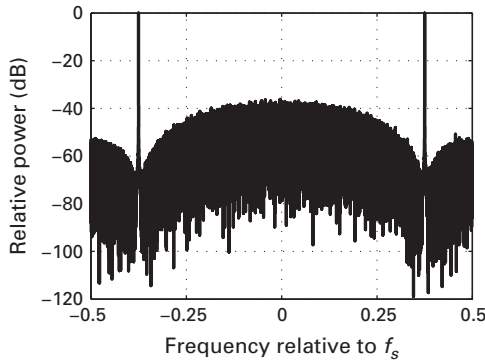


Figure 14.9 Simulated output spectrum for a BP $\Sigma\Delta$ modulator with a sine-wave input and an AWGN modeled quantization error.

As in the case of the lowpass modulator, a simulation with sine-wave excitation confirms the noise shaping effect. The output spectrum is given in Figure 14.9, verifying the NTF behavior demonstrated in Figure 14.8. The full-band SNR, defining thus the noise variance, is again equivalent to the theoretical 7.76 dB.

14.1.3 Quadrature bandpass modulator

The basic principle of quadrature $\Sigma\Delta$ conversion is presented in [5] and [22]. Analysis of [22] is extended in [24] and circuit implementation is discussed therein. Considerations and results in [24] are complemented with the mismatch analysis of [23] and more thorough background theory in [21]. A conventional quadrature $\Sigma\Delta$ converter consists of an input branch, a loop filter(s), and a feedback branch. The main difference between the real and quadrature modulators is that the latter operates on complex input samples and similarly the output is given in complex form.

This complex processing enables asymmetric noise shaping around DC for quadrature modulators [5], [22], [52]. It also implies that the poles and the zeros of the modulator can be asymmetric about the real axis, i.e., they do not need to be complex conjugates. These characteristics are demonstrated with the zero-pole plot and amplitude response in Figure 14.10. The zero-pole plot in Figure 14.10(b) shows that the modulator zero is situated on a positive frequency around 0.375 relative to the sampling frequency and lies on the unit circle. The pole is in origin.

The zero seen in the zero-pole plot in Figure 14.10(b) makes the NTF amplitude response notch on that frequency. This single zero results in a single frequency-asymmetric notch in contrast to the symmetric twin-notches in the case of the real bandpass modulator. If the desired complex signal is situated solely at the frequency of the notch, it is more efficient to have the notch only on that frequency.

Realizing this kind of behavior can be done with a complex transfer function [24]. Therefore, quadrature $\Sigma\Delta$ modulators (Q $\Sigma\Delta$ M) are designed with complex loop filters as is presented in Figure 14.11. A complex modulator has two real inputs and outputs, which describe the real and the imaginary part of the corresponding complex signals [24].

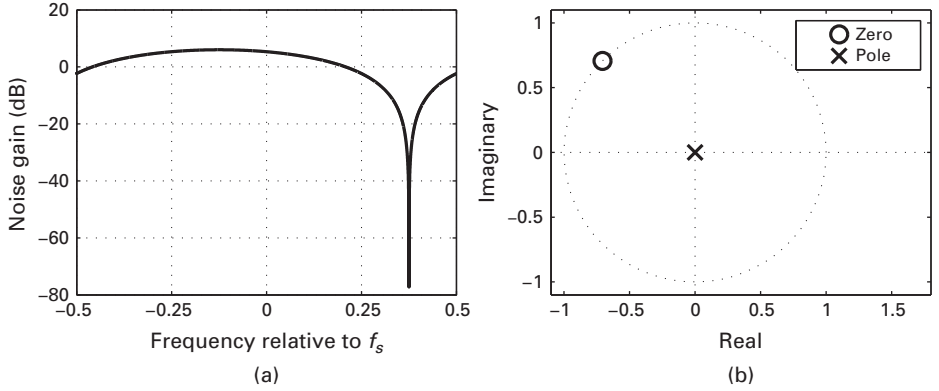


Figure 14.10 Amplitude response (a) and zero-pole plot (b) of the NTF, where the noise notch is at a frequency 0.35 relative to f_s .

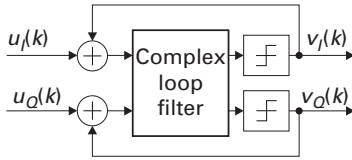


Figure 14.11 Quadrature $\Sigma\Delta$ modulator structure with a complex loop filter.

In principle, the noise shaping of the quadrature $\Sigma\Delta$ modulator can be thought of as a lowpass $\Sigma\Delta$ response with a complex frequency shift to some non-zero center frequency. A first-order quadrature modulator has a noise transfer function notch just as steep and wide as a corresponding lowpass modulator has on DC. This also means that the SNR equation derived for the lowpass modulator can be applied also in the case of the quadrature modulator.

A complex loop filter, employed in the quadrature modulator, can be realized as a complex integrator when considering the first-order modulator. Higher-order systems usually have multiple integrators included. The basic block diagram of a complex integrator is shown in Figure 14.12. The main principle is similar to the real integrator used in lowpass $\Sigma\Delta$ modulators. Input samples are integrated over a unit delay, the difference in a real system being that the samples are complex valued, as is the loop gain.

The complex integrator is graphically presented using complex-valued signals and a complex multiplier M in Figure 14.12(a). In addition, the same structure is given with real signals and real multiplications to demonstrate the actual implementation of the integrator.

The pole of such a complex filter is determined by the feedback multiplier. The structure presented leads to a transfer function $H[z]$ given by

$$H[z] = \frac{1}{z - M} = \frac{1}{z - (m_{re} + jm_{im})}, \quad (14.5)$$

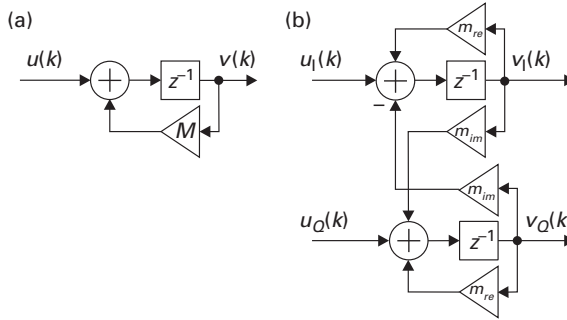


Figure 14.12 Realization of complex integrator with (a) complex signals and complex multiplication and (b) real signals and real multiplications.

where the terms M , m_{re} and m_{im} correspond to the illustration in Figure 14.12 denoting the complex multiplier and the real and imaginary parts of the multiplier, respectively. The delaying nature of the integrator can be highlighted by expanding to the following form

$$H[z] = \frac{z^{-1}}{1 - Mz^{-1}} = \frac{z^{-1}}{1 - (m_{re} + jm_{im})z^{-1}}, \quad (14.6)$$

where the term z^{-1} in the numerator clearly presents unit delay of the system. The pole of the complex integrator defines the zero of the noise transfer function of the whole modulator as will be seen later in the following subsection.

14.1.3.1 First-order quadrature modulator

The principle of quadrature $\Sigma\Delta$ conversion is demonstrated with analysis of the first-order system. In addition, analytical equations will be derived for the NTF and the STF. By definition, the first-order system has one noise shaping stage, which enables one NTF zero to be placed at the frequency band in use. In addition, the STF may have a notch on a chosen frequency as will be shown later.

In order to verify the linearity assumption (the AWGN model for the quantization error), the output spectra of a first-order quadrature modulator with an exponential excitation are given in Figure 14.13. The plot in Figure 14.13(a) is with 1-bit quantization. Figure 14.13(b) has quantization replaced with AWGN corresponding to 1-bit theoretical full-band SNR of 7.76 dB. The signal transfer function is frequency flat with unity gain in both cases. The interference peaks visible in Figure 14.13(a) around the relative frequency of -0.125 match to the aliased third-order nonlinear distortion from the second Nyquist zone with the relative center frequency of $f_{3rd} = -3 \times 0.375f_S + f_S = -0.125f_S$. The correspondence of the AWGN model is considered good enough, for analysis purposes, to be adopted for use. In addition to simplicity, the model enables closed form transfer function analysis due to the linearity assumption [23].

The basic structure of the first-order $\Sigma\Delta$ modulator is presented in Figure 14.14, where $u(k)$, $e(k)$, and $v(k)$ are the complex input, the linearized complex quantization error, and the complex output signals, respectively. The multipliers A , B , and G are complex valued

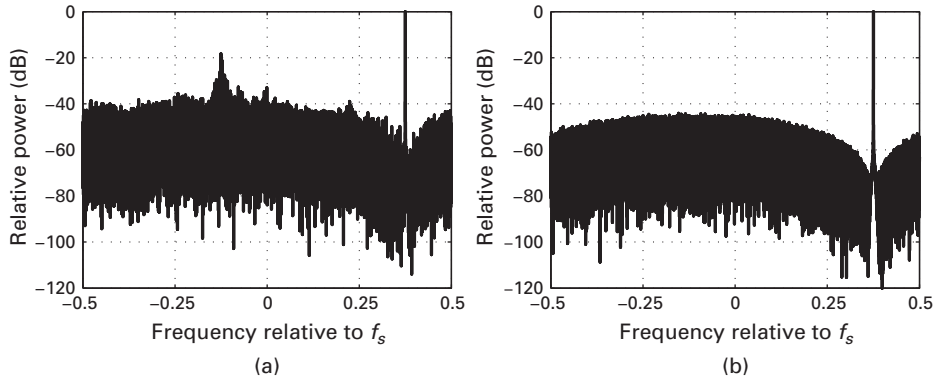


Figure 14.13 Modulator output spectra with exponential input and the NTF notch at 0.375 relative to f_s , simulated with 1-bit quantization (a) and AWGN (b).

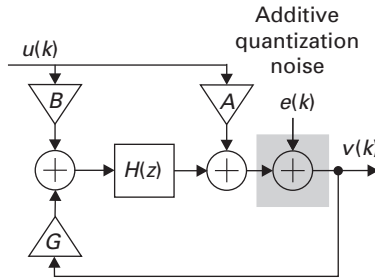


Figure 14.14 First-order quadrature $\Sigma\Delta$ modulator with the loop filter $H[z]$.

and affect the input signal fed straight to quantizer, the input signal fed to the loop filter $H[z]$, and the feedback signal from the output of the modulator, respectively.

Based on the linear model in Figure 14.14, the z -domain transfer function of the modulator can be derived. Thus, the output $V[z]$ of the modulator can be expressed with the input $U[z]$ and the quantization error $E[z]$ as

$$V[z] = \frac{A + BH[z]}{1 - GH[z]} U[z] + \frac{1}{1 - GH[z]} E[z], \quad (14.7)$$

where A , B , G , and $H[z]$ are the same as mentioned before and defined in Figure 14.14.

The loop filter $H[z]$ is a complex filter, which can be realized in either discrete-time switched-capacitor or continuous-time technique. Here, we concentrate mostly on the discrete-time system. The integrator's feedback branch has complex multiplier M (check Figure 14.12), which defines the pole – and thus the transfer function – of the integrator [24].

Based on the complex structure presented in Figure 14.15, the time domain behavior of the modulator can be analyzed. The system results in the discrete output $v(k)$, which can be expressed with help of the input $u(k)$ and the quantization error $e(k)$. The output is described by equation

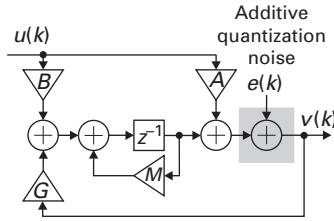


Figure 14.15 First-order quadrature $\Sigma\Delta$ modulator with a complex integrator. [36] © 2010 IEEE.

$$v(k) = Au(k) + (B - MA)u(k-1) + e(k) - Me(k-1) + (G + M)v(k-1). \quad (14.8)$$

Transforming Equation (14.8) to the Z-domain results in the transfer function of the modulator. Output is thus given by an equation

$$V[z] = \frac{A + (B - MA)z^{-1}}{1 - (G + M)z^{-1}} U[z] + \frac{1 - Mz^{-1}}{1 - (G + M)z^{-1}} E[z]. \quad (14.9)$$

This result pronounces the significance of the specified coefficients. On the other hand, it corresponds to Equation (14.7), where the Z-domain output was derived with the loop filter marked with $H[z]$, being now $H[z] = 1/(z - M)$ as given in (14.7). More detailed analysis of the overall transfer function will be given later in this section. However, it can already be noted that both of the transfer functions have a common denominator, which defines the pole of the modulator, giving the same pole for the STF and the NTF.

Implementationwise, the complex structure of Figure 14.15, which was described by Equation (14.8), can be realized with two real inputs and outputs. Complex multipliers also have two separate parts – real and imaginary. Based on this, the structure of a first-order quadrature $\Sigma\Delta$ modulator is presented with parallel real signals in time-domain in Figure 14.16. The figure depicts the coefficients of the structure in case of ideal matching between the I and Q branches.

In the modulator, the signal transfer function shapes the input of the modulator, as already mentioned when discussing the basic principles of $\Sigma\Delta$ modulation. Thus, the output of the modulator – excluding quantization error at this point – is a filtered version of the input. The signal transfer function of the first-order modulator, presented in Figure 14.15, is given by the multiplier of the input $U[z]$ in Equation (14.9). Thus, resulting in a formulation

$$STF[z] = \frac{A + (B - MA)z^{-1}}{1 - (G + M)z^{-1}}. \quad (14.10)$$

From the above equation it can be seen that in case of the first-order modulator one STF zero can be placed on the converted frequency band. The zero of the numerator, given by

$$z_{STF} = -\frac{B - MA}{A}, \quad (14.11)$$

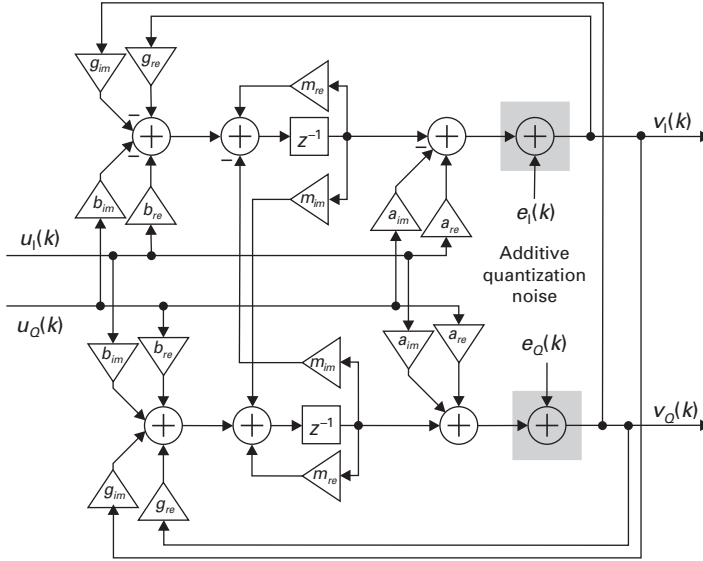


Figure 14.16 First-order quadrature $\Sigma\Delta$ modulator with real coefficients and parallel real signals. Used with permission from [37].

defines the notch in the amplitude response of the STF at a certain frequency. On the other hand, if the zero is placed in the origin, the resulting STF has a unity gain and is thus frequency flat. The coefficients A and B can be used to tune the STF response independent of the NTF (as can be seen from (14.9)). Thus, these complex values define the actual signal transfer function zero $z_{STF} = |z_{STF}|e^{j2\pi\tilde{f}_{int}}$, where $\tilde{f}_{int}$ is the notch frequency relative to the f_s .

At the same time, the noise transfer function describes how the quantization error is shaped inside the modulator. Thereby it can be separated from the overall transfer function as a multiplier of quantization error $E[z]$, resulting in an equation

$$NTF[z] = \frac{1 - Mz^{-1}}{1 - (G + M)z^{-1}}. \quad (14.12)$$

This indeed shows that the NTF design is independent of the input coefficients A and B in the modulator structure described earlier. In the first-order modulator, the feedback branch gain and the loop filter response define the noise shaping characteristics. The NTF zero location is dominated by the integrator pole. The numerator zero confirms this giving the zero location

$$z_{NTF} = M. \quad (14.13)$$

This also gives freedom to the NTF zero design because the loop filter pole M can be selected solely on the noise shaping basis, to give $z_{NTF} = |z_{NTF}|e^{j2\pi\tilde{f}_{pass}}$, where $\tilde{f}_{pass}$ is the noise transfer function notch frequency relative to the f_s . The signal transfer function shape can be designed thereafter using the complex gains of the input branches (A and B).

On the other hand, the transfer functions share the common pole (p_{common}) as was already discussed above. The pole can be solved straightforwardly from the common denominator of (14.10) and (14.12), being defined by

$$p_{common} = G + M. \quad (14.14)$$

As noted above, the parameter M , which is the pole of the complex integrator, is dominated by the desired noise shaping characteristics and the notch frequency, but the feedback gain G can be chosen to tune both the transfer functions. The effects of the zero and pole placements and related optimization trade-offs will be discussed in the following.

The placement of the transfer function zero on the unit circle, i.e., the angle of the zero, affects the frequency at which the notch is set. If the zero is pulled inwards from the unit circle, the notch flattens gradually. At the same time the pole placement does not affect the frequency of the actual notch, but does affect the shape of the frequency responses by giving gain for certain frequencies. This effect is demonstrated for the STF in Figure 14.17 with an amplitude response and a zero-pole plot. For the NTF, similar plots are given in Figure 14.18. When the pole is at the frequency of the zero and just inside the unit circle, amplitude response is flat except at the notch, which is steep and relatively narrow. Pulling the pole towards the origin makes the notch wider. Anyway, if the STF and the NTF notches are placed on separate frequencies, there is always a trade-off between the characteristics of the functions.

A flat STF passband would be desirable but, at the same time, placing the pole at the STF notch frequency gives rise to the noise gain as seen in Figure 14.18. Ideally, this out-of-band noise gain would not drastically affect the system performance, but, e.g., under branch mismatches and resulting imbalance between in-phase and quadrature components, this should be taken into account as will be seen later. On the other hand, STF gain on the NTF notch band would be desirable in order to amplify the desired signal, but this makes the noise shaping less efficient as can be seen from Figure 14.18.

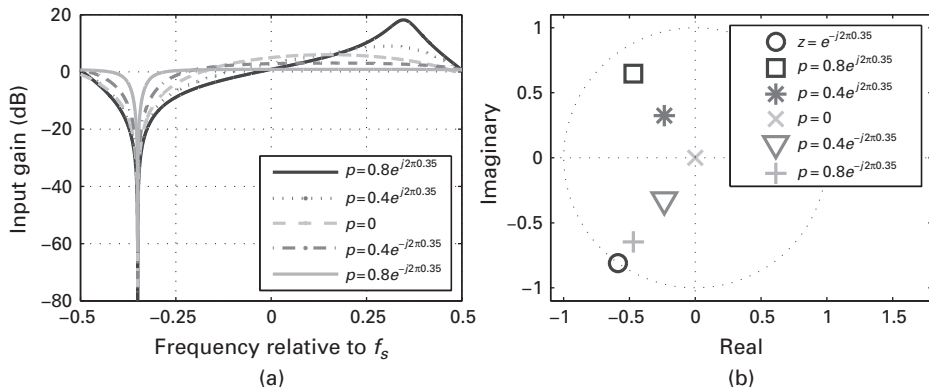


Figure 14.17 (a) An amplitude response of the STF, where the zero is located at a frequency -0.35 relative to f_s , and (b) a corresponding zero-pole plot. The zero and the pole position(s) are marked with z and p , respectively.

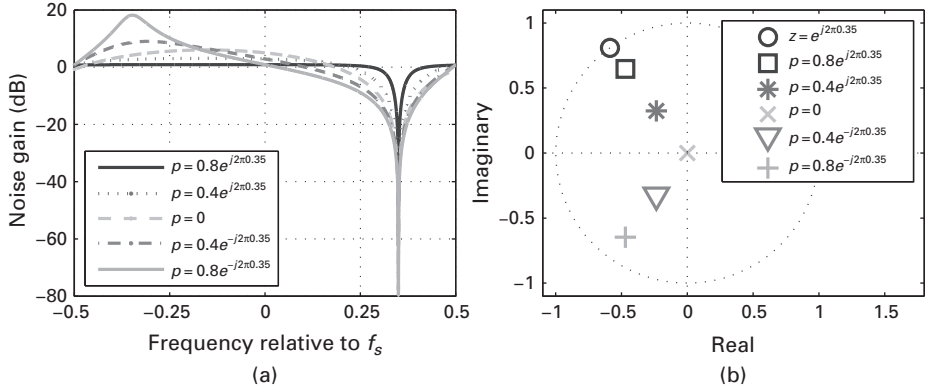


Figure 14.18 (a) An amplitude response of the NTF, where the zero is located at a frequency 0.35 relative to f_s , and (b) a corresponding zero-pole plot. The zero and the pole position(s) are marked with z and p , respectively.

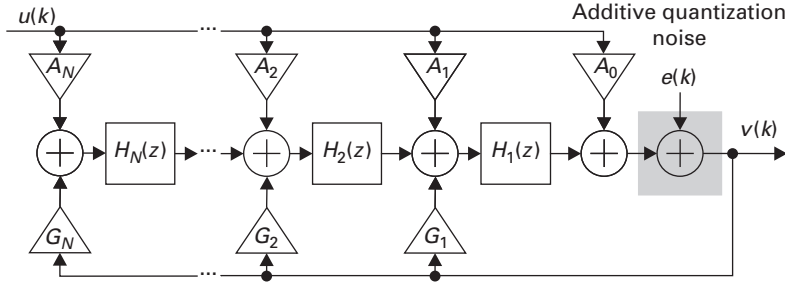


Figure 14.19 N th-order quadrature $\Sigma\Delta$ modulator. Used with permission from [37].

14.1.3.2 Higher-order quadrature modulator

A general model for a quadrature $\Sigma\Delta$ modulator of an arbitrary order can be derived by adding N loop filters (from $H_1[z]$ to $H_N[z]$) with the corresponding input (from A_0 to A_N) and feedback (from G_1 to G_N) branches to the structure. This is illustrated with a block diagram in Figure 14.19. In case of an N th-order modulator, N zeros and poles can be used to design the signal transfer function and the noise transfer function responses.

Next, the overall transfer function is derived for the above N th-order structure. The analysis of the modulator output $v(k)$ and Z -domain version $V[z]$ results in the following Z -domain equation:

$$V[z] = \frac{A_0 + \sum_{i=1}^N A_i \prod_{j=1}^i H_j[z]}{1 - \sum_{k=1}^N G_k \prod_{l=1}^k H_l[z]} U[z] + \frac{1}{1 - \sum_{k=1}^N G_k \prod_{l=1}^k H_l[z]} E[z]. \quad (14.15)$$

Based on the above formula, the STF and the NTF of the N th-order modulator are given by

$$STF[z] = \frac{A_0 + \sum_{i=1}^N A_i \prod_{j=1}^i H_j[z]}{1 - \sum_{k=1}^N G_k \prod_{l=1}^k H_l[z]} \quad (14.16)$$

and

$$NTF[z] = \frac{1}{1 - \sum_{k=1}^N G_k \prod_{l=1}^k H_l[z]}, \quad (14.17)$$

respectively.

As mentioned above, this kind of structure allows theoretically N separate zeros for the NTF and the STF. However, the practical order of the modulator is limited, e.g., by the stability of the whole structure. The highest well-accepted order in true circuit implementation proposed in the literature has been four [21], [24].

14.2 Considerations on further modulator concepts

In this section, the modulator concepts presented above are developed further. The selected, most practical topics, are highlighted in the following, based on [35]. For more detailed description in different $\Sigma\Delta$ implementation variants readers are advised to check [2], [4], [9], [21], [37], [47], [52], [55].

14.2.1 Single-bit vs. multi-bit

Either a 1-bit or a multi-bit quantizer can be employed in the $\Sigma\Delta$ architecture [2], [4], [15]. The number of bits is effectively changed in the forward-line ADC and the feedback DAC. The 1-bit converter has an inherent linearity but a limited resolution, which demands high oversampling ratios in order to gain resolution of numerous effective bits. The multi-bit quantization can be seen as an attractive solution especially for the conversion of wideband information signals when the achievable sampling rates limit the possible oversampling ratios [45].

On the other hand, a multi-bit converter has better resolution, but element mismatches in the feedback DAC cause a nonideal transfer function and nonlinear behavior, limiting linearity of the whole converter [27], [52]. A few methods have been presented in the literature to compensate for nonlinearities in DAC, e.g., [13], [49] and [64]. The role of the nonlinearities will be discussed in more detail in [Subsection 14.4.1](#).

14.2.2 Modulator order

The effects of the feedback order variations on a lowpass modulator have been documented quite extensively and analysis fit for purpose can be found from, e.g., [2] and [4]. The following analysis is based on [4]. In the basic $\Sigma\Delta$ principle, the negative feedback signal is fed straight back to a single integrator on the forward branch. At the same time,

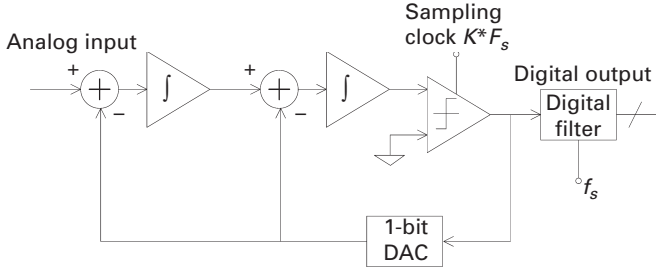


Figure 14.20 Second-order $\Sigma\Delta$ converter.

the noise shaping characteristics of the modulator can be enhanced by increasing the order of the feedback. The second-order structure introduces another integrator with feedback of its own as shown in [Figure 14.20](#).

Following again the analysis of [4] for the second-order modulator, the inband SNR becomes

$$SNR = 10 \log_{10}(\sigma_x^2) - 10 \log_{10}(\sigma_e^2) - 10 \log_{10}\left(\frac{\pi^4}{5}\right) + 50 \log_{10}\left(\frac{f_s}{2f_B}\right). \quad (14.18)$$

Based on this it is clear, that now every doubling of OSR increases SNR by approximately 15 dB. So there is obvious improvement when compared to the findings for the first-order modulator in Equation (14.4). The generalized version of SNR for L th-order modulator in [4] is given by

$$SNR = 10 \log_{10}(\sigma_x^2) - 10 \log_{10}(\sigma_e^2) - 10 \log_{10}\left(\frac{\pi^{2L}}{2L+1}\right) + (20L+10) \log_{10}\left(\frac{f_s}{2f_B}\right). \quad (14.19)$$

Thus, with this kind of modulator, every doubling of the oversampling ratio $f_s/2f_B$ gives $6L+3$ dBs of SNR improvement [4].

In [2] some stability issues with higher than second-order architecture are reported, but [4] and particularly [18] give a more thorough analysis on stability problems caused by raised feedback order. In brief, input values greater than a certain limit can cause the quantizer to overload and thus risk the stability. On the other hand, the limit cycle oscillations may also cause the quantizer to overload [4].

14.2.3 Single-stage vs. multi-stage

A higher-order noise transfer function and thus better noise shaping characteristics can be obtained also by employing cascaded modulators on a single input [52]. The main principle is that the following stage extracts and digitizes the remaining error of the preceding stage(s). After the conversion, this error signal is subtracted from the previous output. If the latter stages operate on $\Sigma\Delta$ principle, which is not obligatory in general, also the noise of these stages is shaped.

On the other hand, a traditional Nyquist ADC, e.g., a flash converter, can also be used as a following stage to simplify the structure. This kind of setup, where the second stage converter has zero-order noise shaping, is usually noted as L -0 cascade, where L denotes the order of the first-stage $\Sigma\Delta$ modulator. The L -0 cascade structure is also called the Leslie–Singh modulator [52]. In case of subsequent non-zero-order noise shaping stages, the structure is named MASH, which stands for Multi-stAge noise SHaping [52]. The principle can be extended to contain more than two stages, e.g., three-stage MASH is proposed in [52].

These kinds of structures have been proposed particularly for wideband systems with limited oversampling possibilities [45]. In addition, reconfigurable multi-stage converters for software defined radio receivers have been studied recently [7], [44], [48]. Reconfigurability means that some parts or stages can be digitally tuned or turned on and off depending on the demands of the communication standard being received at each moment.

The examples in the following subsections consider lowpass $\Sigma\Delta$ modulators for the sake of simplicity. However, the multi-stage concept can be utilized also in case of real [11] or complex [9] bandpass modulators. The quadrature variant of the multi-stage structure will be discussed in Subsection 14.3.1.

14.2.3.1 Leslie–Singh Modulator

An example of a Leslie–Singh structure consisting of a first-order $\Sigma\Delta$ modulator as a first stage is shown in Figure 14.21. Thus, this setup represents 1–0 cascade [52]. The quantization noise of the first stage $e_1(k)$ is extracted by subtracting the nonquantized signal $y_1(k)$ from the output of the first stage $v_1(k)$. This is actually done in the analog domain after digital-to-analog conversion in the feedback branch. Now this error signal is converted in order to obtain the digital version of the error as an output $v_2(k)$ of the second stage.

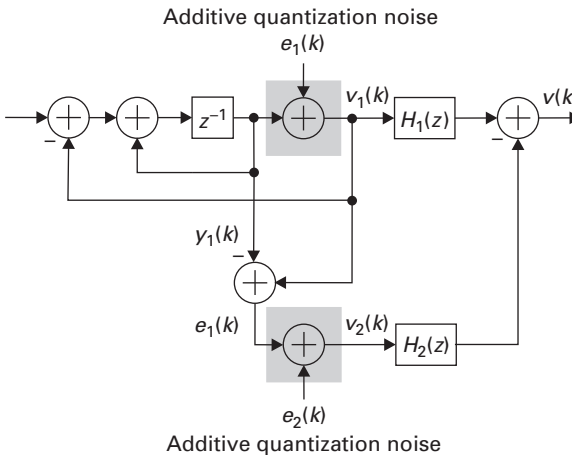


Figure 14.21 Discrete-time linearized model for the 1–0 cascade (Leslie–Singh) multi-stage $\Sigma\Delta$ structure.

To match the outputs of the stages, the output of the first stage must be delayed such that it is equal to the latency of the second conversion stage. This is done with choosing $H_1[z] = z^{-k}$. On the other hand, also the digitized error signal must be matched to the shaped version of the first stage quantization noise in order to successfully remove it from the output signal. This is done digitally with selection of $H_2[z] = NTF_1[z]$. Now subtracting the filtered output of the second stage from the delayed output of the first stage gives the final output

$$\begin{aligned} V[z] &= H_1[z]V_1[z] - H_2[z]V_2[z] \\ &= z^{-k}[STF_1[z]U[z] - NTF_1[z]E_2[z]]. \end{aligned} \quad (14.20)$$

Now, it can be seen that the error in the output signal is the quantization noise of the second stage shaped by the NTF of the first stage. This can be assumed smaller than the original error because the second-stage ADC can easily have more resolution (e.g., multi-bit pipeline converter) [52].

14.2.3.2 Multi-stage noise shaping (MASH)

If the second-stage quantization error $e_2(k)$ is also shaped with a $\Sigma\Delta$ modulator, the structure is known as MASH [52], which is graphically demonstrated in Figure 14.22. This leaves the desired signal band relatively clean of the second stage noise also. The price is increased complexity with the loop filter and the feedback. On the other hand, due to the noise shaping, now also the resolution of the second quantizer can possibly be lowered.

The overall noise shaping done in the stages is of the order of summed orders of the stages. This can be shown, e.g., by assuming purely delaying signal transfer functions with $STF_1[z] = STF_2[z] = z^{-1}$ and lowpass noise shaping modulators with $NTF_1[z] = NTF_2[z] = 1 - z^{-1}$. With this setup, deriving the output of the whole structure gives

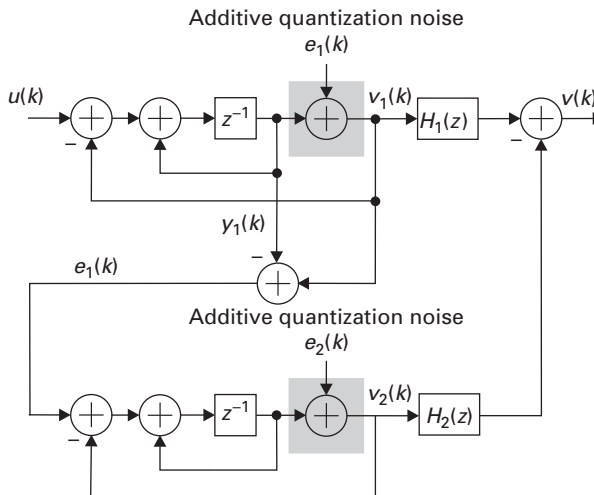


Figure 14.22 Discrete-time linearized model for 1-1 MASH $\Sigma\Delta$ structure.

$$\begin{aligned}
 V[z] &= H_1[z]V_1[z] - H_2[z]V_2[z] \\
 &= STF_1[z]STF_2[z]U[z] - NTF_1[z]NTF_2[z]E_2[z] \\
 &= z^{-2}U[z] - (1 - z^{-1})^2E_2[z],
 \end{aligned} \tag{14.21}$$

if the condition

$$H_1[z]NTF_1[z] - H_2[z]STF_2[z] = 0 \tag{14.22}$$

holds. Thus, the noise shaping performance of the 1–1 MASH is similar to the second-order modulator, but the stability is of the first order, because both of the stages have only one feedback.

One practical solution for satisfying the condition in (14.22) is to choose $H_1[z] = STF_2[z]$ and $H_2[z] = NTF_1[z]$, $STF_2[z]$ being often only a delay. This leads again to the matching of the digital $H_2[z]$ and the analog $NTF_1[z]$. Mismatches between the two transfer functions mentioned have been analyzed to cause leakage of the first-stage quantization noise to the output of the whole modulator structure. Several methods have been developed to prevent this, e.g., adaptive filtering [28], [29], [49], [64].

14.3 Advanced modulators

The following subsections discuss carefully chosen advanced QΣΔM structures, aimed towards emerging software and cognitive radio concepts and related radio receivers [1], [6], [7], [31], [34], [44], [46], [62], [63], [66]. First, the multi-stage modulator presented in Subsection 14.2.3 is extended to cover the quadrature variant of the modulator. Thereafter, a multiband modulator aimed to receive multiple information signals on distinct center frequencies is introduced. Finally, in Subsection 14.3.3 these concepts are combined by realizing multiband transfer functions with a multi-stage structure.

14.3.1 Quadrature multi-stage modulator

Real multi-stage structures presented in Subsection 14.2.3 can be extended to use quadrature signal processing. The resulting system has the degrees of freedom provided by the quadrature structure and, in addition, the multi-stage properties described earlier for the real lowpass modulator (Subsection 14.2.3) [9], [30], [55]. A block diagram of the quadrature MASH (real MASH discussed in Subsection 14.2.3.2) structure is given in Figure 14.23. The coefficients are marked as they were in the first-order single-stage quadrature ΣΔ modulator in Figure 14.15 and the coefficients of the stages are separated with the superscripts. On the other hand, if only the first stage has noise shaping properties, the system corresponds to the Leslie–Singh topology (real Leslie–Singh discussed in Subsection 14.2.3.1).

The overall output derived for the real MASH modulator in (14.21) holds also for the quadrature variant with the difference of transfer functions not being restricted to

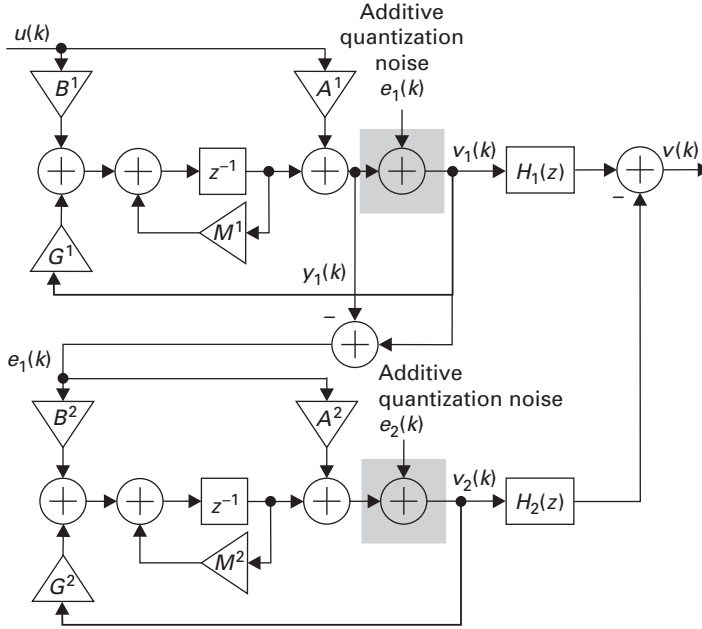


Figure 14.23 Quadrature multi-stage $\Sigma\Delta$ modulator with first-order noise shaping on both stages. Modulator coefficients of the stages are separated with superscript-markings (1 and 2 for the first and the second stage, respectively).

frequency-symmetry, i.e., the STF_s and NTF_s can be designed more freely. Thus, the (generally complex valued) output of the quadrature MASH modulator is

$$V[z] = STF_1[z]STF_2[z]U[z] - NTF_1[z]NTF_2[z]E_2[z]. \quad (14.23)$$

However, the condition of the analog and the digital transfer functions in (14.22) should still be valid.

14.3.2 Multiband quadrature modulator

The multiband $\Sigma\Delta$ principle discussed here, was introduced in [37] by the authors and the following presentation is based on the article. The quadrature $\Sigma\Delta$ modulator gives an extra degree of freedom for the design of the modulator transfer functions (STF and NTF). This can be exploited, e.g., by designing multiple parallel noise shaping notches, which can be tuned independently of each other.

With emerging cognitive radio and spectrum-sensing based communication systems, it is essential to be able to adapt to varying waveforms, especially on the receiver side, in terms of the used radio spectrum. The Q $\Sigma\Delta$ M offers a great possibility in this respect, especially in cases with scattered spectrum use, since the design of the transfer functions is not restricted to any symmetry with respect to zero frequency. With an N th-order modulator structure, it is possible to design N separate notches in the NTF, allowing reception of signals on N separate frequency slices. On the other hand, if wideband

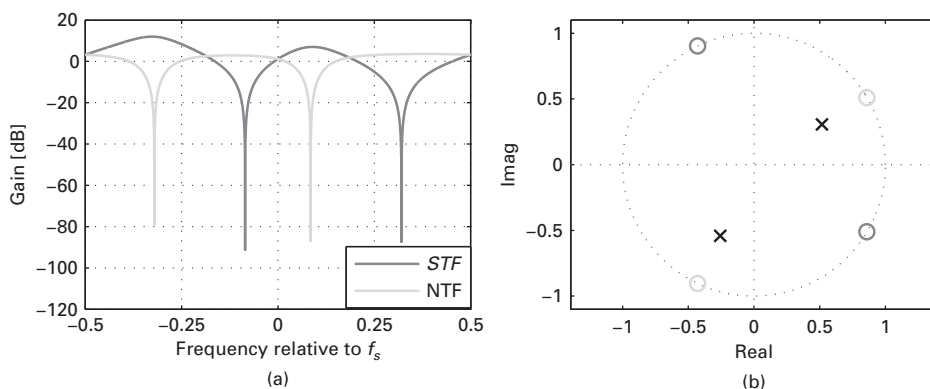


Figure 14.24 Example multiband transfer functions with the desired signal spectrum slices at normalized frequencies of 0.321 and -0.085 . NTF notches are placed at those frequencies and STF notches at the corresponding mirror frequencies. (a) Amplitude response of the transfer functions and (b) related zero-pole plot, where common poles are marked with black.

waveforms (with overall bandwidth in the order of tens of MHz) are deployed, the NTF zeros can also be placed next to each other forming a wide uniform notch. This is actually the traditional way of exploiting higher-order modulators and has been discussed extensively in the literature, e.g., in [24], [50], [52].

For the sake of simplicity, the multiband quadrature $\Sigma\Delta$ modulator design is demonstrated here with a second-order quadrature modulator. One possible transfer function setup is illustrated graphically in Figure 14.24, where the NTF notches are placed on the assumed desired signal frequencies (here 0.085 and -0.321 relative to sampling rate). The STF, in turn, is used to create selectivity for the overall modulator response. More specifically, the zeros of the signal transfer function are placed at the mirror frequencies of the desired signals (-0.085 and 0.321) to prevent I/Q mismatch originated interference. Such a design principle was concluded earlier to be an effective way to mitigate the mirror-frequency interference caused by the mismatches located in the modulator feedback branch [36], [37]. The poles of the transfer functions, in turn, are placed at the NTF notch frequencies in order to elevate the STF response at those bands.

A more concrete simulation example and details on the transfer function design are given in Subsection 14.5.1. Therein, the multiband $\Sigma\Delta$ system is simulated with true-like radio signals and a practical receiver structure.

14.3.3 Multiband quadrature multi-stage modulator

Both concepts presented in this section can be combined by realizing multiband noise and signal shaping with a multi-stage structure. In this way, higher-order transfer functions can be achieved without risking the stability of the modulator by increasing the order of a single stage [52].

Based on (14.23), the overall noise shaping response of a two-stage quadrature $\Sigma\Delta$ modulator is

$$NTF[z] = NTF_1[z]NTF_2[z], \quad (14.24)$$

and thus the product of the noise transfer functions of the stages. Therein, the $NTF_1[z]$ part is implemented by a digital filter $H_2[z]$ in the output of the second stage. When considering the notches of the overall response, it can be straightforwardly concluded that notching either of the stage-NTFs ($NTF_1[z]$ or $NTF_2[z]$) does notch the overall NTF. Thus, the multiband principle discussed in [Subsection 14.3.2](#) can be implemented in such a way, that the first stage realizes zeros (and thus notches) on some chosen frequencies and the latter stage on the rest. Therefore, the overall response is not dependent on the stage where the zero and the notch are implemented.

The same holds also for the STFs, the overall response being

$$STF[z] = STF_1[z]STF_2[z], \quad (14.25)$$

where the $STF_2[z]$ part is implemented by the digital filter $H_1[z]$ in the output of the first stage. However, $STF_2[z]$ is usually chosen to be a pure delay for the sake of simplicity, as discussed in [Subsection 14.2.3.2](#) for the real MASH. This concept can be, nonetheless, extended to include also signal selectivity in both the stages, which will be an interesting topic for future analysis.

A computer simulation based example on the concept is given in [Subsection 14.5.2](#). It is also shown with implemented I/Q mismatches, that the principle is applicable under nonideal conditions and can clearly improve the quality of the received information signals.

14.4 Nonidealities in quadrature sigma-delta modulators

In the following, the most important circuit nonidealities concerning $\Sigma\Delta$ modulator implementations are discussed. First, the role of nonlinearities inside the modulator, in either forward or feedback branch, is introduced. Thereafter, jitter effects in $\Sigma\Delta$ modulation are presented. Finally, there will be some discussion on I/Q imbalance in quadrature $\Sigma\Delta$ modulators. The focus is on I/Q imbalance, being a proven concern in quadrature structures [1], [23], [40], [41], [59], [60] and thus, an important issue in modulator systems like the ones discussed in [Section 14.3](#) and other related chapters in this book.

14.4.1 Nonlinearities

The linearity of the receiver is crucial in modern communication receivers, when multiple blocking signals are present with tens of dBs differences in power levels [53], [66]. In $\Sigma\Delta$ modulators, there are two main sources for the nonlinearities. The first one is the loop quantizer performing the nonlinear quantization process [2], [3], [27], [52]. The other nonlinearity source is a nonideal multi-bit DAC in the feedback branch of the modulator [13], [16], [17], [28], [29], [52]. The main difference between these error sources is that forward-branch errors, including the nonlinearities, are shaped by the NTF. At the same

time, nonlinearities in the feedback are not shaped and the errors appear directly in the output, limiting the linearity of the modulator.

In addition to nonlinearities in A/D and D/A conversion stages, the amplifiers used in the integrator stages generate nonlinear distortion in $\Sigma\Delta$ modulators. Nonidealities in operational amplifiers include noise, finite DC gain, limited bandwidth, finite slew rate, and saturation of voltage levels [33]. Despite the noise in the amplifier, the other nonidealities mentioned generate nonlinear distortion, corrupting the signal quality [33].

The main concern as a nonlinearity source has been the multi-bit DAC because those errors are not shaped by the NTF, as mentioned above. To mitigate this error, several algorithms have been presented in the literature. These include, e.g., randomization of the DAC elements [16], [17] and digital post-correction algorithms [13], [28], [29], [52]. In addition, [13] gives a thorough discussion on the current state of the art in DAC nonlinearity correction, and is thus recommended to interested readers.

14.4.2 Jitter

Nowadays, it is common to sample a received signal before the final conversion to baseband, e.g., on some chosen intermediate frequency [12]. This is also the case when discussing real or quadrature bandpass $\Sigma\Delta$ converters (Subsections 14.1.2 and 14.1.3, respectively). Thus, when higher frequencies are sampled, the nonidealities and uncertainty of the right sampling moment come to play an important role. When discussing the sampling jitter in $\Sigma\Delta$ converters, it is important to note a difference between switched-capacitor and continuous-time implementations. In a SC system, the sampling is done before the modulator loop and the modulator itself is implemented with discrete-time processing [52]. On the other hand, in a continuous-time modulator, the sampling circuit is placed right before the quantizer, and is thus inside the $\Sigma\Delta$ loop [52]. This results in a situation where the jitter noise of the continuous-time $\Sigma\Delta$ modulator is also shaped by the NTF of the modulator [42], [58].

In addition, in continuous-time modulators timing of the feedback signal becomes critical, creating another timing jitter source besides the sampling of the input signal. The severity of the DAC jitter in the feedback branch depends on the used DAC waveform and number of digital signal levels (number of bits in ADC and DAC). If the DAC pulse has (low-pass) filtered shape without steep transitions, small timing errors generate less error than with traditional non-return-to-zero or return-to-zero pulses [57], [58].

For more details on the jitter effects in $\Sigma\Delta$ modulators, it is recommended to check [42] and [58]. Specific features related to jitter in sampling in bandpass $\Sigma\Delta$ modulators are discussed in [56].

14.4.3 I/Q imbalance in quadrature $\Sigma\Delta$ modulators

In an ideal quadrature system, the component values of the parallel I- and Q-rails are matched precisely. This, however, is not possible in a true circuit implementation and always gives rise to I/Q imbalance and subsequent mirror frequency interference [1], [23], [40], [41], [59], [60]. The severity of the interference is dependent on the

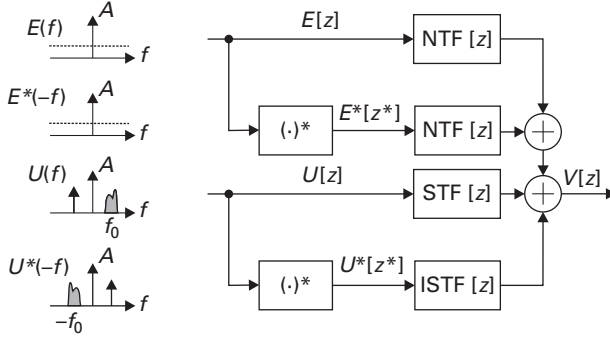


Figure 14.25 A principal block diagram of the output of mismatched quadrature $\Sigma\Delta$ modulator with principal spectra illustrating the signal components. [36] © 2010 IEEE.

accuracy of the component matching. The presentation in this subsection is based on [35], [36], [37].

In $\Sigma\Delta$ modulators, I/Q mismatch results generally in mirror transfer functions for both signal and noise components. Thus, the overall output of a mismatched modulator [23] is then presented by

$$V[z] = STF[z]U[z] + ISTF[z]U^*[z^*] + NTF[z]E[z] + INTF[z]E^*[z^*], \quad (14.26)$$

where $ISTF[z]$ and $INTF[z]$ are the image signal transfer function and image noise transfer function, respectively. The ISTF response shapes the conjugate version of the original input whereas the INTF response shapes the conjugate of the quantization noise. This principle and the impact of separate transfer functions are demonstrated in Figure 14.25.

The image rejection ratio (IRR) of the modulator can be defined based on the separate transfer functions for the signal and the noise components [23]. Frequency domain responses are attained with substitution $z \leftarrow e^{j2\pi f T_s}$, where f and T_s are the frequency in hertz and the sampling time, respectively, giving

$$IRR_{STF}(e^{j2\pi f T_s}) = 10 \log_{10} \left(\frac{|STF(e^{j2\pi f T_s})|^2}{|ISTF(e^{j2\pi f T_s})|^2} \right) \quad (14.27)$$

for the signal component and

$$IRR_{NTF}(e^{j2\pi f T_s}) = 10 \log_{10} \left(\frac{|NTF(e^{j2\pi f T_s})|^2}{|INTF(e^{j2\pi f T_s})|^2} \right) \quad (14.28)$$

for the quantization noise. These IRR quantities describe the relation of the original input signal and noise energy to the respective mismatch-induced mirror-frequency interference at the output signal. As an example, $IRR_{NTF}(e^{j2\pi f T_s}) = 20$ dB means that the power of the mismatch induced (mirrored) conjugate noise at a particular frequency f is 20 dB lower than the “original” quantization noise.

In the literature, several methods have been presented in order to improve the IRRs of quadrature $\Sigma\Delta$ modulators. The image rejection of the noise component has been proposed to be improved by placing one NTF zero, and thus generating a noise shaping

notch, on the mirror frequency of the desired signal [21], [23], [24]. Also switching between the components of the I and Q rails has been discussed [8], [39], [43]. Additionally, adaptive digital filtering has been proposed to compensate for the mirror frequency interference effects using insertion of the reference signal in the receiver to track the interference mechanism [67]. Interested readers are encouraged to check the original publications for details on the algorithms.

The actual frequency responses of the ISTF and INTF – and thus IRRs – are dependent on the exact location of the mismatched components inside the modulator. In [36] and [37] four different mismatch sources are analyzed in the case of the first-order quadrature $\Sigma\Delta$ modulator. These error sources are both the input branches, the loop filter, and the feedback branch. Some examples on different IRR behavior are given in the following, considering the input signal IRR ($IRR_{STF}(e^{j2\pi f T_s})$), which is an important concern especially in modern cognitive radio oriented radio receivers. The feasible IRR range in practical quadrature receivers as a whole is typically reported to be around 20–40 dB [12], [40], which considering a gain mismatch between the real coefficients of the branches means values in the order of 1–10%. Because the IRR of a single receiver component should be higher than the IRR of the whole receiver with multiple mirror-frequency interference sources, mismatches from 10^{-1} to 10^{-5} will be considered to cover the realistic range.

First, in the case of a mismatch in the input branch feeding the quantizer, and with unity STF design, the IRR peaks on the assumed desired signal band as shown in Figure 14.26. On the other hand, if the mismatch is located in the modulator feedback, the IRR behavior is frequency flat (with unity STF) as seen in Figure 14.27. This shows that the location of the mismatch inside the modulator has a big impact on the achieved image attenuation at the target band. More comprehensive presentation of the I/Q imbalance effects related to different mismatch scenarios, along with thorough analysis, is given in [35], [36], [37].

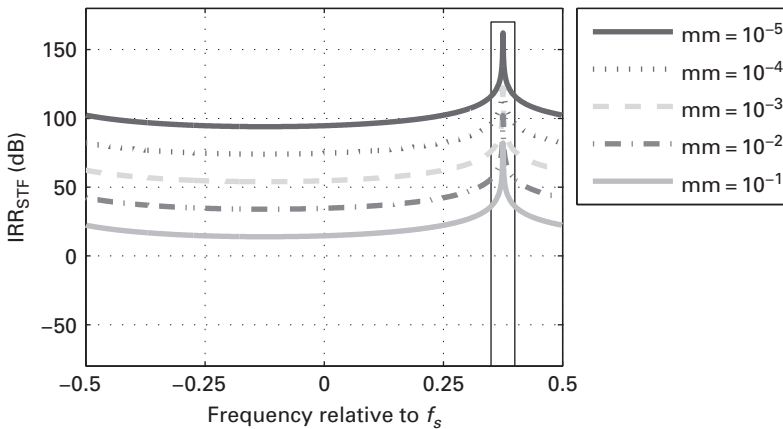


Figure 14.26 Signal IRR as a function of normalized frequency with mismatch in the input branch feeding the quantizer (unity STF design). Assumed desired band marked with black line. [36] © 2010 IEEE.

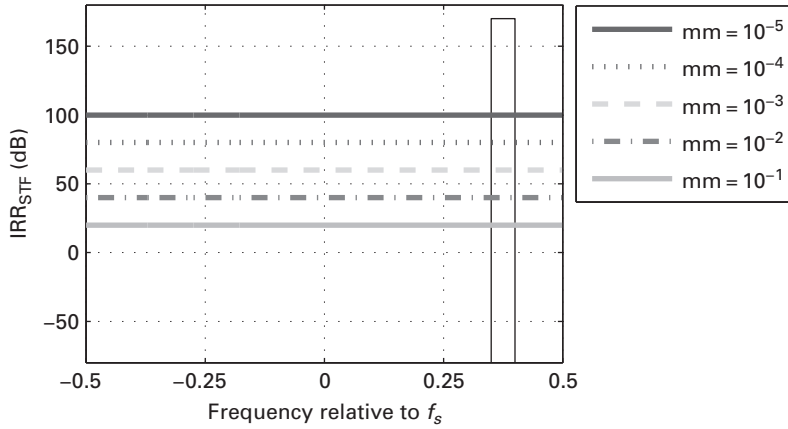


Figure 14.27 Signal IRR as a function of normalized frequency with mismatch in the modulator feedback branch of the modulator (unity STF design). Assumed desired band marked with black line. [36] © 2010 IEEE.

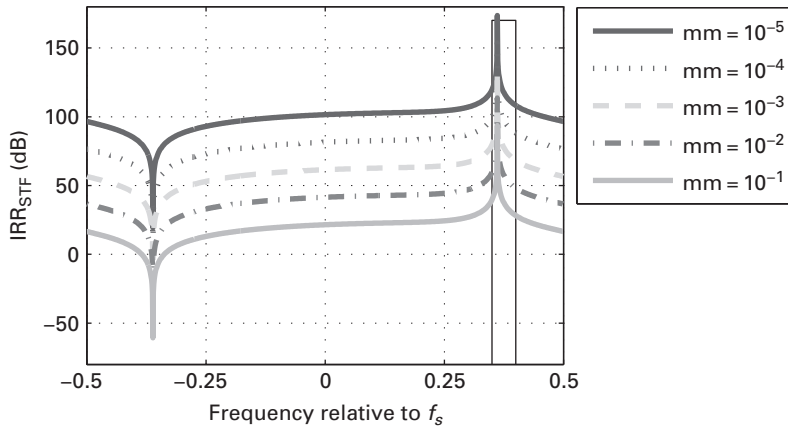


Figure 14.28 Signal IRR as a function of normalized frequency with mismatch in the feedback branch of the modulator (mirror frequency rejecting STF design). Assumed desired band marked with black line. [36] © 2010 IEEE.

Next, the idea of a modified STF design, with a zero on the assumed mirror frequency interference band (-0.361 relative to f_s for distinguishability) is examined. The resulting IRRs with such a STF design are given in Figure 14.28 showing a significant improvement in the desired band IRR. The peak IRRs seen in Figure 14.28 at 0.361 relative to f_s (mirror frequency of the STF notch) are in the order of 70 dB better than the ones achieved in Figure 14.27 with similar I/Q matching levels. Thus, the harmful input mirroring, originating from the feedback branch mismatch, can be practically cancelled with sophisticated STF design. This idea will be demonstrated with computer simulations on true-like radio signals in Subsection 14.5.1.

14.5 Examples on complex multiband transfer function design

The purpose of the following studies is to illustrate I/Q imbalance effects in multiband quadrature $\Sigma\Delta$ modulator with receiver simulations. At the same time, it is demonstrated that the mirror frequency rejecting STF design can be used to mitigate the interference effects (Subsection 14.5.1) [36], [37] and on the other hand, that the quadrature multi-stage principle can be exploited in similar scenarios (Subsection 14.5.2) [37], [55]. The basics of multi-stage and multiband systems were introduced in Subsections 14.3.1 and 14.3.2, respectively.

The cognitive radio receiver principle used in the following is shown with a block diagram in Figure 14.29. In the principal spectra of the receiver stages, the two light gray channels are the desired information signals on distinct center frequencies. Finally, digital filtering is done with a pair of complex BP filters, whereafter both the information signals are downconverted and detected separately.

The receiver principle presented above is next simulated with true-like radio signals in order to verify the applicability. The spectrum of the signal composition used in the simulation is given in Figure 14.30, assuming intermediate frequency A/D conversion. The overall desired signal consists of two sub-waveforms, a 64-QAM signal around 22 MHz center frequency with 12 MHz bandwidth and a 8-PSK signal around -82 MHz center frequency with 16 MHz bandwidth. There are also two blocking signals consisting of band-limited Gaussian noise at center frequencies of -82 MHz and 22 MHz. The relative power levels of the 64-QAM desired signal, the blocker around $+82$ MHz and the blocker around -22 MHz are +5 dB, +15 dB, and +25 dB, respectively, when compared to the 8-PSK signal at -82 MHz. The quantization noise is assumed Gaussian, white, uncorrelated with the input and having a variance equal to 1-bit quantization error. In addition, I/Q mismatch of 1% is introduced in both the modulator feedback branches,

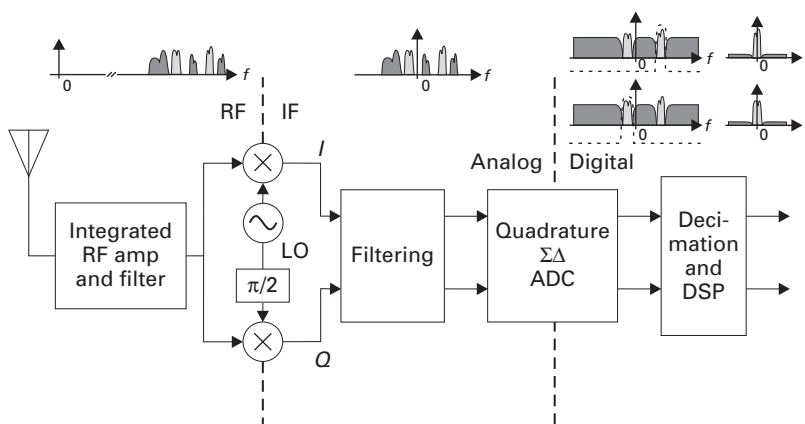


Figure 14.29 A block diagram of a $Q\Sigma\Delta M$ -based *multiband* low-IF quadrature receiver with principal spectra, where the two light gray signals are the desired ones, illustrating the signal composition at each stage used with permission from [37].

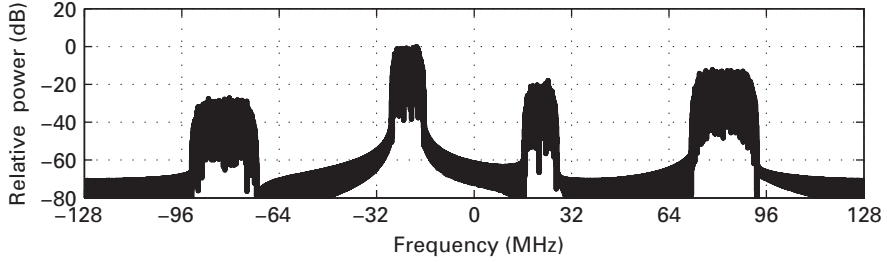


Figure 14.30 Input spectrum of the multiband quadrature $\Sigma\Delta$ modulator, consisting of the desired 64-QAM signal (roughly 22 MHz center frequency), the desired 8-PSK signal (roughly -82 MHz center frequency), and the two Gaussian blocking signals at the mirror bands.

meaning the complex gains G_1 and G_2 (check block diagram of the N th-order modulator in Figure 14.19).

In the following, I/Q mismatch is introduced between the rails in the feedback branch of the modulator. The principle of implementing the mismatches is illustrated in the case of first-order Q $\Sigma\Delta$ in Figure 14.31. Ideally, the complex gain value is $G = -M + e^{-j2\pi\bar{f}_{pass}/2}$, where $\bar{f}_{pass}$ is the normalized frequency of the NTF notch being implemented. With mismatch Δ_G , the implementation values are as follows: $g_{re,1} = (1 + \Delta_G)g_{re}$, $g_{re,2} = (1 - \Delta_G)g_{re}$, $g_{im,1} = (1 + \Delta_G)g_{im}$ and $g_{im,2} = (1 - \Delta_G)g_{im}$ (see Figure 14.31). With the second-order modulator, the same applies for both of the feedback branches (complex gains G_1 and G_2). Symmetric mismatches in both real and imaginary parts were chosen for simplicity of analysis and because mirror transfer functions depend only on the amount of difference between the branches, not on the possible symmetry of the deviation [23]. Moreover, it was confirmed in simulations that no significant differences appear between the mirroring effects with symmetric and asymmetric mismatches. In addition, in a two-stage system, similar mismatches are assumed in both the stages.

14.5.1 Single-stage multiband with I/Q imbalance

In this subsection, a single-stage second-order quadrature $\Sigma\Delta$ modulator is used in the receiver composition described above (block diagram given in Figure 14.29). Two separate cases are discussed. Namely, frequency flat STF design without any filtering effects and mirror frequency rejecting STF design, which is shown to be effective in the presence of blocking signals and mismatch between I and Q rails of the modulator.

The frequency flat STF is designed with the following (second-order) modulator coefficient values in the input branches (see Figure 14.19): $A_0 = 1$, $A_1 = M_1 + M_2$, $A_2 = M_2A_1 - M_1M_2$ and in the feedback branches: $G_1 = -M_1 - M_2$, $G_2 = M_1M_2 + M_2G_1$. The coefficients M_1 and M_2 are the complex-valued feedback gains of the two loop integrators ($H_1[z]$ and $H_2[z]$) giving the zeros of the NTF. The zeros are defined as $M_1 = e^{j2\pi\bar{f}_{pass1}}$ and $M_2 = e^{j2\pi\bar{f}_{pass2}}$, having $\bar{f}_{pass1} = 0.085$ and $\bar{f}_{pass2} = -0.321$ (relative center frequencies of the 64-QAM and 8-PSK signals,

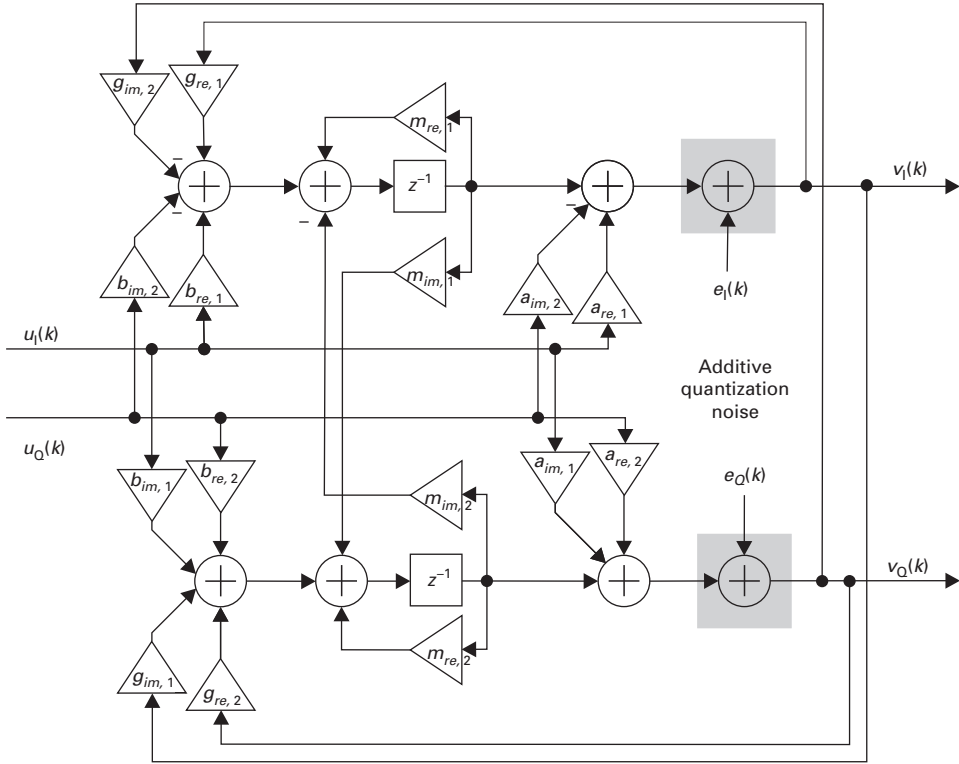


Figure 14.31 First-order quadrature $\Sigma\Delta$ modulator with real coefficients and parallel real signals and taking implementation I/Q mismatches into account. [36] © 2010 IEEE.

respectively) for this particular case. This creates two distinct noise shaping notches, which are placed on the center frequencies of the information signals of interest. This noise shaping effect can also be seen in the output spectrum in case of the flat STF design case, which is given in Figure 14.32. It is also worth noting that the blocking signals are also present in the output without any attenuation. The related detected information signal constellations are shown in Figure 14.33, from where it is visible that the quantization noise and generated mirror frequency interference have corrupted the signals. In addition, the bit error ratios (BER) are given in the caption.

Next, the effect of the mirror frequency rejecting second-order STF design is studied. This particular setup is achieved with the following complex gains on the input branches (see again Figure 14.19): $A_0 = 1$, $A_1 = M_2 + M_1 - (e^{j2\pi\bar{f}_{int1}} + e^{j2\pi\bar{f}_{int2}})$, and $A_2 = M_2A_1 - M_1M_2 + (e^{j2\pi\bar{f}_{int1}} + e^{j2\pi\bar{f}_{int2}})$, while the corresponding gains on the feedback branches have values $G_1 = -M_1 - M_2 + (0.6e^{j2\pi\bar{f}_{pass1}} + 0.6e^{j2\pi\bar{f}_{pass2}})$ and $G_2 = M_1M_2 + M_2G_1 - (0.6e^{j2\pi\bar{f}_{pass1}} + 0.6e^{j2\pi\bar{f}_{pass2}})$, where $\bar{f}_{int1} = -0.085$, $\bar{f}_{int2} = 0.321$, $\bar{f}_{pass1} = 0.085$, and $\bar{f}_{pass2} = -0.321$ denote the normalized frequencies of the two STF notches and the two NTF notches, respectively. Above M_1 and M_2 are defined by the STF zeros, as was the case with the flat STF design earlier. At the same time, the terms $0.6e^{j2\pi\bar{f}_{pass1}}$ and $0.6e^{j2\pi\bar{f}_{pass2}}$ define the positions of the two common poles

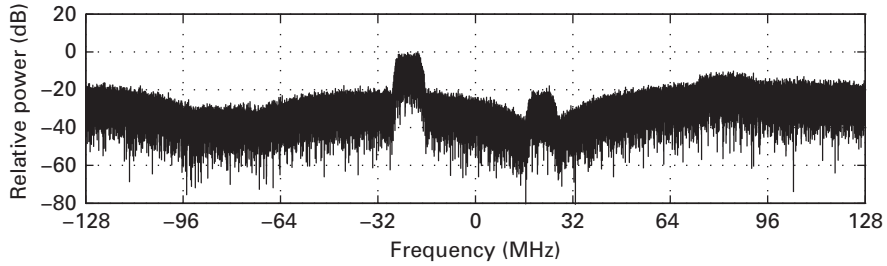


Figure 14.32 Output spectrum of the multiband quadrature $\Sigma\Delta$ modulator with frequency flat STF design, consisting of the desired 64-QAM signal (roughly 22 MHz center frequency), the desired 8-PSK signal (roughly -82 MHz center frequency) and the two Gaussian blocking signals at the mirror bands.

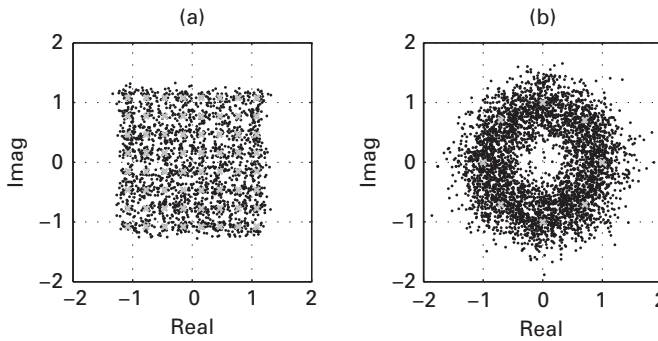


Figure 14.33 Received constellations in multiband setup with frequency flat STF design for (a) 64-QAM signal around 22 MHz (BER 4.28%) and (b) 8-PSK signal around -82 MHz (BER 12.0%). Ideal symbol locations marked with crosses.

for the STF and the NTF. It should also be noticed that the previous frequency flat STF design is achieved with setting the pole ($0.6e^{j2\pi f_{pass1}}$ and $0.6e^{j2\pi f_{pass2}}$ above) and the STF zero ($e^{j2\pi f_{int1}}$ and $e^{j2\pi f_{int2}}$ above) terms to zero.

Now, with the mirror frequency rejecting STF design, the blockers are attenuated in the output of the modulator, as can be seen from the spectrum shown in Figure 14.34. The blocking signals are pushed to the noise floor. It is also clear that the desired 64-QAM signal on 22 MHz center frequency is more above the noise floor compared to the flat STF design. The constellation plots in Figure 14.35 confirm this observation. The amount of noise seen in the constellations is significantly less and in both the signals (64-QAM and 8-PSK constellations) the bit error ratios have dropped below 1%. In the spectrum plot, the weak 8-PSK signal appears to be buried in the noise, but the noise shaping applied is pushing the noise level down exactly at that frequency band, thus enabling reliable enough detection of the signal. Also the mirror-frequency interference is kept at minimum, despite the I/Q mismatches, due to selective signal transfer function design.

All the NTF and the STF notches can be tuned independently of each other, giving the needed degree of freedom for the design of the cognitive receiver. In the second-order

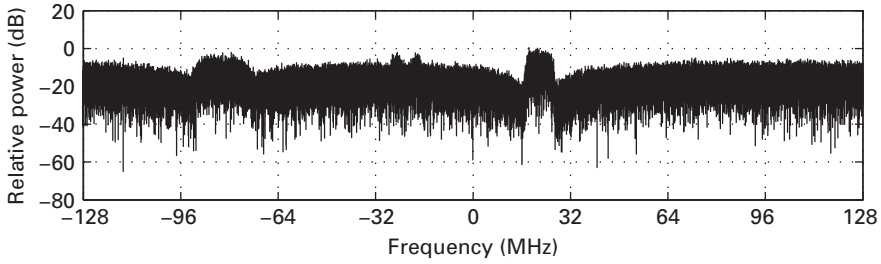


Figure 14.34 Output spectrum of the multiband quadrature $\Sigma\Delta$ modulator with mirror frequency rejecting STF design, consisting of the desired 64-QAM signal (roughly 22 MHz center frequency), the desired 8-PSK signal (roughly -82 MHz center frequency) and the attenuated Gaussian blocking signals at the mirror bands.

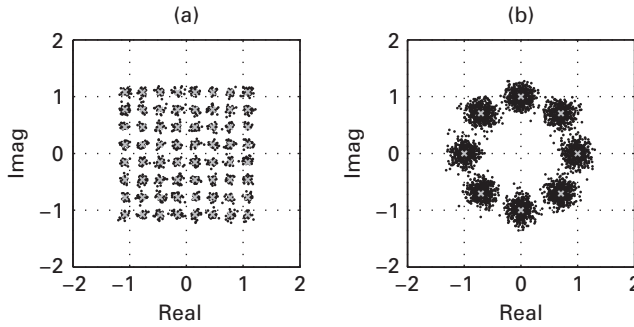


Figure 14.35 Received constellations in multiband setup with mirror frequency rejecting STF design for (a) 64-QAM signal around 22 MHz (BER 0.003%) and (b) 8-PSK signal around -82 MHz (BER 0.008%). Ideal symbol locations marked with crosses.

case, modifying the aforementioned frequency values ($\bar{f}_{int1}$, $\bar{f}_{int2}$, $\bar{f}_{pass1}$ and $\bar{f}_{pass2}$) based on the signal composition at hand (spectrum sensing), allows straightforward reconfigurability of the system.

Based on these findings the complex multiband principle is found to be valid. In this way, it is possible to design a highly reconfigurable and efficient A/D stage with built-in selectivity and robustness against I/Q mismatches. The Q $\Sigma\Delta$ M offers higher flexibility and degree of freedom when compared to real $\Sigma\Delta$ modulators and thus the multiband structure can be implemented more efficiently.

14.5.2 Multi-stage multiband with I/Q imbalance

Multi-stage $\Sigma\Delta$ modulation (real version discussed in [Subsection 14.2.3](#) and quadrature variant in [Subsection 14.3.2](#)) allows the precision of the related A/D converter to be enhanced. In 2-stage structure, the overall NTF is a product of the NTFs of the two stages, thus allowing more sophisticated noise shaping [52]. In this subsection, this property is exploited by generating four noise shaping notches with a 2-stage quadrature $\Sigma\Delta$

modulator in which both the stages are of second order. This kind of system has, of course, the drawback of additional complexity in the circuitry to be implemented, but is able to improve the performance significantly [55].

The receiver principle implemented in the simulation is similar to that described above and again, 1% I/Q mismatches are implemented in the modulator feedback (now both the first and the second stage), as was the case also in Subsection 14.5.1. For simplicity, the STFs of the stages are implemented as frequency flat here. Digital filters $H_1[z]$ and $H_2[z]$ are matched to the transfer functions of the stages (described in Subsection 14.3.1). This is done by having $H_1[z] = STF_2[z]$, which in this case means pure delay (frequency flat $STF_2[z]$) and $H_2[z] = NTF_1[z]$ implementing the noise shaping of the first stage for the quantization error of the second stage.

The NTFs of the multi-stage system are designed in such a way that both stages realize two separate noise shaping notches on distinct frequencies. The input signal composition is the same as used in the single-stage system in Subsection 14.5.1 and is described at the beginning of this section. When comparing the spectra in Figure 14.36, the noise shaping effects of the stages can be seen. It can be seen, e.g., around 22 MHz (center frequency of the desired 64-QAM signal), that the first stage is shaping the noise away from the lower frequencies of the desired signal band (above the spectrum in the figure) and the second stage is clearing the higher end of the band (below the spectrum in the figure), making reliable detection possible. It is also noted in [50] that optimal selection of the NTF zeros (and thus notch frequencies) is this kind of non-overlapping composition.

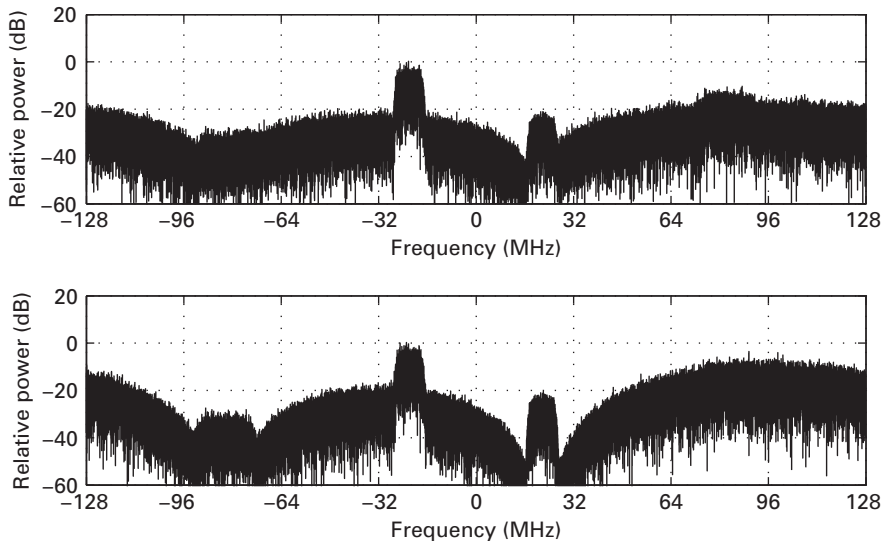


Figure 14.36 Output spectra (output of the *first stage* above and output of the *whole structure* below) of the multiband quadrature $\Sigma\Delta$ modulator with 2-stage structure, consisting of the desired 64-QAM signal (roughly 22 MHz center frequency), the desired 8-PSK signal (roughly -82 MHz center frequency) and the Gaussian blocking signals at the mirror bands.

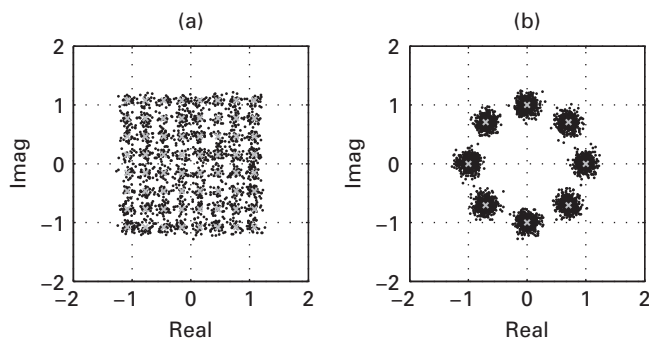


Figure 14.37 Received constellations in multiband setup with 2-stage structure for (a) 64-QAM signal around 22 MHz (BER 1.29%) and (b) 8-PSK signal around -82 MHz (no bit errors). Ideal symbol locations marked with crosses.

The noise shaping allows reliable detection of the desired information signals despite the blocking signals present in the input of the modulator and 1% I/Q mismatches implemented in the feedback branches of both the stages. The constellation plots of the information signals are given in Figure 14.37. From the figure it can be seen that the signals are not badly corrupted and noise levels are quite low. Due to the efficient noise shaping, only a small amount of errors are present. The 8-PSK signal is detected without errors and also 64-QAM signal BER is at a feasible level when channel coding is not considered.

14.6 Conclusion

In this chapter the basic theory of $\Sigma\Delta$ modulation was presented in Section 14.1, including, e.g., lowpass, real bandpass and complex bandpass variants of the $\Sigma\Delta$ modulator. Based on that, selected further concepts were emphasized in Sections 14.2 and 14.3, focusing on solutions applicable in modern cognitive type of radio receivers. Thereafter, related inevitable implementation nonidealities were discussed in Section 14.4. In Section 14.5, the advanced structures discussed earlier were concretized with illustrative simulation examples, in which also circuit mismatches were considered. The simulation results validated the concepts, which could be highly exploited in reconfigurable receivers.

Traditionally, $\Sigma\Delta$ modulator based A/D converters have been directed toward narrow-band solutions, e.g., in audio solutions. However, more advanced structures, like the ones described in Section 14.3, are making more versatile exploitation of the $\Sigma\Delta$ principle attainable. Particularly the quadrature variant of the modulator offers interesting possibilities, e.g., in radio receivers directed to emerging cognitive radio networks. Frequency agility is an important feature in this kind of system, which is exactly what quadrature $\Sigma\Delta$ modulator offers – including the multiband principle discussed in Sections 14.3 and 14.5. Straightforward reconfigurability of the transfer functions – and thus modulator

functionalities – is achieved with simple parameterization. In addition, the STF and the NTF are not limited to frequency symmetric responses, giving an additional degree of freedom for the design. The presented methods (mirror frequency rejecting STF design and multi-stage quadrature modulator) offer also robustness against inevitable mismatches in implementation of quadrature circuitry, making the realization of such systems even more interesting.

References

- [1] A. A. Abidi, "Direct-conversion radio transceivers for digital communications," *IEEE J. Solid-State Circuits*, vol. **30**, no. 12, pp. 1399–1410, Dec. 1995.
- [2] Analog Devices, *The Data Conversion Handbook*, W. Kester, ed. Burlington, MA: Newnes, 2004.
- [3] S. H. Ardalan and J. J. Paulos, "Analysis of nonlinear behavior in delta-sigma modulators," *IEEE Trans. Circuits Syst.*, vol. **34**, no. 3, pp. 593–603, June 1987.
- [4] P. A. Aziz, H. V. Sorensen, and J. van der Spiegel, "An overview of sigma-delta converters: how a 1-bit ADC achieves more than 16-bit resolution," *IEEE Signal Processing Mag.*, vol. **13**, no. 1, pp. 61–84, Sept. 1996.
- [5] P. A. Aziz, H. V. Sorensen, and J. van der Spiegel, "Performance of complex noise transfer functions in bandpass and multi band sigma delta systems," *IEEE Int. Symp. Circuits and Syst.*, Seattle, WA, 1995, pp. 641–644.
- [6] R. Bagheri, A. Mirzaei, M. E. Heidari, *et al.*, "Software-defined radio receiver: dream to reality," *IEEE Commun. Mag.*, vol. **44**, no. 8, pp. 111–118, Aug. 2006.
- [7] A. Baschiroto, F. Campi, R. Castello, *et al.*, "Baseband analog front-end and digital back-end for reconfigurable multi-standard terminals," *IEEE Circuits Syst. Mag.*, vol. **6**, no. 1, pp. 8–28, 1st quarter 2006.
- [8] L. J. Breems, E. C. Dijkmans, and J. H. Huijsing, "A quadrature data-dependent DEM algorithm to improve image rejection of a complex $\Sigma\Delta$ modulator," *IEEE J. Solid-State Circuits*, vol. **36**, no. 12, Dec. 2001.
- [9] L. J. Breems, R. Rutten, R. H. M. van Veldhoven, and G. van der Welde, "A 56 mW continuous-time quadrature cascaded $\Sigma\Delta$ modulator with 77 dB DR in a near zero-IF 20 MHz band," *IEEE J. Solid-State Circuits*, vol. **42**, no. 12, pp. 2696–2705, Dec. 2007.
- [10] L. Cardelli, L. Fanucci, Y. Kempe, E. Mannozi, and D. Strle, "Tunable bandpass sigma delta modulator using one input parameter," *Electronics Lett.*, vol. **39**, no. 2, pp. 187–189, Jan. 2003.
- [11] T.-H. Chang and L.-R. Dung, "Fourth-order cascaded $\Sigma\Delta$ modulator using tri-level quantization and bandpass noise shaping for broadband telecommunication applications," *IEEE Trans. Circuits Systems I, Reg. Papers*, vol. **55**, no. 6, pp. 1722–1732, July 2008.
- [12] J. Crols and M. S. J. Steyaert, "Low-IF topologies for high-performance analog front ends of fully integrated receivers," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. **45**, no. 3, pp. 269–282, Mar. 1998.
- [13] G. Gagnon, "Continuous compensation of binary-weighted DAC nonlinearities in bandpass delta-sigma modulators," Ph.D. dissertation, Dept. Electronics, Carleton Univ., Ottawa, Canada, 2008.
- [14] R. M. Gray and D. Neuhoff, "Quantization," *IEEE Trans. Information Theory*, vol. **44**, no. 6, pp. 2325–2383, Oct. 1998.

- [15] R. M. Gray, "Quantization noise in $\Delta\Sigma$ A/D converters," in *Delta-Sigma Data Converters*. Hoboken, NJ: Wiley-IEEE Press, 1995, ch. 2, pp. 44–74.
- [16] A. K. Gupta, E. Sanchez-Sinencio, S. Karthikeyan, W. M. Koe, and Y.-I. Park, "Second order dynamic element matching technique for low oversampling delta sigma ADC," *IEEE Int. Symp. Circuits and Syst.* Island of Kos, Greece, 2006.
- [17] A. A. Hamoui and K. W. Martin, "High-order multibit modulators and pseudo data-weighted-averaging in low-oversampling $\Delta\Sigma$ ADCs for broad-band applications," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. **51**, no. 1, pp. 72–85, Jan. 2004.
- [18] S. Hein and A. Zakhor, "On the stability of sigma delta modulators," *IEEE Trans. Signal Process.*, vol. **41**, no. 7, pp. 2322–2348, July 1993.
- [19] H. Inose and Y. Yasuda, "A unity bit coding method by negative feedback," *Proc. IEEE*, vol. **51**, pp. 1524–1535, Nov. 1963.
- [20] D. Jakonis, K. Folkesson, J. Dabrowski, P. Erikson and C. Svensson, "A 2.4-GHz RF sampling receiver front-end in 0.18- μm CMOS," *IEEE J. Solid-State Circuits*, vol. **40**, no. 6, pp. 1265–1277, June 2005.
- [21] S. Jantzi, "Quadrature bandpass delta-sigma modulation for digital radio," Ph.D. dissertation, Dept. Electrical and Computer Eng., University of Toronto, Toronto, Canada, 1997.
- [22] S. Jantzi, K. W. Martin, M. Snelgrove, and A. S. Sedra, "A complex bandpass $\Delta\Sigma$ converter for digital radio," *IEEE Int. Symp. Circuits and Syst.*, London, UK, 1994, pp. 453–456.
- [23] S. Jantzi, K. W. Martin, and A. S. Sedra, "The effects of mismatch in complex bandpass $\Delta\Sigma$ modulators," *IEEE Int. Symp. Circuits and Syst.*, Atlanta, GA, 1996, pp. 227–230.
- [24] S. Jantzi, K. W. Martin, and A. S. Sedra, "Quadrature bandpass $\Delta\Sigma$ modulation for digital radio," *IEEE J. Solid-State Circuits*, vol. **32**, no. 12, pp. 1935–1950, Dec. 1997.
- [25] S. Jantzi, M. Snelgrove, and R. Schreier, "Bandpass sigma-delta analog-to-digital conversion," *IEEE Trans. Circuits Syst.*, vol. **38**, no. 11, pp. 1406–1409, Nov. 1991.
- [26] D. Johns and K. Martin, *Analog Integrated Circuit Design*. Crawfordsville, IN: John Wiley & Sons, 1997.
- [27] J. G. Kenney and L. R. Carley, "Design of multibit noise-shaping data converters," *Analog Int. Circuits Signal Processing J.*, vol. **3**, pp. 259–272, May, 1993.
- [28] P. Kiss, "Adaptive digital compensation of analog circuit imperfections for cascaded delta-sigma analog-to-digital converters," Ph.D. dissertation, Elect. Eng. and Comput. Sci., Oregon State University, Corvallis, 1999.
- [29] P. Kiss, J. Silva, A. Wiesbauer, *et al.*, "Adaptive correction of analog errors in MASH ADCs – part II. Correction using test-signal injection," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. **47**, no. 7, pp. 629–638, July, 2000.
- [30] Y. B. N. Kumar, S. Talay, and F. Maloberti, "Complex cascaded bandpass $\Sigma\Delta$ ADC design," *IEEE Int. Symp. Circuits and Syst.*, Taipei, Taiwan, 2009, pp. 3118–3121.
- [31] B. Le, T. W. Rondeau, J. H. Reed, and C. W. Bostian, "Analog-to-digital converters," *IEEE Signal Processing Mag.*, vol. **22**, no. 6, pp. 69–77, Nov. 2005.
- [32] B. Li, "Design of multi-bit sigma delta modulators for digital wireless communications," Ph.D. dissertation, Dept. Microelectronics & Inf. Technology, Royal Institute of Technology, Stockholm, Sweden, 2003.
- [33] P. Malmovati, S. Brigati, F. Francesconi, *et al.*, "Behavioral modeling of switched-capacitor sigma-delta modulators," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. **50**, no. 3, pp. 352–364, March 2003.
- [34] P.-I. Mak, S.-P. U, and R. P. Martins, "Transceiver architecture selection: review, state-of-the-art survey and case study," *IEEE Circuits Syst. Mag.*, vol. **7**, no. 2, pp. 6–25, 2nd quarter 2007.

- [35] J. Marttila, "Quadrature sigma-delta ADCs: modeling and signal processing," M.S. thesis, Faculty Computing and Elect. Eng., Tampere University of Technology, Tampere, Finland, 2010.
- [36] J. Marttila, M. Allén, and M. Valkama, "I/Q imbalance effects in quadrature $\Sigma\Delta$ modulators – analysis and signal processing," in *IEEE Int. Microwave Workshop Series RF Front-ends for Software Defined and Cognitive Radio Solutions*, Aveiro, Portugal, 2010, pp. 1–4.
- [37] J. Marttila, M. Allén, and M. Valkama, "Quadrature $\Sigma\Delta$ modulation for cognitive radio – I/Q imbalance analysis and complex multiband principle," *Circuits, Syst. and Signal Process.: Special Issue Embedded Signal Process. Circuits and Syst. Cognitive Radio-based Wireless Comm. Devices*, (in press, available online: www.springerlink.com/content/c6813177233w1370/).
- [38] S. Mirabbasi and K. Martin, "Classical and modern receiver architectures," *IEEE Commun. Mag.*, vol. **38**, no. 11, pp. 132–139, Nov. 2000.
- [39] K.-P. Pun, C.-S. Choy, C.-F. Chan, and J. E. da Franca, "An I/Q mismatch-free switched-capacitor complex sigma-delta modulator," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. **51**, no. 5, May 2004.
- [40] B. Razavi, "Design considerations on for direct-conversion receiver," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. **44**, no. 6, pp. 428–435, June 1997.
- [41] B. Razavi, *RF Microelectronics*. Upper Saddle River, NJ: Prentice Hall, 1998.
- [42] K. Reddy and S. Pavan, "Fundamental limitations of continuous-time delta-sigma modulators due to clock jitter," *IEEE Trans. Circuits Syst. I, Reg. Papers*, pp. 2184–2194, vol. **54**, no. 10, Oct. 2007.
- [43] S. Reekmans, P. Rombouts, and L. Weyten, "Mismatch insensitive double-sampling quadrature bandpass $\Sigma\Delta$ modulation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. **54**, no. 12, Dec. 2007.
- [44] A. Rusu, D. Rodríguez de Llera González, and M. Ismail, "Reconfigurable ADCs enable smart radios for 4G wireless connectivity," *IEEE Circuits Devices Mag.*, vol. **22**, no. 3, pp. 6–11, May–June 2006.
- [45] A. Rusu, D. Rodríguez de Llera González, M. Ismail, and H. Tenhunen, "The design of a low-distortion sigma-delta ADC for WLAN standards," *Int. Symp. Signals, Circuits and Syst.*, Lasi, Romania, 2005, pp. 151–154.
- [46] A. Rusu, B. Dong, and M. Ismail, "Putting the 'flex' in flexible mobile wireless radios," *IEEE Circuits Devices Mag.* vol. **22**, no. 6, pp. 24–30, Nov.–Dec. 2006.
- [47] T. Salo, "Bandpass delta-sigma modulators for radio receivers," Ph.D. dissertation, Dept. Electrical and Communications Eng., Helsinki University of Technology, Helsinki, Finland, 2003.
- [48] A. Silva, J. Guilherme, and N. Horta, "Reconfigurable multi-mode sigma-delta modulator for 4G mobile terminals," *Integration, The VLSI J.*, vol. **42**, no. 1, pp. 34–46, Jan. 2009.
- [49] J. Silva, "High performance delta-sigma analog-to-digital converters," Ph.D. dissertation, *Elect. Eng. and Comput. Sci.*, Oregon State University, Corvallis, 2005.
- [50] R. Schreier, "An empirical study of high-order single-bit delta-sigma modulators," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. **40**, no. 8, pp. 461–466, Aug. 1993.
- [51] R. Schreier and M. Snelgrove, "Bandpass sigma-delta modulation," *Electronics Lett.*, vol. **25**, no. 23, pp. 1560–1561, Nov. 1989.
- [52] R. Schreier and G. C. Temes, *Understanding Delta-Sigma Data Converters*. Hoboken, NJ: John Wiley & Sons, 2005.
- [53] C. Svensson, "The blocker challenge when implementing software defined radio receiver RF frontends," *Analog Integrated Circuits and Signal Process.*, vol. **64**, no. 2, pp. 81–89, Dec. 2009.

- [54] A. Swaminathan, "A single-IF receiver architecture using a complex sigma-delta modulator," M.S. thesis, Dept. Electronics, Carleton University, Ottawa, Canada, 1997.
- [55] Y. Tang, K.-W. Cheng, S. Gupta, J. Paramesh, and D. J. Allstot, "Cascaded complex ADCs with adaptive digital calibration for I/Q mismatch," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. **55**, no. 3, pp. 817–827, Apr. 2008.
- [56] H. Tao, L. Tóth, and J.M. Khoury, "Analysis of timing jitter in bandpass sigma-delta modulators," *IEEE Trans. Circuits and Syst. II, Analog and Digital Signal Process.*, vol. **46**, no. 8, pp. 991–1001, Aug. 1999.
- [57] K.-T. Tiew and Y. Chen, "DAC compensation for continuous-time delta-sigma modulators," *IEEE Int. Symp. Circuits and Syst.*, Kobe, Japan, 2005, pp. 3680–3683.
- [58] R. Tortosa, J. M. de la Rosa, F. V. Fernández, and A. Rodríguez-Vázquez, "Clock jitter error in multi-bit continuous-time sigma-delta modulators with non-return-to-zero feedback waveform," *Microelectronics J.*, vol. **39**, no. 1, pp. 137–151, Jan. 2008.
- [59] M. Valkama, "Advanced I/Q signal processing for wideband receivers: models and algorithms," Ph.D. dissertation, Dept. Information Technology, Tampere University of Technology, Tampere, Finland. 2001.
- [60] M. Valkama, J. Pirskanen, and M. Renfors, "Signal processing challenges for applying software radio principles in future wireless terminals: an overview," *Int. J. Commun. Syst.*, vol. **15**, no. 8, pp. 741–769, Sep. 2002.
- [61] R. J. van de Plassche, "A sigma-delta modulator as an A/D converter," *IEEE Trans. Circuits Syst.*, vol. **25**, no. 7, pp. 510–514, July 1978.
- [62] N. Vun and A. B. Premkumar, "ADC systems for SDR digital front-end," *Proc. 9th Int. Symp. Consumer Electronics*, Macau, Hong Kong, 2005, pp. 359–363.
- [63] R.H. Walden, "Analog-to-Digital Converter Survey and Analysis," *IEEE J. Sel. Areas Commun.*, vol. **17**, no. 4, pp. 539–550, Apr. 1999.
- [64] X. Wang, "A fully digital technique for the estimation and correction of the DAC error in multi-bit delta-sigma ADCs," Ph.D. dissertation, Elect. Eng. and Comput. Sci., Oregon State University, Corvallis, OR, 2003.
- [65] R. Winoto, "Downconverting sigma-delta A/D converter for a reconfigurable RF receiver," Ph.D. dissertation, Elect. Eng. and Comput. Sciences, University of California, Berkeley, 2009.
- [66] J. Yang, R. W. Brodersen, and D. Tse, "Addressing the dynamic range problem in cognitive radios," *IEEE Int. Conf. Commun.*, Glasgow, Scotland, 2007, pp. 5183–5188.
- [67] L. Yu and W. M. Snelgrove, "A novel adaptive mismatch cancellation system for quadrature IF radio receivers," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. **46**, no. 6, pp. 789–801, Jun. 1999.

15 Digital suppression of A/D interface nonlinearities

Mikko Valkama, Markus Allén, and Jaakko Marttila

The analog-to-digital converter (ADC) is one of the key components in modern radio front-end design. Real-world ADC components always have certain trade-offs in performance which have to be taken into account. One reason for that is the rather slow development of ADCs compared to other technical achievements in radio technology [22], [33]. The fundamental trade-offs are illustrated in Figure 15.1. Power dissipation is a very important aspect especially in mobile devices, but low-power ADCs tends to have low resolution and sampling rate. On the other hand, if high resolution is required, it usually means that the sampling rate of that high-precision ADC is not very high.

Very high requirements for both the sampling rate and resolution are set by software defined radios, where most of the selectivity and other functionalities are implemented with digital signal processing [21], [32]. In other words, ADCs have to digitize high-bandwidth signals with large dynamic range. Figure 15.2 illustrates why high resolution is needed to cope with high signal dynamics. Before analog-to-digital conversion the signal has to be scaled properly to avoid overshooting the voltage range of the ADC [4]. If the overall waveform to be digitized consists of several signal bands with different power levels, strong signals result in fewer quantization levels that can be used for weak signals. This means that the weak signals suffer from quantization noise more than they would if there were no strong signals present at the same time.

Quantization is a nonlinear operation and thus even an ideal ADC always distorts the signal. However, real-world ADCs are not ideal and they have additional sources of nonlinearities due to their practical implementation. The nonlinearities appear as harmonic and intermodulation distortion, which can be severe especially for wideband signals. Figure 15.3 shows a principal spectral illustration of how a strong blocker signal can mask a significantly weaker signal due to ADC nonlinearities. This chapter is devoted to ADC nonlinearities and their digital suppression. The rest of the chapter is organized as follows. Some significant sources of ADC nonlinearities are discussed in Section 15.1. Digital signal processing enables development of new ways to suppress these nonlinearities. Basic approaches such as look-up tables, dithering, and model inversion are covered in Sections 15.2, 15.3, and 15.4, respectively. After that,

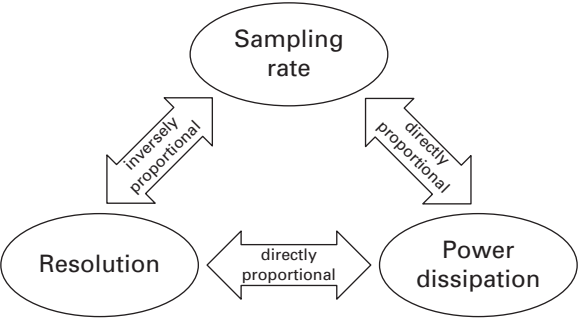


Figure 15.1 Illustration of trade-offs between three fundamental performance parameters in analog-to-digital converters.

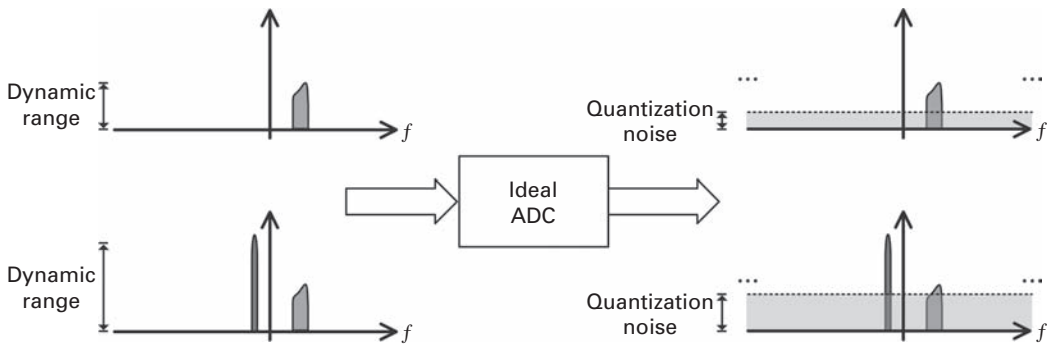


Figure 15.2 Spectrum illustrations of the upper part show how an ideal ADC induces quantization noise to the signal due to finite resolution. In the lower part, the dynamic range is increased due to a strong blocking signal and hence the weaker signal is more affected by the quantization noise when the same ideal ADC is used.

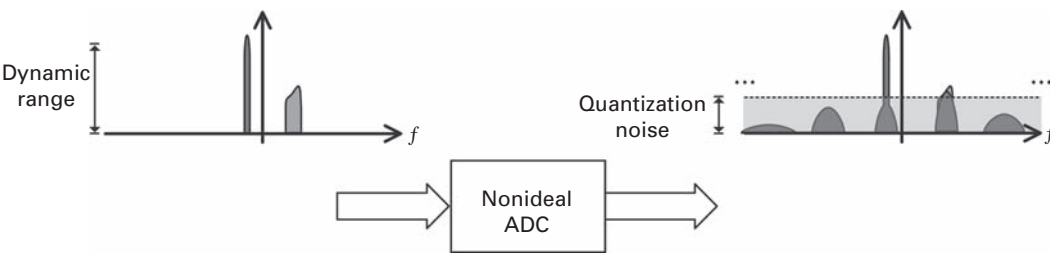


Figure 15.3 Spectral illustrations from the input and output of a nonideal ADC. The weak signal band is heavily affected by the nonlinear distortion stemming from the strong blocker signal.

Sections 15.5 and 15.6 propose adaptive interference cancellation and interpolation techniques in detail. Furthermore, some performance examples are given in Section 15.7. In the final section, we will make some conclusions and there will be further discussion. In general, this chapter partly follows the presentation given in [5].

15.1 Introduction to ADC nonlinearities

Real-life ADCs have various unwanted nonlinearities. Gain and offset errors are significant but their effects are rather trivial to mitigate. Other salient nonlinearities are differential nonlinearity (DNL) and integral nonlinearity (INL) [9], [25], [26]. When the nonlinearities are described in ADC specifications, the definitions of the performance metrics vary slightly from one manufacturer to another. Therefore, checking the definitions of the specific data sheet is of utmost importance. In addition, the clipping phenomenon is discussed in this section due to its importance in modern wireless communications.

15.1.1 Basics

One basic error in analog-to-digital converters is an offset error, which is defined as a constant shift from the zero input. An example of this is illustrated in Figure 15.4(a) in the case of a unipolar converter. Another fundamental error in ADCs is gain error, which is illustrated in Figure 15.4(b). When fitting a straight line through the transfer function of the ADC, the gain error is the difference in the slope of the line compared to the ideal one.

In general, IEEE standard 1241–2000 for analog-to-digital converters [19] defines the transfer function of an ADC with the equation

$$G \times T[k] + V_{os} + \varepsilon[k] = Q \times (k - 1) + T_1, \quad (15.1)$$

where G is the gain, $T[k]$ is the actual input value corresponding to the k th transition, V_{os} is the output offset, $\varepsilon[k]$ is the residual error of the k th transition, Q is the ideal code width and T_1 is the ideal value of $T[1]$. The IEEE standard [19] describes two ways of identifying offset and gain errors based on Equation (15.1). In the independently based method, G and V_{os} get values that minimize the mean squared value of $\varepsilon[k]$ over all k . Another method is called terminal based, which defines that the offset and the gain are the

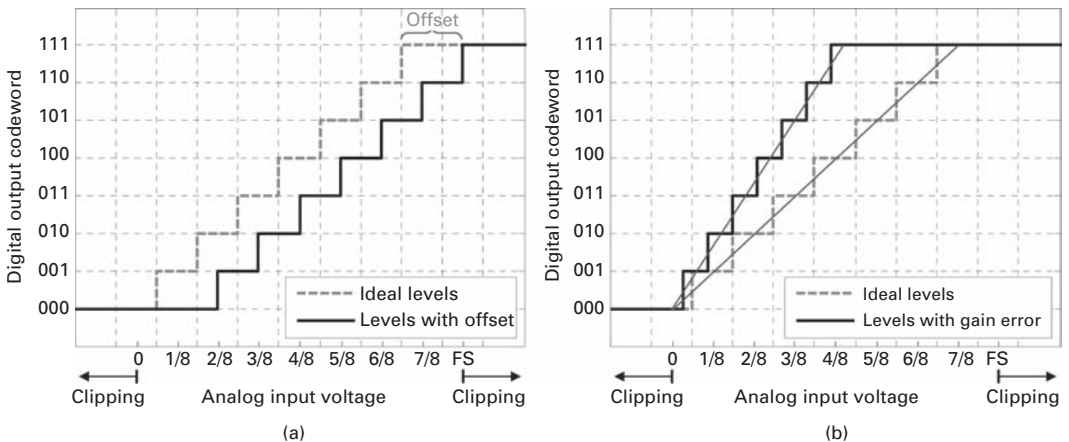


Figure 15.4 Example of (a) offset and (b) gain errors in a unipolar 3-bit ADC.

values which cause the deviations of the first and last codes to be zero. In other words, the offset and the gain are evaluated when $\varepsilon[1] = 0$ and $\varepsilon[2^b - 1] = 0$, where b is the number of bits in quantizer.

The INL and DNL errors can be calculated only after the correction of offset and gain errors. Differential nonlinearity is the relative difference between the actual and ideal code widths compared to the ideal code width. It is defined as

$$DNL[k] = \frac{W[k] - Q}{Q} \quad (15.2)$$

in which $W[k]$ is the k th actual code width [19]. In a similar manner, INL is the difference between the actual and the ideal code transition threshold. Another way of interpreting the INL is thinking of it as a cumulative sum of the DNL. According to [19], INL can be defined as

$$INL[k] = 100\% \times \frac{T[k] - T_{ideal}[k]}{2^b \times Q}, \quad (15.3)$$

where $T_{ideal}[k]$ is the ideal value for the k th transition. It should be taken into account that Equation (15.2) gives DNL as a multiple of least significant bit (LSB), but INL in (15.3) is given in relation to the full scale. In some cases, a more convenient equation is

$$INL[k] = \frac{T[k] - T_{ideal}[k]}{Q}, \quad (15.4)$$

which gives INL as a multiple of LSB. For instance, ADC data sheets typically use the LSB unit. In addition to that, usually when DNL or INL is expressed as a single value it is the maximum of (15.2) or (15.4) over all k .

An example of DNL and INL errors is presented in Figure 15.5 so that a more concrete conception can be attained. The dashed gray line indicates the ideal transfer function and

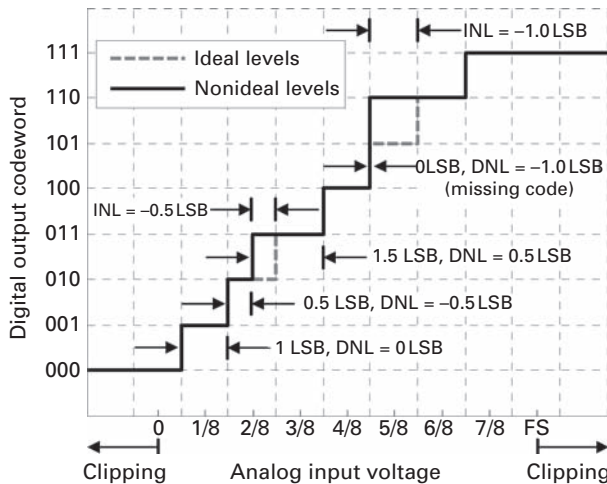


Figure 15.5 Example of DNL and INL errors in a unipolar 3-bit ADC [2] © 2009 IEEE.

the black line represents an example of nonideal transfer function for a 3-bit unipolar ADC. The corresponding DNL and INL errors have been marked in Figure 15.5. There is an example of missing code, which means a situation where the DNL error is less than or equal to -1 LSB and therefore two quantization levels are fully overlapped. In the case of Figure 15.5, the codeword *101* doesn't correspond to any input voltage. In some situations the errors can make the transfer function non-monotonic (not illustrated in Figure 15.5). It means that there are two separate ranges of input voltages that correspond to the same digital output codeword. The non-monotonicity is not possible in every ADC architecture, but it can happen, for instance, in pipelined ADCs [9].

A practical way to visualize INL error is to present it as a function of digital output codes. From a curve like that it is easy to see what kind of correlation the INL error has between different quantization levels. It is a common practice to include a typical INL error curve in the data sheet of the ADC. Figure 15.6 presents an example of an INL error curve for a commercial 10-bit ADC. Although the structure of the INL error curve may vary considerably between different types of converters, it is never purely random. For example in Figure 15.6, the correlation between consecutive output codes is clearly visible.

For more detailed analysis, N. Björnsell and P. Händel [12] have modeled the INL error as consisting of three parts. The model is

$$INL(T[k]) = INL_{HCF}(T[k]) + INL_{LCF}(T[k]) + INL_{noise}(T[k]), \quad (15.5)$$

where $T[k]$ is the k th code transition level. The low code frequency (LCF) component $INL_{LCF}(T[k])$ is the slowly varying fluctuation that can be seen in Figure 15.6. Correspondingly the high code frequency (HCF) component $INL_{HCF}(T[k])$ describes the rapid variations (architecture dependent) on top of the LCF component. These are

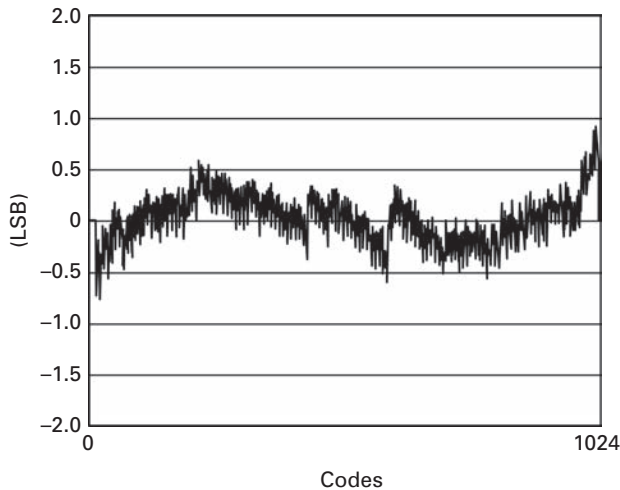


Figure 15.6 Typical measured INL error curve as a function of digital output code for the 10-bit analog-to-digital converter AD9218. Used with permission from [6].

usually modeled as piecewise linear. The rest of the INL error is random variations and this is modeled by a noise component called $INL_{noise}(T[k])$.

15.1.2 Clipping phenomenon

In general, clipping stands for limiting signal amplitude to a certain maximum value. This can be either intentional or unintentional depending on the application. There are basically three different types of clipping: a limiter, hard clipping, and soft clipping. An ideal limiter essentially outputs only the sign of the input signal. For the input signal $v_{in}(t)$ the ideal limiter is defined as

$$v_{lim}(t) = V_0 \text{sgn}\{v_{in}(t)\} = \begin{cases} +V_0, & \forall t : v_{in}(t) > 0 \\ -V_0, & \forall t : v_{in}(t) < 0, \end{cases} \quad (15.6)$$

where $v_{lim}(t)$ is the clipped output signal and V_0 is the clipping level [14]. It is also intuitive to present the transfer characteristics of the limiter in a graphical form as has been done in Figure 15.7(a). Hard clipping differs from the limiter so that it also has a linear region. Therefore symmetric hard clipping is formulated to be

$$v_{cl}(t) = \begin{cases} v_{in}(t), & \forall t : |v_{in}(t)| < V_0 \\ +V_0, & \forall t : v_{in}(t) \geq V_0 \\ -V_0, & \forall t : v_{in}(t) \leq -V_0, \end{cases} \quad (15.7)$$

where $v_{cl}(t)$ represents the clipped signal. Figure 15.7(b) presents these input–output characteristics for symmetric hard clipping. Furthermore, soft clipping is like hard clipping but it has smoother behavior near the saturation region as illustrated in Figure 15.7(c) [31]. This can be modeled, for instance, by the arcus tangent function but naturally this is highly application dependent. It is worth noticing that various names are used in the literature for different clipping types. There is a possibility of confusion, e.g., when the limiter is called hard clipping or hard clipping is called soft clipping.

When an analog signal with amplitude higher than the full-scale range is fed to the input of an ADC, the converter saturates and the signal is clipped. It is a rather realistic assumption that the behavior of the ADC can be modeled by symmetric hard clipping. The clipping is an instantaneous phenomenon which causes nonlinear distortion. In case of zero-symmetric hard clipping, only odd-order distortion exists. Symmetry is a

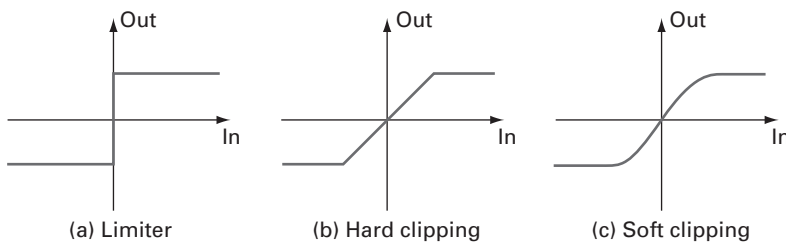


Figure 15.7 Transfer functions for (a) a limiter, (b) hard clipping, and (c) soft clipping.

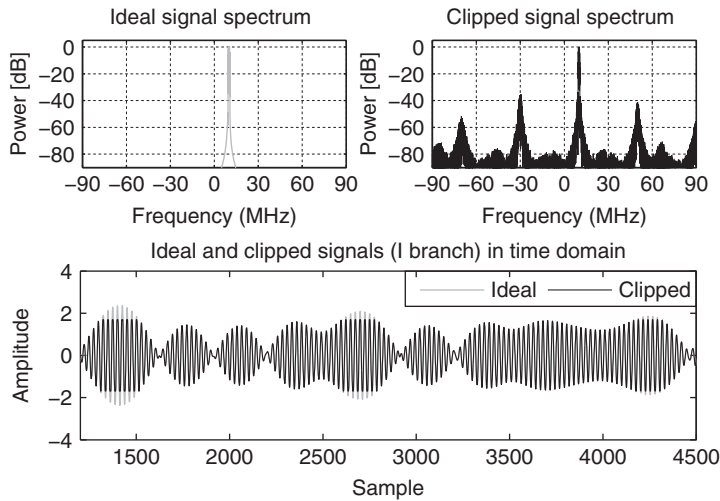


Figure 15.8 Upper part: frequency domain illustrations for a complex bandpass QPSK signal with center frequency of 10 MHz and the symmetrically clipped version of the same signal. Lower part: piece of the I branch signal presented in time domain before and after clipping.

fair assumption at least when the DC offset has been corrected. It is important to avoid or mitigate the clipping because even light clipping in the time domain can have severe effects on the frequency content.

The instantaneous nature of clipping can be clarified through an example. Figure 15.8 shows the spectrum for both unclipped and clipped band-pass quadrature phase shift keying (QPSK) signal with 10 MHz center frequency and oversampling factor of 256. A sampling frequency of 256 MHz is used. In this simulation example, the signal is ideally clipped without quantization or any other additional distortion to better emphasize the effect of clipping. The clipping level is identical for in-phase (I) and quadrature (Q) branches. From the power spectrum it is easy to see the odd-order nonlinear distortion due to the clipping around 10 MHz, -30 MHz, 50 MHz etc. The lower part of Figure 15.8 shows a piece of the I branch signal in the time domain. At first glance the distortion in the time domain may look small, but it really has significant effect on frequency content as can be seen from the spectrum figure.

15.2 Look-up table

Look-up table (LUT) methods are based on precalculated estimation values which are stored in a table. The output samples of an ADC are used for indexing to point out corresponding values from a look-up table. The LUT carries values which are meant either for correction or replacement [23]. In the correction scheme, the LUT value is added to the output sample of the ADC. The replacement scheme refers to replacing the output sample of the ADC with the corresponding value from the LUT.

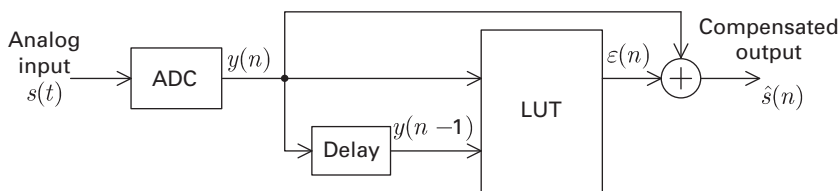


Figure 15.9 Look-up table with state-space indexing.

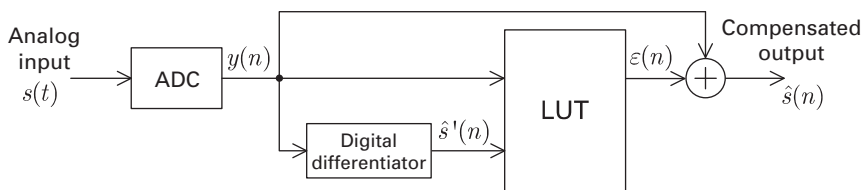


Figure 15.10 Look-up table with phase-plane indexing.

The indexing is one of the most important parts of the LUT method and it can be done in several different manners. The simplest one is static correction where the same output value of the ADC is every time mapped into the same index [16]. Here, two more advanced methods are discussed, which are called state-space indexing and phase-plane indexing [29]. Both take the memory of the nonlinearity into account. The state-space indexing exploits both the current output sample $y(n)$ and the previous sample $y(n-1)$ for creating the index. Therefore, the LUT can be considered as two-dimensional. This principle is illustrated in Figure 15.9. The continuous-time input signal $s(t)$ is converted with the nonideal ADC, and then output samples $y(n)$ and $y(n-1)$ are used to retrieve the corresponding value $\varepsilon(n)$ from the LUT. This value $\varepsilon(n)$ is then used to create the compensated output sample $\hat{s}(n)$. The state-space indexing approach can be extended to use more than just one previous sample. Naturally, this increases the effectiveness of the compensation but, on the other hand, the number of dimensions in the LUT also increases. Consequently, the LUT then requires significantly more memory. One solution for this problem is to apply more quantization for delayed samples, i.e. to use less bits, so that the size of the index space can be reduced [23].

Another dynamic indexing method is phase-plane indexing [23], [29]. It is based on the present output sample $y(n)$ and the estimate of the slope (or derivative) of the input signal $\hat{s}'(n)$. The principle is visualized in Figure 15.10. As in the previous method, the continuous-time input signal $s(t)$ is first converted with the ADC. Then a digital filter, for example a finite impulse response (FIR) differentiator, is used to produce the slope estimate $\hat{s}'(n)$. The corresponding LUT value $\varepsilon(n)$ is, once again, added to the output sample $y(n)$ to create a compensated output sample $\hat{s}(n)$. Instead of using the digital filter, the slope can also be acquired by measuring the analog signal.

Apart from the indexing, a further challenge is to properly calibrate the correction (or replacement) values in the LUT. Basically, a calibration signal is fed to the ADC under test and the output is compared with the digital version of the calibration signal. Unfortunately, the process is never ideal because only an estimate of the calibration signal can be used. The LUT calibration can be performed, for example, by using an additional very accurate ADC to acquire the digital version of the calibration signal fed to the ADC under test. Another option is to generate a digital calibration signal and use a DAC to feed the signal for the ADC under test [23]. More advanced calibration methods proposed in the literature include additional signal processing, e.g., optimal filtering schemes [24] or blind calibration based on probability density functions [17] in order to estimate the calibration signal.

15.3 Dithering

The essential idea of dithering is to intentionally add noise to the ADC input signal before quantization and thus to reduce distortion induced by the quantization. The dithering reduces correlation between the input signal and the quantization error. For this reason, a pseudo quantization noise model can be made applicable. This model is widely used because of its simplicity. It models the quantization error as an additive, uniformly distributed, white and zero-mean noise which is independent of the input signal [23]. The dithering can also reduce distortion caused by quantizer nonidealities by randomizing the DNL error pattern of a non-ideal uniform quantizer [10]. Furthermore, dithering is able to increase the ADC resolution for slowly varying signals [11]. It is typical that a quantized slowly varying signal has several consecutive output samples with the same value and thus quantization error cannot be reduced by averaging, i.e., low-pass filtering. By adding a proper dither signal prior to the quantization, the output values are not all the same. Therefore, in this case the averaging may yield better results and increase resolution.

There are two types of dithering methods, which are called subtractive and non-subtractive dithering [23]. Both principles are illustrated in Figure 15.11. In subtractive dithering, the dither noise is added to the input signal before the quantizer and the same noise is subtracted from the output after the quantization in the digital domain. However,

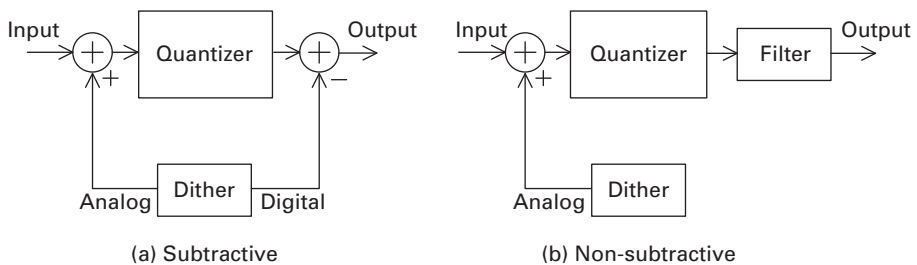


Figure 15.11 Basic structures of (a) subtractive and (b) non-subtractive dithering principles.

it is worth noticing that the digital version of the dither noise is not always known explicitly and thus the subtraction is not ideal. On the contrary, the non-subtractive dithering method adds the dither noise prior to the quantizer but is not even trying to remove all the noise after quantization. The method is based on the fact that the out-of-band noise can be filtered out in the digital domain. Hence, only a part of the noise still exists in the final output signal. However, the non-subtractive method can never be expected to perform as well as ideal subtractive dithering.

15.4 Model inversion

The fundamental principle of model inversion is to identify a mathematical system model which approximates the input–output relationship of the ADC. Then the inverse system, or an approximation of it, is calculated to mitigate nonidealities from the output signal [23]. These two stages are depicted in Figure 15.12. The model identification is performed by feeding a continuous-time input signal $s(t)$ to the ADC under test and to the ideal ADC. Then the test ADC output $y(n)$ and the model output $\hat{y}(n)$ are compared to each other in order to tune the model H to match better with the test ADC. As presented in Figure 15.12(b), the post-distortion stage compensates ADC non-idealities by using the inverse of the model H , denoted by H^{-1} . There are several alternatives for the system modeling such as Chebyshev polynomials [1], but most of the proposed methods in the literature are based on the Volterra model [23], [29]. Instead of the two-stage process illustrated in Figure 15.12, it is also possible to identify the inverse model H^{-1} directly. In practice, this means tuning H^{-1} using the error signal $\hat{s} - s$ [23].

Volterra theory is a mathematical approach for describing causal nonlinear time-invariant systems with memory. The theory is usable for systems where the linear term is dominant compared to nonlinearities. The downside of the Volterra model is the computational complexity due to the large number of parameters. Therefore, simplified models such as Hammerstein, Wiener and Kautz–Volterra models have been introduced. Their number of parameters is lower but they do not have the generality of Volterra series [13].

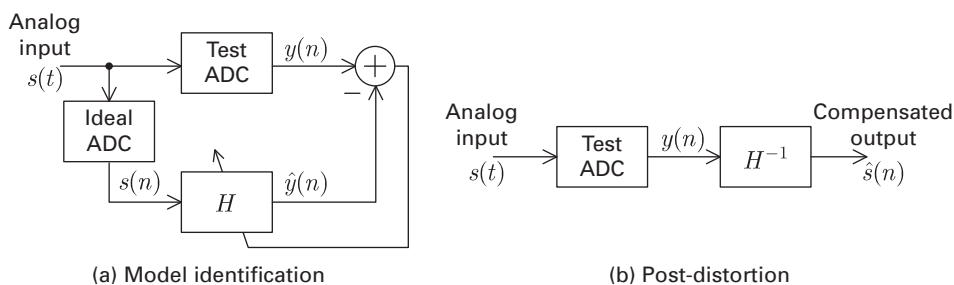


Figure 15.12 Stages of (a) model identification and (b) post-distortion for ADC correction in the model inversion method.

15.5 Adaptive interference cancellation

The basic idea of adaptive interference cancellation (AIC) is grounded on the study by M. Valkama *et al.* [30]. They considered AIC for compensating mixer and low-noise amplifier (LNA) nonlinearities in wideband multi-carrier radio receivers. This section describes how the AIC principle can be applied for the compensation of ADC nonlinearities. In addition, a modified approach called enhanced adaptive interference cancellation (E-AIC) is presented to overcome some limitations of the original design when using it for clipping compensation. The AIC and E-AIC approaches are originally published in [2], [3] and [4]. Some performance examples are given in Section 15.7.

15.5.1 Basic principle

An interference cancellation method discussed here relies on a presumption that ADC nonlinearities can be modeled by a memoryless nonlinearity and an ideal quantizer. The cascade of these two is presented in Figure 15.13 for input signal s and output signal y . One very general approach is to model the signal after the nonlinearity by a polynomial, which can be defined as

$$\tilde{s} = g(s) = c_0 + c_1s + c_2s^2 + c_3s^3 + \dots, \quad (15.8)$$

where the c 's are weights for the different powers of s . This kind of model is useful particularly from the interference cancellation point of view as is shown later.

The AIC principle is used with a wideband ADC that converts a wide frequency band consisting of several signals with different center frequencies. The overall signal may have dynamic range of tens of dBs and therefore nonlinearities in the ADC can have severe effects when the distortion from strong blocker signals fall at a weak signal band. The AIC system tries to reduce the interference on that weak signal.

The block diagram of the AIC principle is shown in Figure 15.14 accompanied by very simplified stage-by-stage spectrum examples labeled with capital letters from A to F . After the ADC, the input signal (spectrum A) is nonlinearly distorted which is illustrated in spectrum B in Figure 15.14. For the sake of simplicity, only a single intermodulation frequency component originating from two strong blocker signals is drawn within the band of the weak signal of interest. The band-splitting stage uses digital filters to divide the distorted signal into two branches. The lower branch should contain only the interfering signals as is shown in spectrum figure C . From now on, the lower branch signal

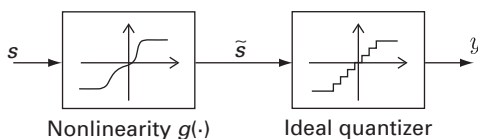


Figure 15.13 Model for ADC nonlinearities consisting of a memoryless nonlinearity and an ideal quantizer [2] © 2009 IEEE.

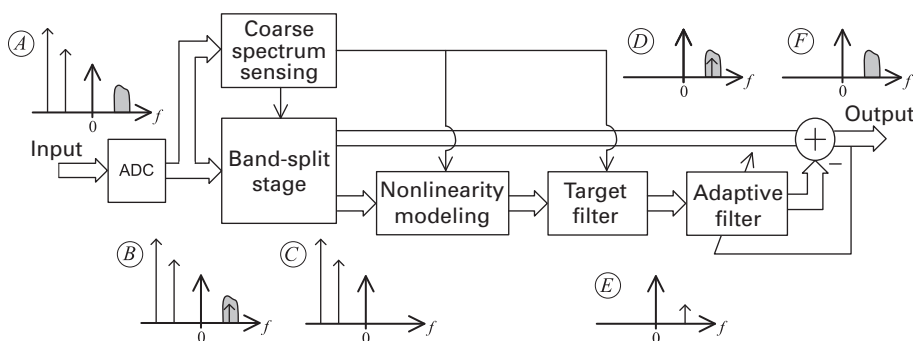


Figure 15.14 Adaptive interference cancellation principle for reducing nonlinear distortion with digital post-processing [2] © 2009 IEEE.

is called a reference signal. Correspondingly, the upper branch should contain only the distorted version of the desired weak signal band. This is illustrated in spectrum figure *D*.

After the band-split stage, the target is to regenerate the interference at the weak signal band and in the end subtract it from the actual weak signal band (the upper branch signal shown in spectrum *D*). This is done by modeling the nonlinearity using, e.g., polynomial signal processing for the reference signal containing the interfering signal components. The equation for the polynomial or other nonlinear function highly depends on the application. The most important things to consider here are the form of the interference and from which frequencies the interference is originating. These affect the powers used in the polynomial model. For instance, if third-order distortion needs to be removed, then the reference signal is raised to the power of three. Correspondingly, whatever more complex signal alteration can be used to recreate the distortion and achieve more accurate results. The target filter mentioned in Figure 15.14 is used after the nonlinearity modeling to extract only the distortion at the weak signal band. Spectrum figure *E* illustrates the reference signal after the target filter and only the regenerated intermodulation component is seen there.

Before the adaptive filter depicted in Figure 15.14, there is the weak signal with interference in the upper branch and only the interference in the reference signal of the lower branch. Now, the coefficients of the adaptive filter are tuned in such a manner that the interference is subtracted from the weak signal band as perfectly as possible. The most apparent solution for coefficient optimization is to use the least mean square (LMS) algorithm for minimizing the interference power at the weak signal band. Due to the simple implementation of LMS, it is an adequate choice for real-time systems. When the LMS algorithm has converged and the interference has been subtracted, there is ideally only the desired weak signal without interference in the output. This is illustrated with spectrum figure *F* in Figure 15.14. The suitable number of adaptive filter coefficients depends on the application. If the distortion signal is generated accurately enough, it is usually adequate to use a single-tap adaptive filter in ADC applications. That is because the proper scaling of the reference signal is sufficient. Using more coefficients can help shape the reference signal to match better with the distortion on the weak signal band.

Another example of using several coefficients is to compensate for memory effects in power amplifiers [30].

In practice, either parallel or serial processing can be used in the reference generation and adaptive filtering stages. Parallel processing requires more computational resources since it uses several processing branches at the same time. Each branch processes different order distortion and has its own adaptive filter coefficients. In the serial version, only a single processing branch is used but there can be several processing cycles to cancel the different distortion orders. This may decrease the need for computational resources but increases the processing delay. Generally, the performance difference between the parallel and serial implementations is dependent on the computer architecture used.

An outer-loop control mechanism called coarse spectrum sensing is used for the overall system management. It measures the spectrum density of the digital signal from the ADC. In practice, this can be performed, e.g., with the fast Fourier transform (FFT). The information is then used to control band-split filter properties so that the weakest signal band or bands are selected for the interference cancellation. Furthermore, the spectrum sensing is needed to locate the strongest blocker signals. It can be then used to select distortion orders used in the polynomial signal processing stage. Naturally, the target filter properties have to be controlled also according to the location of the weak signal band(s).

15.5.2 Implementation example

One possible option for the exact implementation of the proposed AIC method for clipping mitigation is illustrated in Figure 15.15. Zero-symmetric signal clipping in the ADC essentially produces odd-order nonlinear distortion and hence only odd powers in

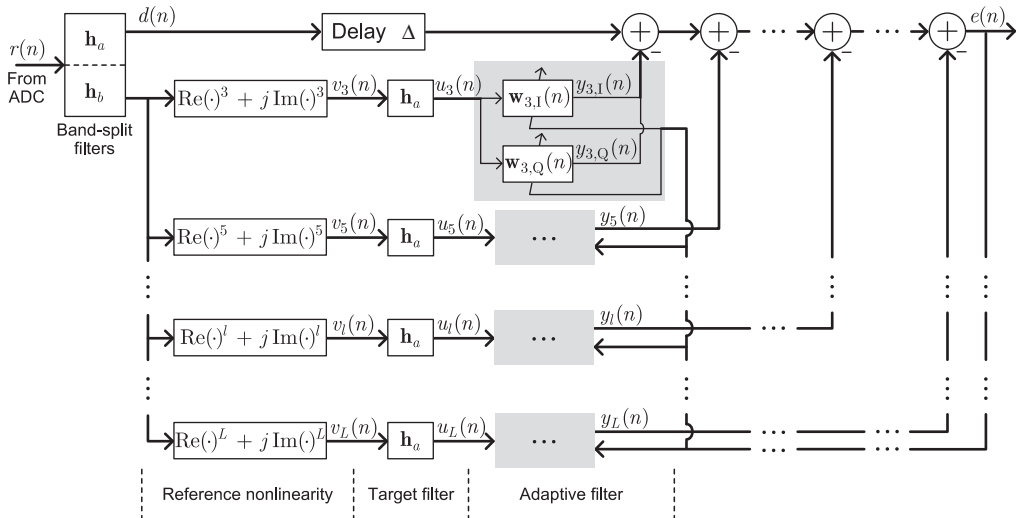


Figure 15.15 Mathematical notation for illustrating one possible implementation of the AIC method for clipping compensation. The digital input signal $r(n)$ is coming from the ADC and the enhanced output signal after post-processing is denoted with $e(n)$ [4].

Equation (15.8) are taken into account here. The digitized signal from the ADC is denoted with $r(n)$, where the index n refers to samples inside one packet. Generally, the mathematical notation here is that the real and the imaginary parts of the vector $\mathbf{x}$ are marked with $\mathbf{x}_I$ and $\mathbf{x}_Q$, respectively. The lengths of filters $\mathbf{h}_a$, $\mathbf{h}_b$ and $\mathbf{w}_l$ are M_a , M_b , and M_w , respectively. The coarse spectrum sensing block tunes the filter $\mathbf{h}_a$ so that the weak signal band is separated in the band-split filtering stage. After that, the signal containing the weak signal band is

$$d(n) = \mathbf{h}_a^T \mathbf{r}_a(n), \quad (15.9)$$

where $\mathbf{r}_a(n) = [r(n), r(n-1), \dots, r(n-M_a)]^T$. The signal containing all the spectral content outside the weak signal band is created using the filter $\mathbf{h}_b$. Then this signal is used to generate L separate reference signals for removing distortion of different orders. Therefore, the l th-order distortion signal is defined as

$$v_l(n) = \text{Re}[\mathbf{h}_b^T \mathbf{r}_b(n)]^l + j \text{Im}[\mathbf{h}_b^T \mathbf{r}_b(n)]^l, \quad (15.10)$$

where $\mathbf{r}_b(n) = [r(n), r(n-1), \dots, r(n-M_b)]^T$. Only the distortion at the weak signal band is needed and thus the distortion signal is filtered with the target filter $\mathbf{h}_a$. For the l th-order branch, the signal is then

$$u_l(n) = \mathbf{h}_a^T \mathbf{v}_l(n), \quad (15.11)$$

where $\mathbf{v}_l(n) = [v_l(n), v_l(n-1), \dots, v_l(n-M_a)]^T$. After that, the real and the imaginary parts of the distortion signal $u_l(n)$ are filtered with the adaptive filters $\mathbf{w}_{l,I}$ and $\mathbf{w}_{l,Q}$, respectively. This is performed to match the regenerated distortion to the distortion at the weak signal band. The overall complex output of the adaptive filter stage for the l th branch is described by

$$y_l(n) = \mathbf{w}_{l,I}^T(n) \mathbf{u}_{l,I}(n) + j \mathbf{w}_{l,Q}^T(n) \mathbf{u}_{l,Q}(n), \quad (15.12)$$

where $\mathbf{u}_{l,I}(n) = [u_{l,I}(n), u_{l,I}(n-1), \dots, u_{l,I}(n-M_w)]^T$ and $\mathbf{u}_{l,Q}(n) = [u_{l,Q}(n), u_{l,Q}(n-1), \dots, u_{l,Q}(n-M_w)]^T$.

The LMS algorithm for finding the filter coefficients of the adaptive filters is described next. Only the I branch of the l th-order distortion filter is considered here, because the algorithm for the Q branch and all the other distortion orders can be realized in a similar manner. In the initial phase, all the filter coefficients are set to zero, i.e.,

$$\mathbf{w}_{l,I}(0) = \mathbf{0}. \quad (15.13)$$

The overall output $e(n)$ is a common input for all the reference branches and it is used for adapting the filter coefficients. The I branch of the overall output signal is

$$e_I(n) = d_I(n - \Delta) - \sum_{i=2}^{\frac{L+1}{2}} y_{2i-1,I}(n), \quad (15.14)$$

where Δ is the delay required for the reference signal processing. In the LMS algorithm, the actual coefficient update for the I branch of the l th-order distortion filter is defined as

$$\mathbf{w}_{l,1}(n+1) = \mathbf{w}_{l,1}(n) + \mu_l \mathbf{u}_{l,1}(n) e_1(n), \quad (15.15)$$

where the LMS step-size parameter for the l th-order distortion branch is denoted by μ_l .

Adaptation speed of the adaptive filtering algorithm is important in radio applications. There is always a trade-off between adaptation speed, accuracy, and computational complexity. Therefore, the LMS algorithm may not always be the best choice. For example, recursive least squares (RLS) and normalized LMS provide considerably better adaptation speed [18], [27]. In addition, methods such as Gauss–Newton RLS [27] and fast approximate RLS [15] are proposed in the literature as well.

15.5.3 Enhanced adaptive interference cancellation

The adaptive interference cancellation principle relies on the assumption that the interference part of the signal (from where the reference signal is generated) is not distorted or contains only mild distortion. Unfortunately, this is not the case for a heavily clipped signal. Therefore the generated reference signal would contain additional interference that does not match with the interference to be removed from the weak signal band.

One powerful solution for bypassing the reference signal distortion problem in AIC is presented in Figure 15.16. This E-AIC principle exploits two analog-to-digital converters. The first one is the main ADC and the second one is used to create the reference signal. The latter converter can have lower resolution than the main ADC and therefore it should be rather inexpensive to include. The idea is to avoid clipping in the second ADC by using constant attenuation (with respect to the main ADC) when feeding the input signal. The appropriate attenuation can be, e.g., between 10 and 20 dB so that it is possible to digitize strong blocker signals intact.

The overall process goes as described in Section 15.5.1 and also in 15.5.2. Furthermore, the spectrum illustrations in Figure 15.16 are comparable with the ones in Figure 15.14. The only difference in E-AIC compared to the AIC method is that a properly attenuated version of the input signal is exploited in the reference signal

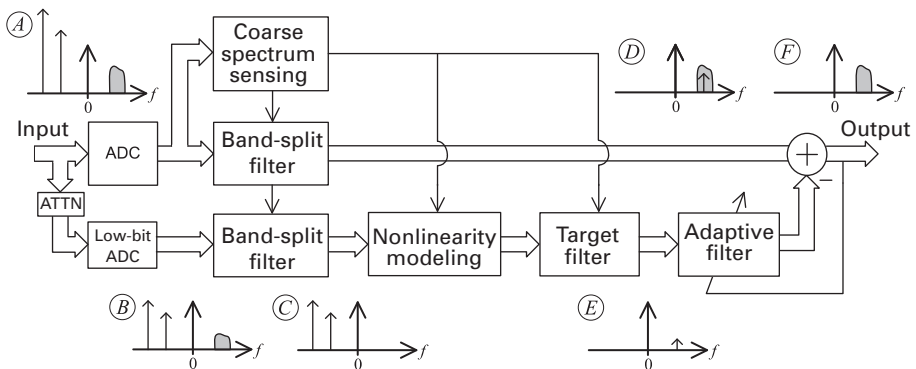


Figure 15.16 Adaptive interference cancellation principle using two ADCs. Attenuated version of the input signal is fed to the low-bit ADC to avoid clipping and thus non-distorted version of the blocker signal can be acquired for post-processing purposes [3] © 2010 IEEE.

generation branch and hence also this secondary ADC is required. From the reference signal generation point of view, the most important thing is that the strong blocker signals are digitized as perfectly as possible (clipping avoided) by the secondary ADC. That is important because the reference signal is generated based on the strong blockers, which causes most of the interference. The attenuation in the input of the secondary ADC also means that the weak signal band of interest is converted with considerably lower resolution. However, this is not a problem, because the weak signal band is also digitized by the main ADC using better resolution. In fact, the low-resolution version of the weak signal band is filtered out in the band-split filtering stage, because otherwise it would ruin the reference signal generation.

15.6 Clipping compensation using interpolation

Interpolation can be used for enhancing a clipped waveform because all the clipped samples can be thought of as lost samples and interpolation provides one way to estimate their original values. Especially if the clipping happens in a radio receiver, the information about clipping distortion is very limited. For example, the exact clipping level may not be known. Due to the lack of information, there are various interpolation approaches whose accuracy is more or less the same. The real differences between the approaches appear when the computational complexity is taken into account. This section presents the maximum selection interpolation (MSI) method, which was originally proposed in [3] and is partially based on the interpolation method proposed by T. Tomioka *et al.* [28]. The MSI method exploits multirate signal processing and polyphase structures of which the basics are covered, e.g., in [20].

Figure 15.17 illustrates a polyphase decomposition for a clipped waveform with oversampling factor of four. Samples in different polyphase branches are denoted with different shades and markers. The essential idea of the MSI is that the successive samples, i.e., different polyphase branches have redundant information. Therefore, a clipped sample in one branch can be recovered, or at least estimated, with a fractional delay filter using corresponding samples in other branches.

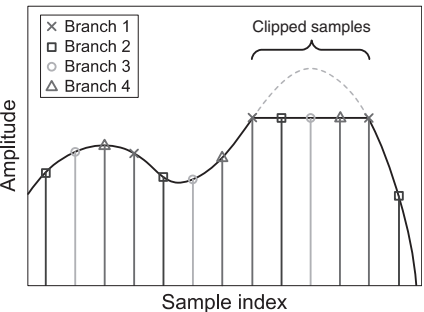


Figure 15.17 A clipped waveform with oversampling factor of four and its every polyphase branch illustrated with different symbol and shade [3] © 2010 IEEE.

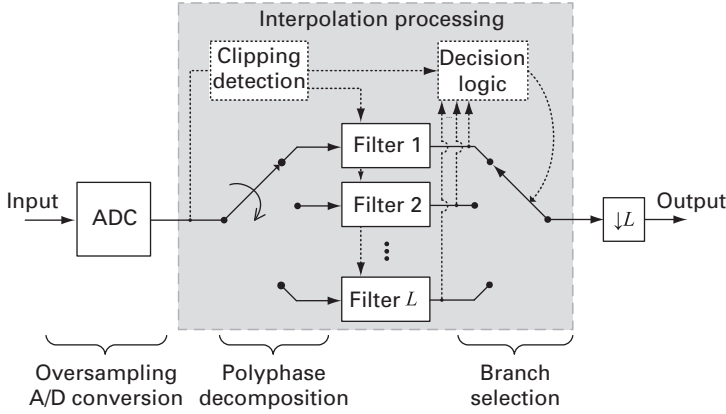


Figure 15.18 Block diagram of the MSI method for a signal with the oversampling factor L . After the oversampling ADC, a polyphase decomposition of the signal is created and then, once filtered, the decision logic chooses, sample by sample, the best branch for the output. The clipping detection control changes the order of the branch filters according to the position of the clipped sample [3] © 2010 IEEE.

The MSI method assumes that the signal is band-limited when fed to the input of an analog-to-digital converter. The ADC itself must oversample the signal and the oversampling factor is here denoted by L . After the analog-to-digital conversion, the oversampled signal is uniformly divided to L different branches to form the polyphase decomposition. This is illustrated in Figure 15.18 with a block diagram. Once the polyphase decomposition is formed, the signal is filtered using the branch filters to get estimates for the clipped sample. There are many suitable options for designing these branch filters and a simple but rather efficient solution is presented in the following.

One way to design the branch filters is creating a low-pass FIR filter with a passband corresponding to the bandwidth of the useful signal. So, this is essentially the same approach as is commonly used in designing a basic decimation filter [20]. The branch filters are created by simply making a polyphase decomposition of the impulse response of the FIR filter. If the overall impulse response is denoted with $h(n)$, then the branch filters are

$$\begin{aligned}
 h_0(n) &= \{h(0), h(L), h(2L), \dots\} \\
 h_1(n) &= \{h(1), h(L+1), h(2L+1), \dots\} \\
 &\vdots \\
 h_{L-1}(n) &= \{h(L-1), h(2L-1), h(3L-1), \dots\}.
 \end{aligned} \tag{15.16}$$

These branch filters are assigned to different branches according to the knowledge about the branch in which the clipped sample to be interpolated is located. For example, if the clipped sample is in the first (top) branch, which is illustrated in Figure 15.18, filters 1, 2, $\dots$, L are $h_0(n), h_1(n), \dots, h_{L-1}(n)$, respectively. Whereas a clipped sample in the second branch would mean the filters 1, 2, $\dots$, L are $h_{L-1}(n), h_0(n), h_1(n), \dots, h_{L-2}(n)$,

respectively. The filter assignment is illustrated in Figure 15.18 with the clipping detection box. At the time instances when there are no clipped samples to be interpolated, the filtering is not necessary and this way the amount of computation can be reduced. It is also worth noticing that the outcome of the filter $h_0(n)$ does not usually have to be calculated when there is a clipped sample, because it is almost always replaced with a better estimate from the other branches. The clipped sample would be the best estimate only if all the other branches provide estimates whose absolute values are lower than the absolute value of the clipped sample.

One of the most important things in the interpolation scheme is the logic for selecting the best estimate for the clipped sample among all the branches. The MSI method always chooses the branch which gives the biggest absolute value for the clipped sample. It requires calculating estimates with $L - 1$ different filters but only one of these estimates is eventually used. On the other hand, this method is likely to provide good estimates without a need for numerous iteration cycles.

In the proposed interpolation scheme, the sampling rate is kept unchanged until the very end. That is because it is possible to iterate the filtering process several times in order to achieve better interpolation results. On the first iteration round there are usually clipped samples involved in the estimation process. On the next round there is some kind of estimate for every clipped sample already and thus it is possible to calculate better estimates based on those initial estimates. In this way more and more information is extracted from the clipped signal. After a sufficient number of iterations have been carried out, the sampling rate of the interpolated signal can be reduced by factor L . This is also illustrated on the right side of Figure 15.18.

15.7 Performance examples of online post-processing methods

This section is devoted to giving some insight into the ADC nonlinearity compensation methods previously discussed. A wideband radio receiver case is considered here because it is used in cognitive radios, which are an attractive development trend for future wireless communications [22], [32]. The first example given compares performance of different AIC and interpolation methods in clipping compensation. All of these methods are designed for online processing, which is an essential requirement in radio applications. The second example provided discusses applying AIC to mitigate INL errors.

15.7.1 Clipping compensation

The test signal used in this example consists of five separate frequency bands. Each frequency band contains a QPSK-modulated single-carrier waveform. Every band has a different power level and bandwidth as shown in the upper part of Figure 15.19(a). The Lower part of the same figure shows the spectrum after the signal is clipped 6 dB over its average power level. This laboratory measurement was carried out using a commercially available 14-bit ADC [7]. It is noticeable that especially the third-order intermodulation

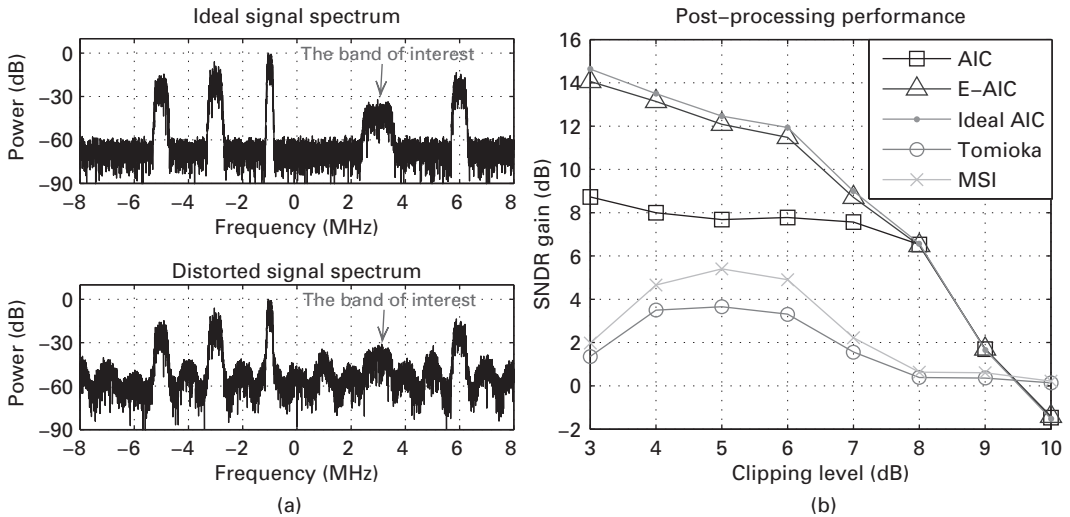


Figure 15.19 (a) Spectra of the digitized waveform in frequency domain before and after clipping. Especially the third-order intermodulation distortion originating from the strong signal at -1 MHz falls on top of the weak signal around $+3$ MHz. (b) SNDR gain comparison between different clipping compensation techniques at the weak signal band ($+3$ MHz). Clipping level is described as a number of dBs above the average power level of the signal [3] © 2010 IEEE.

distortion originating from the strong signal at -1 MHz falls on top of the weak signal around $+3$ MHz.

Figure 15.19(b) compares the performances of the previously presented post-processing methods using signal-to-noise-and-distortion ratio (SNDR) gain. The SNDR gain is here defined as a relation of SNDRs of the weak signal band around $+3$ MHz before and after applying the post-processing method. The interpolation methods, i.e., Tomioka [28] and MSI, use 32-tap polyphase branch filters and after four iteration rounds the presented results are achieved. The MSI method has lower complexity than Tomioka's due to the absence of preliminary polynomial interpolation stage. However, the MSI method performs better because of properly designed FIR filters instead of a truncated sinc-function based filter impulse response as in Tomioka's method. In addition, the decision logic for the branch selection in the MSI method is superior when using a low number of iterations. When comparing interpolation methods to AIC approaches it is clearly visible that the performance of the interpolation is rather limited. That is due to other ADC nonidealities (besides clipping) which reduce the estimation accuracy in the interpolation process.

The AIC and E-AIC approaches use third-, fifth-, and seventh-order nonlinearity modeling in order to obtain the results shown in Figure 15.19(b). These are more robust than interpolation methods because of their ability to remove any kind of nonlinear distortion (of the predefined orders) regardless of its source. The performance of the E-AIC method is very near the ideal AIC performance level when a 7-bit secondary ADC is used with 7-dB input attenuation. Due to laboratory equipment limitations, in this

example the secondary ADC is realized with the same 14-bit ADC as the main converter but the resolution is reduced afterwards by software. Here, the ideal AIC does not mean that all the distortion is perfectly removed but it just gives a performance limit for the AIC approach itself by using a perfect reference signal. In general, the performance of the AIC method is limited by inaccuracies in the nonlinearity modeling and in the adaptive filter coefficients. In addition, it is natural that the SNDR gain is lower in mild clipping cases (see, e.g., clipping levels 8–10 dB in Figure 15.19) because there is not much distortion to remove in the first place.

15.7.2 INL mitigation

The test signal used in this example is similar to the one described in the clipping compensation example, although some minor changes are made to emphasize INL error effects. Figure 15.20 shows spectra of the overall waveform before and after analog-to-digital conversion, which is performed with a 10-bit ADC AD9218 [6]. Because the actual hardware was not available, the conversion was here made with a software simulation using an accurate model for the ADC [8] provided by the manufacturer.

As can be seen in Figure 15.20, INL errors cause nonlinear distortion and especially the weak signal at +3 MHz suffers from it due to the strong blocking signal at –1 MHz. This mainly stems from the LCF part of the INL error. The HCF and noise components of the INL error are essentially seen as a risen noise floor in the spectrum. Most of the noise is due to the quantization and therefore it is inevitable from the quantization theory point of view.

The AIC method is employed in order to reduce nonlinear distortion of orders three, five, and seven. These orders are selected according to the behavior seen in Figure 15.20, which stems from the shape of the INL error curve (see Figure 15.6). It is important to be aware of the fact that other ADCs have different INL error shapes and therefore the dominating distortion orders vary being, e.g., second and fourth. Figure 15.21 illustrates the constellations for the weak QPSK signal demodulated from the signal band with the center frequency of +3 MHz before and after applying the AIC post-processing method.

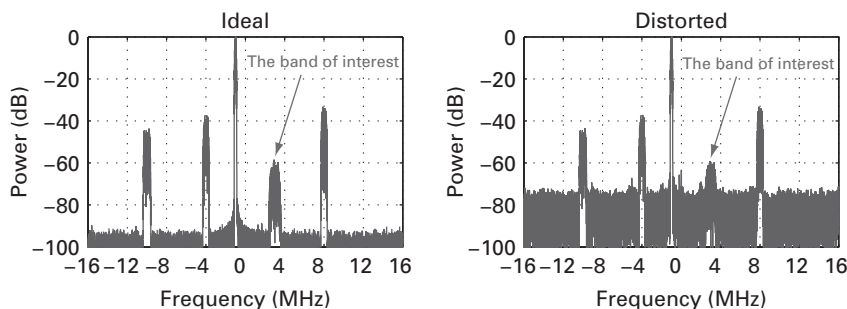


Figure 15.20 Power spectra of the overall waveform before and after analog-to-digital conversion with the 10-bit ADC (AD9218, [6]) illustrating the frequency domain effects of the INL errors.

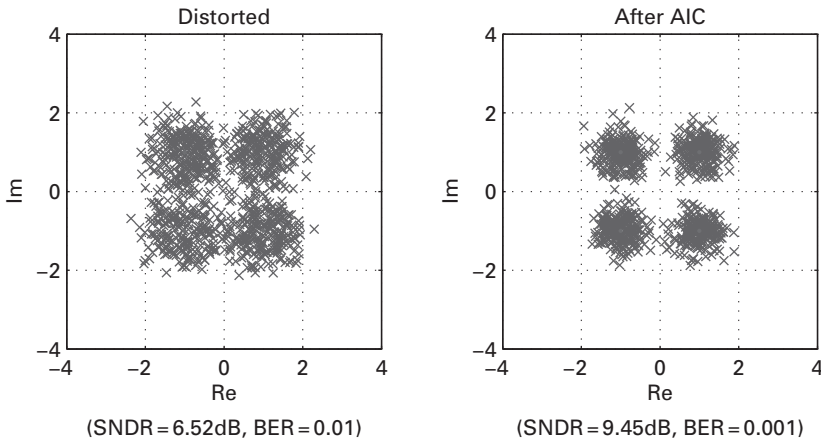


Figure 15.21 Constellations of the weak QPSK signal demodulated from the signal band around +3 MHz without and with the AIC post-processing.

In this example, the SNDR gain is almost 3 dB. The AIC method is only able to remove the LCF part of the INL error and this sets the limit for the performance in practice. It would require an entirely different kind of approach to reduce the HCF part (see, e.g., [12]) and still the performance is limited by the inband noise, which cannot be removed with any post-processing approach.

15.8 Conclusions

This chapter dealt with analog-to-digital converter nonlinearities and their digital mitigation from the wideband radio receiver point of view. First, some essential ADC nonlinearities were discussed and it was shown that individual information signals at separate frequency bands can severely interfere with one another due to these nonlinearities. Therefore, the mitigation of these nonlinearities is very important in wideband radio applications and more than ever in emerging cognitive radio systems. After the introduction, a wide range of different kinds of digital post-processing techniques for ADC correction was discussed in this chapter. However, some of them have limitations in radio applications. For example, look-up table methods require an offline calibration before actual operation.

The AIC and MSI methods were presented in this chapter in order to give some practical examples of ADC nonlinearity mitigation techniques, which are especially designed for wideband radios. In addition, both the simulation and laboratory measurement results were used to present the performance of these methods. The MSI method is particularly intended for mitigating clipping distortion and hence it cannot compensate any other nonlinearities. In practice, its performance may even be limited by the nonlinear distortion from other sources. In this sense, the AIC method is very effective, because it is able to suppress nonlinear distortion in general making good use of

polynomial modeling, which can be done dynamically without concentrating on just one source of nonlinear distortion.

Increased computational resources in radio receivers have enabled the possibility to exploit digital signal processing on the large scale. Despite the extensive research in the field in recent years, there is still much more to be done due to the ever-increasing demands of wireless communications. More advanced post-processing algorithms are needed, so that all the available information at the radio transceiver can be cleverly exploited to overcome the unavoidable imperfections in physical circuits.

References

- [1] F. Adamo, F. Attivissimo, N. Giaquinto, and M. Savino, "FFT test of A/D converters to determine the integral nonlinearity," *IEEE Trans. Instrum. Meas.*, vol. **51**, no. 5, pp. 1050–1054, Oct. 2002.
- [2] M. Allén, J. Marttila, and M. Valkama, "Digital post-processing for reducing A/D converter nonlinear distortion in wideband radio receivers," in *Proc. Forty-Third Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2009.
- [3] M. Allén, J. Marttila, and M. Valkama, "Digitally-enhanced wideband analog-digital interfaces for future cognitive radio devices," in *8th IEEE International NEWCAS Conference (NEWCAS2010)*, Montréal, Canada, June 2010.
- [4] M. Allén, J. Marttila, and M. Valkama, "Modeling and mitigation of nonlinear distortion in wideband A/D converters for cognitive radio receivers," *European Microwave Assoc. Int. J. Microwave and Wireless Technologies*, vol. **2**, no. 2, pp. 183–192, Apr. 2010.
- [5] M. Allén, "Nonlinear distortion studies in wideband analog-to-digital converters," M.Sc. thesis, Dept. Commun. Eng., Tampere University of Technology, Tampere, Finland, 2010.
- [6] Analog Devices Inc., *AD9218 Data Sheet*, rev. C, Dec. 2006. Available: www.analog.com
- [7] Analog Devices Inc., *AD9248 Data Sheet*, rev. A, Mar. 2005. Available: www.analog.com
- [8] Analog Devices Inc., *How ADIsimADC models an ADC*, application note AN-737, rev. B. Available: www.analog.com
- [9] Analog Devices Inc., *The Data Conversion Handbook*. W. Kester, ed. Burlington, MA: Newnes, 2004.
- [10] Analog Devices Inc., *Overcoming Converter Nonlinearities with Dither*, application note AN-410, B. Brannon, 1995. Available: www.analog.com
- [11] O. Aumala and J. Holub, "Dithering design for measurement of slowly varying signals," *Measurement*, vol. **23**, no. 4, pp. 271–276, Jun. 1998.
- [12] N. Björssell and P. Händel, "Dynamic behavior models of analog to digital converters aimed for post-correction in wideband applications," in *XVIII IMEKO World Congress 11th Workshop on ADC Modelling and Testing*, Rio de Janeiro, Brazil, 2006.
- [13] N. Björssell, "Modeling analog to digital converters at radio frequency," Ph.D. dissertation, School Elect. Eng., Royal Inst. Technology (KTH), Stockholm, Sweden, 2007.
- [14] A. B. Carlson, *Communication Systems: an Introduction to Signals and Noise in Electrical Communication*, 4th edn. New York: McGraw-Hill, 2002.
- [15] M. M. Chansarkar, U. B. Desai, "A fast approximate RLS algorithm," *IEEE Region 10 Conf. on Computer, Communication, Control and Power Engineering (TENCON'93)*, Beijing, China, 1993.

- [16] A. C. Dent and C. F. N. Cowan, "Linearization of Analog-to-Digital Converters," *IEEE Trans. Circuits Syst.*, vol. **37**, no. 6, pp. 729–737, June 1990.
- [17] J. Elbornsson, "Equalization of distortion in A/D converters," Licentiate thesis, Dept. Elect. Eng., Linköping University, Linköping, Sweden, 2001.
- [18] S. Haykin, *Adaptive Filter Theory*, 3rd edn., Upper Saddle River, NJ: Prentice-Hall, 1996.
- [19] IEEE Standard for Terminology and Test Methods for Analog-to-Digital Converters, IEEE Standard 1241–2000, 2001.
- [20] E. C. Ifeachor and B. W. Jervis, *Digital Signal Processing: A Practical approach*, 2nd edn. Harlow: Pearson Education, 2002.
- [21] M. Ismail and D. González, eds., *Radio Design in Nanometer Technologies*. Dordrecht: Springer, 2006.
- [22] B. Le, T. W. Rondeau, J. H. Reed, and C. W. Bostian, "Analog-to-digital converters," *IEEE Signal Processing Mag.*, vol. **22**, pp. 69–77, Nov. 2005.
- [23] H. Lundin, "Characterization and correction of analog-to-digital converters," Ph.D. dissertation, School Elect. Eng., Royal Inst. Technology (KTH), Stockholm, Sweden, 2005.
- [24] H. Lundin, M. Skoglund, and P. Händel, "On external calibration of analog-to-digital converters," in *Proc. 11th IEEE Signal Processing Workshop on Statistical Signal Processing*, Singapore, 2001.
- [25] F. Maloberti, *Data Converters*. Dordrecht: Springer, 2007.
- [26] Maxim Integrated Products Inc., *The ABCs of ADCs: Understanding how ADC errors affect system performance*, application note 748, July, 2002. Available: www.maxim-ic.com/appnotes.cfm/appnote_number/748/
- [27] A. H. Sayed, *Adaptive Filters*, Hoboken, NJ: John Wiley & Sons, 2008.
- [28] T. Tomioka, R. Sakata, T. Horiguchi, T. Tomizawa, and K. Inoue, "A/D converter clipping noise suppression for high-sensitivity carrier-sensing of cognitive radio transceiver," in *IEEE Global Telecommunications Conf.* 2007, Washington, DC, pp. 4170–4174.
- [29] J. Tsimbinos, "Identification and compensation of nonlinear distortion," Ph.D. dissertation, School Elect. Eng., University of South Australia, The Levels, Australia, 1995.
- [30] M. Valkama, A. S. H. Ghadam, L. Anttila, and M. Renfors, "Advanced digital signal processing techniques for compensation of nonlinear distortion in wideband multicarrier radio receivers," *IEEE Trans. Microw. Theory Tech.*, vol. **54**, no. 6, pp. 2356–2366, June 2006.
- [31] J. H. Van Vleck and D. Middleton, "The spectrum of clipped noise," *Proc. IEEE*, vol. **54**, no. 1, pp. 2–19, Jan. 1966.
- [32] N. Vun and A. B. Premkumar, "ADC systems for SDR digital front-end," *Proc. 9th Int. Symp. Consumer Electronics*, Macau, 2005, pp. 14–16.
- [33] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE J. Sel. Areas Commun.*, vol. **17**, no. 4, pp. 539–550, Apr. 1999.

Part IV

Digital calibration, imbalance compensation, and error corrections

16 Digital compensation and calibration of I/Q gain and phase imbalances

Lauri Anttila, Yaning Zou, and Mikko Valkama

16.1 Introduction

The implementation challenges in building compact and low-cost radios for future wireless systems are continuously growing. This is partially due to the introduction of multi-antenna transmission techniques as well as the use of wideband communication waveforms and high-order symbol alphabets, in addition to the increasing demands for more efficient radio spectrum utilization through e.g. carrier aggregation and scattered spectrum use. In general, implementations of several parallel radios with wide operating bandwidth and high performance, in terms of linearity and spurious free dynamic range, are required in a single device. Then, to keep the overall implementation costs and size feasible, simplified radio architectures and lower-cost radio electronics are typically used. This in turn implies that various nonidealities in the deployed analog radio frequency (RF) modules, stemming from the unavoidable physical limitations of the electronics used, are expected to play a critical role in future radio devices.

Good examples of the above “dirty-RF” paradigm [1], [2] are, e.g., oscillator phase noise, power amplifier (PA) nonlinearities, imperfections of the sampling and analog-to-digital (A/D) interface, in-phase/quadrature (I/Q) branch amplitude and phase mismatches, as well as nonlinearities of receiver small signal components like low-noise amplifiers (LNAs) and mixers. In this chapter, we will focus on the behavioral modeling and digital signal processing (DSP) based mitigation of I/Q imbalances and the resulting mirror-frequency interference in direct-conversion type radio transmitters and receivers. For generality, in most of the developments, the I/Q imbalances are assumed to be frequency-dependent within the processed bandwidth, which is then built in to both modeling as well as mitigation algorithms. Also an extensive list of state-of-the-art literature is given.

The rest of this chapter is organized into the following five sections. In [Section 16.2](#), we will shortly review I/Q modulation and direct-conversion type radio architectures. [Section 16.3](#) is devoted to behavioral modeling of the I/Q imbalance problem in radio transmitters and receivers, covering both frequency-independent and frequency-dependent I/Q imbalance cases. In [Section 16.4](#), various approaches for imbalance estimation and calibration are presented, covering both digital pre-distortion type techniques on the transmitter side as well as digital post-correction methods on the receiver

side. Also complete link models, including imbalanced transmitter, multipath radio channel, and imbalanced receiver are given. [Section 16.5](#) is devoted to design and testing examples, while in [Section 16.6](#), we will give a concluding summary and there will be some further discussion.

16.2 I/Q modulation and direct-conversion radio architecture

One of the key functionalities of any radio transceiver is to up-convert and down-convert the low-frequency message signals to and from radio frequencies (RF). In this context, complex or I/Q mixing is of great practical importance (compared to more traditional real mixing) since, e.g., the so-called mirror-frequency problem is in theory avoided in the resulting frequency translations. Consequently, for example, the RF filtering requirements are relaxed and thus the transceiver implementation is simplified. Practical examples of the I/Q mixing principle are the direct-conversion and low-IF radio architectures [3]–[9].

Embedded with the simplest frequency translation idea, the direct-conversion topology is considered as one of the most promising radio architectures for developing future wireless transceivers. Unlike the more traditional superheterodyne radio, the direct-conversion architecture directly down-converts RF signals to baseband (BB) or up-converts BB signals to RF without any intermediate frequency (IF) stages. Fittingly, it is also referred to as homodyne or zero-IF architecture in the literature. Then, fewer analog components and blocks are needed compared to the superheterodyne architecture. This inherent simplicity offers the direct-conversion radio important advantages over the heterodyne in, e.g., used silicon area, implementation cost, and power consumption. In addition, the location of channel selection filters at baseband enables the possibility of implementing multiple or adjustable filter bandwidths more easily, without consuming extra silicon area. This is considered one of the key elements towards building ever more flexible multimode receivers in the future. The low-IF radio, in turn, also uses I/Q mixing but the frequency conversion carried out by the I/Q mixer is between the selected IF and RF. This has some benefits over the plain direct-conversion radio, like the reduced effects of DC offsets and second-order intermodulation on the receiver side, but also some relative disadvantages, such as an increased sensitivity to I/Q imbalance and the resulting mirror-frequency interference. This is the general problem discussed in detail in this chapter.

Though simple in theory, the implementation of direct-conversion and low-IF radios faces a number of problems and technical challenges. As shown in [Figure 16.1](#), I/Q mixing is applied in the down-conversion and up-conversion stages. Ideally, this mixing approach builds on two local oscillator (LO) signals with exactly 90° phase difference and equal amplitudes, and also contains two independent but identical signal paths in the circuits. However, in practice, even with state-of-the-art advanced RF integrated circuit (IC) technology, the nominal 90° phase shift and the equal amplitudes of the I and Q signal paths can only be realized up to finite accuracy. Furthermore, the differences in the frequency responses of the I and Q branch low-pass filters (LPFs), data converters, and amplifiers also contribute to effective overall amplitude and phase mismatches. In general, total effective amplitude and phase imbalances in the order of 1%–5% and 1°–5°

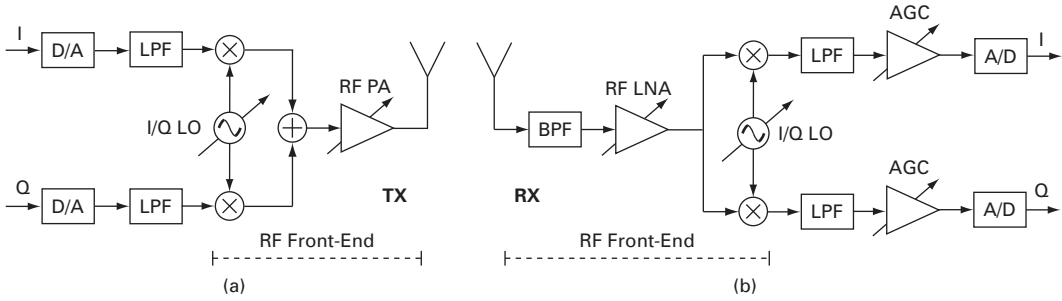


Figure 16.1 Conceptual radio transmitter (a) and radio receiver (b) block-diagrams using quadrature or I/Q mixing. In plain direct-conversion radio, the I and Q signal are baseband signals while in the low-IF radio, the I and Q signals are intermediate frequency signals.

are typically stated feasible with careful design [3]–[7]. The resulting corruption on the down-converted or up-converted signal waveform can easily degrade the system performance and raise, e.g., the symbol error rate (SER) or bit error rate (BER) in the detection. This is the I/Q imbalance problem which is the central theme of this chapter.

In the more general context of the multiband or low-IF radio transceiver, the above-mentioned I/Q imbalances cause interference between mirror-frequency bands. This is potentially even a bigger problem, compared to the plain single-channel zero-IF case, due to possibly different signal dynamics at different bands. For generality, all the modeling in the forthcoming sections is carried out for a general wideband I/Q mixing based radio transceiver, which then covers the direct-conversion and low-IF radios as special cases.

16.3 I/Q imbalance signal modeling

Physically the amplitude and phase mismatches between the transceiver I and Q signal branches stem from the relative differences between all the analog components of the I/Q front-end. On the transmitter side, this includes the actual I/Q up-conversion stage as well as the I and Q branch analog filters and digital-to-analog (D/A) converters. On the receiver side, in turn, the I/Q down-conversion as well as the I and Q branch analog filtering, amplification, and sampling stages contribute to the overall effective I/Q imbalances.

16.3.1 Frequency-independent imbalance modeling

As the first modeling example, we begin with the narrowband signal assumption, and refer all the mismatches to the I/Q up- and down-conversion stages. Conceptual illustrations of such modulators and demodulators are given in Figure 16.2. Considering then the implications at waveform level, we first write the corresponding complex LO signals as

$$\begin{aligned} x_{LO}^{TX}(t) &= \cos(\omega_{LO}t) + jg_{TX} \sin(\omega_{LO}t + \phi_{TX}) \\ &= K_{1,TX}e^{j\omega_{LO}t} + K_{2,TX}^*e^{-j\omega_{LO}t} \end{aligned} \quad (16.1)$$

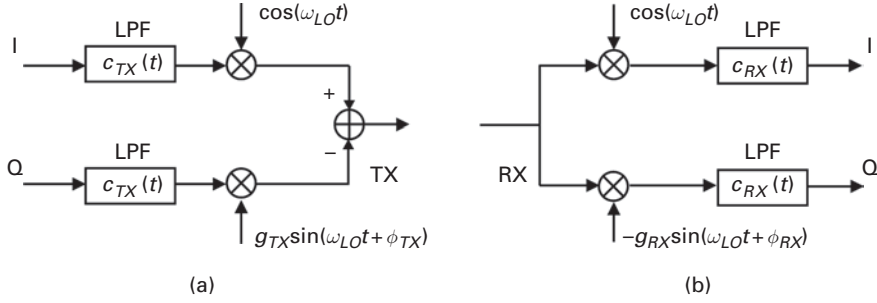


Figure 16.2 Frequency-independent I/Q imbalance in TX (a) and RX (b). The parameters g and ϕ denote the I/Q mixer amplitude and phase mismatches while the impulse response $c(t)$ models the common signal response of the I and Q rails.

$$\begin{aligned} x_{LO}^{RX}(t) &= \cos(\omega_{LO}t) - jg_{RX}\sin(\omega_{LO}t + \phi_{RX}) \\ &= K_{1,RX}e^{-j\omega_{LO}t} + K_{2,RX}e^{j\omega_{LO}t} \end{aligned} \quad (16.2)$$

where $\omega_{LO} = 2\pi f_{LO}$ denotes the LO angular frequency, $[g_{TX}, \phi_{TX}]$ and $[g_{RX}, \phi_{RX}]$ represent the total effective amplitude and phase imbalances of the TX and the RX, respectively, and the coefficients $K_{1,TX}$, $K_{2,TX}$, $K_{1,RX}$, and $K_{2,RX}$ are of the form

$$\begin{aligned} K_{1,TX} &= (1 + g_{TX}e^{j\phi_{TX}})/2, & K_{2,TX} &= (1 - g_{TX}e^{j\phi_{TX}})/2 \\ K_{1,RX} &= (1 + g_{RX}e^{-j\phi_{RX}})/2, & K_{2,RX} &= (1 - g_{RX}e^{j\phi_{RX}})/2. \end{aligned} \quad (16.3)$$

Then from the individual transmitter and receiver point of views, the above I/Q imbalance models result in transformations of the effective baseband equivalent signals, given by

$$\begin{aligned} \bar{z}_{TX}(t) &= K_{1,TX}(c_{TX}(t) * z(t)) + K_{2,TX}(c_{TX}(t) * z(t))^* \\ &= c_{TX}(t) * [K_{1,TX}z(t) + K_{2,TX}z^*(t)] \end{aligned} \quad (16.4)$$

$$\begin{aligned} \bar{z}_{RX}(t) &= c_{RX}(t) * [K_{1,RX}z(t) + K_{2,RX}z^*(t)] \\ &= K_{1,RX}(c_{RX}(t) * z(t)) + K_{2,RX}(c_{RX}(t) * z(t))^* \end{aligned} \quad (16.5)$$

where $z(t)$ denotes the ideal complex baseband equivalent under perfect I/Q matching and the real-valued impulse responses $c_{TX}(t)$ and $c_{RX}(t)$ denote the common responses of the transmitter and receiver I and Q branch filtering. Based on (16.4) and (16.5), the main effect of I/Q imbalance at complex baseband or low-frequency signal level is that a conjugated version of the ideal signal is showing up. The common responses $c_{TX}(t)$ and $c_{RX}(t)$ do not contribute to the relative strengths of the two signal components ($z(t)$ and $z^*(t)$), and are typically dropped, yielding more simple models of the form

$$\bar{z}_{TX}(t) = K_{1,TX}z(t) + K_{2,TX}z^*(t) \quad (16.6)$$

$$\bar{z}_{RX}(t) = K_{1,RX}z(t) + K_{2,RX}z^*(t). \quad (16.7)$$

These are the typical models used in the literature, e.g., [10]–[14]. In link-level developments, on the other hand, the common responses $c_{TX}(t)$ and $c_{RX}(t)$ can also be considered part of the radio channel linking the transmitter and receiver.

In the frequency domain, the distortion due to the conjugate signal term corresponds to *mirror-frequency interference*. This can be seen by taking Fourier transforms (FT) of (16.6) and (16.7) as

$$\bar{Z}_{TX}(f) = K_{1,TX}Z(f) + K_{2,TX}Z^*(-f) \quad (16.8)$$

$$\bar{Z}_{RX}(f) = K_{1,RX}Z(f) + K_{2,RX}Z^*(-f). \quad (16.9)$$

The corresponding mirror-frequency attenuations L_{TX} and L_{RX} of the individual front-ends are then given by

$$L_{TX} = \frac{|K_{1,TX}|^2}{|K_{2,TX}|^2}, \quad L_{RX} = \frac{|K_{1,RX}|^2}{|K_{2,RX}|^2} \quad (16.10)$$

which typically range in the order of 25–40 dB with imbalances in the order of 1–5% and 1–5 degrees.

16.3.2 Frequency-dependent I/Q imbalance modeling

In the context of wideband radio and waveform, the overall effective I/Q imbalances can easily vary as a function of frequency within the system band, due to e.g. frequency-response differences between the I and Q branch filtering, data conversion and amplification stages [15]–[18]. This should also be reflected in the imbalance modeling as well as in imbalance compensation. Using the frequency-independent I/Q imbalance modeling in (16.1)–(16.9) as a starting point, the frequency-response differences between I and Q branches are modeled here as branch *mismatch filters* $b_{TX}(t)$ and $b_{RX}(t)$, on the transmitter and receiver sides, respectively, as shown in Figure 16.3. Then if $z(t)$ denotes again the ideal (perfect I/Q balance) complex baseband equivalent signal, the overall baseband equivalent I/Q imbalance models for individual transmitters and receivers appear as

$$\begin{aligned} \bar{z}_{TX}(t) &= g_{1,TX}(t) * (c_{TX}(t) * z(t)) + g_{2,TX}(t) * (c_{TX}(t) * z(t))^* \\ &= c_{TX}(t) * [g_{1,TX}(t) * z(t) + g_{2,TX}(t) * z^*(t)] \end{aligned} \quad (16.11)$$

$$\begin{aligned} \bar{z}_{RX}(t) &= c_{RX}(t) * [g_{1,RX}(t) * z(t) + g_{2,RX}(t) * z^*(t)] \\ &= g_{1,RX}(t) * (c_{RX}(t) * z(t)) + g_{2,RX}(t) * (c_{RX}(t) * z(t))^*, \end{aligned} \quad (16.12)$$

where $c_{TX}(t)$ and $c_{RX}(t)$ denote again the common response filtering and the effective impulse responses $g_{1,TX}(t)$, $g_{2,TX}(t)$, $g_{1,RX}(t)$, and $g_{2,RX}(t)$ depend on the actual imbalance properties as

$$\begin{aligned} g_{1,TX}(t) &= (\delta(t) + b_{TX}(t)g_{TX}e^{j\phi_{TX}})/2, & g_{2,TX}(t) &= (\delta(t) - b_{TX}(t)g_{TX}e^{j\phi_{TX}})/2 \\ g_{1,RX}(t) &= (\delta(t) + b_{RX}(t)g_{RX}e^{-j\phi_{RX}})/2, & g_{2,RX}(t) &= (\delta(t) - b_{RX}(t)g_{RX}e^{j\phi_{RX}})/2 \end{aligned} \quad (16.13)$$

where $\delta(t)$ denotes a Dirac impulse function. Similarly as in the [previous section](#), the common response does not contribute to the relative strengths of the two signal components, and thus simplified models of the form

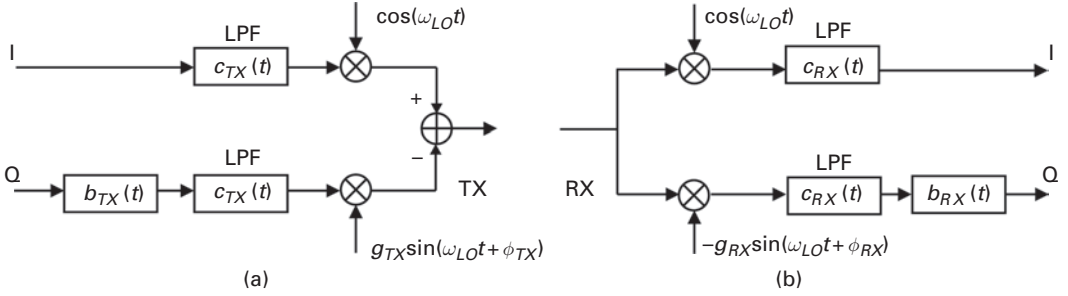


Figure 16.3 Frequency-dependent I/Q imbalance in TX (a) and RX (b). The parameters g and ϕ denote the I/Q mixer amplitude and phase mismatches while the mismatch between the signal path I and Q rails is modeled with filter $b(t)$. The impulse response $c(t)$ models the common response of the I and Q rails.

$$\bar{z}_{TX}(t) = g_{1,TX}(t) * z(t) + g_{2,TX}(t) * z^*(t) \quad (16.14)$$

$$\bar{z}_{RX}(t) = g_{1,RX}(t) * z(t) + g_{2,RX}(t) * z^*(t) \quad (16.15)$$

can be and typically are used. Notice that the earlier frequency-independent (instantaneous) I/Q imbalance models of the form $\bar{z}_{TX}(t) = K_{1,TX}z(t) + K_{2,TX}z^*(t)$ and $\bar{z}_{RX}(t) = K_{1,RX}z(t) + K_{2,RX}z^*(t)$ are obtained as special cases of (16.4) and (16.15) when $b_{TX}(t) = \delta(t)$ and $b_{RX}(t) = \delta(t)$.

Based on the models (16.14) in and (16.15), when viewed in the frequency domain, the distortion due to frequency-dependent I/Q imbalance corresponds now to *mirror-frequency interference whose strength varies as a function of frequency*. This can be seen by taking FT of and (16.14) and (16.15), yielding

$$\bar{Z}_{TX}(f) = G_{1,TX}(f)Z(f) + G_{2,TX}(f)Z^*(-f) \quad (16.16)$$

$$\bar{Z}_{RX}(f) = G_{1,RX}(f)Z(f) + G_{2,RX}(f)Z^*(-f) \quad (16.17)$$

in which the transfer functions are of the form

$$\begin{aligned} G_{1,TX}(f) &= (1 + B_{TX}(f)g_{TX}e^{j\phi_{TX}})/2, & G_{2,TX}(f) &= (1 - B_{TX}(f)g_{TX}e^{j\phi_{TX}})/2 \\ G_{1,RX}(f) &= (1 + B_{RX}(f)g_{RX}e^{-j\phi_{RX}})/2, & G_{2,RX}(f) &= (1 - B_{RX}(f)g_{RX}e^{j\phi_{RX}})/2. \end{aligned} \quad (16.18)$$

Thus, the corresponding mirror-frequency attenuations or image rejection ratios (IRRs) of the individual radio front-ends are now frequency-dependent functions and given by

$$L_{TX}(f) = \frac{|G_{1,TX}(f)|^2}{|G_{2,TX}(f)|^2}, \quad L_{RX}(f) = \frac{|G_{1,RX}(f)|^2}{|G_{2,RX}(f)|^2}. \quad (16.19)$$

With practical analog front-end electronics, these mirror-frequency attenuations are typically in the range of 25–40 dB and vary as a function of frequency, when bandwidths in the order of several MHz are considered. An example is given below in Figure 16.4 which shows the measured mirror-frequency attenuation characteristics, obtained in a laboratory test setup, of a state-of-the-art wireless receiver RF-IC operating at 2 GHz.

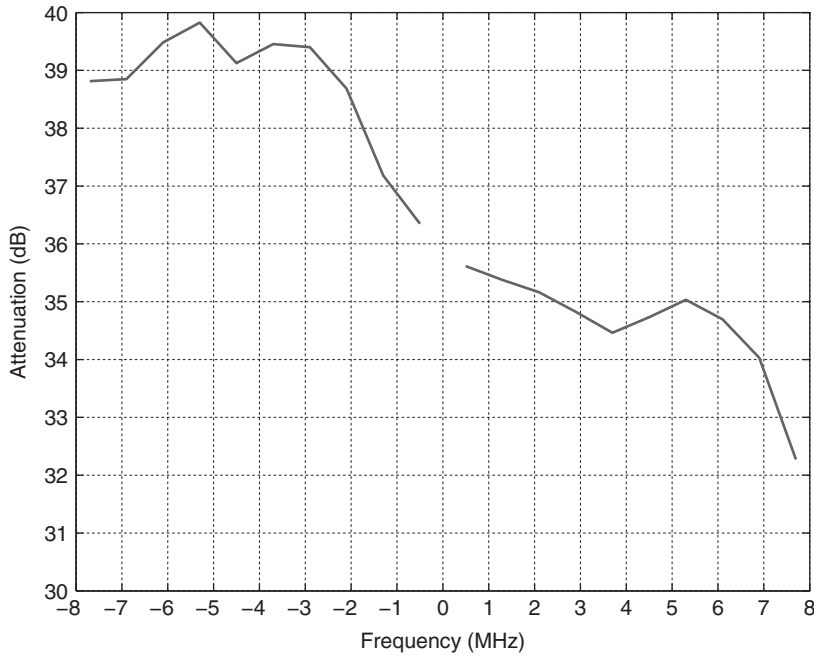


Figure 16.4 Measured mirror-frequency attenuation of state-of-the-art I/Q receiver RF-IC operating at 2 GHz RF. The x -axis refers to frequencies of the down-converted complex (I/Q) signal, or equivalently, to the frequencies around the LO frequency at RF.

Clearly, for bandwidths in the order of 1–10 MHz, the mirror-frequency attenuation or IRRs (and thus, the effective I/Q imbalances) indeed depend on frequency. Thus, the use of frequency-dependent I/Q imbalance modeling is necessary.

16.3.3 Alternative imbalance models based on parallel I/Q signals and illustrations

The previous imbalance models build heavily on complex-valued signal notations. The practical implementation of complex-valued signals, on the other hand, relies on parallel real-valued (I and Q) signals. Thus also the imbalance modeling can, of course, be formulated using parallel real-valued signals.

The previous frequency-independent complex signal model of the form $\bar{z}(t) = K_1 z(t) + K_2 z^*(t)$ corresponds to the following parallel real signal model

$$\begin{aligned}\bar{z}_I(t) &= z_I(t) \\ \bar{z}_Q(t) &= g \cos(\phi) z_Q(t) - g \sin(\phi) z_I(t)\end{aligned}\tag{16.20}$$

where g and ϕ model the amplitude and phase imbalances, referred here to the Q branch. The corresponding frequency-dependent case, corresponding to a complex-valued model $\bar{z}(t) = g_1(t) * z(t) + g_2(t) * z^*(t)$, is then given by

$$\begin{aligned}\bar{z}_I(t) &= z_I(t) \\ \bar{z}_Q(t) &= g \cos(\phi) b(t) * z_Q(t) - g \sin(\phi) b(t) * z_I(t)\end{aligned}\tag{16.21}$$

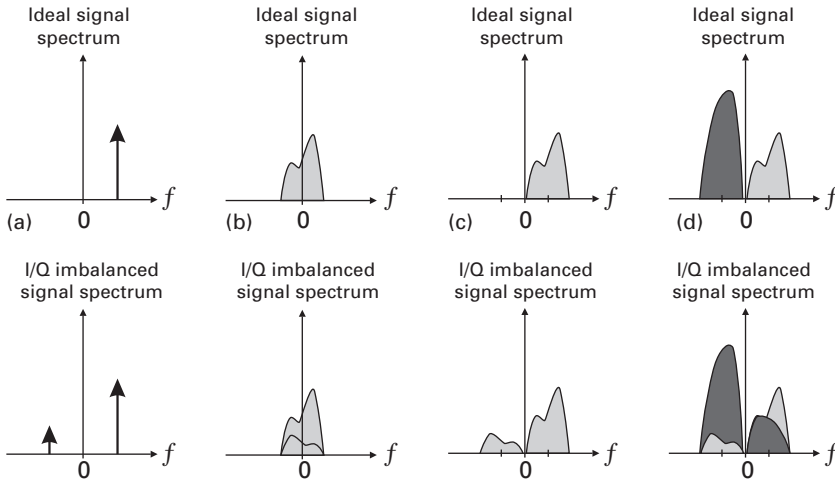


Figure 16.5 Conceptual illustration of mirror-frequency cross-talk due to I/Q imbalance with different complex envelopes.

where again the imbalances are referred to the Q branch. Such parallel real signal models are utilized in some compensation developments, e.g., in [19]–[21].

One example where the earlier complex-valued signal models are particularly handy is spectral interpretations of I/Q imbalances. Based on the complex-valued models in (16.6)–(16.7) and (16.14)–(16.15), the basic effect is that in addition to the “direct” signal term (relative to $z(t)$), also the conjugated signal term (relative to $z^*(t)$) shows up. When interpreted in the frequency domain, this means cross-talk between frequency components located symmetrically around zero frequency (or around LO frequency if interpreted for the corresponding RF signals). This is exactly where the mirror-frequency or image-frequency terminology is stemming from. Conceptual illustrations of the phenomenon with different complex envelopes (different types of $z(t)$) are given in Figure 16.5.

16.4 Compensation and calibration methods

16.4.1 Principle of TX calibration using pre-distortion

For generality, based on the imbalance modeling in the [previous section](#), we assume the general frequency-dependent I/Q imbalance model of the form $\bar{z}_{TX}(t) = g_{1,TX}(t) * z(t) + g_{2,TX}(t) * z^*(t)$. Then, stemming directly from the structure of this model, a natural form for the *pre-distortion processing* is given by

$$z_{pre}(t) = z(t) + w_{TX}(t) * z^*(t) \quad (16.22)$$

where $w_{TX}(t)$ denotes the pre-distortion filter impulse response and $z_{pre}(t)$ is the pre-distorted complex envelope. Then, the combined baseband equivalent model for the overall TX waveform, including the pre-distortion, can be directly written as

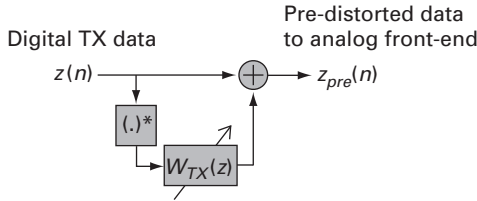


Figure 16.6 Principle of transmitter pre-distortion for suppressing mirror-frequency interference.

$$\bar{z}_{TX,pre}(t) = g_{1,TX}(t) * z_{pre}(t) + g_{2,TX}(t) * z_{pre}^*(t) \quad (16.23)$$

or

$$\begin{aligned} \bar{z}_{TX,pre}(t) = & [g_{1,TX}(t) + g_{2,TX}(t) * w_{TX}^*(t)] * z(t) \\ & + [g_{2,TX}(t) + g_{1,TX}(t) * w_{TX}(t)] * z^*(t) \end{aligned} \quad (16.24)$$

which follows by combining (16.22) and (16.23). Then to remove the mirror-frequency interference (conjugate signal term in (16.24)), and thus to restore perfect I/Q balance, the pre-distortion filter $w_{TX}(t)$ should be selected such that $[g_{2,TX}(t) + g_{1,TX}(t) * w_{TX}(t)] = 0 \forall t$. In terms of frequency response, this corresponds to the following optimum (reference) pre-distortion filter

$$W_{TX}^{OPT}(f) = -\frac{G_{2,TX}(f)}{G_{1,TX}(f)}. \quad (16.25)$$

A conceptual illustration of the pre-distortion processing is given in Figure 16.6. In the special case of frequency-independent imbalances, this boils down to a one tap filter with complex-valued coefficient $w_{TX}^{OPT} = -K_{2,TX}/K_{1,TX}$.

In practice, the exact imbalance characteristics defining the above optimum solution are naturally unknown, and thus some imbalance parameter estimation mechanisms are needed. These will be addressed later in Section 16.4.4. Notice also that practical pre-distortion implementations are typically based on digital signal processing, so the frequency response of the digital pre-distorter used should then approximate the one in (16.25) as closely as possible.

16.4.2 Principle of RX calibration using post-distortion

Next, we turn the focus to receiver I/Q imbalance mitigation. The idea is to do some *post-processing* on the observed signal of the form $\bar{z}_{RX}(t) = g_{1,RX}(t) * z(t) + g_{2,RX}(t) * z^*(t)$, in order to push down the mirror-frequency interference in the receiver digital front-end. Here $z(t)$ denotes the received and I/Q down-converted complex-valued signal under perfect I/Q balance. Thus conceptually, the mitigation task in the receiver is more complicated (compared to transmitter calibration) since the exact waveform structure of the perfectly balanced signal $z(t)$ is unknown due to e.g. multipath, noise and synchronization errors. This has an impact especially on imbalance parameter estimation, which is described later in Section 16.4.5.

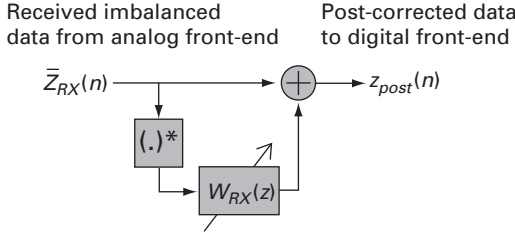


Figure 16.7 Principle of receiver post-processing for suppressing mirror-frequency interference.

Formally, the compensation structure to be deployed is similar to the transmitter case, and is given by

$$z_{post}(t) = \bar{z}_{RX}(t) + w_{RX}(t) * \bar{z}_{RX}^*(t). \quad (16.26)$$

This is illustrated at structural level in Figure 16.7. Now, a direct substitution of the imbalanced signal $\bar{z}_{RX}(t)$ into the above post-processing model yields

$$\begin{aligned} z_{post}(t) = & [g_{1,RX}(t) + w_{RX}(t) * g_{2,RX}^*(t)] * z(t) \\ & + [g_{2,RX}(t) + w_{RX}(t) * g_{1,RX}^*(t)] * z^*(t) \end{aligned} \quad (16.27)$$

Thus, the optimum solution for the post-processing coefficients is the one which sets $[g_{2,RX}(t) + w_{RX}(t) * g_{1,RX}^*(t)] = 0 \forall t$, or in terms of frequency response

$$W_{RX}^{OPT}(f) = -\frac{G_{2,RX}(f)}{G_{1,RX}^*(-f)}. \quad (16.28)$$

In the special case of frequency-independent imbalances, this reduces again to a one tap filter with complex-valued coefficient $w_{RX}^{OPT} = -K_{2,RX}/K_{1,RX}^*$.

Similar to the one on the transmitter side, the optimum solution in (16.28) depends on the receiver imbalance characteristics which are by default unknown. Thus efficient imbalance parameter estimation mechanisms are needed. These will be addressed in Sections 16.4.4 and 16.4.5.

16.4.3 Some waveform and/or radio architecture-specific solutions and link-level aspects

OFDM link models

Many of the emerging or existing radio systems build on multi-carrier modulation methods, and especially OFDM. Interpreting the previous imbalance models with OFDM waveforms, the basic signal model is of the form $\bar{Z}_k = G_{1,k}Z_k + G_{2,k}Z_{-k}^*$ where Z_k and Z_{-k} denote the frequency-domain signal samples at subcarriers k and $-k$ under perfect I/Q balance. This is one form of intercarrier interference (ICI), namely cross-talk between mirror-symmetric subcarriers. Given a typical link-model of the form $Z_k = H_k X_k + N_k$ where X_k , H_k , and N_k denote the transmit symbol, channel response, and channel noise, respectively, all at subcarrier k , and further assuming receiver imbalances only, the imbalanced received signal can also be written as

$$\begin{aligned}\bar{Z}_k &= G_{1,k}Z_k + G_{2,k}Z_{-k}^* \\ &= G_{1,k}H_kX_k + G_{2,k}H_{-k}^*X_{-k}^* + G_{1,k}N_k + G_{2,k}N_{-k}^*.\end{aligned}\quad (16.29)$$

This demonstrates the link-level mirror-subcarrier effect from the transmit data point of view.

At conceptual level, based on (16.29), a fairly obvious structure for the compensator in the OFDM radio receiver is to combine the observations at mirror subcarriers k and $-k$, with proper complex-valued weight W_k , to form an output signal of the form

$$\begin{aligned}Z_{post,k} &= \bar{Z}_k + W_k\bar{Z}_{-k}^* \\ &= (G_{1,k} + W_kG_{2,-k}^*)Z_k + (G_{2,k} + W_kG_{1,-k}^*)Z_{-k}^*.\end{aligned}\quad (16.30)$$

Now, it follows directly that the optimum compensator coefficient for subcarrier k is given by

$$W_{OPT,k} = -\frac{G_{2,k}}{G_{1,-k}^*}.\quad (16.31)$$

Notice that if the imbalances are frequency-independent, each mirror-carrier pair can be processed with the same compensator coefficient $W_{OPT,k} = w_{OPT} = -K_2/K_1^*$.

The above formulation in (16.29) assumes an ideal transmitter (i.e., imbalances affecting only in the receiver). Given that both transmitter and receiver have their own imbalances, a similar link-level model as in can be developed. This is formulated below for the whole link as

$$\begin{aligned}\bar{Z}_k &= G_{1,RX,k}[H_k(G_{1,TX,k}X_k + G_{2,TX,k}X_{-k}^*) + N_k] \\ &\quad + G_{2,RX,k}[H_{-k}^*(G_{1,TX,-k}^*X_{-k}^* + G_{2,TX,-k}^*X_k) + N_{-k}^*] \\ &= (G_{1,TX,k}G_{1,RX,k}H_k + G_{2,TX,-k}^*G_{2,RX,k}H_{-k}^*)X_k \\ &\quad + (G_{2,TX,k}G_{1,RX,k}H_k + G_{1,TX,-k}^*G_{2,RX,k}H_{-k}^*)X_{-k}^* \\ &\quad + G_{1,RX,k}N_k + G_{2,RX,k}N_{-k}^* \\ &= H_{TOT,1,k}X_k + H_{TOT,2,k}X_{-k}^* + G_{1,RX,k}N_k + G_{2,RX,k}N_{-k}^*\end{aligned}\quad (16.32)$$

where the subscripts TX and RX denote the transmitter and receiver side imbalance coefficients, respectively. Compared to (16.29), and when interpreted from the transmit data point of view, a similar mirror-subcarrier interference model is observed but obviously with more complicated parameterization. With knowledge of the overall link coefficients $H_{TOT,1,k}$ and $H_{TOT,2,k}$, a structurally similar compensator (as before) of the form $Z_{post,k} = \bar{Z}_k + W_k\bar{Z}_{-k}^*$ can be used. It is straightforward to show that the corresponding optimum solution is now

$$W_{OPT,k} = -\frac{H_{TOT,2,k}}{H_{TOT,1,-k}^*}\quad (16.33)$$

MIMO-OFDM link models

The above modeling can also be generalized to multi-antenna or multiple-input multiple-output (MIMO) radio communications, which is an essential ingredient in modern radio

system developments. Given N_T transmit and N_R receive antennas, and assuming direct spatial multiplexing without any transmitter pre-coding for simplicity, the ideal link model under perfect I/Q balance in all the individual radios is of the form $\mathbf{Z}(k) = \mathbf{H}(k)^T \mathbf{X}(k) + \mathbf{N}(k)$ as is well known [22]. Here $\mathbf{Z}(k) = [Z_1(k), Z_2(k), \dots, Z_{N_R}(k)]^T$, $\mathbf{X}(k) = [X_1(k), X_2(k), \dots, X_{N_T}(k)]^T$, and $\mathbf{N}(k) = [N_1(k), N_2(k), \dots, N_{N_R}(k)]^T$ denote the received signal, transmit signal, and noise vectors, respectively, and the $N_T \times N_R$ matrix $\mathbf{H}(k)$ contains the channel responses where $[\mathbf{H}(k)]_{j,i} = H_{j,i}(k)$ denotes the channel frequency response from transmitter j to receiver i , all at the k th subcarrier. Assuming then that the individual radios (N_T transmitters and N_R receivers) all contain their own imbalances, the resulting overall link-level model can be rewritten as

$$\begin{aligned} \bar{\mathbf{Z}}(k) &= [\mathbf{R}_1(k)\mathbf{H}(k)^T\mathbf{T}_1(k) + \mathbf{R}_2(k)\mathbf{H}(-k)^H\mathbf{T}_2^*(-k)]\mathbf{X}(k) \\ &\quad + [\mathbf{R}_1(k)\mathbf{H}(k)^T\mathbf{T}_2(k) + \mathbf{R}_2(k)\mathbf{H}(-k)^H\mathbf{T}_1^*(-k)]\mathbf{X}^*(-k) \\ &\quad + \mathbf{R}_1(k)\mathbf{N}(k) + \mathbf{R}_2(k)\mathbf{N}^*(-k) \\ &= \mathbf{H}_{\text{TOT},1}(k)\mathbf{X}(k) + \mathbf{H}_{\text{TOT},2}(k)\mathbf{X}^*(-k) + \mathbf{R}_1(k)\mathbf{N}(k) + \mathbf{R}_2(k)\mathbf{N}^*(-k) \end{aligned} \quad (16.34)$$

where

$$\begin{aligned} \mathbf{H}_{\text{TOT},1}(k) &= \mathbf{R}_1(k)\mathbf{H}(k)^T\mathbf{T}_1(k) + \mathbf{R}_2(k)\mathbf{H}(-k)^H\mathbf{T}_2^*(-k) \\ \mathbf{H}_{\text{TOT},2}(k) &= \mathbf{R}_1(k)\mathbf{H}(k)^T\mathbf{T}_2(k) + \mathbf{R}_2(k)\mathbf{H}(-k)^H\mathbf{T}_1^*(-k) \end{aligned} \quad (16.35)$$

and

$$\begin{aligned} \mathbf{R}_m(k) &= \text{diag}\{G_{m,RX(1)}(k), G_{m,RX(2)}(k), \dots, G_{m,RX(N_R)}(k)\} \\ \mathbf{T}_m(k) &= \text{diag}\{G_{m,TX(1)}(k), G_{m,TX(2)}(k), \dots, G_{m,TX(N_T)}(k)\} \end{aligned}, m = 1, 2. \quad (16.36)$$

Above, $G_{m,RX(i)}(k)$ and $G_{m,TX(j)}(k)$ represent the I/Q imbalance responses at the k th subcarrier of the i th receiver and the j th transmitter, respectively, with $m = 1, 2$. Based on (16.34), the observed signal vector at subcarrier k is interfered due to I/Q imbalance by the conjugate of the data vector at the corresponding mirror carrier $-k$. When this is interpreted from the individual streams' point of view, this means increased interstream interference, compared to the perfect I/Q balance case, coming from the streams transmitted at mirror subcarriers. The above modeling gives directly a basis for the corresponding receiver processing to suppress the interference. This leads to the following multi-dimensional equalizer or detector of the form

$$\begin{aligned} \mathbf{Z}_{\text{post}}(k) &= \bar{\mathbf{Z}}(k) + \mathbf{W}(k)\bar{\mathbf{Z}}^*(-k) \\ &= [\mathbf{H}_{\text{TOT},1}(k) + \mathbf{W}(k)\mathbf{H}_{\text{TOT},2}^*(-k)]\mathbf{X}(k) \\ &\quad + [\mathbf{H}_{\text{TOT},2}(k) + \mathbf{W}(k)\mathbf{H}_{\text{TOT},1}^*(-k)]\mathbf{X}^*(-k) + \mathbf{V}(k) \end{aligned} \quad (16.37)$$

where $\mathbf{V}(k)$ denotes the post-processed noise terms. To fully suppress the mirror-stream interference, the detector is designed such that $\mathbf{H}_{\text{TOT},2}(k) + \mathbf{W}_{\text{OPT}}(k)\mathbf{H}_{\text{TOT},1}^*(-k) = \mathbf{0}$ or

$$\mathbf{W}_{\text{OPT}}(k) = -\mathbf{H}_{\text{TOT},2}(k)(\mathbf{H}_{\text{TOT},1}^*(-k))^+ \quad (16.38)$$

where in general $\mathbf{A}^+$ denotes the pseudo-inverse of matrix $\mathbf{A}$. This is the multi dimensional version of the earlier single-antenna optimum solution in (16.33).

It should be noted that both the MIMO-OFDM and OFDM link models described above are still simplified models. Certain other physical impairments, namely carrier frequency offsets (CFO), phase noise, and Doppler effects of the mobile radio channel, can create additional ICI, breaking the exact mirror-carrier structure in (16.39) and (16.32). More complex modeling and mitigation approaches may be needed in these cases, depending on the severity of these additional impairments in the system under study. The reader is referred to [23]–[27] for approaches to CFO, to [28] for Doppler in conjunction with I/Q imbalance in OFDM systems, to [29], [30] for details on CFO, and to [31], [32] for phase noise in conjunction with I/Q imbalance in MIMO-OFDM systems.

Low-IF RX-specific models

Most of the above developments have been assuming, explicitly or implicitly, direct-conversion type radio. If, on the other hand, low-IF type radio architecture is used, the observed signals at positive and negative IFs will interfere with each other. As shown originally in [10], [16], [33], assuming the frequency-independent imbalance case for simplicity, the resulting complex baseband observations of the positive and negative IFs, denoted here by $\bar{z}_{pos}(t)$ and $\bar{z}_{neg}(t)$, are of the form

$$\begin{aligned}\bar{z}_{pos}(t) &= K_1 s(t) + K_2 i^*(t) \\ \bar{z}_{neg}(t) &= K_2^* s(t) + K_1^* i^*(t)\end{aligned}\quad (16.40)$$

where $s(t)$ and $i(t)$ denote the baseband equivalents of the original RF carriers, centered at $f_{LO} + f_{IF}$ and $f_{LO} - f_{IF}$, respectively. Given the imbalance knowledge, these signals can then be separated as

$$\begin{aligned}\hat{s}(t) &= \frac{1}{|K_1|^2 - |K_2|^2} [K_1^* \bar{z}_{pos}(t) - K_2 \bar{z}_{neg}(t)] \\ \hat{i}(t) &= \frac{1}{|K_1|^2 - |K_2|^2} [K_1^* \bar{z}_{neg}^*(t) - K_2 \bar{z}_{pos}^*(t)]\end{aligned}\quad (16.41)$$

as is easy to verify by a direct substitution of (16.40) into (16.41). It is also interesting to note that given a relatively big power difference between the two signals, say $i(t)$ is much stronger compared to $s(t)$, and assuming practical imbalance levels for which $|K_1|^2 \gg |K_2|^2$, then it follows that $\bar{z}_{neg}(t) \simeq K_1^* i^*(t)$ and thus

$$\hat{s}(t) \simeq \bar{z}_{pos}(t) - \frac{K_2}{K_1^*} \bar{z}_{neg}(t) = \bar{z}_{pos}(t) + w_{OPT} \bar{z}_{neg}(t) \quad (16.42)$$

where $w_{OPT} = -K_2/K_1^*$. Such developments can be found, e.g., in [33].

16.4.4 Training or pilot-signal based imbalance estimation methods

RX-based estimation techniques

Many of the practical compensation developments in the literature focus on training signal based estimation of the previously defined optimum compensators, especially in the case of OFDM waveforms with built-in pilot subcarriers and assuming receiver I/Q imbalances

only. In such a scenario, one possible approach to carry out the estimation is to view the basic signal model $\bar{\mathbf{Z}}_k = \mathbf{G}_{1,k}\mathbf{Z}_k + \mathbf{G}_{2,k}\mathbf{Z}_{-k}^*$ as an augmented linear model of the form $\bar{\mathbf{Z}}_k = \mathbf{G}_k^T \mathbf{Z}_k$ where $\mathbf{G}_k^T = [\mathbf{G}_{1,k} \ \mathbf{G}_{2,k}]$ and $\mathbf{Z}_k^T = [\mathbf{Z}_k \ \mathbf{Z}_{-k}^*]$. Utilizing pilot data, the imbalance parameters can be identified jointly with the channel response using, e.g., least squares type model fitting or minimum mean square error (MMSE) estimation. Such work has been reported, e.g., in [34]–[39] for estimation of RX induced I/Q imbalances and the channel, and for example in [40]–[42] for joint estimation of both TX and RX induced imbalances along with the channel. Extensions to these approaches, taking into account also CFO, are detailed in [23]–[27], [43]. Extensions to multi-antenna MIMO-OFDM systems are then considered in [29]–[32], [44]–[46], focusing on the spatial multiplexing case along with different mixtures of RX and TX I/Q imbalances, CFO, and phase noise, as well as in [47]–[49] considering the space-time coded (STC) multi-antenna OFDM case. More general pilot-based formulations in the *time-domain* have been proposed in [12], [50] for the single-antenna case, and in [30] for a general MIMO receiver.

TX-internal imbalance estimation

The above developments have mostly focused on the receiver side of the radio link. Another interesting ingredient is to consider the estimation of the transmitter calibration parameters formulated in (16.25) inside the transmitting device. Compared to receiver processing, this is a different type of estimation problem since inside a transmitter the ideal complex envelope is known since it is being generated in the device itself. For parameter estimation purposes, however, some reference to the actual generated RF waveform is needed, since the impairments at hand are related to the analog/RF modules. One possible approach is then to have some kind of feedback chain from the RF stage back to lower frequencies, and thereon back to the transmitter digital front-end, to estimate the calibration parameters. Such approaches have been described e.g. in [14], [51]–[53]. Given that the feedback loop is free from “excess” I/Q imbalances, this feedback signal is structurally of the form [52]

$$\begin{aligned} z_{fb}(t) &= G_{fb} e^{j\theta_{fb}} h_{fb}(t) * \bar{\mathbf{z}}_{TX}(t - \tau_{fb}) \\ &= G_{fb} e^{j\theta_{fb}} h_{fb}(t) * [g_{1,TX}(t) * z(t - \tau_{fb}) + g_{2,TX}(t) * z^*(t - \tau_{fb})] \\ &= \tilde{g}_{1,TX}(t) * z(t - \tau_{fb}) + \tilde{g}_{2,TX}(t) * z^*(t - \tau_{fb}) \end{aligned} \quad (16.43)$$

where G_{fb} , θ_{fb} , $h_{fb}(t)$, and τ_{fb} denote the gain, phase, impulse response, and delay of the feedback path, respectively, relative to the main transmit path. In (16.43), $\tilde{g}_{1,TX}(t) = G_{fb} e^{j\theta_{fb}} h_{fb}(t) * g_{1,TX}(t)$ and $\tilde{g}_{2,TX}(t) = G_{fb} e^{j\theta_{fb}} h_{fb}(t) * g_{2,TX}(t)$ denote the *observable imbalance filters* which contain also the feedback loop response. Now, as described in more detail in [52], certain cross-correlations between the feedback signal $z_{fb}(t)$ and the ideal transmit data $z(t)$ give a basis for identifying the observable imbalance model filters $\tilde{g}_{1,TX}(t)$ and $\tilde{g}_{2,TX}(t)$, which in turn can be used to solve for the optimum solution in (16.25). This is formulated here as

$$\begin{aligned} R_{fb,1}(\lambda) &= E[z_{fb}(t)z^*(t - \lambda)] = \tilde{g}_{1,TX}(\lambda) * R_z(\lambda - \tau_{fb}) \\ R_{fb,2}(\lambda) &= E[z_{fb}(t)z(t - \lambda)] = \tilde{g}_{2,TX}(\lambda) * R_z(\tau_{fb} - \lambda) \end{aligned} \quad (16.44)$$

where $R_z(\lambda) \triangleq E[z(t)z^*(t-\lambda)]$ denotes the ordinary auto-correlation function of the ideal transmit signal. These two correlations can then be used to solve for the optimum solution in (16.25) as described in [52]. Notice that $R_z(\lambda)$ is basically known, being determined by the type of used waveforms (modulation, etc.). Above, we have further assumed that the ideal transmit signal $z(t)$ is a circular random signal for which the *complementary auto-correlation* vanishes, i.e. $\forall \lambda: C_z(\lambda) \triangleq E[z(t)z(t-\lambda)] = 0$, as defined in [54]–[56]. This assumption can be shown to hold for most complex communications waveforms of practical interest, such as M -QAM and M -PSK signals with $M > 2$, and general multi-carrier signals (including OFDM) with individually circular carrier signals. In practice, the above correlations are evaluated using sample correlations over a processing block of L samples.

Another interesting, and possibly yet more effective way, to utilize the feedback signal in for parameter estimation is to do least squares (LS) type model fitting. Such model-fitting methods for pre-distorter coefficient estimation are described in detail in [52], [53]. Notice that the generalization of such methods for joint estimation and digital pre-distortion of I/Q modulator and power amplifier impairments are also described in Chapter 17 of this book by Anttila and Valkama.

Now, considering I/Q modulator pre-distortion only, we first switch to discrete-time signals and vector-matrix notation, for better describing the least squares parameter estimation. The feedback signal vector is now of the form

$$\mathbf{z}_{fb}(t) = \mathbf{Z}(t)\tilde{\mathbf{g}}_{1,TX} + \mathbf{Z}^*(t)\tilde{\mathbf{g}}_{2,TX} = [\mathbf{Z}(t) \quad \mathbf{Z}^*(t)] \begin{bmatrix} \tilde{\mathbf{g}}_{1,TX} \\ \tilde{\mathbf{g}}_{2,TX} \end{bmatrix} = \mathbf{Z}_{block}(t) \begin{bmatrix} \tilde{\mathbf{g}}_{1,TX} \\ \tilde{\mathbf{g}}_{2,TX} \end{bmatrix}. \quad (16.45)$$

In the above, $\tilde{\mathbf{g}}_{1,TX}$ and $\tilde{\mathbf{g}}_{2,TX}$ are the $N_g \times 1$ vector representations of the observable (but unknown) imbalance filters, $\mathbf{Z}(t)$ is the convolution matrix formed from the transmit data samples (samples of $z(t)$), $\mathbf{z}_{fb}(t) \triangleq [z_{fb}(t), z_{fb}(t-1), \dots, z_{fb}(t-L+1)]^T$, and L is the length of the data block used. Now the observable imbalance filters can be estimated using the least squares model fitting principle as

$$\begin{bmatrix} \hat{\tilde{\mathbf{g}}}_{1,TX} \\ \hat{\tilde{\mathbf{g}}}_{2,TX} \end{bmatrix} = \mathbf{Z}_{block}^+(t) \mathbf{z}_{fb}(t), \quad (16.46)$$

where $\mathbf{Z}_{block}^+(t)$ represents the pseudo-inverse of $\mathbf{Z}_{block}(t)$, which is given by $\mathbf{Z}_{block}^+(t) = (\mathbf{Z}_{block}^H(t)\mathbf{Z}_{block}(t))^{-1}\mathbf{Z}_{block}^H(t)$ assuming $\mathbf{Z}_{block}(t)$ has linearly independent columns such that the inverse of $\mathbf{Z}_{block}^H(t)\mathbf{Z}_{block}(t)$ exists. The convolution matrix of the input data $\mathbf{Z}(t)$, in turn, has the following form

$$\mathbf{Z}(t) = \begin{bmatrix} z(t-N_g+1) & z(t-N_g+2) & \dots & z(t) \\ z(t-N_g) & z(t-N_g+1) & \dots & z(t-1) \\ \vdots & \vdots & \ddots & \vdots \\ z(t-L+1) & z(t-L) & \dots & z(t-L-N_g) \end{bmatrix}. \quad (16.47)$$

Finally, the pre-distortion filter is obtained by solving $\hat{\tilde{\mathbf{g}}}_{1,TX}(t) * w(t) + \hat{\tilde{\mathbf{g}}}_{2,TX}(t) = 0$. In vector-matrix notation, the solution is given by

$$\mathbf{w} = - \left(\hat{\mathbf{G}}_{1,TX}^H \hat{\mathbf{G}}_{1,TX} \right)^{-1} \hat{\mathbf{G}}_{1,TX}^H \hat{\mathbf{g}}_{2,TX}^0, \quad (16.48)$$

where $\hat{\mathbf{G}}_{1,TX}$ is the convolution matrix formed from $\hat{\mathbf{g}}_{1,TX}$ and $\hat{\mathbf{g}}_{2,TX}^0 = [\hat{\mathbf{g}}_{2,TX}^T, 0, \dots, 0]^T$ is a zero-padded version of $\hat{\mathbf{g}}_{2,TX}$, with $N_g - 1$ zeros attached to its end. The final pre-distortion filter (vector), of say length N_w , is obtained as the first N_w elements of (16.48).

16.4.5 Blind or non-data-aided imbalance estimation methods

Compared to the previous data-aided techniques, a completely different approach to imbalance estimation and compensation is to utilize the statistical properties of the communications waveforms used. In other words, the objective is to devise non-data-aided or blind imbalance estimation methods where no explicit knowledge of the transmit waveform samples is needed. This has been a topic under active research during the last 5–10 years, and two different general themes have been established – one based on the statistics of the I and Q signals and another one based on the statistics of the corresponding complex signals. These are addressed shortly below, with emphasis on the so-called re-circularization techniques developed by the authors of this chapter in [13], [57]–[59]. The work in [17] is related to these techniques by implicitly relying on the circularity feature. In general, such statistical methods have most deployment scenarios on the receiver side of the link where, by default, the exact received waveform is unknown due to channel noise, multipath, and lack of synchronization information, etc.

The techniques utilizing parallel I/Q signal models given in (16.20)–(16.21), combined with the corresponding I/Q signal statistics, build on the general fact that I/Q imbalances create correlation (or statistical dependence more generally) between the observed I and Q signals. Different second-order statistics based techniques to remove this correlation are proposed in [19], [21], shown to be robust also against for example CFO. Higher-order statistics based techniques in the form of blind I/Q signal separation (BSS) are in turn proposed in [20], also taking into account the CFO and channel equalization issues.

At complex signal level, various statistical techniques have been proposed for both low-IF and direct-conversion type radio architectures. In the low-IF case, one well-established approach is to down-convert the signals from both positive and negative IFs to baseband in the receiver digital front-end, as discussed earlier in this chapter. Using the IF observations given in (16.40), some signal estimation type blind imbalance compensation schemes stemming from adaptive interference cancellation (AIC) or blind signal separation (BSS) principles are proposed, e.g., in [10], [16], [18], and [33] for both frequency-independent and frequency-dependent imbalance scenarios. An alternative approach is described in [60] where the cross-correlation of $\bar{z}_{pos}(t)$ and $\bar{z}_{neg}(t)$ together with the power of $\bar{z}_{pos}(t) + \bar{z}_{neg}(t)$ are used in estimating the optimum receiver coefficient $w_{RX}^{OPT} = -K_{2,RX}/K_{1,RX}^*$. Similar techniques are also proposed in the OFDM direct-conversion receiver context in [61] where the cross-correlation of the mirror-carrier observations $\bar{Z}_k = K_1 Z_k + K_2 Z_{-k}^*$ and $\bar{Z}_{-k}^* = K_2^* Z_k + K_1^* Z_{-k}^*$ combined with the power of $\bar{Z}_k + \bar{Z}_{-k}^*$ are utilized to estimate $w_{RX}^{OPT} = -K_{2,RX}/K_{1,RX}^*$.

The above techniques utilizing the cross-correlations of the mirror-frequency components can in general be seen as special cases of the so-called circular or proper nature of

complex communications waveforms, being then degraded by I/Q imbalances. This approach is taken explicitly in [57], [13] assuming frequency-independent I/Q imbalances and then generalized to cover the frequency-dependent I/Q imbalance cases and also both direct-conversion and low-IF radio architectures in [58], [59]. In general, as briefly discussed earlier in this chapter, a circular or proper complex random signal is defined as a signal whose complementary auto-correlation $C_z(\tau)$ equals zero at all time-lags, i.e., [54]–[56]

$$\forall \tau: C_z(\tau) \stackrel{\Delta}{=} E[z(t)z(t-\tau)] = 0 = E[z(t)(z^*(t-\tau))^*]. \quad (16.49)$$

It is well established in the literature that this feature holds for most complex communications waveforms, but only under perfect I/Q balance. Based on the expression on the right-hand side of (16.49), circularity can also be interpreted as $z(t)$ being uncorrelated against its own complex conjugate which obviously (at least at zero lag) can hold only for complex-valued quantities in general. Now, based on the general receiver imbalance model given in (16.15), the ordinary and complementary correlation functions $R_{\bar{z},RX}(\tau)$ and $C_{\bar{z},RX}(\tau)$ for the imbalanced signal $\bar{z}_{RX}(t)$ read

$$\begin{aligned} R_{\bar{z},RX}(\tau) &\stackrel{\Delta}{=} E[\bar{z}_{RX}(t)\bar{z}_{RX}^*(t-\tau)] \\ &= g_{1,RX}(\tau) * g_{1,RX}^*(-\tau) * R_z(\tau) + g_{2,RX}(\tau) * g_{2,RX}^*(-\tau) * R_z(-\tau) \\ C_{\bar{z},RX}(\tau) &\stackrel{\Delta}{=} E[\bar{z}_{RX}(t)\bar{z}_{RX}(t-\tau)] \\ &= g_{1,RX}(\tau) * g_{2,RX}(-\tau) * R_z(\tau) + g_{1,RX}(-\tau) * g_{2,RX}(\tau) * R_z(-\tau). \end{aligned} \quad (16.50)$$

Thus since $C_{\bar{z},RX}(\tau) \neq 0$, at least for some lags around and including zero, the imbalanced signal is in general non-circular. To give a concrete example, consider the frequency-independent imbalance case for which $C_{\bar{z},RX}(\tau) = 2K_{1,RX}K_{2,RX}\sigma_z^2 \neq 0$ where σ_z^2 denotes the power of $z(t)$. Now, as shown in detail in [59], the output signal of a general post-distortion compensator $z_{post}(t) = \bar{z}_{RX}(t) + w_{RX}(t) * \bar{z}_{RX}^*(t)$ can only fulfill the circularity condition $C_{z,post}(\tau) = 0$ if $W_{RX}(f) = W_{RX}^{OPT}(f) = -G_{2,RX}(f)/G_{1,RX}^*(-f)$ (the desired solution) or $W_{RX}(f) = [1/W_{RX}^{OPT}(f)]^*$ (the “mirror” solution). Thus compensation can be based on “projecting” the received signal back to the circular domain.

Different adaptive filtering as well as block-processing based solutions to do the re-circularization in practice are described in detail in [13], [57]–[59]. We will only briefly review the algorithms proposed in [58] and [59], and the reader is referred to the original articles for more complete descriptions and discussions.

Moment-based estimator for frequency-dependent RX I/Q imbalance compensation utilizing the circularity condition

Now, two solutions for frequency-dependent RX I/Q imbalance compensation through the re-circularization of the received, imbalanced, signal $\bar{z}_{RX}(t)$ are given. They were originally proposed by the authors of this chapter in [58] and [59]. The first detailed method is a block-processing approach based on estimation of the ordinary and complementary autocorrelations of $\bar{z}_{RX}(t)$, while the second one is an adaptive algorithm based on the explicit circularization of the compensator output signal.

For notational convenience, we will use vector-matrix notations, and write the compensator output signal as $z_{post}(t) = \bar{z}_{RX}(t) + \mathbf{w}^T \bar{\mathbf{z}}_{RX}^*(t)$, where $\mathbf{w} \triangleq [w_1, w_2, \dots, w_{N_W}]^T$ denote the N_W tap coefficients of the compensator filter and $\bar{\mathbf{z}}_{RX}(t) \triangleq [\bar{z}_{RX}(t) \ \bar{z}_{RX}(t-1) \ \dots \ \bar{z}_{RX}(t-N_W+1)]^T$ denote the corresponding signal samples. The objective here is to null the complementary autocorrelation of the compensator output $z_{post}(t)$ for the span of the compensation filter (N_W samples), i.e., $E[\mathbf{z}_{post}(t)\mathbf{z}_{post}(t)] = \mathbf{0}$. Expanding this set of equations, and writing it as a function of the *compensator input statistics*, we obtain a new set of equations

$$\mathbf{c}_{\bar{z}} + \mathbf{R}_{\bar{z}}\mathbf{w} + \bar{\mathbf{R}}_{\bar{z}}\mathbf{w} + \mathbf{W}\mathbf{C}_{\bar{z}}^*\mathbf{w} = \mathbf{0}, \quad (16.51)$$

where $\mathbf{c}_{\bar{z}} \triangleq E[\bar{\mathbf{z}}_{RX}(t)\bar{\mathbf{z}}_{RX}(t)] = [C_{\bar{z},RX}(0) \ C_{\bar{z},RX}(1) \ \dots \ C_{\bar{z},RX}(N_W-1)]^T$, with $\bar{z}_{RX}(t) = [\bar{z}_{RX}(t) \ \bar{z}_{RX}(t-1) \ \dots \ \bar{z}_{RX}(t-N_W+1)]^T$, and

$$\mathbf{R}_{\bar{z}} \triangleq \begin{bmatrix} R_{\bar{z},RX}(0) & R_{\bar{z},RX}(1) & \dots & R_{\bar{z},RX}(N_W-1) \\ R_{\bar{z},RX}^*(1) & R_{\bar{z},RX}^*(0) & \dots & R_{\bar{z},RX}^*(N_W-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_{\bar{z},RX}^*(N_W-1) & R_{\bar{z},RX}^*(N_W-2) & \dots & R_{\bar{z},RX}^*(0) \end{bmatrix}, \quad (16.52)$$

$$\bar{\mathbf{R}}_{\bar{z}} \triangleq \begin{bmatrix} R_{\bar{z},RX}(0) & R_{\bar{z},RX}(1) & \dots & R_{\bar{z},RX}(N_W-1) \\ R_{\bar{z},RX}(1) & R_{\bar{z},RX}(2) & \dots & R_{\bar{z},RX}(N_W) \\ \vdots & \vdots & \ddots & \vdots \\ R_{\bar{z},RX}(N_W-1) & R_{\bar{z},RX}(N_W) & \dots & R_{\bar{z},RX}(2N_W-2) \end{bmatrix}, \quad (16.53)$$

$$\mathbf{C}_{\bar{z}} \triangleq \begin{bmatrix} C_{\bar{z},RX}(0) & C_{\bar{z},RX}(1) & C_{\bar{z},RX}(2) & \dots & C_{\bar{z},RX}(N_W-1) \\ C_{\bar{z},RX}(1) & C_{\bar{z},RX}(0) & C_{\bar{z},RX}(1) & \dots & C_{\bar{z},RX}(N_W-2) \\ C_{\bar{z},RX}(2) & C_{\bar{z},RX}(1) & C_{\bar{z},RX}(0) & \dots & C_{\bar{z},RX}(N_W-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{\bar{z},RX}(2N_W-2) & C_{\bar{z},RX}(2N_W-3) & C_{\bar{z},RX}(2N_W-4) & \dots & C_{\bar{z},RX}(N_W-1) \end{bmatrix}, \quad (16.54)$$

and

$$\mathbf{W} \triangleq \begin{bmatrix} \mathbf{w}^T & 0 & \dots & 0 \\ 0 & \mathbf{w}^T & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{w}^T \end{bmatrix}. \quad (16.55)$$

The matrix $\mathbf{C}_{\bar{z}}$ has dimensions $(2N_W-1) \times N_W$, while matrix $\mathbf{W}$, which is constructed from the compensator filter coefficients, is of size $N_W \times (2N_W-1)$. Thus, altogether $2N_W-1$ values of the ordinary and complementary correlation functions of the compensator input signal $\bar{z}_{RX}(t)$ are needed to form the matrices.

In general, there is no closed-form solution to (16.51), but an *approximate* closed-form solution can be obtained by ignoring the last term on the left-hand side of $(\mathbf{W}\mathbf{C}_{\bar{z}}^*\mathbf{w})$, which

is, with any practical imbalance values, very small compared to the other three terms. Doing this, we can obtain the compensator filter as

$$\mathbf{w} = -(\mathbf{R}_{\bar{\mathbf{z}}} + \bar{\mathbf{R}}_{\bar{\mathbf{z}}})^{-1} \bar{\mathbf{c}}_{\bar{\mathbf{z}}}. \quad (16.56)$$

The accuracy of this approximate solution depends essentially on the magnitude of $\mathbf{W}\mathbf{C}_{\bar{\mathbf{z}}}^*\mathbf{w}$ relative to the magnitude of the other terms, and is thus directly dependent on the severity of the I/Q imbalance. If we consider the frequency-independent case with $g_{1,RX}(t) = K_{1,RX}\delta(t)$, $g_{2,RX}(t) = K_{2,RX}\delta(t)$, and $N_W = 1$, we can easily conclude from (16.51) and the discussion following it that $\|\mathbf{W}\mathbf{C}_{\bar{\mathbf{z}}}^*\mathbf{w}\|$ grows *cubically* with $|K_{2,RX}|$, while the magnitudes of the other three terms only grow *linearly* with $|K_{2,RX}|$. Thus, the smaller $|K_{2,RX}|$ is (the higher the IRR), the better the approximate solution in (16.56). In fact, it was shown in [13] that for the frequency-independent case there is an exact relation between the front-end IRR and the IRR obtained with the approximate solution, namely the dB value of the front-end IRR will be tripled, assuming the statistics are estimated correctly.

The statistics needed for (16.51)–(16.56) will in practice be estimated from the incoming data. The most straight forward approach is to use the traditional sample estimates of the needed moments as $\hat{R}_{\bar{\mathbf{z}},RX}(k) = \frac{1}{L}\sum_{l=1}^L \bar{z}_{RX}(l)\bar{z}_{RX}^*(l-k)$ and $\hat{C}_{\bar{\mathbf{z}},RX}(k) = \frac{1}{L}\sum_{l=1}^L \bar{z}_{RX}(l)\bar{z}_{RX}(l-k)$, $k = 0, 1, \dots, 2N_W - 2$, where L denotes the estimator block-size. The finite accuracy of the sample statistics will affect the achievable performance to some extent, depending essentially on the block-size L . Numerical examples will be given in Section 16.5.2.

Frequency-dependent RX I/Q imbalance compensation based on iterative re-circularization

A practical blind adaptive algorithm for re-circularizing the received and imbalanced signal is given as follows. We denote by $\mathbf{w}_t \triangleq [w_{t,1}, w_{t,2}, \dots, w_{t,N_W}]^T$ the N_W tap coefficients of the compensator filter at time instant t . The algorithm is iterated for $t = 0, 1, 2, \dots$ as

$$\begin{aligned} z_{post}(t) &= \bar{z}_{RX}(t) + \mathbf{w}_t^T \bar{\mathbf{z}}_{RX}^*(t) \\ \mathbf{w}_{t+1} &= \mathbf{w}_t - \lambda \mathbf{z}_{post}(t) \mathbf{z}_{post}^*(t) \end{aligned} \quad (16.57)$$

where λ denotes the adaptation step-size and the compensator output vector is $\mathbf{z}_{post}(t) \triangleq [z_{post}(t) \ z_{post}(t-1) \ \dots \ z_{post}(t-N_W+1)]^T$ [58], [59]. This algorithm converges towards the circular output signal by design, and by initializing the compensator coefficients as $\mathbf{w}_0 = \mathbf{0}$ guarantees that the iteration converges towards $W_{RX}^{OPT}(f)$ instead of the mirror solution $[1/W_{RX}^{OPT}(f)]^*$ [59]. The exact convergence speed and the obtained steady-state estimation performance depend on the applied step-size value. Further improvements to the steady-state operation (accuracy) can be obtained by building coefficient smoothing or averaging inside the basic algorithm in (16.57), as also described in [59]. Numerical illustrations will again be given in Section 16.5.2.

16.5 Design and testing examples

16.5.1 Direct-conversion transmitter I/Q calibration with widely linear least squares model fitting based method

In this section, we demonstrate the performance of the compensation algorithm with computer simulations and with a practical laboratory radio signal. Both zero-IF and low-IF input signals are considered. In all the simulations and measurements, we keep the imbalance estimator filter length N_g fixed to 3-taps and evaluate the results for pre-distortion filter of lengths (N_W) 1, 2, and 3-taps.

The simulation example corresponds to a zero-IF transmitter with 64-QAM single-carrier signal with 9 MHz bandwidth and 40 MHz sample rate in the digital front-end. The I/Q mixer imbalance values used in the simulations are $g = 1.03$ and $\phi = 3^\circ$, and the impulse responses of frequency selective I and Q branches are $\mathbf{h}_I = [0.99, -0.1]^T$ and $\mathbf{h}_Q = [0.98, -0.07]^T$, respectively. This represents a practical case and corresponds to an overall IRR of approximately 25–35 dB without compensation, as shown in Figure 16.8 (front-end). The achieved IRR including pre-distortion is also plotted in the figure as a function of frequency after averaging 100 independent simulation runs. In this simulation, in the feedback loop, we use 10 000 samples of the original and imbalanced signals to calculate the predistortion filter coefficients using the least squares method described earlier. The simulation results indicate a clear improvement in the obtained mirror frequency attenuation for a pre-distortion filter of 3-taps, reaching an IRR level in the order of 75–80 dB, which is surely enough for any practical transmitter deployment.

In the measurement example, we use standard state-of-the-art signal generation and analysis devices in a radio signal laboratory environment. The transmitter topology used

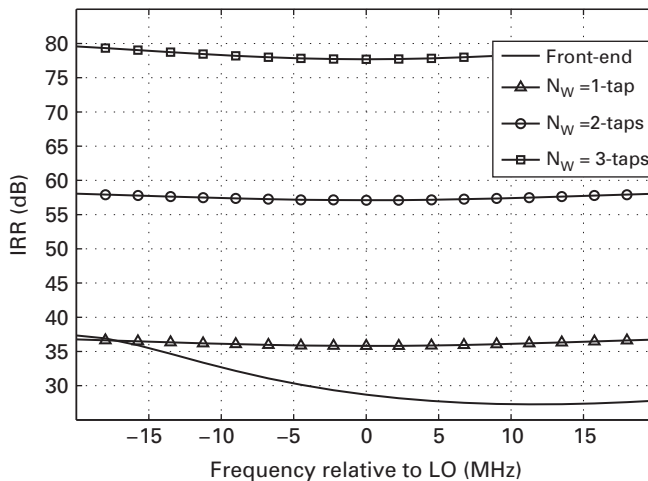


Figure 16.8 Simulated IRR vs. frequency curves for the direct-conversion transmitter case with different pre-distortion filters using the widely linear least squares model fitting approach.

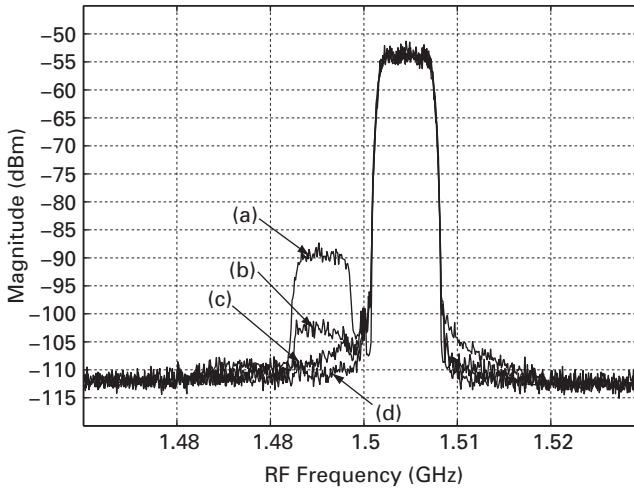


Figure 16.9 Measured I/Q modulator output without and with I/Q pre-distortion, (a) without pre-distortion (b) with 1-tap pre-distortion filter, (c) with 2-tap pre-distortion filter, (d) with 3-tap pre-distortion filter.

is low-IF transmitter with 7.5 MHz waveform bandwidth and 4.5 MHz transmitter IF. The modulation method used is again 64-QAM with root-raised cosine pulse-shaping. The basic sample rate in the transmitter digital front-end is 72 MHz. The actual RF frequency range is 1.5 GHz. A total of 20 000 samples of the original and feedback data are used for pre-distortion coefficient estimation. The obtained transmitter RF spectrum is depicted in Figure 16.9. Clearly, the mirror-frequency component is well-attenuated, reaching 60 dB measured IRR level with the 3-tap pre-distorter.

16.5.2 Blind circularization-based methods in a dual-carrier direct-conversion receiver

Now, simulation examples are given for the blind circularity-based algorithms presented in Section 16.4.5. These results are reproduced from [59]. The simulation scenario assumed here is a 2-carrier low-IF receiver operating at 2 GHz RF range. Both carriers have 16-QAM data modulation, ordinary raised-cosine pulse-shaping, and 25 percent roll-off. A symbol rate of 3.84 MHz complying with the 3GPP specifications is assumed, yielding individual waveform bandwidths of roughly 5 MHz. In the receiver analog front-end, the two incoming carriers are I/Q downconverted symmetrically around DC, being then located at ± 3 MHz IFs, and sampled at a sampling rate of 4×3.84 MHz. The simulation model contains also independent fading multipath channels for the two carrier signals, with their power delay profiles following the Extended Vehicular A model discussed in [62]. A mobility of 120 km/h is assumed as a challenging example case. The channel model also includes additive white Gaussian noise (AWGN), with the weaker carrier having an effective in-band signal-to-noise ratio (SNR) of 10 dB. The receiver analog front-end model has frequency-dependent I/Q imbalances, with the resulting IRR varying smoothly between 27 and 33 dB. For comparison purposes, both 1-tap and 3-tap versions of the algorithms are implemented.

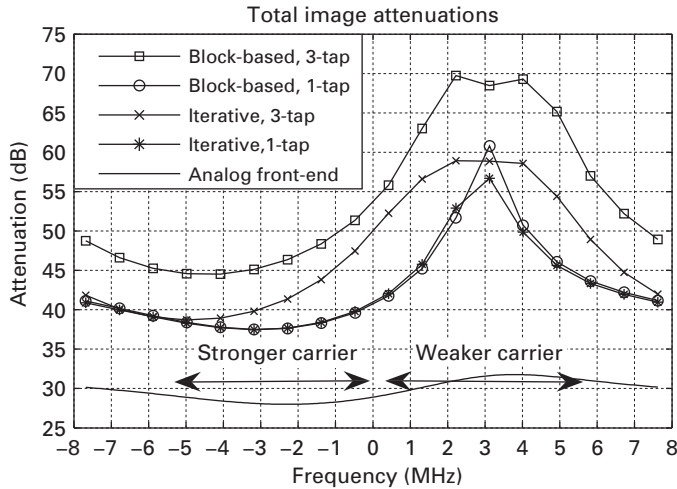


Figure 16.10 Total image attenuations of the circularity-based iterative and block algorithms, with 1-tap and 3-tap compensation filters, in a two-carrier low-IF receiver. Both signals are 16-QAM, and they have an RF power difference of 30 dB. Extended Vehicular A (120 km/h) channel and received SNR = 10 dB. Frequency-dependent analog front-end I/Q imbalances, resulting in IRR varying between 27–33dB. © 2008 IEEE.

Figure 16.10 shows the obtained image attenuation curves of the iterative compensator in (16.57) and the block-compensator in (16.56), plotted as a function of frequency, and assuming 30 dB RF power difference between the two carriers. Coefficient smoothing for the iterative algorithm is also included to illustrate its effects (see [59] for details). Also shown is the corresponding IRR curve of the analog front-end alone. The IRR curves shown represent an average value obtained in 100 independent simulation runs (ensemble averaging). All algorithms utilize 50 000 samples for estimation, and the step-sizes of the iterative algorithms are chosen such that the coefficients have converged in this time. The first important observation is that the obtained image attenuation is clearly better with 3-tap compensators, compared to the 1-tap algorithms. This is indeed due to the frequency-selective nature of the imbalances, which the 1-tap compensators do not account for. A very desirable feature of the compensators is that they are tuning most of the attenuation to those frequencies which experience the *most* mirror-frequency interference, which in this case is the band of the weaker carrier. Another key observation is that while both the iterative and block approaches are providing very high IRR at the weaker carrier band, the performance of the block algorithm is still considerably better. Thus in this sense, the block algorithm is using the available data more efficiently. Even though this is somewhat dependent on the step-size of the adaptive algorithm, these kinds of conclusions are rather typical in the estimation literature, when the parameters (here imbalances) are *not* changing in time.

Next, we simulate the 3-tap iterative and block-based algorithms, and vary the RF power difference of the two carriers between 0 dB and 50 dB. Otherwise, the simulation scenario is identical to the previous one. Again, all the step-sizes of the adaptive

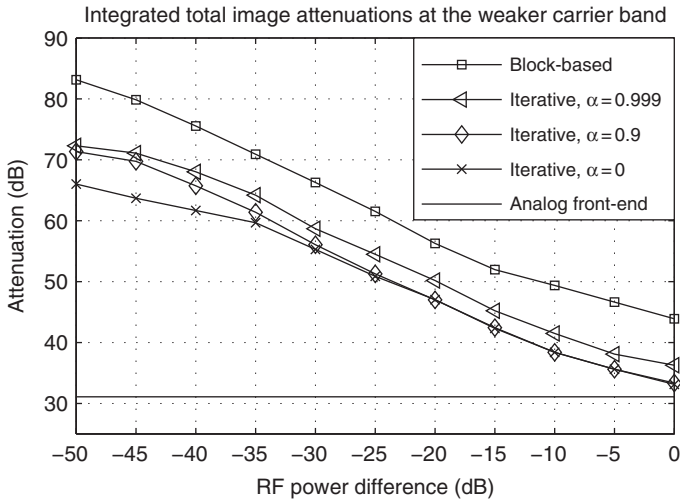


Figure 16.11 Integrated total in-band image attenuations at the weaker carrier band in a two-carrier receiver with 16-QAM data modulation, Extended Vehicular A (120 km/h) channel model and received SNR = 10 dB. The iterative and block-processing based methods (3-tap compensators) are simulated with varying the RF power difference between -50 and 0 dBs; α denotes the coefficient smoothing parameter in the iterative solutions. Frequency-dependent analog front-end I/Q imbalances, resulting in 31 dB integrated image attenuation at the weaker carrier band. © 2008 IEEE.

algorithms are selected such that similar convergence properties are obtained, and the overall number of received samples used for the adaptation and/or estimation is 50 000. The obtained results in terms of the *total integrated IRR* within the weaker carrier band are shown in Figure 16.11. The most important observation is that the overall IRR is increasing with increasing power difference between the signals. Thus essentially, the algorithms are able to keep a relatively constant *signal-to-interference ratio* (SIR) at the compensator output. Again, the iterative algorithm gives a consistently lower level of performance compared to the block algorithm. The effect of coefficient smoothing (α , see [59] for details) is also clearly observed in Figure 16.11. With $\alpha = 0.999$, the integrated IRR is improved by 3–5 dB's compared to $\alpha = 0$ (the basic algorithm in (16.57)), depending on the RF power difference.

16.6 Conclusions

In this chapter, an overview of one central analog/RF impairment, called I/Q imbalance, encountered in radio transceiver design and implementation was given. Essential behavioral models, including both frequency-independent as well as frequency-dependent I/Q imbalance cases, were given. Frequency-dependent characteristics were shown to be essential in many emerging wideband radio developments where processed bandwidths are in the order of several tens of MHz. Stemming from the developed behavioral models, efficient digital signal processing based imbalance estimation and calibration methods

were then presented, with most focus on the frequency-dependent case. Both digital pre-distortion type methods, applicable on the transmitter side, as well as digital post-correction type methods, applicable on the receiver side, were developed and reviewed. The presentation also included both data-aided as well as non-data-aided or blind imbalance estimation techniques. Furthermore, some link-level models and processing principles were also highlighted, incorporating imbalanced transmitter, multipath radio channel, and imbalanced receiver. Also multi-antenna extensions implementing spatial multiplexing under imbalanced radios were covered. The chapter also presented some practical design and testing examples for both transmitter calibration and receiver post-compensation, including simulated as well as measured radio signals. Altogether these examples demonstrated the power and potential that efficient DSP solutions can offer for relaxing and relieving the analog/RF circuit design and implementation burden. One interesting alternative, in addition to pure digital solutions, is digitally assisted hybrid analog/digital calibration with information exchange and control mechanisms between the radio digital front-end and analog circuit modules. It is envisaged that such work will make significant advances in the coming years and this work forms an interesting and important topic for future research.

References

- [1] G. Fettweis, M. Löhning, D. Petrovic, *et al.*, “Dirty RF: a new paradigm,” *Springer Int. J. Wireless Inf. Netw.*, vol. **14**, pp. 133–148, June 2007.
- [2] M. Valkama, A. Springer, and G. Hueber, “Digital signal processing for reducing the effects of RF imperfections in radio devices – An overview,” *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS’10)*, Paris, France, May–June 2010, pp. 813–816.
- [3] P.-I. Mak, S.-P. U, and R. P. Martins, “Transceiver architecture selection: review, state-of-the-art survey and case study,” *IEEE Circuits Syst. Mag.*, vol. **7**, pp. 6–25, second quarter 2007.
- [4] S. Mirabbasi and K. Martin, “Classical and modern receiver architectures,” *IEEE Commun. Mag.*, vol. **38**, pp. 132–139, Nov. 2000.
- [5] B. Razavi, “Design considerations for direct-conversion receivers,” *IEEE Trans. Circuits Syst. II*, vol. **44**, pp. 428–435, June 1997.
- [6] J. Crols and M. S. J. Steyaert, *CMOS Wireless Transceiver Design*, Dordrecht: Kluwer, 1997.
- [7] X. Li and M. Ismail, *Multi-Standard CMOS Wireless Receivers*. Norwell, MA: Kluwer, 2002.
- [8] W. Tuttlebee, ed., *Software Defined Radio: Enabling Technologies*. Chichester: Wiley, 2002.
- [9] M. Brandolini *et al.*, “Toward multistandard mobile terminals – fully integrated receivers requirements and architectures,” *IEEE Trans. Microwave Theory Tech.*, vol. **53**, pp. 1026–1038, March 2005.
- [10] M. Valkama, M. Renfors, and V. Koivunen, “Advanced methods for I/Q imbalance compensation in communications receivers,” *IEEE Trans. Signal Process.*, vol. **49**, pp. 2335–2344, Oct. 2001.
- [11] J. Tubbax *et al.*, “Compensation of IQ imbalance and phase noise in OFDM systems,” *IEEE Trans. Wireless Comm.*, vol. **4**, pp. 872–877, May 2005.
- [12] I.-H. Sohn, E.-R. Jeong, and Y. H. Lee, “Data-aided approach to I/Q mismatch and DC-offset compensation in communication receivers,” *IEEE Comm. Lett.*, vol. **6**, pp. 547–549, Dec. 2002.

- [13] L. Anttila, M. Valkama, and M. Renfors, "Blind moment estimation techniques for I/Q imbalance compensation in quadrature receivers," *Proc. IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications (PIMRC'06)*, Helsinki, Finland, Sept. 2006.
- [14] M. Windisch and G. Fettweis, "Adaptive I/Q imbalance compensation in low-IF transmitter architectures," *Proc. IEEE Veh. Technol. Conf. (VTC2004-Fall)*, Los Angeles, CA, Sept. 2004, pp. 2096–2100.
- [15] J. Tuthill and A. Cantoni, "Efficient compensation for frequency-dependent errors in analog reconstruction filters used in IQ modulators," *IEEE Trans. Communications*, vol. **53**, pp. 489–496, March 2005.
- [16] L. Yu and W. M. Snelgrove, "A novel adaptive mismatch cancellation system for quadrature IF radio receivers," *IEEE Trans. Circuit Syst. II*, vol. **46**, pp. 789–801, June 1999.
- [17] K. P. Pun *et al.*, "Correction of frequency-dependent I/Q mismatches in quadrature receivers," *Electronics Letters*, vol. **37**, pp. 1415–1417, Nov. 2001.
- [18] M. Valkama, M. Renfors, and V. Koivunen, "Compensation of frequency-selective I/Q imbalances in wideband receivers: Models and algorithms," *Proc. Third IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications (SPAWC'01)*, Taoyuan, Taiwan, R.O.C., Mar. 2001, pp. 42–45.
- [19] F. Harris, "Digital filter equalization of analog gain and phase mismatch in I-Q receivers," *Proc. IEEE Int. Conf. Universal Personal Communications*, Cambridge, MA, Oct. 1996, pp. 793–796.
- [20] M. Valkama, M. Renfors, and V. Koivunen, "Blind I/Q signal separation based solutions for receiver signal processing," *EURASIP JASP Special Issue on DSP Enabled Radios*, vol. **2005**, no. 16, pp. 2708–2718, Sept. 2005.
- [21] M. Valkama, M. Renfors, and V. Koivunen, "Blind I/Q imbalance compensation in OFDM receivers based on adaptive I/Q signal decorrelation," *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS'05)*, Kobe, Japan, May 2005, pp. 2611–2614.
- [22] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bölcskei, "An overview of MIMO communications – A key to gigabit wireless," *Proc. IEEE*, vol. **92**, pp. 198–218, Feb. 2004.
- [23] S. Simoens, M. de Courville, F. Bourzeix, and P. de Champs, "New I/Q imbalance modeling and compensation in OFDM systems with frequency offset," *Proc. IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications (PIMRC'02)*, Lisboa, Portugal, Sept. 2002, pp. 561–566.
- [24] S. Fouladifard and H. Shafiee, "Frequency offset estimation in OFDM systems in presence of IQ imbalance," *Proc. IEEE Int. Conf. Communications (ICC'03)*, Anchorage, AK, May 2003, pp. 2071–2075.
- [25] D. Tandur and M. Moonen, "Joint adaptive compensation of transmitter and receiver IQ imbalance under carrier frequency offset in OFDM-based systems," *IEEE Trans. Signal Process.*, vol. **55**, pp. 5246–5252, Nov. 2007.
- [26] J. Tubbax *et al.*, "Joint compensation of IQ imbalance and frequency offset in OFDM systems," *IEEE Proc. IEEE Radio Wireless Conference (RAWCON'03)*, Boston, MA, Aug. 2003, pp. 39–42.
- [27] E. Tsui and J. Lin, "Adaptive IQ imbalance correction for OFDM systems with frequency and time offsets," *Proc. IEEE Global Telecommunications Conference (Globecom'04)*, Dallas, TX, Nov. 2004, pp. 4004–4010.
- [28] B. Narasimhan, D. Wang, S. Narayanan, H. Minn, and N. Al-Dhahir, "Digital compensation of frequency-dependent joint Tx/Rx I/Q imbalance in OFDM systems under high mobility," *IEEE J. Select Topics Signal Process.*, vol. **3**, pp. 405–417, June 2009.
- [29] D. Tandur and M. Moonen, "Compensation of RF impairments in MIMO OFDM systems," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'08)*, Las Vegas, NV, USA, April 2008.

- [30] C.-J. Hsu, R. Cheng, and W.-H. Sheen, "Joint least squares estimation of frequency, DC offset, I-Q imbalance, and channel in MIMO receivers," *IEEE Trans. Veh. Technol.*, vol. **58**, pp. 2201–2213, June 2009.
- [31] R. M. Rao and B. Daneshrad, "Analog impairments in MIMO-OFDM systems," *IEEE Trans. Wireless Communications*, vol. **5**, pp. 3382–3387, Dec. 2006.
- [32] Q. Zou, A. Tarighat, and A. H. Sayed, "On the joint compensation of IQ imbalances and phase noise in MIMO-OFDM systems," *Proc. Int. Symp. Circuits Syst. (ISCAS'07)*, New Orleans, LA, May 2007.
- [33] M. Valkama and M. Renfors, "Advanced DSP for I/Q imbalance compensation in a low-IF receiver," *Proc. IEEE Int. Conf. Commun. (ICC'00)*, New Orleans, LA, June 2000, pp. 768–772.
- [34] A. Schuchert, R. Hasholzner, and P. Antoine, "A novel IQ imbalance compensation scheme for the reception of OFDM signals," *IEEE Trans. Consumer Elec.*, vol. **47**, pp. 313–318, Aug. 2001.
- [35] A. Tarighat, R. Bagheri, and A. H. Sayed, "Compensation schemes and performance analysis of IQ imbalances in OFDM receivers," *IEEE Trans. Signal Process.*, vol. **53**, pp. 3257–3268, Aug. 2005.
- [36] J. Tubbax *et al.*, "Compensation of IQ imbalance in OFDM systems," *Proc. IEEE Int. Conf. Commun. (ICC'03)*, Anchorage, AK, May 2003, pp. 3403–3407.
- [37] P. Rykaczewski, V. Blaschke, and F. Jondral, "I/Q imbalance compensation for software defined radio OFDM based direct conversion receivers," *Proc. 8th IEEE Int. OFDM Workshop (InOw'03)*, Hamburg, Germany, Sept. 2003, pp. 279–283.
- [38] E. Lopez-Estraviz *et al.*, "Optimal training sequences for joint channel and frequency-dependent IQ imbalance estimation in OFDM-based receivers," *IEEE Int. Conf. Communications (ICC'06)*, Istanbul, Turkey, June 2006, pp. 4595–4600.
- [39] L. Anttila, M. Valkama, and M. Renfors, "Efficient mitigation of frequency-selective I/Q imbalance in OFDM receivers," *Proc. IEEE Veh. Technol. Conf. (VTC2008-Fall)*, Calgary, Canada, Sept. 2008.
- [40] A. Tarighat and A. H. Sayed, "Joint compensation of transmitter and receiver impairments in OFDM systems," *IEEE Trans. Wireless Communications*, vol. **6**, pp. 240–247, Jan. 2007.
- [41] J. Lin and E. Tsui, "Joint adaptive transmitter/receiver IQ imbalance correction for OFDM systems," *Proc. IEEE Int. Symp. Personal, Indoor and Mobile Radio Communications (PIMRC'04)*, Barcelona, Spain, Sept. 2004, pp. 1511–1516.
- [42] E. Lopez-Estraviz and L. Van der Perre, "EM based frequency-dependent transmit/receive IQ imbalance estimation and compensation in OFDM-based transceivers," *Proc. IEEE Global Telecommunications Conference (GLOBECOM'07)*, Washington DC, Nov. 2007, pp. 4274–4279.
- [43] D. Brady and J. Feigin, "Joint transmitter/receiver I/Q imbalance compensation for direct conversion OFDM in packet-switched multipath environments," *IEEE Trans. Signal Process.*, vol. **57**, pp. 4588–4593, Nov. 2009.
- [44] T. C. W. Schenk, RF Impairments in Multiple Antenna OFDM – Influence and Mitigation. Ph.D. Thesis, TU Eindhoven, The Netherlands, Nov. 2006.
- [45] T. C. W. Schenk, P. F. M. Smulders, and E. R. Fledderus, "Estimation and compensation of frequency selective TX/RX IQ imbalance in MIMO OFDM systems," *Proc. IEEE Int. Conf. Communications (ICC'06)*, Istanbul, Turkey, June 2006, pp. 251–256.
- [46] A. Tarighat and A. H. Sayed, "MIMO OFDM receivers for systems with IQ imbalances," *IEEE Trans. Signal Process.*, vol. **53**, pp. 3583–3596, Sept. 2005.

- [47] Y. Zou, M. Valkama, and M. Renfors, "Digital compensation of I/Q imbalance effects in space-time coded transmit diversity systems," *IEEE Trans. Signal Process.*, vol. **56**, pp. 2496–2508, June 2008.
- [48] Y. Zou, M. Valkama, and M. Renfors, "Analysis and compensation of transmitter and receiver I/Q imbalances in space-time coded multiantenna OFDM systems," *EURASIP Journal on Wireless Communications Networks (Special Issue on Multicarrier Techniques)*, vol. **2008**, Article ID 391025, 2008.
- [49] Y. Zou, Analysis and Mitigation of I/Q Imbalances in Multi-Antenna Transmission Systems. Ph.D. Thesis, Tampere University of Technology, Finland, 2009.
- [50] G. Xing, M. Shen, and H. Liu, "Frequency offset and I/Q imbalance compensation for direct-conversion receivers," *IEEE Trans. Wireless Communications*, vol. **4**, pp. 673–680, March 2005.
- [51] J. K. Cavers, "New methods for adaptation of quadrature modulators and demodulators in amplifier linearization circuits," *IEEE Trans. Veh. Technol.*, vol. **46**, no 3, pp. 707–716, August 1997.
- [52] L. Anttila, M. Valkama, and M. Renfors, "Frequency-selective I/Q mismatch calibration of wideband direct-conversion transmitters," *IEEE Trans. Circuits Syst. II (Special Issue on Multifunctional Circuits and Systems for Future Generations of Wireless Communications)*, vol. **55**, pp. 359–363, Apr. 2008.
- [53] L. Ding, Z. Ma, D. R. Morgan, M. Zierdt, and G. T. Zhou, "Compensation of frequency-dependent gain/phase imbalance in predistortion linearization systems," *IEEE Trans. Circuits Syst.*, vol. **55**, pp. 390–397, Feb. 2008.
- [54] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Inform. Theory*, vol. **39**, pp. 1293–1302, July 1993.
- [55] B. Picinbono and P. Bondon, "Second-order statistics of complex signals," *IEEE Trans. Signal Process.*, vol. **45**, pp. 411–420, Feb. 1997.
- [56] P. J. Schreier and L. L. Scharf, "Second-order analysis of improper complex random vectors and processes," *IEEE Trans. Signal Process.*, vol. **51**, pp. 714–725, March 2003.
- [57] M. Valkama, M. Renfors, and V. Koivunen, "Blind signal estimation in conjugate signal models with application to I/Q imbalance compensation" *IEEE Signal Process. Lett.*, vol. **12**, pp. 733–736, Nov. 2005.
- [58] L. Anttila, M. Valkama, and M. Renfors, "Blind compensation of frequency-selective I/Q imbalances in quadrature radio receivers: Circularity-based approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP '07)*, Hawaii, HI, USA, Apr. 2007.
- [59] L. Anttila, M. Valkama, and M. Renfors, "Circularity-based I/Q imbalance compensation in wideband direct-conversion receivers," *IEEE Trans. Vehi. Technol.*, vol. **57**, pp. 2099–2113, July 2008, 2008.
- [60] M. Windisch and G. Fettweis, "Blind I/Q imbalance parameter estimation and compensation in low-IF receivers," *Proc. First Int. Symp. Control, Communications, and Signal Processing (ISCCSP '04)*, Hammamet, Tunisia, March 2004, pp. 75–78.
- [61] M. Windisch and G. Fettweis, "Standard-independent I/Q imbalance compensation in OFDM direct-conversion receivers," *Proc. Int. OFDM Workshop (InOWo '04)*, Dresden, Germany, Sept. 2004.
- [62] T. B. Sorensen, P. E. Mogensen, and F. Frederiksen, "Extension of the ITU channel models for wideband (OFDM) systems," *Proc. IEEE Veh. Technol. Conf. (VTC'05)*, Dallas, TX, Sept. 2005, pp. 392–396.

17 Joint digital predistortion of I/Q modulator and power amplifier impairments

Mikko Valkama and Lauri Anttila

17.1 Introduction

The analog front-end of a direct-conversion transmitter suffers from several performance-degrading circuit implementation impairments. The main impairments are power amplifier (PA) nonlinear distortion, in-phase/quadrature-phase (I/Q) imbalance, and local oscillator (LO) leakage. Each of these impairments has been treated separately in the literature as well as on the pages of this book. For details on state-of-the-art PA predistortion, the reader is referred to the related chapters in [Part II](#) of this book and the references therein, and for a comprehensive review on I/Q imbalance compensation, to [Chapter 16](#) of this book and the literature cited therein. It has been demonstrated that, when treated separately, each of the impairments can be mitigated by using digital predistortion. What is often overlooked, however, is that in direct-conversion transmitters these impairments interact in a manner that may severely cripple the overall transmitted signal quality. In addition to the obvious effects of I/Q imbalance and LO leakage (mirror-frequency interference (MFI) and spurious signal energy at the LO frequency, respectively), there are several other performance-degrading phenomena arising from their interaction with the nonlinearity that need addressing. First, I/Q imbalance and LO leakage cause extra intermodulation distortion (IMD) products to appear at the PA output [6], [9]. Effectively this means that even with access to ideal PA predistorter (PD) coefficients, spectral regrowth will not be fully mitigated. Second, the extra IMD products at the PA output will interfere with the estimation of an adaptive PA PD [6], [9]. In other words, if the PA PD is trained with no regard for I/Q imbalance and LO leakage, the resulting PD will be biased, and thus the overall transmitted signal quality will be further compromised. Third, PA nonlinearity interferes with the estimation of the I/Q modulator (IQM) predistorter, yielding biased estimates. This makes it difficult to compensate for IQM impairments prior to PA PD estimation. These aspects will be discussed in more detail in [Section 17.2](#).

The interplay between PA nonlinearity and IQM impairments has received relatively little attention in the literature. The first analytical work addressing this issue was by Faulkner and Mattson [9], who studied the effects of IQM impairments on the spectrum of the PA output signal, deriving an approximate expression for the spectrum as a function of IQM impairment parameters, a power series model for the PA, backoff, and

known autocorrelation properties of the modulated signal. For linear modulations, it was found that the power spectral density (PSD) of the out-of-band spectrum increases approximately with the square of the misalignment. The results of [9] will be partially reiterated in Section 17.2. An extended analysis of the problem was carried out by Cavers [6], who analyzed the effects of both I/Q modulator and I/Q demodulator impairments on the estimation of an adaptive PA predistorter in the context of a general memoryless PA, providing expressions for the power and power spectra of intermodulation error at the PA output. Previous work on IQM impairment estimation and mitigation in predistortion circuits is summarized in the following. Throughout, we are assuming an adaptive PD setup, which utilizes an internal feedback loop (FBL) for the purpose of PD parameter estimation (as shown for example in Figure 17.2).

Even though the effects of PA and IQM impairments are nonlinearly coupled, it is possible to estimate these impairments (or the corresponding predistorters) individually with the help of extra RF hardware. The techniques in [10] and [11] assume that a second coupler is implemented (at the IQM output), along with an RF switch to choose between the couplers at the IQM output and the PA output. In [7] and [8] a second, envelope detector based, FBL is needed for the purpose of IQM impairment estimation. Purely algorithmic methods for decoupling the PA and IQM predistorter estimates using a single observation of the PA output only do not exist at the moment, at least to the best of the authors' knowledge. Therefore, if targeting for high-performance digital predistortion, and one wants to avoid the extra hardware and the related component and engineering costs that are associated with the above mentioned decoupled methods, a joint method is the only possibility.

Joint predistorters, needing no extra RF hardware for their training, were developed in [13] and [14]. However, these techniques only considered frequency-independent (memoryless) impairments, limiting their applicability in modern broadband systems. The purpose of this chapter is to introduce a general, feasible, and effective joint PA and IQM predistorter structure for *frequency-dependent* impairments. Altogether handling of frequency-dependent RF impairments can be seen as critical for the emerging broadband wireless systems, like IMT-Advanced, in which radio bandwidths in the order of 10–100 MHz will be deployed.

The new parallel PD structure is theoretically equivalent to a serial PD structure consisting of a serial combination of a truncated Volterra-based PA PD, an LO leakage compensator, and an I/Q imbalance compensator. Such a PD structure compensates for the impairments in reverse order of appearance, and is assumed, e.g., in [10], [11]. Unlike the serial structure, however, the parallel structure enables the joint estimation of all the predistorter parameters from a single observation, and without any extra RF hardware compared to pure PA predistortion. The new PD is composed of two parallel truncated Volterra predistorters with direct and conjugated inputs, plus an LO leakage compensator [31]. A simplified predistortion structure based on parallel Hammerstein (PH) systems (originally proposed by the authors in [16]) is then presented. The PH system has been shown to be a versatile tool for both direct and inverse modeling of PAs [17] – [19]. The PH system is also much less complex than a Volterra system of similar order in terms of the number of model parameters needed.

The estimation of the predistorter parameters is based on the indirect learning architecture (ILA) and least squares (LS) estimation. No specially designed training signals are needed in the parameter estimation stage, meaning that estimation can in principle be done during normal operation of the transmitter. Some of the presented results are reproduced from [16] and [30], where the PH based joint predistorter structure was first introduced by the authors of this chapter.

The rest of this chapter is organized as follows. In Section 17.2, the motivation for using joint processing for PA and IQM impairments is laid out. We couple a simple behavioral model for the PA, based on complex-valued polynomials, with the I/Q mismatch model introduced in Chapter 16 on I/Q imbalance compensation. It is then illustrated that, even with perfect p th-order inverse predistorter for the PA, I/Q imbalance and LO leakage indeed modify the IMD products seen at the PA output. The effect of these extra IMD products on adaptive PD estimation is also discussed and experimented. Section 17.3 presents the joint PA and IQM predistorter structures based on Volterra and parallel Hammerstein nonlinearities. The least squares estimation procedure for the PH based PD, utilizing indirect learning, is introduced in Section 17.4. In Section 17.5, experimental results from both computer simulations and RF laboratory measurements are presented. Finally, Section 17.6 concludes the chapter.

Notation: vectors are in bold lower-case and matrices in bold upper-case. Complex-conjugation, matrix transpose, and conjugate-transpose are denoted by $(\cdot)^*$, $(\cdot)^T$ and $(\cdot)^H$, respectively. The statistical expectation operator is $E[\cdot]$.

17.2 Effects of IQM impairments on PA intermodulation products and predistorter estimation

Now we present the basic RF front-end signal models, including frequency-independent I/Q mismatch, LO leakage, and a polynomial PA model. This is done solely for motivational purposes, to illustrate the effects that the IQM impairments have on the intermodulation products created by the PA, and on the estimation of the PA predistorter parameters. In the sections to follow, we will assume frequency-dependent behavior for both I/Q mismatch and PA nonlinearity.

With reference to Figure 17.1, we denote the baseband I/Q signal, exiting the PA predistorter, as $z(t) \triangleq z_I(t) + jz_Q(t)$. If the IQM is ideal, i.e., it has no impairments, $z(t)$ is also the complex envelope of the IQM output signal. Under I/Q mismatch and LO leakage, the complex envelope of the IQM output is written as

$$r(t) = g_1 z(t) + g_2 z^*(t) + c, \quad (17.1)$$

where

$$g_1 = \frac{1 + g \exp(j\phi)}{|1 + g \exp(j\phi)|}, g_2 = \frac{1 - g \exp(j\phi)}{|1 + g \exp(j\phi)|}$$

are the (complex-valued) baseband I/Q mismatch coefficients, with g and ϕ denoting the relative gain and phase imbalance, and $c = |c| \exp(j\angle c)$ is the LO leakage signal. Notice

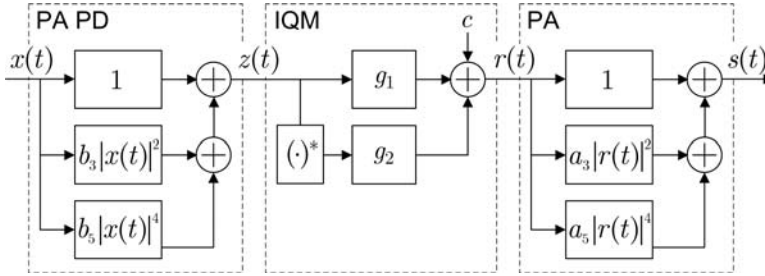


Figure 17.1 Memoryless baseband signal model for a transmitter, with fifth-order polynomial PA predistorter (PA PD), I/Q modulator (IQM) with mismatch and LO leakage, and fifth-order polynomial PA model.

that g_1 and g_2 are scaled versions of the I/Q parameters introduced in Chapter 16 on I/Q imbalance compensation. This scaling simply eliminates the level mismatch caused by the I/Q imbalance model in (17.1), which is known to have its own independent effects on the IMD levels of the PA output [9]. In the following analysis we are interested in the effects of the MFI term $g_2 z^*(t)$, not of the level mismatch. The effects of level mismatch on the IMD products were analyzed in [9].

The PA model assumed in this example is a quasi-memoryless polynomial model of order 5. With the IQM output signal (17.1) as input to the PA, the complex envelope of the PA output signal is expressed as [32]

$$s(t) = H(r(t)) = r(t) + a_3 |r(t)|^2 r(t) + a_5 |r(t)|^4 r(t), \quad (17.2)$$

where the linear gain term has been set to unity. With a perfect IQM ($g_1 = 1$, $g_2 = c = 0$), the PA output takes the form

$$s(t) = H(z(t)) = z(t) + a_3 |z(t)|^2 z(t) + a_5 |z(t)|^4 z(t). \quad (17.3)$$

The PA predistorter that we consider is also a fifth-order polynomial with complex-valued coefficients, given as

$$z(t) = x(t) + b_3 |x(t)|^2 x(t) + b_5 |x(t)|^4 x(t), \quad (17.4)$$

where $x(t)$ is the baseband modulated I/Q signal. We can then obtain the fifth-order inverse solution for the PA predistorter as follows. Substituting (17.4) into (17.3) and expanding, the PA output signal becomes

$$s(t) = x(t) + (b_3 + a_3) |x(t)|^2 x(t) + (b_5 + 2a_3 b_3 + a_3 b_3^* + a_5) |x(t)|^4 x(t) + o(x(t)), \quad (17.5)$$

where the seventh- and higher-order terms have been included in $o(x(t))$. They have no effect on the fifth-order inverse solution. They do produce residual in-band and adjacent

channel interference, but analyzing these effects is out of the scope of this chapter. Now, the fifth-order inverse solution is obtained by setting the third- and fifth-order coefficients in (17.5) to zero, leading to

$$\begin{aligned} b_3 &= -a_3 \\ b_5 &= -a_5 + 2a_3^2 + |a_3|^2 \end{aligned} \quad (17.6)$$

17.2.1 Effect of LO leakage

We can see the effect of LO leakage on the PA output signal as follows. First, plug (17.4) into (17.1), set $g_1 = 1$ and $g_2 = 0$, and then substitute the result into (17.2) to calculate $s(t) = H(r(t))$. Neglecting the terms higher than fifth order, the result is

$$\begin{aligned} s(t) &= A_{c1}x(t) + A_{c2}|x(t)|^2 + A_{c3}x(t)^2 \\ &\quad + A_{c4}|x(t)|^2x(t) + A_{c5}x(t)^3 + A_{c6}|x(t)|^4 \\ &\quad + A_{c7}|x(t)|^2x(t)^2 + A_{c8}|x(t)|^4x(t) + c. \end{aligned} \quad (17.7)$$

Compared to the case of no IQM impairments in (17.3), two second-order products $|x(t)|^2$ and $x(t)^2$, one third-order product $x(t)^3$, and two fourth-order products $|x(t)|^4$ and $|x(t)|^2x(t)^2$, appear. These create additional in-band and adjacent channel interference. The exact expressions for the complex coefficients A_{ci} , $i = 1 \dots 8$ are listed in Table 17.1 (from [9]). The middle column of Table 17.1 shows the PA output distortion coefficients without predistortion ($b_i = 0$), and the rightmost column with the ideal fifth-order inverse predistorter in (17.6). While the fifth-order inverse predistorter is still able to suppress the third- and fifth-order distortion products, the even-order distortion terms are still present after applying the predistortion. Depending on the exact parameter values, the remaining IMD products may be either amplified or attenuated.

17.2.2 Effect of I/Q mismatch

To obtain the PA output signal in the presence of I/Q mismatch, we proceed as in Subsection 17.2.1. First, we plug (17.4) into (17.1), set $c = 0$, and then substitute the result into (17.2) to calculate $s(t) = H(r(t))$. The result is

Table 17.1 PA output distortion terms with LO leakage.

Distortion term	No predistortion	Fifth-order inverse predistorter
A_{c1}	1	1
A_{c2}	$2a_3c$	$2a_3c$
A_{c3}	a_3c^*	a_3c^*
A_{c4}	a_3	0
A_{c5}	a_5c^{*2}	0
A_{c6}	$3a_5c$	$3a_5c - 2 a_3 ^2c - 2a_3^2c$
A_{c7}	$2a_5c^*$	$2a_5c^* - 2a_3^2c^*$
A_{c8}	a_5	0

Table 17.2 PA output distortion terms under I/Q mismatch.

Coefficient	No PD	Ideal fifth-order inverse PD	Estimated fifth-order inverse PD
A_{iq1}	g_1	g_1	g_1
A_{iq2}	g_2	g_2	g_2
A_{iq3}	$a_3 g_1 + 2a_3 g_1 g_2 ^2$	$2a_3 g_1 g_2 ^2$	$a_3 g_1 - a_3 g_1^2$
A_{iq4}	$2a_3 g_2$	$2a_3 g_2 - a_3^* g_2$	$2a_3 g_2 - a_3^* g_1^* g_2$
A_{iq5}	$a_3 g_1^2 g_2^*$	$a_3 g_1^2 g_2^*$	$a_3 g_1^2 g_2^*$
A_{iq6}	$a_3 g_1^* g_2^2$	$a_3 g_1^* g_2^2$	$a_3 g_1^* g_2^2$
A_{iq7}	$a_5 g_1 + 6a_5 g_1 g_2 ^2$	$6a_5 g_1 g_2 ^2 - 2 a_3 ^2 g_1 g_2 ^2$ $- 4a_3^2 g_1 g_2 ^2$	$a_5 g_1 - a_5 g_1^2 + a_3 ^2 g_1 - a_3 ^2$ $+ 2a_3^2 g_1^3 - 2a_3^2 g_1^2$
A_{iq8}	$3a_5 g_2$	$3a_5 g_2 - a_5^* g_2 - 3 a_3 ^2 g_2$ $- 2a_3^2 g_2 + 2a_3^2 g_2$	$3a_5 g_2 - a_5^* g_1^* g_2 + a_3 ^2 g_2 - 4 a_3 ^2 g_1^* g_2$ $+ 2a_3^2 g_1^2 g_2 - 2a_3^2 g_1 g_2$
A_{iq9}	$2a_5 g_1^2 g_2^*$	$2a_5 g_1^2 g_2^* - 3a_3^2 g_1^2 g_2^*$	$2a_5 g_1^2 g_2^* - 3a_3^2 g_1^3 g_2^*$
A_{iq10}	$3a_5 g_1^* g_2^2$	$3a_5 g_1^* g_2^2 - 3 a_3 ^2 g_1^* g_2^2$	$3a_5 g_1^* g_2^2 - 3 a_3 ^2 g_1^2 g_2^2$
A_{iq11}	$2a_5 g_1^3 g_2^{*2}$	$a_5 g_1^3 g_2^{*2}$	$a_5 g_1^3 g_2^{*2}$
A_{iq12}	$a_5 g_1^{*2} g_2^3$	$a_5 g_1^{*2} g_2^3$	$a_5 g_1^{*2} g_2^3$

$$\begin{aligned}
s(t) = & A_{iq1}x(t) + A_{iq2}x(t)^* + A_{iq3}|x(t)|^2x(t) + A_{iq4}|x(t)|^2x(t)^* \\
& + A_{iq5}x(t)^3 + A_{iq6}x(t)^{*3} + A_{iq7}|x(t)|^4x(t) + A_{iq8}|x(t)|^4x(t)^* \\
& + A_{iq9}|x(t)|^2x(t)^3 + A_{iq10}|x(t)|^2x(t)^{*3} + A_{iq11}x(t)^5 + A_{iq12}x(t)^{*5}
\end{aligned} \quad (17.8)$$

where the terms of order higher than 5 have been neglected. Extra intermodulation products again appear compared to (17.3), now due to I/Q mismatch. These are the first-order image interference term $x(t)^*$, three third-order products $|x(t)|^2x(t)^*$, $x(t)^3$ and $x(t)^{*3}$, and five fifth-order products $|x(t)|^4x(t)^*$, $|x(t)|^2x(t)^3$, $|x(t)|^2x(t)^{*3}$, $x(t)^5$, and $x(t)^{*5}$. The exact expressions for the complex coefficients $A_{iqj}, j = 1 \dots 12$ of (17.8) are listed in the second column of Table 17.2. The third column shows the same complex coefficients after ideal fifth-order predistortion for the PA has been applied.¹ It is notable that the fifth-order inverse predistorter is unable to completely suppress the terms proportional to $|x(t)|^2x(t)$ and $|x(t)|^4x(t)$, which would be mitigated without the I/Q mismatch. The main distortion components of these are suppressed, but some residual distortion, proportional to $|g_2|^2$, is left. Also, several IMD products are not attenuated at all, and some may even be amplified, depending on the exact values of the parameters a_3 , a_5 , g_1 , and g_2 . If I/Q mismatch and LO leakage are both considered simultaneously, as in practice happens, the PA output will contain all the distortion terms from (17.7) and (17.8), further complicating matters.

¹ A similar analysis was done in [9], but with a different parameterization for the I/Q imbalance coefficients, thus leading to a slightly different result. Compared to [9], we have also included the nonlinear products $x^{*3}(t)$ [2], $|x(t)|^2x^{*3}(t)$ and $x^{*5}(t)$ for completeness, even though they contribute very little to the total IMD.

17.2.3 Effect of IQM impairments on adaptive predistorter estimation

The [previous subsections](#) dealt with a fixed predistorter that was obtained by calculating the p th-order inverse of a known or pre-estimated nonlinearity. In many cases, the PA coefficients a_i may not be available beforehand, or they may change over time due to component aging or temperature variations. In these cases, an adaptive estimation procedure for the PA model parameters (or for the PA PD coefficients directly) must be adopted. The transmitter then needs to house an internal feedback loop (FBL) from the RF parts back to digital baseband for estimation purposes. One option for realizing the FBL is presented in [Figure 17.2](#), which shows a general I/Q transmitter structure with digital predistortion and the internal FBL. Here, the RF signal is first downconverted to a low intermediate frequency (IF) using a single mixer, then low-pass filtered and converted to digital domain, with the final down-conversion from IF to baseband done in DSP. With this technique, the complex envelope of the PA output can be obtained with high precision. Compared to using I/Q demodulation in the FBL, the possible I/Q demodulator mismatch and LO leakage can be effectively avoided, while also halving the component count. The price paid is the doubling of the sampling rate for the sole ADC.

Now we assume that an observation of the PA output signal is available at digital baseband for estimating the PA model parameters. To see the effect of the IQM impairments on PA predistorter estimation and subsequent predistortion, we can do the following. We assume that the typical least squares identification procedure is able to identify the PA distortion coefficients proportional to $|x(t)|^2 x(t)$ and $|x(t)|^4 x(t)$, i.e., $\hat{A}_{iq3} = a_3 g_1 + 2a_3 g_1 |g_2|^2$ and $\hat{A}_{iq7} = a_5 g_1 + 6a_5 g_1 |g_2|^2$ from the second column of [Table 17.2](#). This is a rather optimistic assumption, since some of the other IM products in (17.8) will affect the LS estimation of these coefficients, as can readily be concluded e.g. from Table II in [9]. Anyway, assuming identifiability for these PA model coefficients, we can then apply (17.6) to obtain the predistorter coefficients as

$$\begin{aligned} \hat{b}_3 &= -\hat{A}_{iq3} \\ &= -a_3 g_1 - 2a_3 g_1 |g_2|^2, \\ \hat{b}_5 &= -\hat{A}_{iq7} + 2\hat{A}_{iq3}^2 + |\hat{A}_{iq3}|^2 \\ &\cong -a_5 (g_1 + 6g_1 |g_2|^2) + 2a_3^2 (g_1^2 + 4g_1^2 |g_2|^2) \\ &\quad + |a_3|^2 (1 + 4|g_2|^2). \end{aligned} \quad (17.9)$$

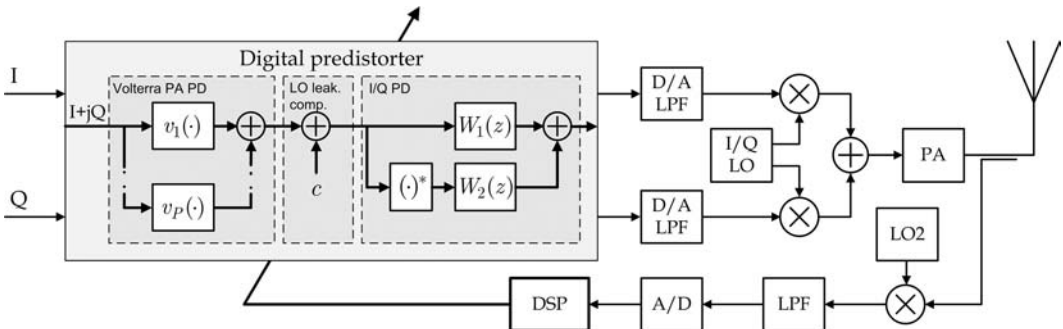


Figure 17.2 General I/Q transmitter structure including the adaptive digital predistorter in serial configuration.

Under I/Q mismatch, using (17.9) as the predistortion coefficients instead of the ideal ones in (17.6) results in a PA output signal of similar form as (17.8), but with modified coefficient values. The coefficients $A_{iqj}, j = 1 \dots 12$ for this biased PD case are listed in the rightmost column of Table 17.2. The coefficients are now much more complex than those in the previous cases, with some IMD products left unchanged by the predistortion, and some amplified or attenuated depending on the exact values of the front-end parameters.

A numerical example is presented next, where a linearly modulated 16-QAM signal with root-raised cosine pulse shape is sent through a mismatched RF front-end. The I/Q modulator has amplitude mismatch of 10% ($g = 1.1$) and phase mismatch $\phi = 10^\circ$, resulting in $g_1 = 0.9958 + j0.0913$ and $g_2 = -0.0398 - j0.0913$. The PA model is a fifth-order polynomial model with coefficients $a_1 = 1$, $a_3 = -0.055607 + j0.012164$, and $a_5 = 0.001860 + j0.000032$. The input backoff is 3 dB. Figure 17.3 shows the spectra of the PA output in cases of a linear PA, nonlinear PA + I/Q imbalance without PD, nonlinear PA with ideal PD but without I/Q imbalance, nonlinear PA + I/Q imbalance with ideal PD, and nonlinear PD + I/Q imbalance with the estimated (biased) PD. In the absence of I/Q imbalance, the ideal fifth-order inverse PD is able to push the adjacent channel interference power down by 15–18 dB. With I/Q imbalance, the improvement with the ideal PD is only about 8 dB. With the biased PD, a further decrease in performance of about 0.5–1 dB is seen. As PD order is increased, the performance gap will grow between the cases without and with I/Q imbalance. Promising new joint predistorter structures, that are able to deal with both PA and IQM impairments, will be reviewed next.

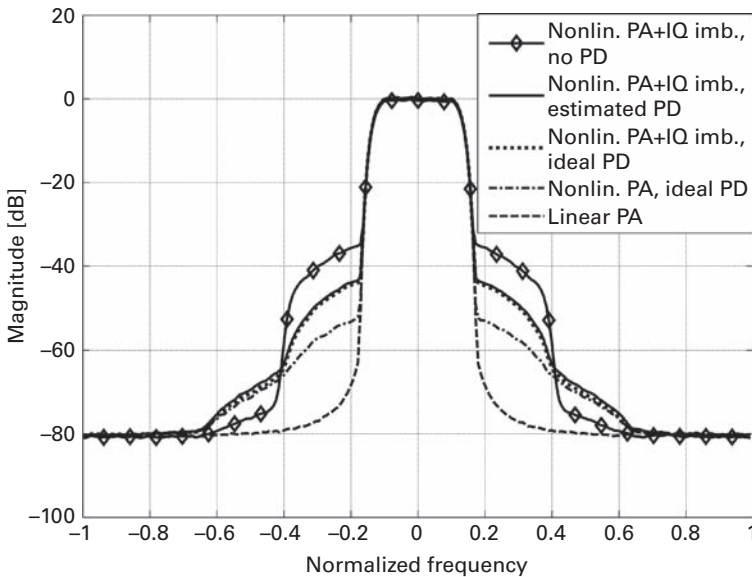


Figure 17.3 Effect of I/Q imbalance and predistortion on the spectral regrowth of the PA output signal.

17.3 Joint PA and IQM predistortion

A general I/Q modulation-based transmitter structure, including a principal digital predistorter in the serial configuration (as in [10], for example), is shown in Figure 17.2. This PD structure compensates for the impairments in the reverse order that they appear, and is used as the starting point here. The PA predistorter that we consider here is a general Volterra PD. The rationale for using the Volterra PD is to obtain a general result, which can then be applied to any simplification of Volterra, such as Wiener, Hammerstein, or parallel Hammerstein. The main details of the Volterra-based joint PD will be given in Subsection 17.3.1, and the parallel Hammerstein-based joint PD will be described in Subsection 17.3.2.

The I/Q predistorter in the serial PD is of the general two-filter type (see, e.g., [10], [11]), where one filter, $W_1(z)$, is filtering the original or non-conjugate signal, and the other filter, $W_2(z)$, is filtering the conjugated signal. The I/Q imbalance compensation filters are complex-valued FIR filters of length L . The actual baseband I/Q imbalance model has a similar, so-called widely linear, form as the I/Q PD, with filters $G_1(z)$ and $G_2(z)$ filtering the non-conjugate and conjugate signals. For a more detailed treatment of mismatch modeling and compensation, the reader is referred to Chapter 16 on I/Q imbalance by Anttila, Zou, and Valkama.

The serial PD structure depicted in Figure 17.2 is conceptually feasible but in practice problematic from the parameter estimation point of view. The filters of the PA PD and I/Q PD are in cascade, making their joint estimation difficult. With the serial structure, one has to train the predistorters separately, requiring additional RF hardware compared to pure PA predistortion. In [16], [30], [31] we have proposed modified PD structures which are completely parallel, enabling one-step estimation of all PD parameters using linear least squares, and without any extra hardware. The following is a review of these joint techniques.

17.3.1 Volterra-based structure

For the PA PD, we first consider a general baseband Volterra series model of a bandpass nonlinearity, given as

$$v(x_n) = \sum_{p=0}^{\frac{P-1}{2}} v_{2p+1}(x_n) \triangleq \sum_{p=0}^{\frac{P-1}{2}} \sum_{d_1=0}^M \cdots \sum_{d_{2p+1}=0}^M h_{2p+1}(\mathbf{d}_{2p+1}) \prod_{i=1}^{p+1} x_{n-d_i} \prod_{i=p+2}^{2p+1} x_{n-d_i}^*, \quad \text{for } P \text{ odd} \quad (17.11)$$

where $\mathbf{d}_l = [d_1, d_2, \dots, d_l]$, $h_l(\mathbf{d}_l)$ is the l th order Volterra kernel, M is the memory depth (assumed equal for all dimensions), and P is the nonlinearity order, assumed odd [32]. The model can be broken down to a sum of Volterra nonlinearities that contain only terms of order $2p+1$, $v_{2p+1}(x_n)$.

In the serial configuration, the LO leakage compensation coefficient c is summed to the output of the Volterra PD, after which the signal goes through the I/Q PD. To modify the structure, we first move the LO leakage compensator c to the end of the PD chain and rename it c' . In this modified structure, the output signal of the Volterra is filtered with the filter $W_1(z)$ and the complex conjugate of the Volterra output is filtered with $W_2(z)$. Through the linearity of summation, we can merge the I/Q imbalance compensation filters with the Volterra kernels, and obtain two modified Volterra series (denoted $f(\cdot)$ and $\bar{f}(\cdot)$), with kernels $h_l^{(1)}(d_1, d_2, \dots, d_l)$ and $h_l^{(2)}(d_1, d_2, \dots, d_l)$, and respective inputs x_n and x_n^* [31]. The output of the non-conjugate signal branch of the I/Q PD is

$$\sum_{p=0}^{\frac{P-1}{2}} \sum_{d_1=0}^N \cdots \sum_{d_{2p+1}=0}^N h_{2p+1}^{(1)}(\mathbf{d}_{2p+1}) \prod_{i=1}^{p+1} x_{n-d_i} \prod_{i=p+2}^{2p+1} x_{n-d_i}^* = \sum_{p=0}^{\frac{P-1}{2}} f_{2p+1}(x_n) \triangleq f(x_n), \quad (17.12)$$

while for the conjugate signal branch we obtain

$$\sum_{q=0}^{\frac{Q-1}{2}} \sum_{d_1=0}^{\bar{N}} \cdots \sum_{d_{2q+1}=0}^{\bar{N}} h_{2q+1}^{(2)}(\mathbf{d}_{2q+1}) \prod_{i=1}^{q+1} x_{n-d_i} \prod_{i=q+2}^{2q+1} x_{n-d_i}^* = \sum_{q=0}^{\frac{Q-1}{2}} \bar{f}_{2q+1}(x_n^*) \triangleq \bar{f}(x_n^*). \quad (17.13)$$

Here, P and Q are the nonlinearity orders, and N and $\bar{N}$ the memory depths of the non-conjugate and conjugate PDs, respectively. Finally, taking into account the LO leakage compensator c' , we can write the output of the complete predistorter as

$$z_n = f(x_n) + \bar{f}(x_n^*) + c'. \quad (17.14)$$

The final PD structure is shown in Figure 17.4. The LO leakage compensator c' in the parallel structure is a transformed version of the original (serial) compensator c , as it is

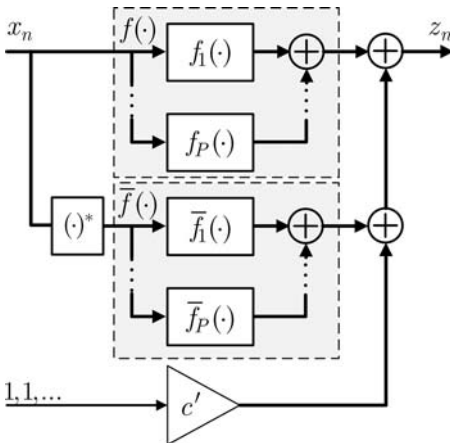


Figure 17.4 Parallel conjugate Volterra-based PD structure for the joint predistortion of frequency-dependent power amplifier and I/Q modulator impairments.

moved to the other side of the I/Q imbalance compensator when deriving the final PD structure. Notice that the nonlinearity orders of the non-conjugate and conjugate Volterra PDs are generally not equal, since in practice we can consider a lower-order PD for the conjugate signal without notable loss in performance [16], [31].

The memory depths are lengthened by $L - 1$ compared to (17.11). If we choose $Q = P$ and $\bar{N} = N = M + L - 1$, the final predistorter of (17.14) is equivalent to the original serial structure in Figure 17.2 in terms of nonlinearity order and memory depth. In the general case (with $Q \neq P$ and $\bar{N} \neq N$), the derived structure gives indeed additional degrees of freedom in the predistorter design, compared to the serial structure, being able to have different processing orders (nonlinearity orders and amount of memory) for the direct and conjugate signal terms.

If we include all the possible kernel coefficients, not taking into account coefficient symmetry, the dimensionality of the serial PD is $2(M + 1)^{(P-1)/2} + 2L$, while for the parallel PD, it is $2(N + 1)^{(P-1)/2} = 2(M + L)^{(P-1)/2}$. Using these expressions, we find that for $L = 1$ (frequency-independent I/Q mismatch), the dimensionality of the parallel PD is approximately doubled compared to the serial PD, while for the general case of $L > 1$ (frequency-dependent I/Q mismatch), the dimensionality will be increased polynomially due to the increased memory depth. However, as discussed in more detail in [31], not all kernel coefficients are needed in the parallel PD for an equivalent description of the serial PD. Coupling this fact with coefficient symmetry and intelligent choice of the orders of the conjugate PD (Q and $\bar{N}$), significant reductions in dimensionality can be obtained compared to $2(M + L)^{(P-1)/2}$ [31]. In spite of these steps, the complexity of the Volterra structure may be prohibitive for practical implementations. It is therefore important to derive alternative but simpler structures that will be more practically useful. The parallel Hammerstein-based joint PD structure from [16] will be reviewed next.

17.3.2 Parallel Hammerstein-based structure

Now, we assume that the original PA predistorter in Figure 17.2 is a parallel Hammerstein (PH) PD with polynomial nonlinearities, instead of a full Volterra series PD. The PH nonlinearity is a parallel connection of P Hammerstein nonlinearities, each of which, in turn, is a cascade of a static (memoryless) nonlinearity ψ_p and a linear time-invariant (LTI) filter $H_p(z)$. The static nonlinearities are given by the polynomials

$$\psi_p(x_n) = \sum_{k \in I_{p,p}} u_{k,p} |x_n|^{k-1} x_n, p \in I_P, \quad (17.15)$$

with $x_n = x_{n,I} + jx_{n,Q}$ denoting the baseband I/Q signal, P the polynomial order, I_P the set of used polynomial orders, and $u_{k,p}$ the polynomial weights [17], [18]. If all polynomial orders up to P are used in (17.15), $I_P = \{1, 2, 3, \dots, P\}$, and if only odd orders are included, $I_P = \{1, 3, 5, \dots, P\}$. The subset of I_P in which orders only up to p are retained, is denoted by $I_{p,p}$. The filters $H_p(z)$, in turn, are usually finite impulse response (FIR) filters for implementation and parameter estimation simplicity.

The polynomial in (17.15) is general and makes it possible to use both even and odd polynomial orders in defining the branch nonlinearities. It also enables the use of either

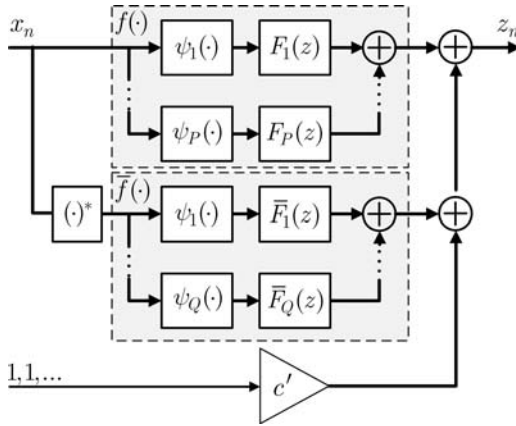


Figure 17.5 Parallel conjugate PH predistorter structure for the joint predistortion of frequency-dependent power amplifier and I/Q modulator impairments. © 2010 IEEE

conventional polynomials (by setting $u_{k,p} = 0$, $k \neq p$) or any type of orthogonal polynomial. In the simulations for this chapter, the *statistically orthogonal* polynomials (meaning that for $p \neq q$ $E[\psi_p(x_n) \psi_q^*(x_n)] = 0$) for complex Gaussian signals from [20] are adopted. Other polynomials, orthogonal or not, can be used as well, but their numerical properties, and hence their performance in fixed-point implementations, may be different. The reader is referred to [20] for a discussion of the possible numerical difficulties encountered in PD parameter estimation.

The joint PA and modulator predistorter structure, proposed in [16], is shown in Figure 17.5. The filters in Figure 17.5 are formally given as $F_p(z) = H_p(z)G_1(z)$ and $\bar{F}_p(z) = H_p^*(z)G_2(z)$, for $p \in I_P$. Thus, the predistorter is a parallel connection of two parallel Hammerstein nonlinearities, given as

$$f(x_n) = \sum_{p \in I_P} f_{p,n} \star \psi_p(x_n) \text{ and} \quad (17.16a)$$

$$\bar{f}(x_n^*) = \sum_{q \in I_Q} \bar{f}_{q,n} \star \psi_q(x_n^*). \quad (17.16b)$$

Here, $f_{p,n}$ and $\bar{f}_{q,n}$ are the (complex-valued) impulse responses of $F_p(z)$ and $\bar{F}_q(z)$, respectively, I_Q is the set of used polynomial orders for the conjugate PD (defined similarly as I_P), and $\star$ denotes convolution. Notice that in (17.16b) the polynomial order is now changed to Q (like in previous Volterra developments), since in general the orders in (17.16a) and (17.16b) need not be the same.

The dimensionality of the parallel Hammerstein-based PD grows linearly with N [16], in contrast with the Volterra-based PD whose dimensionality grows exponentially with N . If we compare the PH-based parallel PD with the PH-based serial PD, the dimensionality of the parallel PD (in terms of complex-valued parameters) is roughly 1.5 to 2-fold [16]. This implies higher computational requirements in both estimation and actual predistortion stages, as discussed in detail in [16]. The number of floating point

operations in the estimation stage is about 4 times greater with the parallel structure. In the predistortion stage, the aforementioned 1.5 to 2-fold increase in complex-valued multiplications compared to the serial structure is seen. However, the serial structure is not identifiable, unless separate training of the two PDs, and the necessary additional hardware, are employed. In contrast, the proposed parallel structure does not need any additional hardware, and is guaranteed to be identifiable (as long as the PA is invertible), due to the linearity with respect to its parameters.

17.4 Predistorter parameter estimation

Now we review the estimation procedure for the parallel Hammerstein-based predistorter. The PD training is based on the indirect learning architecture, combined with linear least squares learning. We present the block-based LS solution first, and then describe the recursive LS (RLS) estimator. We have opted to present only the PH-based predistorter, and not the Volterra-based, because of space limitations and the higher practical value of the PH-based structure. Anyhow, a linear LS estimation procedure for the Volterra-based PD is rather straightforward, as presented for example in [31].

17.4.1 Indirect learning

In the parameter estimation stage we utilize the indirect learning architecture (ILA), which enables the estimation of the predistorter coefficients directly, without the need for finding and inverting a PA model, and by using linear least squares methods. The idea behind the ILA is to find the *post-inverse* of the PA nonlinear function and then, assuming that it equals the *pre-inverse*, use it as the predistorter. The principal operation of the ILA is illustrated in Figure 17.6, where $H(\cdot)$ denotes the PA nonlinear function, and $\hat{P}(\cdot)$ is the estimate of the post-inverse of $H(\cdot)$. Ideally, if $\hat{P}(\cdot)$ is the exact post-inverse of $H(\cdot)$ and measurement noise is absent (i.e., the error e_n vanishes), we will have $\hat{P}(x_n) = \hat{P}(1/G H(z_n))$, where G denotes the intended amplitude gain of the amplifier. Thus, $H(z_n) = Gx_n$, therefore validating the ILA principle in the case of an exact (infinite-order) post-inverse $\hat{P}(\cdot)$.

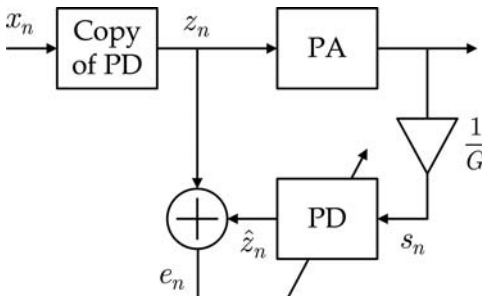


Figure 17.6 The indirect learning architecture operating principle.

Schetzen [21] showed that for an invertible Volterra nonlinearity $H(\cdot)$, the truncated p th-order pre-inverse is equal to the p th-order post-inverse, thus giving a formal justification to the ILA principle in the class of truncated Volterra series nonlinearities. Although here we do not calculate Schetzen's p th-order inverse explicitly, but rather calculate a p th-order inverse which minimizes some error criterion (such as least squares or minimum mean square error between PA input and postdistorter output signals), we assume the commutativity result of Schetzen to hold approximately. This assumption is backed up by the experimental studies that utilize indirect learning [16], [17], [20], [22], [23], [30]. Further discussion and details on indirect learning can be found from these sources and the references therein.

In this chapter, we define an *ILA iteration* as a single cycle of the following operations:

- (1) transmitting a predistorted signal block of length N (in the first ILA iteration the predistorter is turned off),
- (2) measuring the signal in the feedback loop and identifying the post-inverse of the PA, and
- (3) plugging the post-inverse parameter estimates into the predistorter.

The postdistorter coefficients in the identification step (2) are found as a solution to a selected optimization criterion, such as minimizing the least squares error or mean square error between the postdistorter output $\hat{z}_n$ and the reference signal z_n .

A noteworthy property of the ILA is that several such ILA iterations may be needed for the predistorter to fully converge, as noted for example in [20], [23]. This phenomenon is due to the dynamic and nonlinear nature of the estimation problem: the bandwidth and peak-to-average power ratio (PAPR) of the predistorted PA input signal change upon each PD update, until the PD has completely converged [16]. These changes are most notable after the first ILA iteration, when the first estimated parameters are plugged into the PD.

17.4.2 Block least squares estimation

With reference to Figure 17.6, we denote the postdistorter input signal (the complex envelope of the scaled PA output) by s_n , the postdistorter output signal by $\hat{z}_n$, and the training signal length by N . Assuming finite time spans for the filters $f_{p,n}$ and $\bar{f}_{q,n}$, denoted respectively by L_p and $\bar{L}_q$, we can write the postdistorter output signal (17.14) in vector-matrix notation as

$$\hat{z}_n = \sum_{p \in I_p} \psi_{p,n}^T \bar{\mathbf{f}}_p + \sum_{q \in I_Q} \bar{\psi}_{q,n}^T \bar{\mathbf{f}}_q + c'. \quad (17.17)$$

Above, the filter impulse response vectors are given as

$$\bar{\mathbf{f}}_p = [f_{p,0} \ f_{p,1} \ \cdots \ f_{p,L_p-1}]^T, \ p \in I_p \quad (17.18a)$$

$$\bar{\mathbf{f}}_q = [\bar{f}_{q,0} \ \bar{f}_{q,1} \ \cdots \ \bar{f}_{q,\bar{L}_q-1}]^T, \ q \in I_Q, \quad (17.18b)$$

and the input vectors to the filters are obtained from (17.15) as

$$\Psi_{p,n} \triangleq [\psi_p(s_n) \ \psi_p(s_{n-1}) \ \cdots \ \psi_p(s_{n-L_p+1})]^T, \ p \in I_P \quad (17.19a)$$

$$\bar{\Psi}_{q,n} \triangleq [\psi_q(s_n^*) \ \psi_q(s_{n-1}^*) \ \cdots \ \psi_q(s_{n-\bar{L}_q+1}^*)]^T, \ q \in I_Q. \quad (17.19b)$$

Notice that $\mathbf{f}_1$ and $\mathbf{f}_1$ are the *linear filters* of the predistorter, with respective lengths L_1 and $\bar{L}_1$, while $\mathbf{f}_p$ and $\mathbf{f}_p$, $p > 1$ are the *polynomial filters*.

We will now collect a block of N_B samples of the observed signal s_n , starting from sample index $n=0$ for notational convenience. We also assume that the loop delay has been estimated and removed from s_n . We can then write the output vectors of the p th non-conjugate and q th conjugate predistorter branches as

$$\mathbf{z}_p = \Psi_p \mathbf{f}_p \quad (17.20a)$$

$$\bar{\mathbf{z}}_q = \bar{\Psi}_q \bar{\mathbf{f}}_q. \quad (17.20b)$$

Here Ψ_p is the non-conjugate polynomial basis matrix of order p , given as

$$\Psi_p = \begin{bmatrix} \psi_p(s_0) & 0 & 0 & \cdots & 0 \\ \psi_p(s_1) & \psi_p(s_0) & 0 & \cdots & 0 \\ \psi_p(s_2) & \psi_p(s_1) & \psi_p(s_0) & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \psi_p(s_{N_B-1}) & \psi_p(s_{N_B-2}) & \psi_p(s_{N_B-3}) & \cdots & \psi_p(s_{N_B-L_p}) \\ 0 & \psi_p(s_{N_B-1}) & \psi_p(s_{N_B-2}) & \cdots & \psi_p(s_{N_B-L_p+1}) \\ 0 & 0 & \psi_p(s_{N_B-1}) & \cdots & \psi_p(s_{N_B-L_p+2}) \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \psi_p(s_{N_B-1}) \end{bmatrix} \quad (17.21)$$

and $\bar{\Psi}_q$ is the conjugate polynomial basis matrix of order q , constructed in a similar manner. The matrix Ψ_p has dimensions $(N_B + L_p - 1) \times L_p$, and similarly $\bar{\Psi}_q$ has dimensions $(N_B + \bar{L}_q - 1) \times \bar{L}_q$.

Then, we add zero rows to the bottom of all Ψ_p and $\bar{\Psi}_q$ to make them of equal height. By collecting all the polynomial basis matrices into a single block matrix, and appending it with a vector of all-ones (denoted $\mathbf{1}$) to account for the LO leakage compensator, we obtain the complete data matrix

$$\Psi = [\Psi_1 \ \Psi_2 \ \cdots \ \Psi_P \ \bar{\Psi}_1 \ \bar{\Psi}_2 \ \cdots \ \bar{\Psi}_Q \ \mathbf{1}]. \quad (17.22)$$

Finally, stacking the filter impulse responses of (17.18a) and (17.18b) and the LO leakage compensator coefficient into a single vector as

$$\mathbf{f} = [\mathbf{f}_1^T \ \mathbf{f}_2^T \ \cdots \ \mathbf{f}_P^T \ \mathbf{f}_1^T \ \mathbf{f}_2^T \ \cdots \ f_Q^T \ c']^T, \quad (17.23)$$

we can write the complete postdistorter output vector as

$$\hat{\mathbf{z}} = \Psi \mathbf{f}. \quad (17.24)$$

Using (17.18)–(17.24), and denoting the corresponding reference signal vector by $\mathbf{z} = [z_0, z_1, \dots, z_{N_B-1}, 0, \dots, 0]^T$ (appended with $\max(L_p, \bar{L}_q)$ zeros), the LS estimation problem is constructed as finding that parameter vector $\mathbf{f}$ which minimizes the cost function $J(\mathbf{f}) = \|\mathbf{z} - \hat{\mathbf{z}}\|^2$, yielding the well-known LS solution [26]

$$\hat{\mathbf{f}}_{LS} = (\Psi^H \Psi)^{-1} \Psi^H \mathbf{z}. \quad (17.25)$$

In practical implementations, numerical methods like QR decomposition or singular value decomposition (SVD) can be used to implement the pseudo-inverse $((\Psi^H \Psi)^{-1} \Psi^H)$ calculations in (17.25).

For practical purposes, it is important to note that the estimator in (17.25) is in fact *biased* for short observation block lengths. This is due to the employed windowing method that augments the data sequence with zeros outside the interval $[0, N_B - 1]$, known as *autocorrelation windowing* [26]. The *covariance windowing method*, on the other hand, makes no assumptions about the data outside the known interval, and is known to yield unbiased estimates. This approach involves deleting the “incomplete” rows from the data matrix Ψ , and dropping the corresponding samples from the reference signal vector $\mathbf{z}$, before calculating (17.25). Covariance windowing has been used in all simulations in this chapter that involve the block LS estimator.

17.4.3 Recursive least squares solution

The block solution described above is not the best choice for learning time-varying parameters, such as the predistorter parameters for a radio front-end. Time-dependent impairment characteristics may in practice result, e.g., from temperature changes in the operating environment or device ageing. Furthermore, practical FPGA implementations are not well equipped to handle large block LS solutions, especially in terms of matrix inverse or pseudo-inverse calculations, but are more at home with recursive real-time learning algorithms. Thus, the recursive LS (RLS) algorithm is a natural choice for the estimation stage here.

We now follow the notations in Figure 17.6 as well as in the above batch solution, and denote the RLS iteration index by $i = 1, 2, \dots, N_B$. A single update cycle of the recursive learning rule can then be formulated as

$$\begin{aligned} \mathbf{k}(i) &= \frac{\mathbf{M}(i-1)\mathbf{v}^*(i)}{\lambda + \mathbf{v}^T(i)\mathbf{M}(i-1)\mathbf{v}^*(i)} \\ e(i) &= d(i) - \mathbf{f}^T(i-1)\mathbf{v}(i) \\ \mathbf{f}(i) &= \mathbf{f}(i-1) + \mathbf{k}(i)e(i) \\ \mathbf{M}(i) &= \lambda^{-1}[\mathbf{M}(i-1) - \mathbf{k}(i)\mathbf{v}^T(i)\mathbf{M}(i-1)] \end{aligned} \quad (17.26)$$

Here $\mathbf{v}(i) = \Psi(i, :)$ denotes the i th row of the previous data matrix Ψ in (17.22), $\mathbf{f}(i)$ denotes the pre-distortion parameter vector estimate at iteration i , $d(i) = \mathbf{z}(i)$ denotes the reference signal value at iteration i , and finally λ is the internal memory or averaging

parameter of the recursion. For general discussions on recursive learning rules, refer e.g. to [26].

Here, when combining the indirect learning principle and the recursive least squares learning rule, there are basically several different possibilities of how to map the parameters between the post-inverse modeling and the actual predistortion stage. One feasible approach, as was proposed in the first ever paper to consider the ILA in PA PD identification [22], is to estimate *and* update the predistorter continually, in each RLS iteration. However, this approach has some drawbacks. The inevitable delay in updating the PD (resulting from the delay between transmitting the data samples and receiving them in the FBL, including all digital and analog delays) will lead to slower convergence of the algorithm, and can, in principle, even cause instability. The slower convergence will be illustrated in a simulation example in [Section 17.5](#). Another practical difficulty is that a prototype setup for a real-time updating algorithm is challenging to construct.

We propose a block-adaptive approach for learning as follows. First, start with pre-distortion set off ($\mathbf{f}(0) = \mathbf{0}$) and apply the recursive learning approach over a time window of N_B samples. After this, the learned pre-distortion parameters $\hat{\mathbf{f}}$ are deployed to actually pre-distort the transmit (reference) data and a new period (new ILA iteration) with recursive learning is deployed. In each ILA iteration, the parameter vector from the previous ILA iteration is used as the initial point in RLS learning. Furthermore, the matrix $\mathbf{M}(i)$, which is essentially the inverse of the input vector correlation matrix, can be either initialized with a suitable diagonal matrix in each ILA iteration, or retained between successive ILA iterations. Intuitively, it may be advantageous to reset $\mathbf{M}(i)$ between ILA iterations until the parameters have converged, i.e., for the first 2–4 ILA iterations, and after that to keep $\mathbf{M}(i)$ between ILA iterations. These steps effectively combine the indirect learning principle and recursive learning rules such that efficient parameter identification is obtained. This will be demonstrated in [Section 17.5](#) using both computer simulations as well as laboratory radio signal measurements.

17.4.4 Practical aspects

Measurement noise. So far, the feedback signal has been assumed noiseless. In practice, however, there is inevitable circuit and quantization noise in the feedback signal. Contrary to an ordinary LS problem where the noise would lie in the reference vector $\mathbf{z}$, in the current inverse modeling problem the measurement noise is in the data matrix Ψ . This kind of an estimation problem is generally called an errors-in-variables model [28], or a data least squares (DLS) problem [27] in signal processing literature. As is well known in the statistical literature and also shown in [16], applying LS estimation to a DLS problem will induce a *bias* to the parameter estimates. Fortunately, due to the transmitter internal processing, FBL noise levels are low, with signal-to-noise-ratios (SNR) in excess of 50 dB being quite easily obtainable in practice.

Synchronization. One of the main practical problems in implementing any adaptive predistorter is the requirement of very accurate time and frequency synchronization

between the measured and reference signals. Dedicated frequency synchronization is needless when both the up-converting and down-converting LO signals are derived from the same clock, as is usually the case. Time synchronization on the other hand is more problematic, since the loop time delay can only be known nominally beforehand. Thus, an efficient algorithm, capable of providing timing accuracy within about 2 percent of the symbol interval, is required in this application. This is rather challenging, since most delay estimators do not explicitly take into account either the PA nonlinearity or the I/Q mismatch, thus leading to biased estimates and occasional outlier estimates. To our knowledge, timing estimators that can handle these impairments are currently not available in the literature. Developing such an estimator is an important topic for future research in the field. On the other hand, it is possible to relax the demand for highly accurate timing estimation using a practical trick that is described next.

Causality of the PD filters. If there is a *negative delay* between the measured and reference signals, the resulting PD filters should be *anti-causal*, which is an impossible task in practice. A negative delay can arise if the loop time delay, which is removed from the feedback signal before the PD estimation stage, is overestimated. To make sure that the PD filters are causal and implementable, a nominal delay d must be introduced to them during the estimation stage. In practice, this is achieved by making the substitution $s_n \leftarrow s_{n-d}$ in (17.17) and in the estimation procedure following that. In effect, this will make the PD filters longer, but will give robustness against the negative time delay problem described above, and also relax the need for extremely accurate time delay estimation. A simulation example underlining this issue is presented in [Section 17.5](#).

Level mismatch. As discussed in detail in [9], and also mentioned in [Section 17.2](#), a level mismatch in the path between the PD and the PA will result in incomplete IMD mitigation. In the context of adaptive digital predistortion, this means that the gain of the overall transmitter signal path between the PD and the PA must not change, unless the PD coefficients are scaled accordingly or re-estimated.

17.5 Simulation and measurement examples

In the [previous sections](#), we have introduced a novel predistorter structure for the joint compensation of PA and I/Q modulator impairments, and proposed a recursive LS-based learning rule for parameter estimation utilizing the indirect learning architecture. Next, we assess the performance of the proposed solution through extensive computer simulations and laboratory measurements, and compare it with state-of-the-art techniques. The reference techniques used in the simulations and measurements are the plain PH PD without I/Q mismatch compensation (see, e.g., [17]) and the memoryless joint PA/IQ PD from [13]. These essentially represent state-of-the-art indirect learning techniques in the field. Both of these algorithms can be represented by using the formulations of the [previous section](#). The PH PD is obtained simply by setting $Q = 0$, while the memoryless joint PA/IQ algorithm is obtained with $L_p = \bar{L}_q = 1 \ \forall p, q$.

The chosen figure of merit is the adjacent channel power ratio (ACPR), defined as

$$\text{ACPR}_{\text{dB}} = 10 \log_{10} \frac{\int_{\omega_A} S(\omega) d\omega}{\int_{\omega_D} S(\omega) d\omega}, \quad (17.27)$$

with $S(\omega)$ denoting the power spectral density (PSD) of s_n (the complex envelope of the PA output signal), and ω_A and ω_D being the frequency bands of the (worse) adjacent channel and the desired channel, respectively. It is worth noting that there is a difference in adjacent channel content between the direct-conversion and the low-IF transmitters; in direct-conversion, the adjacent channel power consists only of spectral regrowth due to (remaining) PA nonlinearity, while in the low-IF case it also includes the mirror-frequency interference resulting from I/Q imbalance, and possibly also the tone due to LO leakage.

In the following simulations, two different PA models are employed. One is a Wiener system, meaning that there is a linear time-invariant filter followed by a static nonlinearity. The Wiener PA parameters are from [24], and have been reportedly extracted from a real class AB power amplifier. The filter has transfer function $(1 + 0.3z^{-2})/(1 - 0.2z^{-1})$, while the static nonlinearity is a fifth-order polynomial with complex-valued coefficients $a_1 = 14.9740 + 0.0519j$, $a_3 = -23.0954 + 4.9680j$, and $a_5 = 21.3936 + 0.4305j$. The other PA model is a Rapp solid-state amplifier model with smoothness factor $p = 1.5$ [29].

The I/Q imbalance parameters are the same in all simulations, and are as follows: gain imbalance 5 percent, phase imbalance 5 degrees, and impulse response mismatch $h_{TX} = [1, -0.035]^T$ (for more details on the I/Q mismatch model, refer to [Chapter 16](#) of this book, or [11]). These imbalance values represent a realistic practical scenario from the radio electronics point of view, with the resulting IRR varying between 23–30 dB within the transmit band.

17.5.1 Simulation example 1 – OFDM direct-conversion Tx

The example waveform used in the first simulation is an OFDM signal built according to the 3GPP Long-Term Evolution (LTE) specification's [25] 10 MHz mode, with 600 active subcarriers out of 1024, and with subcarrier spacing of 15 kHz. The signal is oversampled 4 times and low-pass filtered to attenuate the spectral sidelobes, before transmission. The PA model is the Wiener model described above. Due to the high PAPR of the OFDM waveform, the PA input power is backed off 5 dB from the PA's 1-dB compression point. As a practical example with feasible implementation complexity, the PD used consists of a ninth-order PD for the non-conjugate and a fifth-order PD for the conjugate signal. Both PDs are PH with branch filter lengths $\bar{L}_p = 5, 5, 4, 4, 3$ for $p = 1, 3, 5, 7, 9$, and $\bar{L}_q = 5, 4, 3$ for $q = 1, 3, 5$, respectively. The orthogonal polynomials from [20] are utilized. Including also the LO leakage compensator, there are altogether 34 complex coefficients to estimate. In each ILA iteration, 500 RLS recursions are used for estimating the coefficients using the recursive learning rule formulated in (17.26).

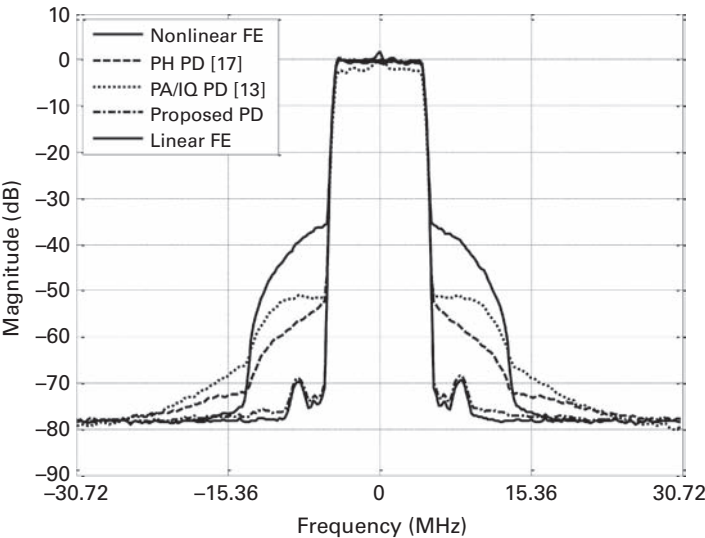


Figure 17.7 Output spectra of the power amplifier with ideal linear amplification, the Wiener PA model without predistortion and with the proposed joint PA and I/Q predistorter. Compared with the plain parallel Hammerstein PD and the memoryless joint PA/IQ PD from [13]. OFDM signal with pulse shaping, and feedback SNR=60 dB. 500 samples used for coefficient estimation in each ILA iteration. Reproduced from [30].

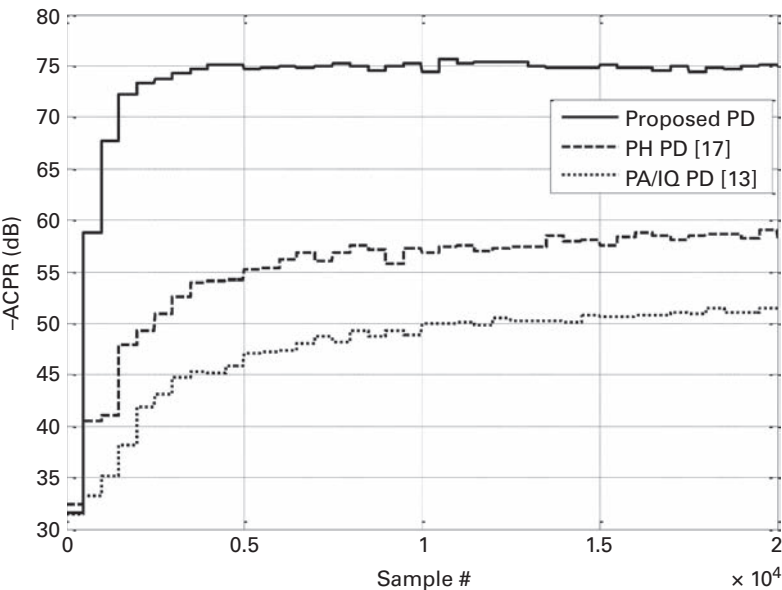


Figure 17.8 Evolution of adjacent channel power ratios over time with the proposed joint PA and I/Q predistorter, the plain parallel Hammerstein PD, and the memoryless joint PA/IQ PD from [13]. OFDM signal with pulse shaping, Wiener PA model, IBO=5 dB, and feedback SNR=60 dB. 500 samples used for coefficient estimation in each ILA iteration. Reproduced from [30].

Figure 17.7 shows the PSD of the PA output without predistortion, with the proposed PD structure, with the reference PD techniques, and with an ideal linear front-end (FE), averaged over 100 realizations. The proposed PD is able to reduce the adjacent channel interference significantly, and clearly outperforms the state-of-the-art reference techniques. Figure 17.8 shows the evolution of the ACPR as a function of sample index for the different algorithms. It can be seen that the proposed PD essentially converges after about 3000–4000 samples of transmitted data, corresponding to 6–8 ILA iterations.

The reference techniques fail to suppress the adjacent channel interference sufficiently. The PH PD does not take into account the I/Q imbalance, and is thus clearly biased. The joint PA and I/Q PD from [13], on the other hand, does not account for memory effects, explaining its poorer performance compared to the proposed PD.

17.5.2 Simulation example 2 – SC-FDMA low-IF Tx

The second simulation example tests the proposed algorithms on the single-carrier frequency division multiple access (SC-FDMA) waveforms of LTE uplink 10 MHz mode [25]. The signal is occupying subcarriers +101 through to +300, thus having a bandwidth of about 3 MHz around an IF of 3 MHz. The amplifier model is the Rapp solid-state amplifier model with smoothness parameter $p = 1.5$, while input back-off is 1 dB. The PD is ninth/fifth order with filter lengths $\bar{L}_p = 4, 4, 4, 3, 3$ and $\bar{L}_q = 4, 3, 3$, and it is again trained with 500 RLS recursions within each ILA iteration. Figure 17.9 shows

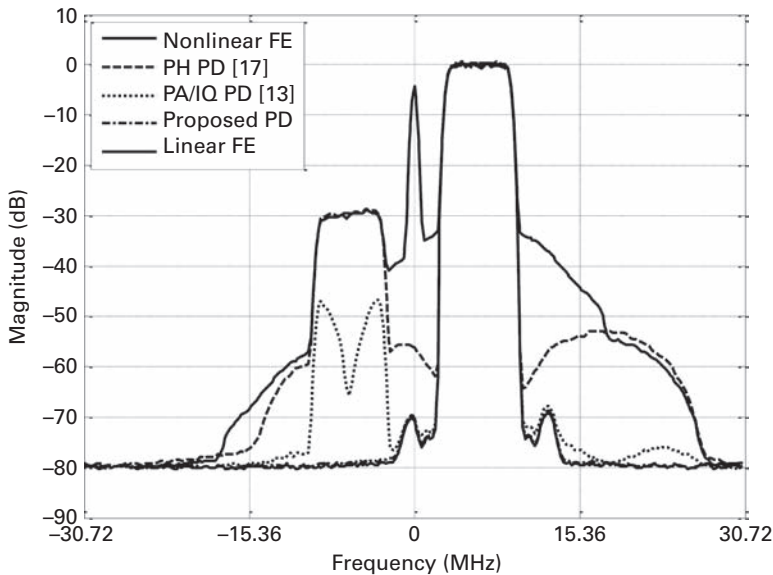


Figure 17.9 Output spectra of the power amplifier with ideal linear amplification, the Rapp PA model ($p=1.5$, IBO=1 dB) without predistortion and with the proposed joint PA and I/Q predistorter. Compared with the plain parallel Hammerstein PD and the memoryless joint PA/IQ PD from [13]. SC-FDMA 16-QAM signal with pulse shaping, and feedback SNR=60 dB. 500 samples used for coefficient estimation in each ILA iteration. Reproduced from [30].

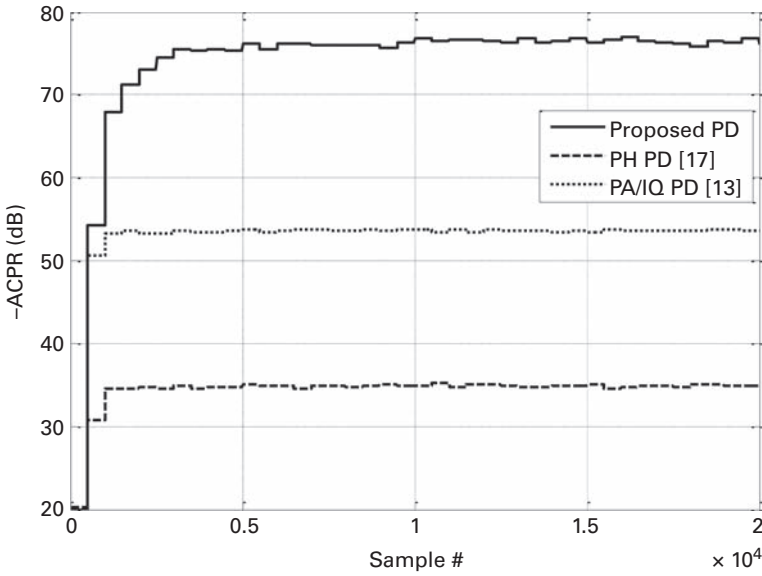


Figure 17.10 Evolution of ACPR over time with the proposed joint PA and I/Q predistorter, the plain parallel Hammerstein PD, and the memoryless joint PA/IQ PD from [13]. SC-FDMA signal with pulse shaping, Rapp PA model with $p=1.5$, IBO=1 dB, and feedback SNR=60 dB. 500 samples used for coefficient estimation in each ILA iteration. Reproduced from [30].

the PSD of the PA output without and with PD, plotted after the tenth ILA iteration, averaged over 100 realizations. The operation of the I/Q imbalance and LO leakage compensators are now seen more clearly than in the previous example since the MFI and LO leakage signal both fall outside the desired signal band. The proposed PD is able to push the spectral regrowth, mirror frequency interference, and LO leakage considerably down, in a way that practically no distortion is visible in the PSD. The memoryless joint PD is able to mitigate the spectral regrowth due to the (memoryless) PA nonlinearity successfully, as well as the LO leakage, but cannot mitigate the MFI fully. The plain PH PD can also mitigate the LO leakage and part of the spectral regrowth, but naturally fails to do anything to the MFI.

The evolution of the ACPR as a function of sample index is shown in Figure 17.10 for the different algorithms. The proposed PD again converges fully after about 3000–4000 samples of transmitted data, corresponding to 6–8 ILA iterations.

17.5.3 Simulation example 3 – comparison between block and recursive algorithms

Now, the different estimator variants of the proposed joint predistorter structure are compared: the block adaptive and sample adaptive RLS algorithms are compared with the block LS estimator. The signal is a 16-QAM single-carrier signal with 7.68 MHz symbol rate, oversampled 10-times, shaped with a square-root raised cosine filter with 22 percent roll-off, and sitting at a low IF of 5 MHz. The PA model is the same Wiener model as was used in the earlier OFDM example in Section 17.4.1. Each estimator uses a total of

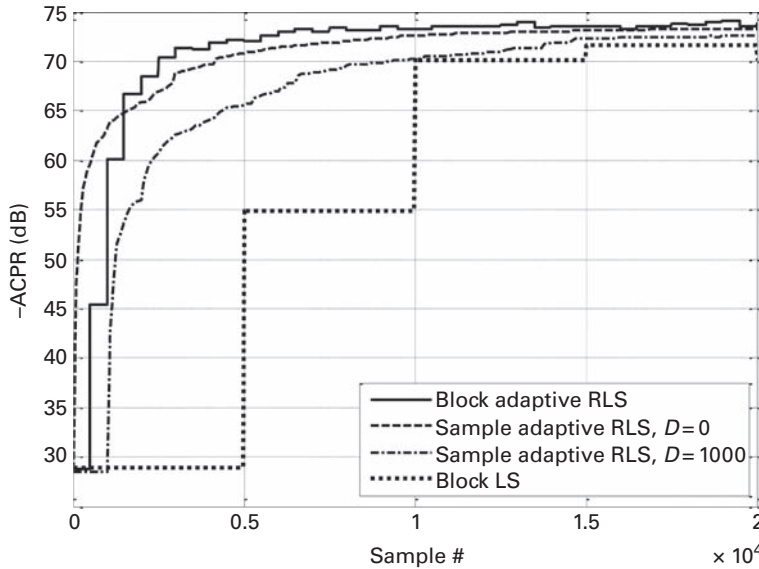


Figure 17.11 Evolution of ACPR vs. sample index with block LS, block adaptive RLS, and sample adaptive RLS estimators. 16-QAM low-IF signals. Wiener PA model with 2 dB input back-off. Reproduced from [30].

20 000 samples for learning. For the block LS estimator, 4 ILA iterations with 5000 samples in each are used for estimation. For the block adaptive RLS estimator, 40 ILA iterations with 500 samples in each are used. Figure 17.11 shows the ACPR as a function of sample index. The ACPR is evaluated in every 100th RLS iteration with a new signal whose length is 20 000 samples. The block LS estimator is the slowest to converge, while the block adaptive RLS is the fastest. The sample adaptive RLS algorithm with zero delay in PD updating is almost as fast as the block adaptive RLS. Having a delay of 1000 samples in updating the PD entails a clear slowing of the convergence. In summary, the block adaptive RLS estimator, as well as the sample adaptive RLS (if loop delay can be kept short), can provide fast learning curves in the proposed PD structure.

17.5.4 Simulation example 4 – effect of fractional delay

We assume a direct-conversion transmitter, 16-QAM single-carrier signal with symbol rate of 7.68 MHz, 8 times oversampling and root raised cosine pulse shape with excess bandwidth of 22 percent. The PA model is the Wiener model, and the I/Q imbalance model is the frequency-selective model described above. The PD used consists of a ninth-order PD for the non-conjugate and a fifth-order PD for the conjugate signal. The residual fractional delay of the FBL signal s_n with respect to the reference signal z_n is varied between $[-12.5\%, +12.5\%]$ of the symbol interval, and the resulting ACPR of the PA output is averaged over 100 independent trials. Figure 17.12 shows the ACPR as a function of the fractional delay for PD filters with $L_p = \bar{L}_q = 5$ and delays $d = 0$

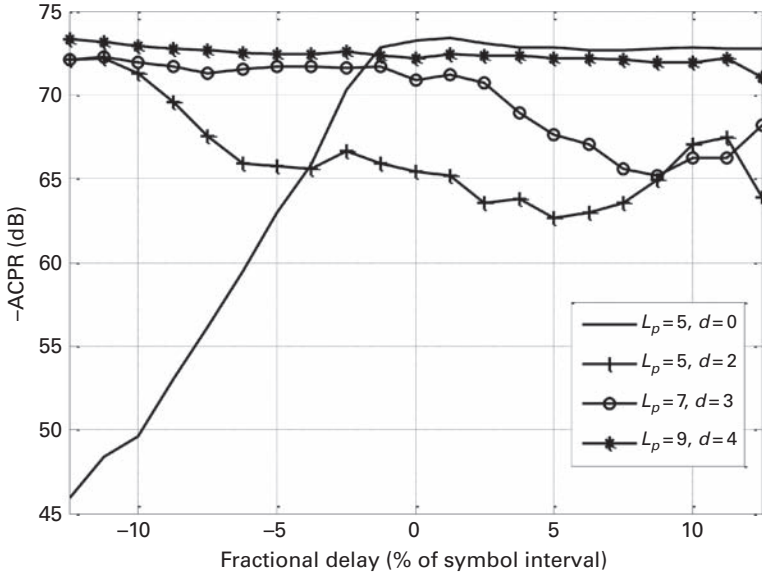


Figure 17.12 ACPR as a function of fractional delay for PD filters with $L_p = \bar{L}_q = 5$ and $d = 0$ or $d = 2$, for PD filters with $L_p = \bar{L}_q = 7$ and $d = 3$, and for PD filters with $L_p = \bar{L}_q = 9$ and $d = 4$. PD orders are $P = 9$, $Q = 5$ for all designs. Block LS estimation with 5000 samples per ILA iteration; result after fourth ILA iteration. No noise.

and $d = 2$, with $L_p = \bar{L}_q = 7$ and $d = 3$, and with $L_p = \bar{L}_q = 9$ and $d = 4$; PD non-linearity orders are $P = 9$, $Q = 5$ for all designs. Clearly, when delay becomes negative, the filter design with $d = 0$ deteriorates quickly. This happens because this design does not (and cannot) take into account the inherent anti-causality of the estimated system. The symmetric designs can tolerate both positive and negative delays, with improved performance as filter length L_p increases, and are very resilient against fractional delay of up to ± 12.5 percent of the symbol interval. Beyond $L_p = \bar{L}_q = 9$ for the symmetric design, and beyond $L_p = \bar{L}_q = 5$ for the zero-delay design, no significant performance improvement was seen, and these cases are therefore not shown.

17.5.5 Measurement results

Finally, results of laboratory radio signal measurements are presented. Here, Rohde&Schwartz (R&S) SMJ Vector Signal Generator (VSG) acts as the transmitter, the amplifier under test is the Mini-Circuits ZJL-3G wideband amplifier, and R&S FSG Spectrum/Signal Analyzer – equipped with a digital IF receiver inside – is used as the FBL receiver. Modulator I/Q imbalance (5 percent gain, 5 deg phase, frequency-independent) and LO leakage are introduced to the signal inside the SMJ transmitter. All other processing is done offline in MATLAB on a PC. Currently an online measurement

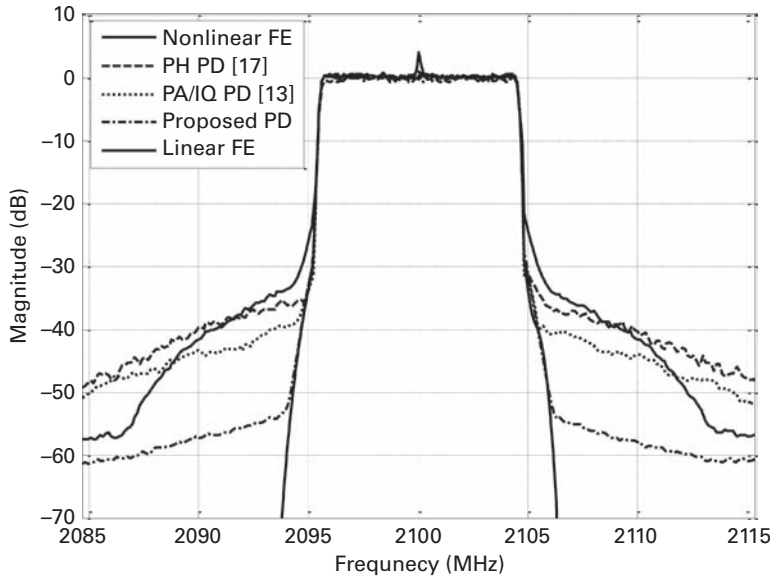


Figure 17.13 Spectra of measured amplifier output signals, averaged over 10 independent measurements. OFDM signals according to LTE downlink specifications, with 600 active subcarriers out of 1024, 15 kHz subcarrier spacing, and 4 times oversampling. PD orders are $P = 9$, $Q = 5$. Reproduced from [30].

demonstration is also under construction using the USRP/GNU radio demonstration environment [33]. These results will hopefully be reported in a future publication by the authors of this chapter.

The first measurement example consists of an OFDM signal with 600 active subcarriers (out of 1024), with a spacing of 15 kHz, thus corresponding to a fully loaded 3GPP LTE downlink signal in 10 MHz mode [25]. The signal is oversampled by 4 and then filtered before upconverting to 2.1 GHz carrier frequency. Predistorter is ninth/fifth order with filter lengths of $L_p = 5, 4, 3, 2, 2$ and $\bar{L}_q = 5, 3, 3$, and it is trained with 1500 RLS recursions in each ILA iteration. Figure 17.13 shows the PSD of the PA output without PD, with the reference techniques, and with the proposed PD structure after the tenth iteration (10 ILA cycles), averaged over 10 independent measurements. Clear performance improvement over the uncompensated case and the reference techniques can be obtained with the proposed method. The new joint PA/IQ PD yields 15–20 dB improvement in ACPR compared to the “nonlinear front-end” case, and 10–15 dB better results than the reference PDs.

In the second example, the signal is an SC-FDMA signal similar to simulation example 2, having altogether 250 active subcarriers out of 1024, occupying subcarriers +51 to +300, and with a spacing of 15 kHz. The effective IF is thus 2.63 MHz. The signal is oversampled 2 times and upconverted to 2.1 GHz for amplification and transmission. Predistorter is again ninth/fifth order with filter lengths of $L_p = 5, 4, 3, 2, 2$ and $\bar{L}_q = 5, 3, 3$. Figure 17.14 shows the measured amplifier output PSD without PD, with

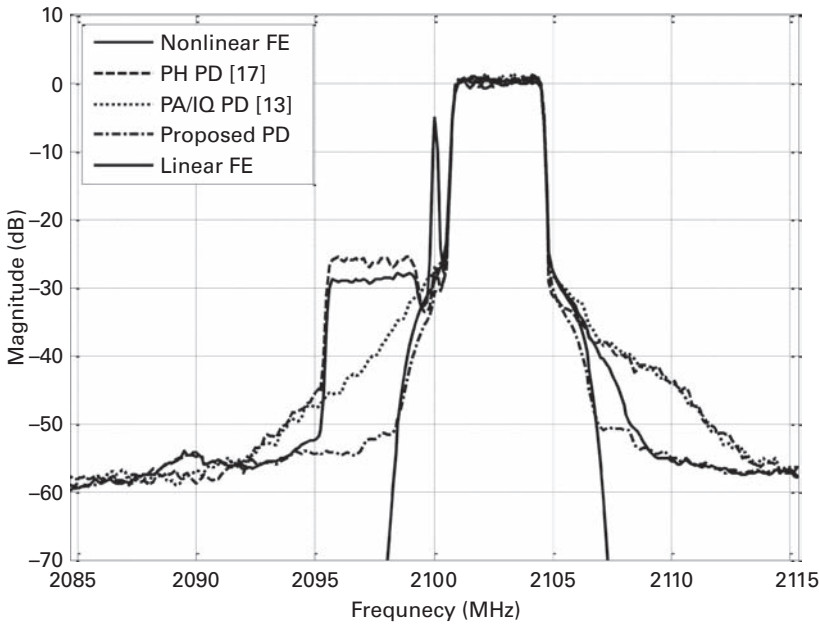


Figure 17.14 Spectra of measured PA output signals, averaged over 10 independent measurements. SC-FDMA signals according to LTE uplink specifications, with 250 active subcarriers out of 1024, 15 kHz subcarrier spacing, and 2 times oversampling. PD orders are $P = 9$, $Q = 5$. Reproduced from [30].

the reference techniques, and with the proposed PD, averaged over 10 measurement realizations. Significant performance improvement is evidenced also with measured SC-FDMA waveforms.

The signal in the third measurement example is a 16-QAM single-carrier signal with a symbol rate of 7.68 MHz, 8 times oversampling, and root-raised cosine pulse shape with 22 percent excess bandwidth. Figure 17.15 shows a representative example of the input–output relationship of the nonlinear front-end and the linearized front-end. The RF front-end, consisting of the R&S SMJ VSG and the Mini-Circuits ZJL-3G amplifier, has clear memory effects, evidenced by the spreading of the original nonlinear input–output plot. The block adaptive RLS is employed, and is updated for 10 ILA iterations with 1500 samples in each. The proposed method is able to reduce both nonlinear distortion and memory effects considerably.

17.6 Conclusions

The problem of joint estimation of IQM and PA predistorters was discussed. It was first illustrated that I/Q mismatch and LO leakage interact with PA nonlinearity such that extra IMD products appear at the PA output. Further, it was shown that these extra IMD products hinder the estimation and performance of an adaptive PA predistorter. A new

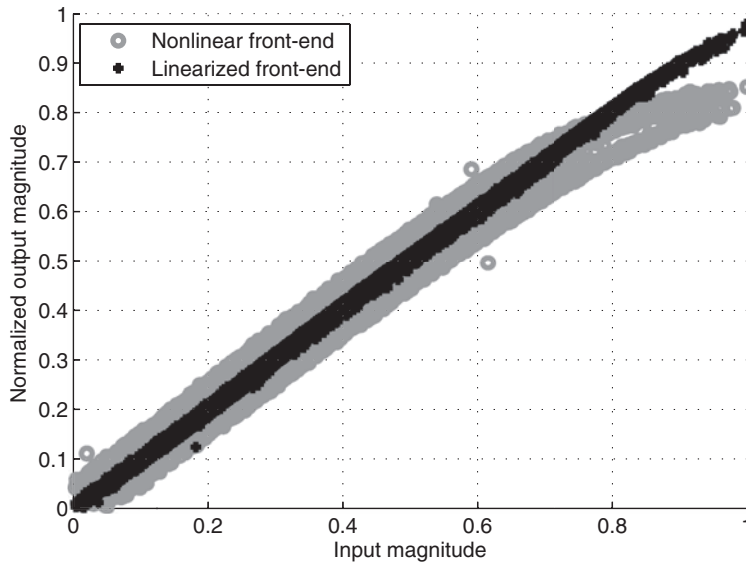


Figure 17.15 Normalized output magnitudes versus input magnitudes for the nonlinear and linearized RF front-end. 16-QAM single-carrier signal with symbol rate of 7.68 MHz, 8-times oversampling, and root-raised cosine pulse shape with 22% roll-off. Reproduced from [30].

predistorter structure was then presented, which is able to jointly mitigate both IQM and PA impairments. The new structure further assumes frequency-dependent behavior (memory) for the impairments, which is considered an important property for the predistortion of wideband waveforms of the emerging radio systems. The discussed PD structures consist of two Volterra-based predistorters plus the LO leakage compensator in parallel, and are thus fully linear with respect to the parameters to be determined, allowing easy estimation with linear least squares techniques. For practical parameter estimation implementations, a recursive learning rule was also presented, combining recursive least squares and indirect learning principles. The simulation and measurement analyses show excellent linearization and calibration performance, clearly outperforming the current state of the art. Future work includes building a real-time measurement setup for algorithm testing purposes, utilizing the USRP/GNU radio combo and commercial real-world PAs.

References

- [1] P.-I. Mak, S.-P. U, R. P. Martins. (2007) Transceiver Architecture Selection: Review, State-of-the-Art Survey and Case Study. *IEEE Circuits and Systems Magazine* 7, second quarter 2007, 6–25.
- [2] G. Fettweis, M. Löhning, D. Petrovic, *et al.* (2007) Dirty RF: A New Paradigm. *Springer International Journal of Wireless Information Networks* 14, 133–148.

- [3] Y. Zou, M. Valkama, M. Renfors. (2009) Performance Analysis of Spatial Multiplexing MIMO-OFDM Systems Under Frequency-Selective I/Q Imbalances. In *Proceedings of International Wireless Communication and Mobile Computing Conference*, Leipzig, Germany, June 2009.
- [4] A. Katz. (2001) Linearization: Reducing Distortion in Power Amplifiers. *IEEE Microwave Magazine*, Dec. 2001, 37–49.
- [5] W.-J. Kim *et al.* (2005) Digital Predistortion Linearizes Wireless Power Amplifiers. *IEEE Microwave Magazine*, Sept. 2005, 54–61.
- [6] J.K. Cavers. (1997) The Effect of Quadrature Modulator and Demodulator Errors on Adaptive Digital Predistorters for Amplifier Linearization. *IEEE Transactions on Vehicular Technology* **46**, 456–466.
- [7] J.K. Cavers. (1997) New Methods for Adaptation of Quadrature Modulators and Demodulators in Amplifier Linearization Circuits. *IEEE Transactions on Vehicular Technology* **46**, 707–716.
- [8] M. Faulkner, T. Mattsson, W. Yates. (1991) Automatic Adjustment of Quadrature Modulators. *Electronics Letters* **27**, 214–216.
- [9] M. Faulkner, T. Mattsson. (1992) Spectral Sensitivity of Power Amplifiers to Quadrature Modulator Misalignment. *IEEE Transactions on Vehicular Technology* **41**, 516–525.
- [10] L. Ding, Z. Ma, D.R. Morgan, M. Zierdt, G. T. Zhou. (2008) Compensation of Frequency-Dependent Gain/Phase Imbalance in Predistortion Linearization Systems. *IEEE Transactions on Circuits and Systems – Part I: Regular Papers* **55**, 390–397.
- [11] L. Anttila, M. Valkama, M. Renfors. (2008) Frequency-Selective I/Q Mismatch Calibration of Wideband Direct-Conversion Transmitters. *IEEE Transactions on Circuits and Systems – Part II: Express Briefs* **55**, 359–363.
- [12] X. Huang, M. Caron. (2007) Efficient Transmitter Self-Calibration and Amplifier Linearization Techniques. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 265–268. New Orleans, LA, May 2007.
- [13] Y.-D. Kim, E.-R. Jeong, Y.H. Lee. (2007) Adaptive Compensation for Power Amplifier Nonlinearity in the Presence of Quadrature Modulation/Demodulation Errors. *IEEE Transactions on Signal Processing* **55**, 4717–4721.
- [14] D.S. Hilborn, S.P. Stapleton, J.K. Cavers. (1994) An Adaptive Direct Conversion Transmitter. *IEEE Transactions on Vehicular Technology* **43**, 223–233.
- [15] H. Cao, A. S. Tehrani, C. Fager, T. Eriksson, H. Zirath. (2009) I/Q Imbalance Compensation Using a Nonlinear Modeling Approach. *IEEE Transactions on Microwave Theory and Techniques* **57**, 513–518.
- [16] L. Anttila, P. Händel, M. Valkama. (2010) Joint Mitigation of Power Amplifier and I/Q Modulator Impairments in Broadband Direct-Conversion Transmitters. *IEEE Transactions on Microwave Theory and Techniques* **58**, 730–739.
- [17] L. Ding, G. T. Zhou, D.R. Morgan, *et al.* (2004) A Robust Predistorter Constructed Using Memory Polynomials. *IEEE Transactions on Communications* **52**, 159–165.
- [18] M. Isaksson, D. Wisell, D. Rönnow. (2006) A Comparative Analysis of Behavioral Models for RF Power Amplifiers. *IEEE Transactions on Microwave Theory and Techniques* **54**, 348–359.
- [19] M. Isaksson, D. Rönnow. (2007) A Parameter-Reduced Volterra Model for Dynamic RF Power Amplifier Modeling Based on Orthonormal Basis Functions. *International Journal of RF and Microwave Computer-Aided Engineering* **17**, 542–551.

- [20] R. Raich, G. T. Zhou. (2004) Orthogonal Polynomials for Complex Gaussian Processes. *IEEE Transactions on Signal Processing* **52**, 2788–2797.
- [21] M. Schetzen. (1976) Theory of p th-Order Inverses of Nonlinear Systems. *IEEE Transactions on Circuits and Systems* **CAS-23**, 285–291.
- [22] C. Eun, E. J. Powers. (1997) A New Volterra Predistorter Based on the Indirect Learning Architecture. *IEEE Transactions on Signal Processing* **45**, 223–227.
- [23] D. R. Morgan *et al.* (2006) A Generalized Memory Polynomial Model for Digital Predistortion of RF Power Amplifiers. *IEEE Transactions on Signal Processing* **54**, 3852–3860.
- [24] L. Ding. (2004) Digital Predistortion of Power Amplifiers for Wireless Applications. Ph.D. dissertation. Georgia Institute of Technology, Atlanta, GA.
- [25] 3GPP Technical Specification Group Radio Access Network, “Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRAN); Overall description; Stage 2,” technical report TS 36.300, V1.0.0, March 2007.
- [26] S. Haykin. (1996) *Adaptive Filter Theory*, 3rd edn. Upper Saddle River, NJ: Prentice-Hall.
- [27] R. D. DeGroat, E. M. Dowling. (1993) The Data Least Squares Problem and Channel Equalization. *IEEE Transactions on Signal Processing* **41**, 407–411.
- [28] Z. Griliches, V. Ringstad. (1970) Errors-in-the-Variables Bias in Nonlinear Contexts. *Econometrica* **38**, 368–370.
- [29] C. Rapp. (1991) Effects of HPA-Nonlinearity on a 4-DPSK/OFDM-Signal for a Digital Sound Broadcasting System. In *Proceedings of the Second European Conference on Satellite Communications*, Liege, Belgium, Oct. 22–24, 1991, pp. 179–184.
- [30] L. Anttila, P. Händel, O. Mylläri, M. Valkama. (2010) Recursive Learning Based Joint Digital Predistorter for Power Amplifier and I/Q Modulator Impairments. *EuMA International Journal of Microwave and Wireless Technologies* **2**(2), 173–182.
- [31] A. Gökçeoglu, L. Anttila, M. Valkama. (2010) Volterra-Based Joint Predistortion of Power Amplifier and I/Q Modulator Impairments (In preparation).
- [32] G. Tong Zhou, H. Qian, L. Ding, R. Raich. (2005) On the Baseband Representation of a Bandpass Nonlinearity. *IEEE Transactions on Signal Processing* **53**, 2953–2957.
- [33] GNU radio website: <http://gnuradio.org/>

18 Diversity and error compensation in OFDM transceivers: principles and implementation

Tomohisa Wada

18.1 Introduction

Digital modulation techniques can be largely divided into two categories. One is single-carrier modulation, which utilizes a single radio frequency (RF) to transmit data. The other is multi-carrier modulation, which utilizes simultaneously modulated multiple RF carriers in order to combat inter symbol interference (ISI) while increasing communication bandwidth. This chapter focuses on a particular type of multi-carrier modulation known as Orthogonal Frequency Division Multiplexing (OFDM). The idea of OFDM [1] was proposed in the 1960s followed a few years later by the Discrete Fourier Transform (DFT) based implementation algorithm [2]. Then, OFDM became practical and has been popular in a number of applications such as asymmetric digital subscriber line (ADSL), wireless local area network (WLAN), digital TV broadcasting (DTV). It also has become a strong candidate for 4th generation cellular land mobile radio system.

It is well known that OFDM modulation and demodulation can be implemented by IDFT and DFT. But in actual implementation, not only those DFTs but also several error compensation mechanisms are indispensable because orthogonality between parallel transmitted subcarrier signals are easily destroyed by synchronization errors such as RF error, clock sampling rate error, and FFT window position shift. When applying OFDM technology for mobile communication, Doppler induced RF errors easily distort a reception performance. Therefore the main contents of this chapter are two fundamental techniques to realize high-performance OFDM communication systems. These are Diversity technologies and Synchronization error Detection and Compensation methods. After this, real hardware implementation examples and their data are summarized.

The rest of this chapter is organized into five sections. [Section 18.2](#) is devoted to presenting the principles and design of OFDM transceivers. [Section 18.3](#) focuses on adaptive antenna array and several diversity techniques. In [Section 18.4](#), we will deal with the detection and compensation of various error sources and practical factors. [Section 18.5](#) presents the hardware implementation and related chip performance. In the [final section](#), we will give a summary and there will be further discussion.

18.2 OFDM transceivers: principles and design

18.2.1 Single-carrier vs. multi-carrier modulation

In single-carrier modulation, data are sent in serial by modulating one single carrier frequency at a baud rate of R symbols per second. Then symbol length ($1/R$) has to be shortened in order to increase data transmission bandwidth as shown in Figure 18.1. However, wireless communication systems usually suffer from multipath fading. When the multipath delay time becomes significant compared with the symbol length, ISI happens and degrades wireless communication performance. Then it is not easy to increase the communication bandwidth by just reducing the symbol length.

The idea of multi-carrier modulation is to use long symbol length against ISI but to use multiple carrier frequencies (called subcarriers) to increase the communication bandwidth. Figure 18.2 shows an example of a Baseband OFDM signal, which consists of $n = 0$ to 6 seven sinusoids of frequency $= n \cdot f_0$ when symbol length $T = 1/f_0$. By increasing the number of subcarriers, one symbol length can be kept long while achieving high bandwidth data transmission.

18.2.2 OFDM signal generation

An OFDM signal consists of plural orthogonal carrier frequencies based on the following three orthogonal relations. In the equations, m and n are integers and $T = 1/f_0$.

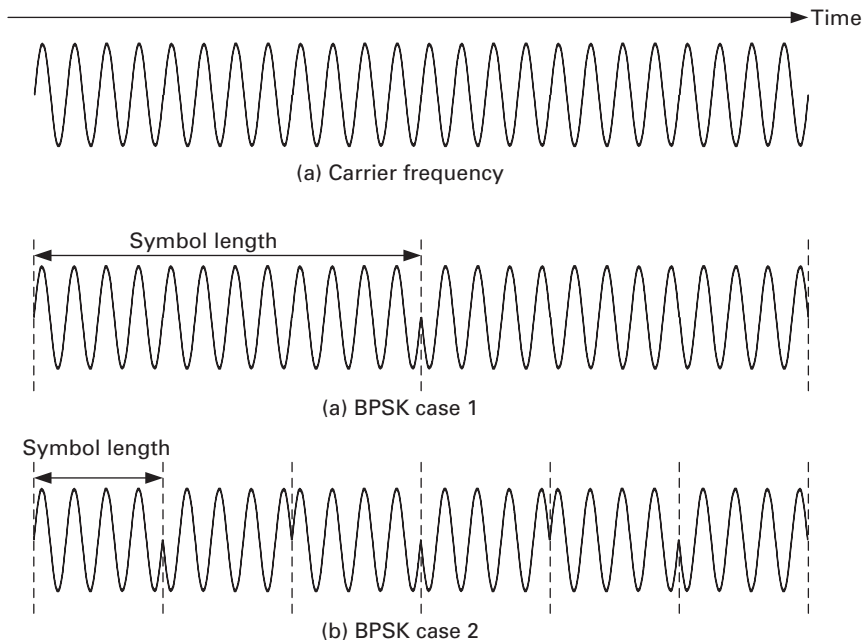


Figure 18.1 Single-carrier modulation.

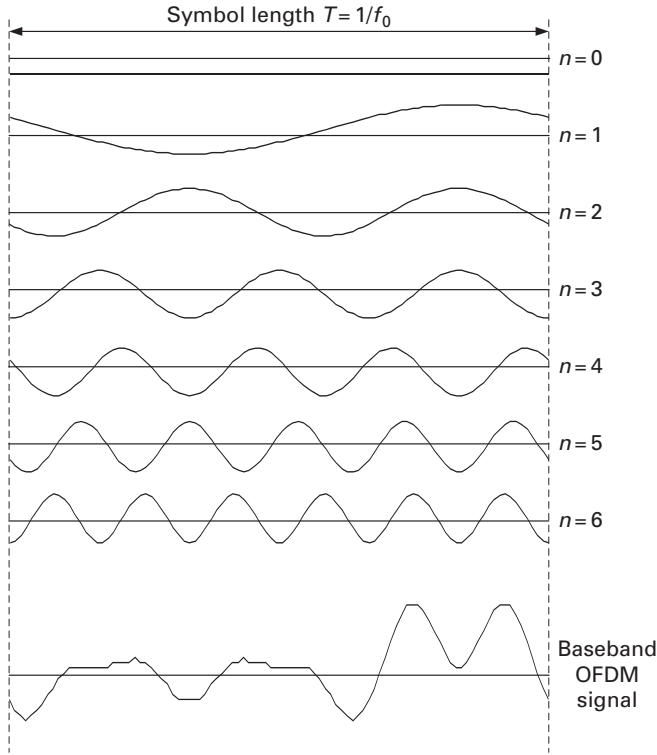


Figure 18.2 Subcarrier and baseband OFDM signal.

$$\int_0^T \cos(2\pi m f_0 t) \cdot \cos(2\pi n f_0 t) dt = \begin{cases} \frac{T}{2} & (m = n) \\ 0 & (m \neq n) \end{cases} \quad (18.1)$$

$$\int_0^T \sin(2\pi m f_0 t) \cdot \sin(2\pi n f_0 t) dt = \begin{cases} \frac{T}{2} & (m = n) \\ 0 & (m \neq n) \end{cases} \quad (18.2)$$

$$\int_0^T \cos(2\pi m f_0 t) \cdot \sin(2\pi n f_0 t) dt = 0. \quad (18.3)$$

One subcarrier signal, which is a basic component of the OFDM signal, in period $[0, T]$ is expressed as Equation (18.4). In one symbol length T , precisely n cycles of sinusoid signal are included. By summing the N component signals, a baseband OFDM signal can be obtained as Equation (18.5). Here, f_0 is neighboring subcarrier spacing:

$$a \cos(2\pi n f_0 t) - b \sin(2\pi n f_0 t) \quad (18.4)$$

$$s_B(t) = \sum_{n=0}^{N-1} \{a_n \cos(2\pi n f_0 t) - b_n \sin(2\pi n f_0 t)\}. \quad (18.5)$$

In practical OFDM transceiver design, a complex baseband OFDM signal is used in signal processing instead of a real baseband OFDM signal $s_B(t)$, which can be shown by using complex baseband OFDM signal $u(t)$ as Equation (18.6). In Equation (18.7), d_n is a complex expression of digital modulation symbols such as $d_n = a_n + jb_n$,

$$s_B(t) = \text{Re}[u(t)] \quad (18.6)$$

$$u(t) = \sum_{n=0}^{N-1} d_n e^{j2\pi n f_0 t}. \quad (18.7)$$

If $u(t)$ is sampled by N points during the one symbol period ($1/f_0$), the time index t can be substituted by $t = (k/Nf_0)$ as shown in Equation (18.8):

$$\begin{aligned} u\left(\frac{k}{Nf_0}\right) &= \sum_{n=0}^{N-1} d_n e^{j2\pi n f_0 \frac{k}{Nf_0}} \\ &= \sum_{n=0}^{N-1} d_n e^{j\frac{2\pi nk}{N}} \\ &= \sum_{n=0}^{N-1} d_n \left(e^{j\frac{2\pi k}{N}}\right)^n \quad (k = 0, 1, 2, \dots, N-1) \end{aligned} \quad (18.8)$$

According to Equation (18.8), it is clear that the complex baseband OFDM signal can be generated by an N -points Inverse Discrete Fourier Transform (IDFT) of complex digital modulation symbols d_n ($n = 0, 1, \dots, N-1$).

The passband signal $s(t)$ which is actually transmitted from an antenna is shown in Equation (18.9). Here, f_c is the minimum frequency of passband. Then, the passband signal consists of $[f_c, f_c + 1f_0, f_c + 2f_0, \dots, f_c + (N-1)f_0]$ frequencies,

$$\begin{aligned} s(t) &= \sum_{n=0}^{N-1} [a_n \cos\{2\pi(f_c + nf_0)t\} - b_n \sin\{2\pi(f_c + nf_0)t\}] \\ &= \text{Re} \left[\sum_{n=0}^{N-1} d_n e^{j2\pi(f_c + nf_0)t} \right] \\ &= \text{Re} \left[e^{j2\pi f_c t} \sum_{n=0}^{N-1} d_n e^{j2\pi n f_0 t} \right] \\ &= \text{Re} \left[(\cos 2\pi f_c t + j \sin 2\pi f_c t) \sum_{n=0}^{N-1} d_n e^{j2\pi n f_0 t} \right] \\ &= \cos 2\pi f_c t \cdot \text{Re} \left[\sum_{n=0}^{N-1} d_n e^{j2\pi n f_0 t} \right] + (-\sin 2\pi f_c t) \text{Im} \left[\sum_{n=0}^{N-1} d_n e^{j2\pi n f_0 t} \right]. \end{aligned} \quad (18.9)$$

According to Equation (18.9), an OFDM transmitter block diagram is shown in Figure 18.3. First, bit information is mapped onto complex modulation symbols d_n . Then the serial-to-parallel converter makes an N -points symbol vector d_n ($n = 0, \dots, N-1$). By Inverse Discrete Fourier Transform, an OFDM complex baseband signal

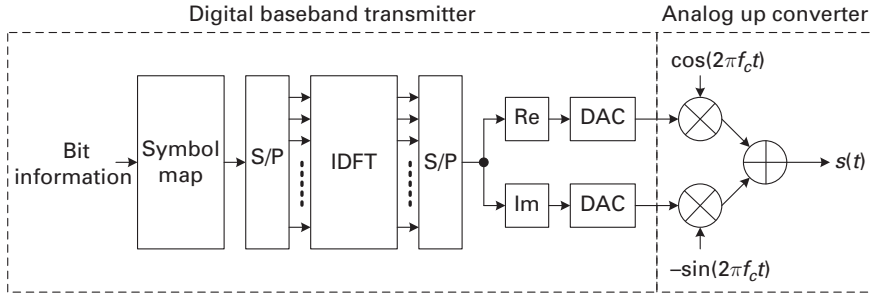


Figure 18.3 OFDM transmitter block diagram.

$u(k/Nf_0)$ ($k=0, \dots, N-1$) is generated. After converting the digital baseband signal to analog, a passband OFDM signal $s(t)$ is generated by some up-converters.

18.2.3 OFDM signal reception

In this section, OFDM demodulation from the passband signal in Equation (18.10) is explained. Since the passband signal is high frequency, an analog down-conversion process is needed before the OFDM digital demodulation process

$$s(t) = \sum_{n=0}^{N-1} [a_n \cos\{2\pi(f_c + nf_0)t\} - b_n \sin\{2\pi(f_c + nf_0)t\}]. \quad (18.10)$$

As shown in the following, a typical direct conversion performs the multiplication of f_c frequency sinusoids to passband signal and then passes a low-pass filtering.

$$\begin{aligned} s(t)2\cos(2\pi f_c t) &= 2 \sum_{n=0}^{N-1} [a_n \cos\{2\pi(f_c + nf_0)t\} \cos(2\pi f_c t) \\ &\quad - b_n \sin\{2\pi(f_c + nf_0)t\} \cos(2\pi f_c t)] \\ &= \sum_{n=0}^{N-1} \{a_n \cos(2\pi n f_0 t) - b_n \sin(2\pi n f_0 t)\} \\ &\quad + \sum_{n=0}^{N-1} [a_n \cos\{2\pi(2f_c + nf_0)t\} - b_n \sin\{2\pi(2f_c + nf_0)t\}] \end{aligned} \quad (18.11)$$

Then, by applying the low-pass filter to Equation (18.11), the $2f_c$ component can be removed and the $s_I(t)$ signal is obtained as Equation (18.12).

$$s_I(t) = \sum_{n=0}^{N-1} \{a_n \cos(2\pi n f_0 t) - b_n \sin(2\pi n f_0 t)\}. \quad (18.12)$$

This signal is equivalent to the baseband OFDM signal in Equation (18.5). In this process, a precise f_c frequency sinusoid is used. In a general down-converter the local

oscillator's frequency includes frequency error or frequency drift on time. Then additional RF error detection and compensation is required. These topics on RF error detection and compensation will be explained in the following related sections.

In order to obtain the complex baseband OFDM signal as Equation (18.7), another down-conversion process is needed:

$$\begin{aligned}
 s(t)2\{-\sin(2\pi f_c t)\} \\
 &= \sum_{n=0}^{N-1} \{a_n \sin(2\pi n f_0 t) + b_n \cos(2\pi n f_0 t)\} \\
 &+ \sum_{n=0}^{N-1} [-a_n \sin\{2\pi(2f_c + n f_0)t\} - b_n \cos\{2\pi(2f_c + n f_0)t\}].
 \end{aligned} \tag{18.13}$$

Again, by low-pass filtering, $s_Q(t)$ signals are obtained as Equation (18.14):

$$s_Q(t) = \sum_{n=0}^{N-1} \{a_n \sin(2\pi n f_0 t) + b_n \cos(2\pi n f_0 t)\}. \tag{18.14}$$

By combining $s_I(t)$ and $s_Q(t)$, a complex baseband signal $u(t)$ can be obtained:

$$\begin{aligned}
 u(t) &= s_I(t) + js_Q(t) = \sum_{n=0}^{N-1} (a_n + jb_n) \{\cos(2\pi n f_0 t) + j \sin(2\pi n f_0 t)\} \\
 &= \sum_{n=0}^{N-1} d_n e^{j2\pi n f_0 t}.
 \end{aligned} \tag{18.15}$$

By a sampling process such as N -points during the one symbol period ($1/f_0$), this equation can be converted as follows:

$$u\left(\frac{k}{Nf_0}\right) = \sum_{n=0}^{N-1} d_n \left(e^{j\frac{2\pi}{N}}\right)^{nk} \quad (k = 0, 1, 2, \dots, N-1) \tag{18.16}$$

As described in the [previous section](#), this is an IDFT operation. Then the complex digital modulation symbol d_n ($n=0, 1, \dots, N-1$) can be calculated by Inverse IDFT operation, that is, Discrete Fourier Transform (DFT) as shown in Equation (18.17):

$$d_l = \frac{1}{N} \sum_{k=0}^{N-1} u\left(\frac{k}{Nf_0}\right) \left(e^{-j\frac{2\pi}{N}}\right)^{lk} \quad (l = 0, 1, 2, \dots, N-1) \tag{18.17}$$

In [Figure 18.4](#), a simple OFDM receiver block diagram is shown. Here, it is assumed that the received OFDM signal is completely the same as the transmitted OFDM signal. However, in real wireless communication, the received signal is distorted. Then the OFDM receiver needs an additional signal processing block to remove the distortion such as an equalizer.

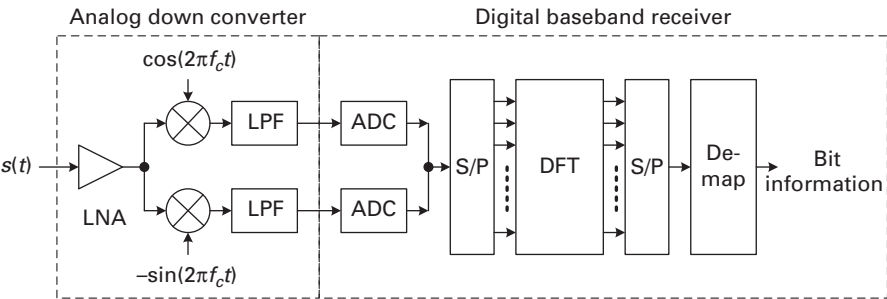


Figure 18.4 Simple OFDM receiver block diagram.

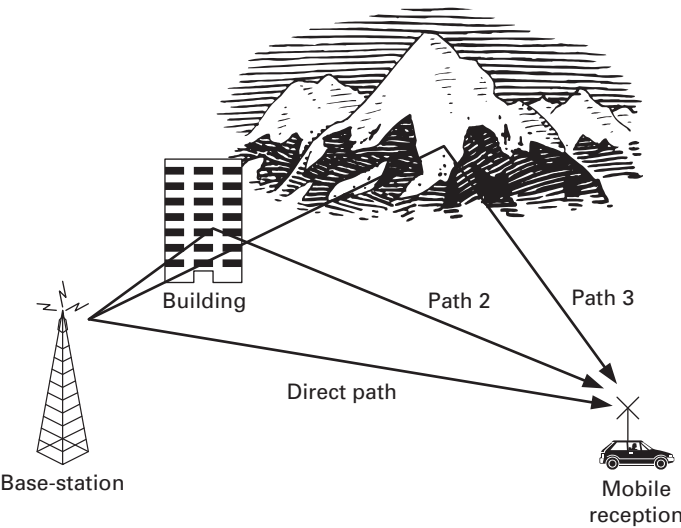


Figure 18.5 Multipath channel.

18.2.4 Multipath channel

In the real wireless communication channel, the RF wave propagates in space where refraction and reflection exist. Multiple propagating paths (multipath) exist between transmitter antenna and receiver antenna as shown in Figure 18.5. Reflecting path2 and path3 are longer propagation paths compared with the direct path. Then at the receiver antenna, multiple delayed and phase-rotated signals are overlapped and cause ISI. Since OFDM is especially sensitive to ISI, a Guard Interval (GI) is usually inserted between OFDM symbols.

A typical guard interval between OFDM symbols is a copy of the tail part of the OFDM symbol as shown in Figure 18.6. Please note that the connection point between the GI and OFDM symbol is smoothly connected because all the components of the OFDM symbol have cycle times of $1/(nf_0)$. This type of guard interval is called cyclic prefix (CP). Since the insertion of the guard interval T_g reduces data rate, a shorter GI length is preferable. Typical GI length is 1/16 to 1/4 of OFDM symbol length $T=1/f_0$.

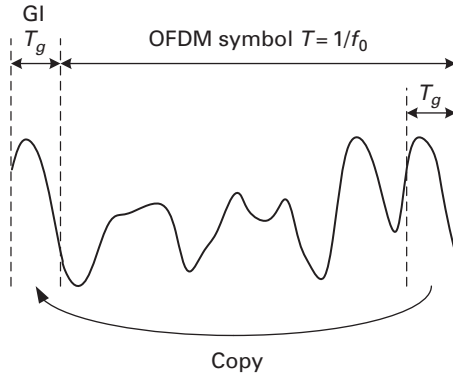


Figure 18.6 Guard interval.

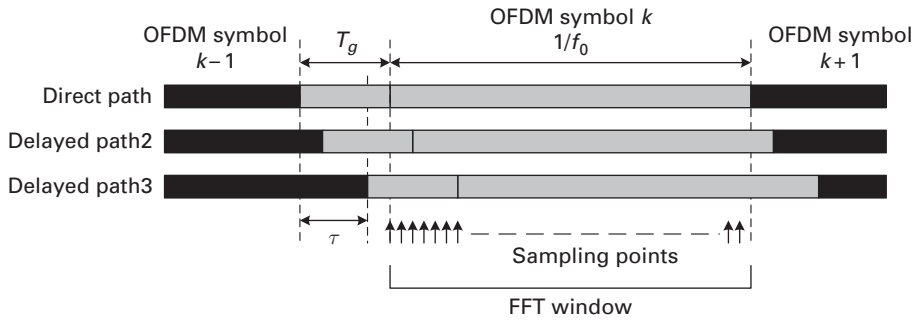


Figure 18.7 Multipath with GI

Figure 18.7 shows how the GI mitigates the ISI. An OFDM receiver performs FFT operation at the FFT window position. If a CP type GI is used and the maximum multipath delay path τ is shorter than T_g , ISI can be removed perfectly and the orthogonality between subcarriers is maintained. Then the equalizer which removes the multipath distortion becomes very simple. However if zero signals (zero padding) are used in the GI period instead of CP, the orthogonality between subcarriers is destroyed and inter carrier interference (ICI) is generated. Then a complicated equalizer is required [3]. In this chapter, only CP type GI is assumed.

18.2.5 Channel estimation and equalizer

Assuming that $h(t)$ is the impulse response of a transmission channel for an OFDM signal, the reception signal $s_R(t)$ can be expressed as follows using convolution:

$$s_R(t) = \int_{-\infty}^{\infty} h(u)s(t-u)du + n_c(t). \quad (18.18)$$

The first term is distortion caused by the multipath channel and the second term is noise addition in the OFDM communication frequency band. If the width of the channel impulse response $h(t)$ is shorter than GI length T_g , the channel transfer function $H(f)$ is given as:

$$H(f) = \int_0^{T_g} h(t) e^{-j2\pi f t} dt \quad (18.19)$$

Since a convolution operation can be replaced by multiplication after DFT, the following equation can be obtained:

$$x(k, l) = H(f_c + lf_0) d(k, l) + N_B(k, l), \quad (18.20)$$

where, k and l are symbol number and subcarrier number, respectively. The $d(k, l)$ and $x(k, l)$ are the complex digital modulation symbols of transmitter and receiver at symbol number k and subcarrier number l ; $N_B(k, l)$ is the frequency domain complex noise. Since $H(f_c + lf_0)$ is a complex number, Equation (18.20) means that received symbol $x(k, l)$ is distorted in amplitude and phase from $d(k, l)$ and noise is added. Once the channel transfer function $H(f_c + lf_0)$ is obtained, the original transmission symbol can be estimated by the following equation. This process is called equalization. Therefore, if the noise is enough small, the equalization process can recover the original symbol $d(k, l)$ from the received symbol $x(k, l)$:

$$\begin{aligned} \hat{d}(k, l) &= \frac{x(k, l)}{H(f_c + lf_0)} \\ &= d(k, l) + \frac{N_B(k, l)}{H(f_c + lf_0)} \end{aligned} \quad (18.21)$$

18.2.6 OFDM communication and broadcasting systems

Table 18.1 summarizes the state-of-the-art OFDM or OFDMA based communication and broadcasting systems. A wide variety of symbol lengths is used from 3.2 μ s to 1008 μ s for various applications. As explained in the [previous section](#), GI length is related to the multipath delay profile. Then, the larger the service area becomes, the longer an OFDM

Table 18.1 OFDM applications and effective symbol length.

Application of OFDM	Effective OFDM symbol length $1/f_0$
WLAN (802.11 g/a/n)	3.2 μ s
XGP (eXtended Global Platform)	26.67 μ s
3.9G mobile phone LTE	66.67 μ s
WiMAX (802.16e)	102.4 μ s
3.9G mobile phone LTE (Extended mode)	133.33 μ s
China Digital TV Broadcasting DTMB	500.0 μ s
Europe Digital TV Broadcasting DVB-T	896 μ s (8K mode)
Japan Digital TV Broadcasting ISDB-T	1008 μ s (8K mode)

symbol length is used. WLAN and three digital TV broadcasting systems are one-to-one data transfer at a time based on OFDM. XGP [4], LTE[5], and WiMAX [6] are one base-station to multiple user terminal communication based on Orthogonal Frequency Division Multiple Access (OFDMA).

A more detailed OFDM communication and broadcasting system block diagram is shown in Figure 18.8. For the transmitter side, GI insertion is added to Figure 18.3. Since the receiver side needs synchronization to the transmitter, more blocks are added. The resampler represents the sampling frequency error compensation block. The de-rotator rotates the baseband OFDM signal to compensate for radio frequency error. The sampling frequency error is caused by system sampling clock accuracy. Then, in the case that enough high accuracy crystal oscillators are used in the system design, the resampler might not be needed. However, radio frequency error is caused by the frequency difference between up-conversion and down-conversion. Since the conversion frequency is as

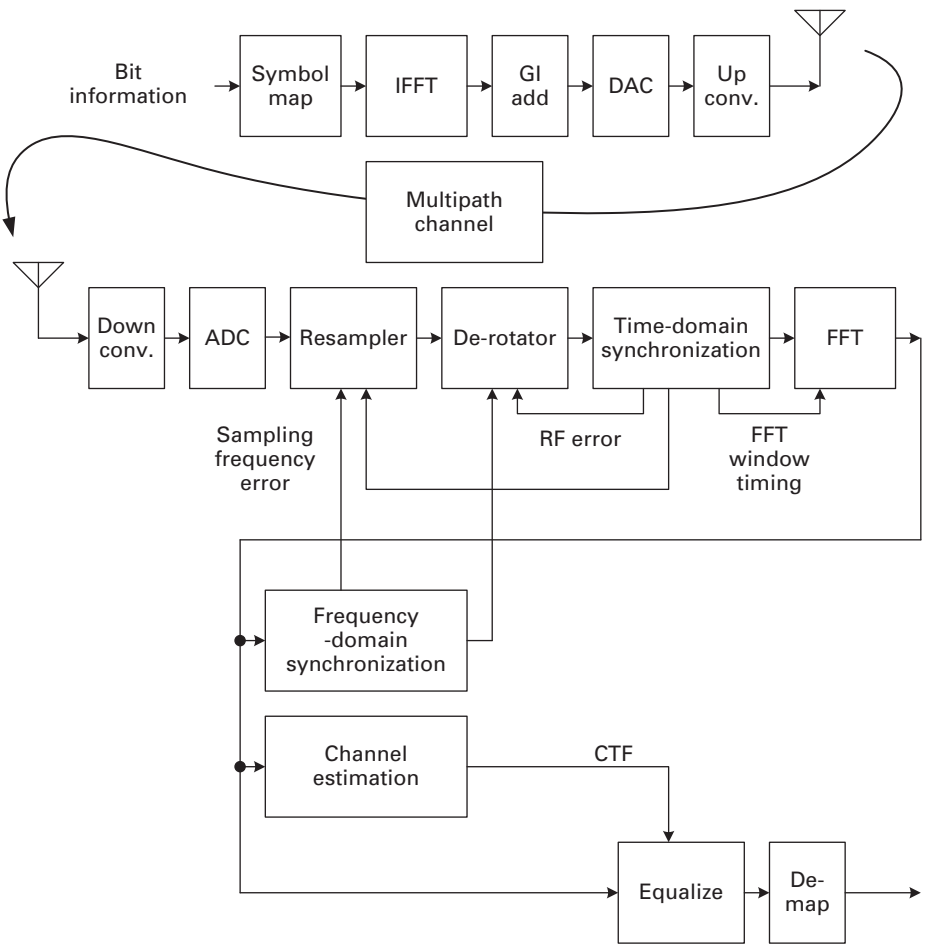


Figure 18.8 OFDM communication and broadcasting system.

large as mega to giga Hz order, the frequency error usually exists and the de-rotator is always needed. Those synchronization error compensation issues will be covered in later sections. The FFT block needs to know the FFT window position as shown in [Figure 18.7](#). Then the synchronization block also generates FFT window position timing:

- radio frequency (RF) synchronization
- sampling rate synchronization
- symbol timing (FFT window position) synchronization.

These three synchronizations have to be considered to implement a real OFDM system. The channel transfer function (CTF) has to be estimated to remove the distortion by multipath channel. In [Figure 18.8](#), the following synchronization functions are assumed as

- time domain RF and sampling rate error detection
- time domain symbol timing (FFT window position) detection
- frequency domain RF and sampling rate error detection
- frequency domain channel estimation.

In addition, time domain channel estimation and frequency domain symbol timing (FFT window position) detection are also possible. Suitable detection and estimation algorithms have to be applied depending on the applications.

18.2.7 OFDM WLAN design

18.2.7.1 WLAN system overview

The IEEE 802.11 standard WLAN is successfully utilized in the market. The first generation IEEE 802.11 standard was made at the end of 1996 to provide 1 or 2 Mbps in the 2.4 GHz Industry-Science-Medical (ISM) band. Later, in order to provide a higher data rate, the OFDM-based physical layer standard was established as IEEE 802.11a/g. Since 802.11a/g is based on distributed carrier sense multiple accesses with collision avoidance (CSMA/CA), in the wireless channel only one packet of OFDM signal is transmitted from one to another. [Table 18.2](#) shows IEEE 802.11a/g OFDM based WLAN parameters. It supports from 6Mbps to 54Mbps data rates. The effective OFDM symbol

Table 18.2 OFDM WLAN parameters (802.11a/g).

Data rate (Mbps)	6	9	12	18	24	36	48	54
Modulation	BPSK	BPSK	QPSK	QPSK	16QAM	16QAM	64QAM	64QAM
Code rate	1/2	3/4	1/2	3/4	1/2	3/4	2/3	3/4
# of subcarriers					52			
# of pilots					4			
OFDM symbol length				4 μ s (effective symbol = 3.2 μ s)				
Guard Interval length					0.8 μ s			
Subcarrier spacing				312.5 kHz = 1/3.2 μ s				
Sampling rate				20 MHz (50ns)				

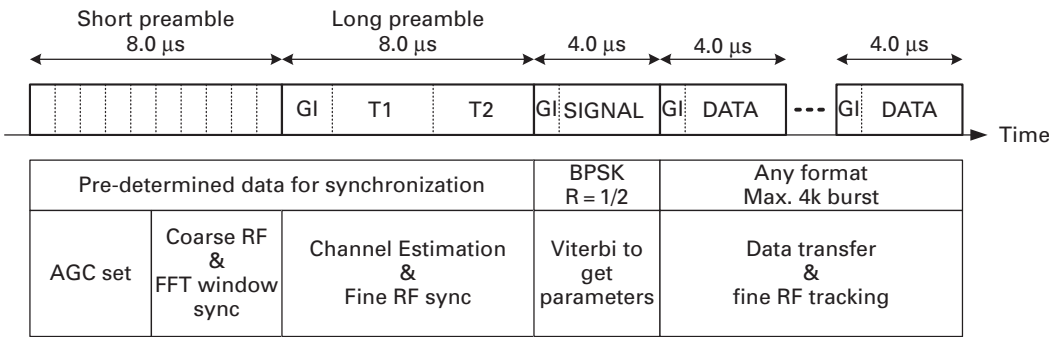


Figure 18.9 WLAN frame format and receiver tasks.

length is 3.2 μ s. Since the sampling frequency is 20MHz, 64 sampling points exist in one effective OFDM symbol. Then the demodulation process needs 64 point FFT.

18.2.7.2 WLAN system architecture

Figure 18.9 shows the frame format of a transmission packet. In addition, receiver tasks are indicated in the lower side of the figure. The first portion of the frame is a short preamble, which is composed of 0.8 μ s pre-determined signals, repeated 10 times. During the former half periods of the short preamble, the auto gain control (AGC) value to the low-noise amplifier (LNA) in the analog front-end has to be determined. Since channel estimation is performed only once in the long preamble period in the WLAN system, the AGC control value has to be fixed in the short preamble period and has to be kept until the end of frame. Then during the latter half period of the short preamble, the RF error has to be measured and compensated by the de-rotator. Details of the RF error detection will be explained in Section 18.4. Then the end point of the short preamble has to be detected to get FFT window position for later OFDM symbols. In the long preamble, there are two identical 3.2 μ s OFDM symbols T1, T2 and 1.6 μ s GI. Using T1 to T2 rotation, a more precise RF error can be estimated (see Section 18.4.1). Since T1 and T2 are predetermined pilot OFDM symbols, CTF can be estimated from the FFT output of T2. This CTF value will be used in the following OFDM symbol equalization until the end of frame since WLAN assumes a NOT moving application such as fixed channel.

As for sampling frequency error, it is not easy to detect the error and compensate for it during the preambles since the preamble duration is limited. However, if less than 100 ppm error crystal is used, a maximum $64 \times 100 \text{ ppm} = 0.0064$ points sampling error is caused during one 3.2 μ s OFDM symbol. This 0.0064 points error in one OFDM symbol is sufficiently smaller than 1 percent error, then this sampling clock error does not cause an ICI problem. The 0.0064 points error will be accumulated in all successive OFDM data symbols with 4K being the maximum number of these symbols for accumulation. Then the accumulation of point shift

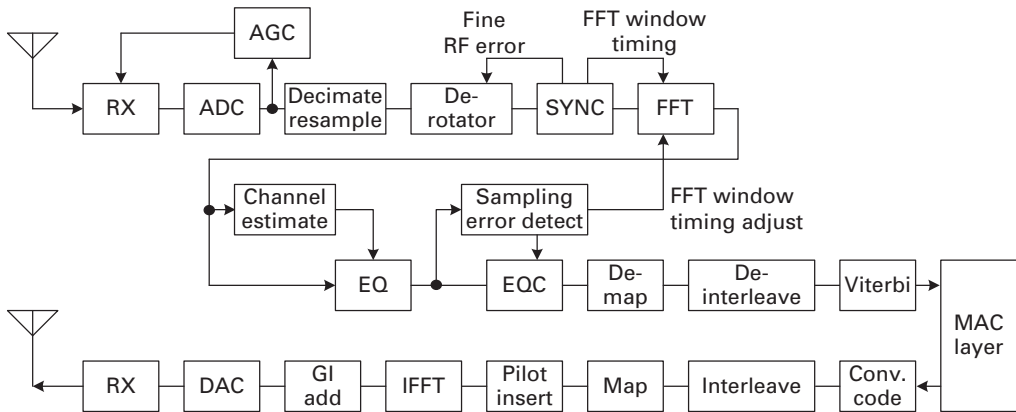


Figure 18.10 WLAN.

will be $4K \times 0.0064 = 26$ points. Since the GI length is 16 points, this FFT window shift has to be taken care of. Therefore, the receiver needs to detect FFT window shift and compensate for it. Since the CTF value is estimated by T1, T2 at the long preamble and is used in the equalization process, the equalized output will be rotated by the FFT window position shift. Then, after a conventional equalizer (EQ), a FFT window shift detector and equalizer output compensator (EQC) is needed as shown in Figure 18.10. If high precision crystal is used, this special detector and compensator are not needed. However, analog down-conversion LO frequency might drift during the 4K OFDM symbols, and then the RF error tracking mechanism would be needed.

Figure 18.10 shows the block diagram of a WLAN transceiver. The RX is an analog front-end. By calculating the short preamble power, AGC feedback value to the LNA is determined. In the diagram, the decimate and resample block is assumed to be fixed rate. Since FFT size is as small as 64, the sampling clock frequency error does not have significant impact on ICI in one OFDM symbol. But an accumulated sampling error causes FFT window position shift, which causes equalizer output constellation rotation depending on the subcarrier number. Therefore in this design, the rotation compensation block EQC is included. The EQC process is further explained in Section 18.4.6. After the de-mapping process, forward error correction for convolutional code is performed using the famous Viterbi algorithm.

18.2.8 OFDMA transceiver design

18.2.8.1 OFDMA transceiver system overview

The abbreviation OFDMA represents Orthogonal Frequency Division Multiple Access. Then multiple users have to share OFDM subcarriers in a symbol. Figure 18.11 shows an example of wireless multiple access for mobile phones. First, Up-Link (UL) and Down-Link (DL) have to be separated. This separation is usually implemented by Frequency

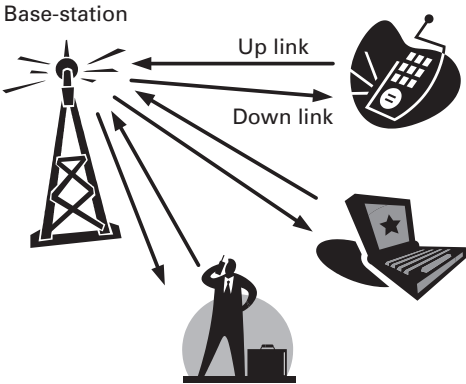


Figure 18.11 Multiple access.

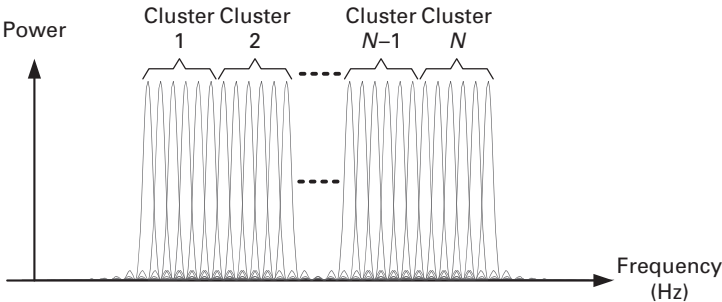


Figure 18.12 OFDMA power spectra.

Division Duplex (FDD) or Time Division Duplex (TDD). The FDD needs two separate frequency bands although TDD has to share one frequency band by time slot. Secondly, multiple user communications have to be separated. This second separation can also be done by FDD or TDD method. However, this inter-user separation can be realized by OFDM technology, which is called OFDMA.

In an OFDMA communication system, many subcarriers are divided into plural clusters as shown in Figure 18.12. Those clusters are assigned to each user. During DL communication, all clusters are transmitted from base-station to users. Each user receives all clusters and demodulates assigned clusters. But in UL communication each user generates pre-assigned cluster subcarriers and the base station receives summed cluster subcarriers from all users.

Figure 18.13 shows a TDD Frame structure for one base-station (BS) and two user terminals U1, U2. Delay time1 and 2 are wireless propagation delays between BS and U1, U2 respectively. Since the UL OFDM symbol has to reach BS with same timing, each user has to transmit the UL signal in advance timing as shown in the figure. The BS measures the UL arrival signal timing from each user and gives feedback to each user. This timing measurement will be further explained in Section 18.4.5.

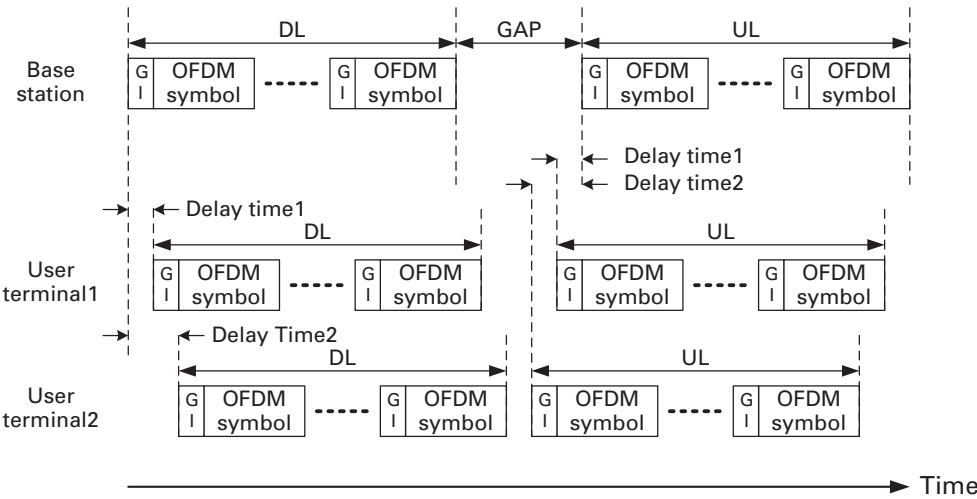


Figure 18.13 TDD frame and GAP.

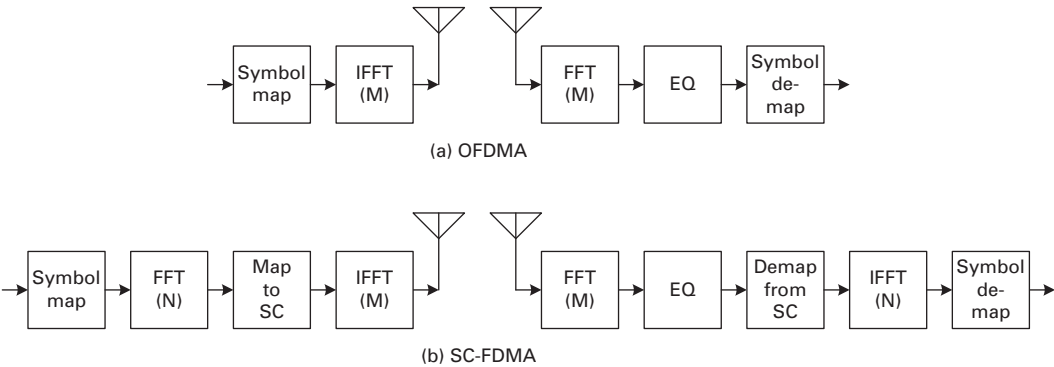


Figure 18.14 Simplified OFDMA and SC-FDMA.

18.2.8.2 SC-FDMA transceiver overview

One well-known problem of OFDM/OFDMA is large peak-to-average power ratio (PAPR), since the transmit waveform is generated by IFFT of many complex digital modulation symbols. The large PAPR needs a wide range of linearity in the transmission power amplifier and causes higher power dissipation at the transmitter. Especially, this will be a problem in battery operated handsets. A SC-FDMA (Single-Carrier Frequency Division Multiple Access) system tries to reduce PAPR while using the same OFDMA frequency domain equalization technique. Figure 18.14 shows a comparison between OFDMA and SC-FDMA systems. In the OFDMA system, bit information is mapped to complex digital modulation symbols such as QPSK/16QAM/64QAM constellation. Then the output of IFFT might have large peak value. In the SC-FDMA system, bit

information is mapped to complex digital modulation symbols and the N symbols are processed by N -points FFT. The N -points FFT outputs are bundled to a block. Then the K blocks are connected and mapped to M -points IFFT inputs. Here, assume $M = N \times K$. In this mapping process, each block corresponds to different subcarrier frequencies. Consequently, at the output of M -points IFFT is a sum of frequency shifted N -FFT input sets. Since the input of N -FFT is determined by complex digital modulation symbols, the PAPR of M -IFFT output is reduced. Please note that GI is also used in SC-FDMA although it is not indicated in Figure 18.14.

In other words, SC-FDMA uses Spread Spectrum method by N -FFT. Then the spread spectrum is assigned onto many subcarrier frequencies and transmitted parallel in channel. At the receiver, frequency domain channel estimation and equalization is performed. Then the spread spectrum is re-converted to original complex digital modulation symbols by N -IFFT. The mapping process between N -FFT and M -IFFT is equivalent to assigning frequency bands to the spread spectrums.

18.2.8.3 OFDMA transceiver architecture

In this section, OFDMA downlink and SC-FDMA uplink are assumed since a SC-FDMA transmitter is good for battery-operated mobile terminals. Figure 18.15 shows a Base-Station Transmitter (TX) and Receiver (RX) block diagram. In the TX, an OFDMA transmitter is used and in RX, a SC-FDMA receiver is used. Since all user terminals communicating with the same BS have to synchronize to the Base-Station, a Base-Station receiver is not required to have sampling clock frequency error compensation and RF error compensation. Since the FFT window position timing is determined in a BS, all user terminals have to adjust the UL transmission timing by the control of the BS. Then the BS needs to measure the UL arrival signal timing from each user and gives feedback to each user. This timing measurement will be further explained in Section 18.4.5.

Figure 18.16 shows a user terminal transmitter and receiver block diagram. In the receiver side, the AGC block controls the LNA in the analog front-end. There are two

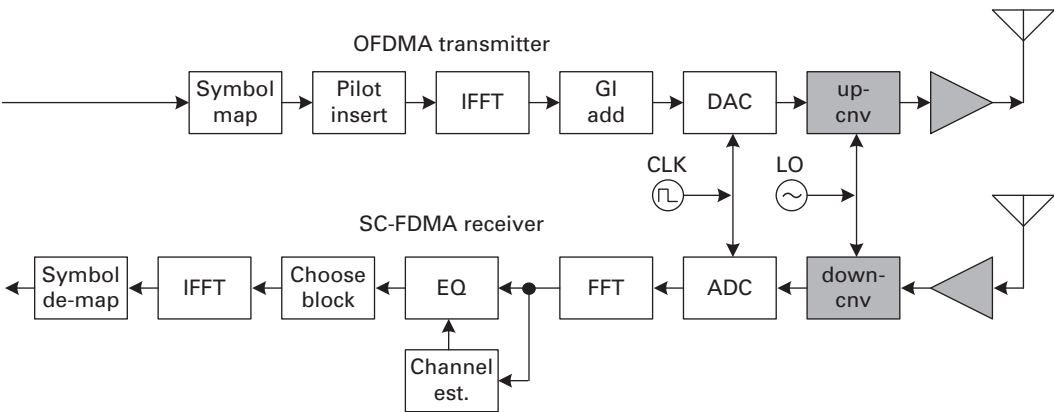


Figure 18.15 Base-station TX and RX.

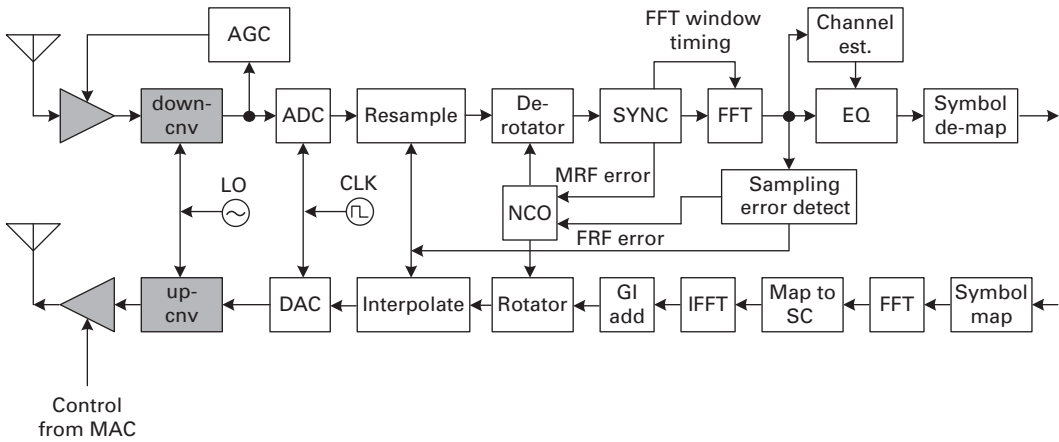


Figure 18.16 User terminal TX and RX.

local oscillators. One is for RF conversion and the other is for ADC/DAC. In order to compensate for the RF error and sampling clock frequency error, a resampler and de-rotator are required. In the figure, NCO represents a Numerical Control Oscillator, which is the digital form of a Voltage Controlled Oscillator (VCO). The NCO frequency corresponds to RF error. In the transmitter side, the reverse operation of the resampler and the de-rotator are required and these are the interpolator and rotator, respectively. Control signals for those blocks are generated on the receiver side. Power amplifier (PA) gain in the analog front-end is controlled by the Base-Station. Another Base-Station control item is UL start timing.

We have discussed various OFDM-based transceivers and design. In the following sections, related error detection and compensation ([Section 18.4](#)) together with diversity techniques ([Section 18.3](#)) will be addressed.

18.3 Adaptive antenna array and diversity

18.3.1 Introduction to multiple antenna receivers

When an OFDM receiver is used in moving vehicles such as automobiles etc., the radio frequency wave suffers from Doppler frequency shift as the automobile is moving as shown in [Figure 18.17](#). The Doppler shift destroys the orthogonality between OFDM subcarrier signals and increases the inter carrier interference. A longer multipath might cause inter symbol interference and weak reception signal power might decrease signal-to-noise ratio (SNR) so as to degrade the reception performance. Therefore, it is a severe challenge to maintain the reception quality in mobility circumstances.

One well-known way to improve the performance of an OFDM receiver is to exploit a spatial diversity by utilizing multiple antenna elements. There are two approaches to utilizing array antenna for the multi-carrier transmission method, which are pre-FFT

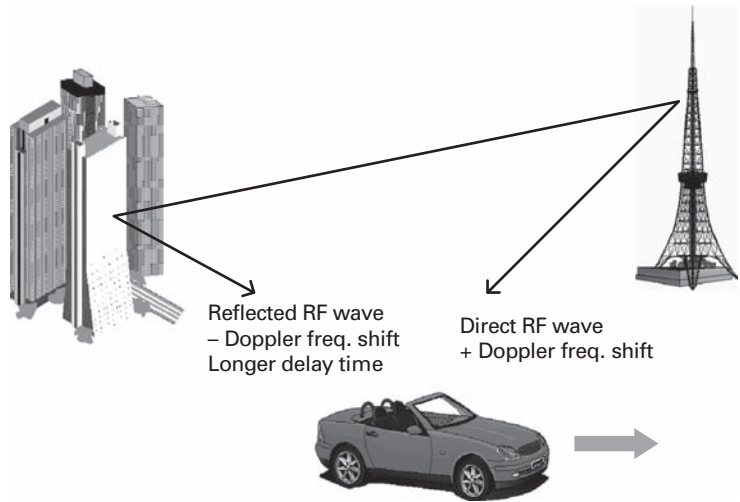


Figure 18.17 Problems in mobile reception.

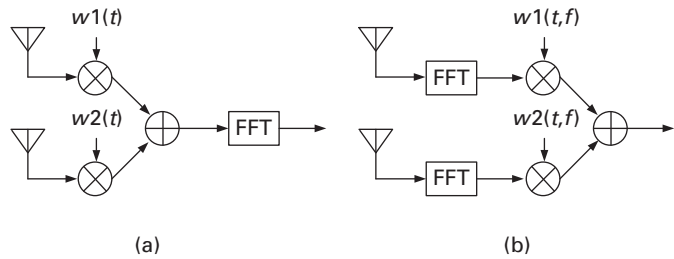


Figure 18.18 Two types of space diversity for OFDM communication.

adaptive array (AA) antenna and post-FFT carrier diversity (CD) combining scheme. The AA scheme is a conventional method to employ the array antenna in which inputs from the array antenna are combined before OFDM demodulation [7–9] as shown in Figure 18.18(a). Since this approach uses one set of coefficients such as $w1(t)$ and $w2(t)$ for each OFDM symbol, it is an attractive solution in term of low computation requirement. However, in order to achieve a good performance, received antenna signals should have higher correlation.

On the other hand, the post-FFT scheme is an advanced method to utilize the array antenna for a multi-carrier system as shown in Figure 18.18(b). Instead of combining before OFDM demodulation, inputs are demodulated using multiple OFDM demodulations. Subcarriers are then combined accordingly in the frequency domain using a diversity combining method [10]. The post-FFT scheme therefore is an optimum approach in term of maximizing SNR for each OFDM subcarriers. However, its computation complexity grows drastically as more antenna elements are associated. The features of the two diversity methods are summarized in Table 18.3.

Table 18.3 Features of two diversities.

Type	Circuit size	Reception performance	Restriction
Adaptive Array (AA)	Small	Moderate	High antenna signal correlation
Carrier Diversity (CD)	Large	High	NONE

18.3.2 Time-domain diversity

18.3.2.1 OFDM-directed array antenna system

An adaptive array antenna is considered as an attractive solution to suppress interferences and to enhance the quality of reception in the multipath environment [11, 12]. It usually requires in advance the knowledge of the characteristics of the desired signals to distinguish between them and undesired signals. However, due to their high complexity, these approaches are costly in terms of computation effort and hardware implementation. In this section, OFDM-directed two array antenna Digital Beam-Forming (DBF) algorithms with lower computation requirements are introduced. The DBF algorithms make use of the cyclic prefix's periodic property. Figure 18.19(a) and (b) illustrate the periodic property of the OFDM signal and the principle of the time-domain adaptive array antenna system. For convenience, here, the GI with duration of T_g is referred to as “HeadGI” and the last part of the effective OFDM symbol with duration of T_g is referred to as “TailGI.”

Suppose that an array antenna is equipped with M elements. The vector input signals of the array antenna can be expressed as

$$\begin{aligned}\mathbf{X}(t) &= [x_1(t)x_2(t) \dots x_M(t)]^T \\ &= \mathbf{V}_s s(t) + \mathbf{V}_i i(t) + \mathbf{N}(t)\end{aligned}\quad (18.22)$$

where superscript $[\cdot]^T$ denotes a transpose of the matrix; $s(t)$ and $i(t)$ are the desired signal and interference, respectively; $\mathbf{V}_s$ and $\mathbf{V}_i$ are the array response vectors of the desired signal and interference, respectively; $\mathbf{N}(t)$ is the AWGN (additive white Gaussian noise) vector of the array antenna. Hence, the output of the array antenna system is calculated as follows

$$\begin{aligned}y(t) &= \mathbf{W}^H \mathbf{X}(t) \\ &= \mathbf{W}^H \mathbf{V}_s s(t) + \mathbf{W}^H \mathbf{V}_i i(t) + \mathbf{W}^H \mathbf{N}(t) \\ &= a_s s(t) + a_i i(t) + n(t)\end{aligned}\quad (18.23)$$

where superscript $[\cdot]^H$ denotes Hermitian; $\mathbf{W} = [w_1 \ w_2 \ \dots \ w_M]^T$ is the coefficient vector of the array antenna; $a_s = \mathbf{W}^H \mathbf{V}_s$ and $a_i = \mathbf{W}^H \mathbf{V}_i$ are constant during one OFDM symbol.

In references [13–14], the conventional Maximum Ratio Combining (MRC) algorithm is derived directly from the cross-correlation of inputs and outputs of the array antenna:

$$\begin{aligned}\mathbf{W}_{convMRC} &= E[\mathbf{X}(t)y^*(t)] \\ &= E[(\mathbf{V}_s s(t) + \mathbf{V}_i i(t) + \mathbf{N}(t))(a_s s(t) + a_i i(t) + n(t))^*] \\ &= a_s^* P_s \mathbf{V}_s + a_i^* P_i \mathbf{V}_i\end{aligned}\quad (18.24)$$

where $P_s = E[|s(t)|^2]$ and $P_i = E[|i(t)|^2]$. Superscript $(.)^*$ denotes complex conjugate. Apparently from Equation (18.24), the array antenna utilizing the conventional MRC combines not only the desired signal but also the interference. Under the circumstances that the power of interference is stronger than that of the desired signal, the interference will be emphasized at the output of the array antenna.

18.3.2.2 OFDM selective Maximum Ratio Combining (MRC)

In order to overcome the limitation of the conventional MRC method described above, by making use of HeadGI of input and feedback of TailGI of output, the cross-correlation of input and output is derived as

$$\begin{aligned} \mathbf{r}_{xy} &= E[\mathbf{X}_h(t)y_t^*(t)] \\ &= E[(\mathbf{V}_s s_h(t) + \mathbf{V}_i i_h(t) + \mathbf{N}_h(t))(a_s s_t(t) + a_i i_t(t) + n_t(t))^*] \\ &= a_s^* \mathbf{V}_s E[s_h(t)s_t^*(t)] + a_i^* \mathbf{V}_i E[i_h(t)i_t^*(t)] \end{aligned} \quad (18.25)$$

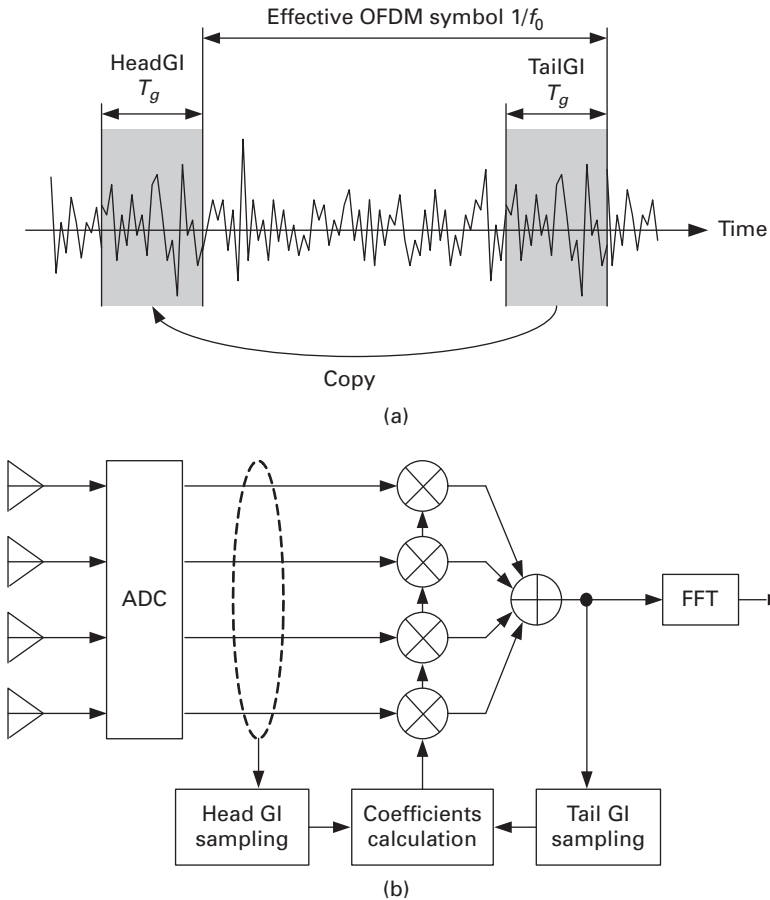


Figure 18.19 (a) Cyclic properties of OFDM signal and (b) Adaptive array using GI.

where subscripts $(\cdot)_h$ and $(\cdot)_t$ denote HeadGI and TailGI of the signal, respectively. Since $i(t)$ is interference and does not have cyclic prefix GI, the second term of Equation (18.25) can be neglected. Employing the periodic property of OFDM signal, Equation (18.25) can be expressed as follows:

$$\mathbf{r}_{xy} = a_s^* \mathbf{P}_s \mathbf{V}_s. \quad (18.26)$$

Hence, Equation (18.26) presents the information of AOA (angle of arrival) of the desired signal without the reference signals. Therefore, the coefficient vector of the array antenna using MRC algorithm is derived as below. In other words, this OFDM selective MRC algorithm is beam-forming only to OFDM signals

$$\mathbf{W}_{MRC} = \text{normalize}(\mathbf{r}_{xy}). \quad (18.27)$$

18.3.2.3 Minimum Mean Square Error (MMSE) and Sampled Matrix Inversion (SMI)

Although the OFDM selective MRC algorithm emphasizes the desired OFDM signal, it does not minimize the interference. In this section, the Minimum Mean Square Error (MMSE) algorithm is derived. Assume a reference signal is $r(t)$. The error signal is given as:

$$e(t) = r(t) - \mathbf{W}^H \mathbf{X}(t). \quad (18.28)$$

The MSE (mean square error) between the output of the array antenna and the reference signal is given as:

$$\text{MSE} = E[|e(t)|^2] = E[|e(t)r(t) - \mathbf{W}^H \mathbf{X}(t)|^2]. \quad (18.29)$$

By minimizing the MSE with the well-known Wiener–Hoff equation, the optimum coefficients are derived as

$$\mathbf{W}_{MMSE} = \text{normalize}(\mathbf{R}_{xx}^{-1} \mathbf{r}_{xy}) \quad (18.30)$$

where $\mathbf{R}_{xx} = E[\mathbf{X}_h \mathbf{X}_h^H]$ is the autocorrelation matrix of inputs. In real applications, $\mathbf{R}_{xx}$ can be obtained by a simple averaging scheme [15]. According to Equation (18.30), the additional calculation of the inverse of $\mathbf{R}_{xx}$ makes MMSE possible and then this algorithm is also called Sampled Matrix Inversion (SMI).

18.3.3 Frequency-domain diversity

18.3.3.1 MRC for uniformly distributed branch noise

Figure 18.20 illustrates the post-FFT carrier diversity (CD) combiner for an OFDM receiver. This system assumes there are M antennas and M OFDM demodulations. The combined symbol at the k th subcarrier of the i th OFDM symbol is given as

$$Z(i, k) = \sum_{l=1}^M C_l(i, k) Y_l(i, k) \quad (18.31)$$

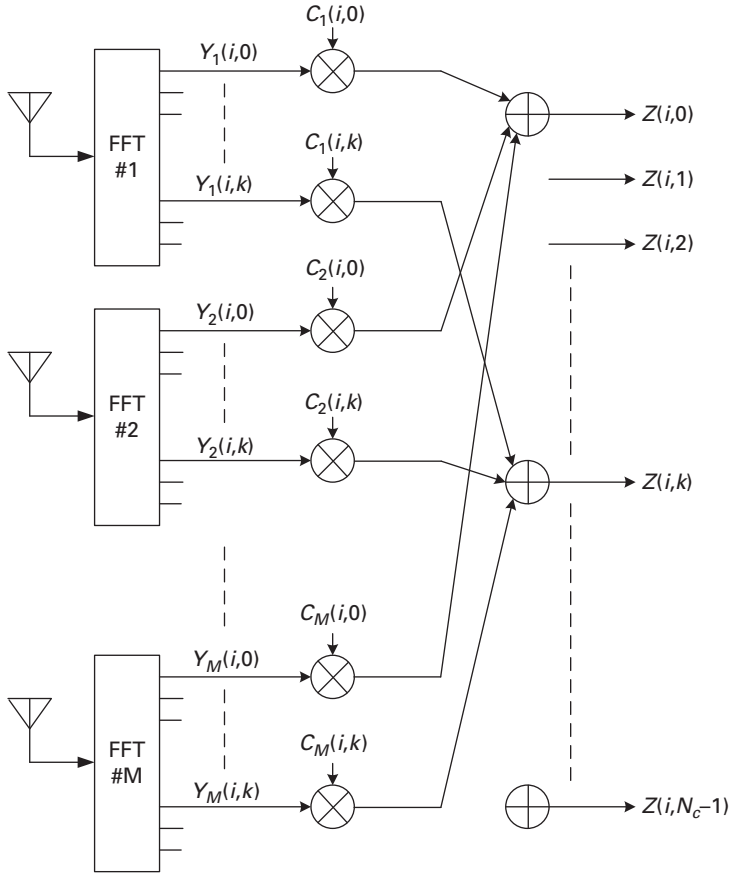


Figure 18.20 Carrier diversity combiner for OFDM receiver.

where $C_l(i,k)$ and $Y_l(i,k)$ are the combining coefficient and the demodulated symbol corresponding to the l th branch of the combiner at the k th subcarrier of the i th OFDM symbol, respectively. Here, the following equation is assumed

$$Y_l(i,k) = H_l(i,k)d(i,k) + N_{Bl}(i,k) \quad (18.32)$$

where $H_l(i,k)$, $N_{Bl}(i,k)$ are the channel transfer function and the additive noise corresponding to the l th branch of the combiner at the k th subcarrier of the i th OFDM symbol, respectively and $d(i,k)$ is the transmitted complex digital modulation symbol at the k th subcarrier of the i th OFDM symbol. According to Reference [10], the combining coefficients for the MRC algorithm can be obtained as follows

$$C_l(i,k) = \frac{H_l^*(i,k)}{\sum_{l=1}^M |H_l(i,k)|^2} \quad (18.33)$$

where assuming that the mean of $N_{Bl}(i,k)=0$ for all l, i, k and the deviation of $N_{Bl}(i,k)$ are all the same for all l, i, k . By using the MRC algorithm, the

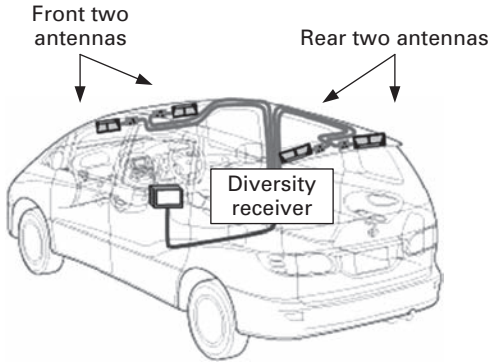


Figure 18.21 Multiple antennas of a car.

optimum combined output SNR can be achieved for each subcarrier by post-FFT CD scheme.

18.3.3.2 MRC for independently distributed branch noise

When one antenna of the receiver is broken or hidden, the noise of the branch increases and has impact on the combined output. Figure 18.21 shows a typical 4 space diversity antenna configuration for an automobile. Two antenna elements are set in the front and two are set in the rear. However as the body of the automobile is made from metal, the directional characteristics of the front antenna is obviously distorted and concentrates in the forward direction. Then, the uniform noise power distribution assumed in the previous section sometimes cannot be applied to these kinds of unbalanced antenna situations.

Equation (18.33) can be rewritten as:

$$\begin{aligned}
 C_l(i, k) &= \frac{H_l^*(i, k)}{|H_l(i, k)|^2} \cdot \frac{|H_l(i, k)|^2}{\sum_{l=1}^M |H_l(i, k)|^2} \\
 &= \frac{1}{H_l(i, k)} \cdot \frac{|H_l(i, k)|^2}{|H_1(i, k)|^2 + |H_2(i, k)|^2 + \cdots + |H_M(i, k)|^2}
 \end{aligned} \tag{18.34}$$

The first division by $H_l(i, k)$ corresponds to the equalizing operation. The second multiplication term means how much branch l contributes to combining. Since $H_l(i, k)$ is CTF, then it corresponds to the signal gain in the transmission channel. If the average noise power of each branch ($l = 1 \dots M$) is denoted by σ_{nl}^2 , the above equation can be modified to the following equation (18.35), which is taking noise distribution between branches into account.

$$C_l(i, k) = \frac{1}{H_l(i, k)} \cdot \frac{|H_l(i, k)|^2 / \sigma_{nl}^2}{|H_1(i, k)|^2 / \sigma_{n1}^2 + |H_2(i, k)|^2 / \sigma_{n2}^2 + \cdots + |H_M(i, k)|^2 / \sigma_{nM}^2} \tag{18.35}$$

According to Equation (18.35), when noise power at branch 1 increases, the combining coefficients $C_{\lambda}(i,k)$ become small and will contribute less to the combining. When using Equation (18.35) in real applications, additional noise power estimation is needed.

18.3.4 Hybrid diversity

Figure 18.22 shows a block diagram of a four antenna diversity combining receiver LSI for Japan OFDM-based Digital TV [16]. In the first stage, two sets of two antenna MRC combiners are used. Then there are two sets of FFT and Equalizer circuits. In the second stage, the two EQ outputs are combined as a carrier diversity method. The final combined output is de-mapped and error correction processing is applied. The two stage diversity combiners halve the number of FFT/EQ circuits from four to two and sustain a similar mobile reception performance comparing with the four FFT/EQs case. In the second CD stage, another LSI interface is prepared and then in total four kinds of diversity configurations are possible as shown in Figure 18.23.

Mobile reception performance was measured by field experiment at three severe test courses in Osaka Japan for 4CD, 4AA, 2CD configurations. The experimented UHF channel is 13ch which is broadcasted from the top of IKOMA Mountain between Osaka and Nara prefecture. 64QAM, code rate of 3/4, 12 segments HDTV broadcasting service are used. The three test courses are a suburban area 50 km from IKOMA broadcasting station (BS), a mountain area 40 km from BS, and an urban area but in the shadow of a hill 20 km from BS. Figure 18.24 shows the reception rate (the rate of the error free reception duration to the total experiment duration) of those test courses. One chip solution 4AA mode showed a little lower performance than the two chip solution 4CD

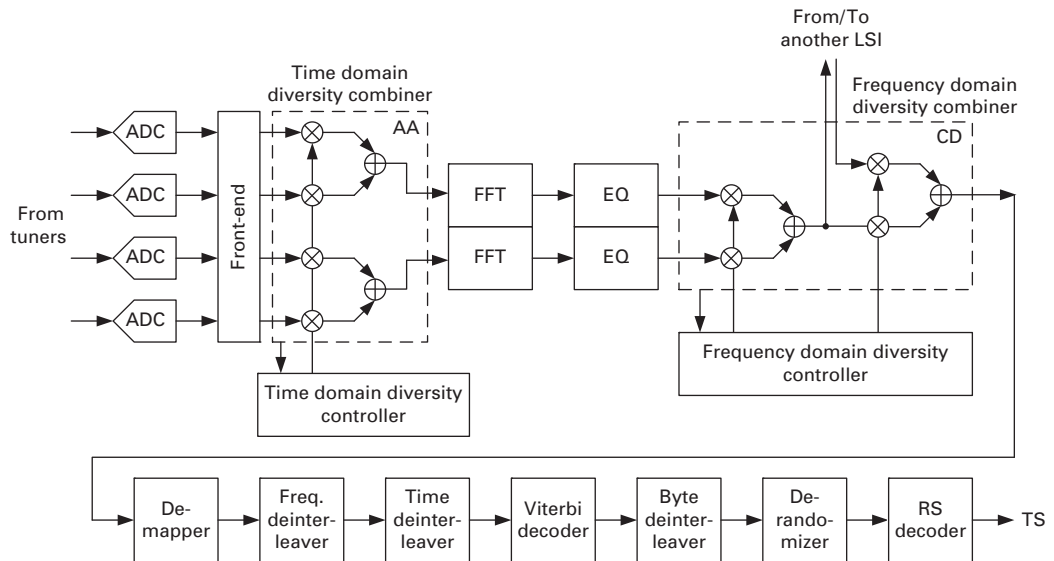


Figure 18.22 Four antenna hybrid diversity combining receiver.

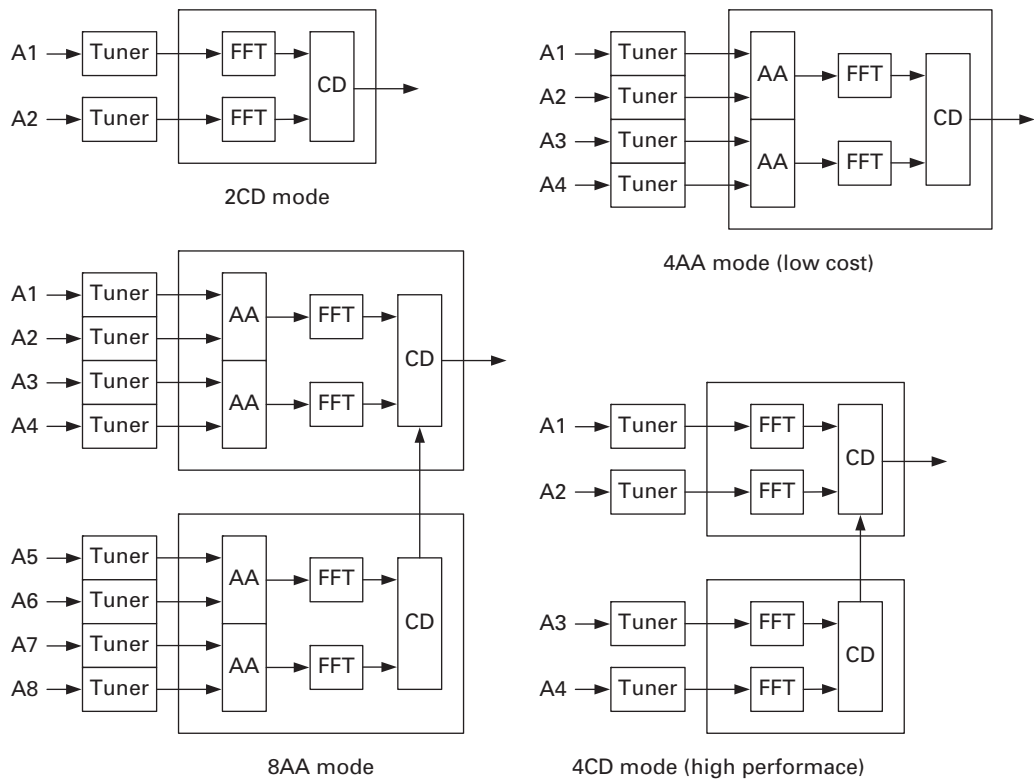


Figure 18.23 Four possible diversity configurations.

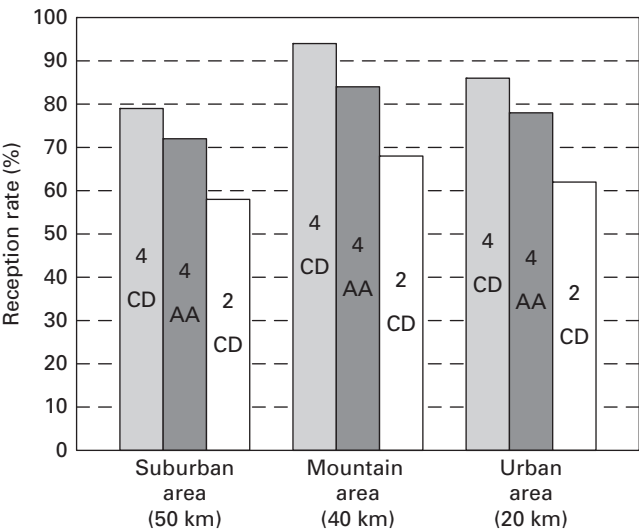


Figure 18.24 Field experimental results.

mode. The hybrid method shows a little lower reception performance with lower system cost. More comparison results are available in [16].

Various diversity techniques for OFDM receivers are addressed in this section. In the next section, related error detection and compensation techniques for OFDM-based transceivers will be discussed.

18.4 Error detection and compensation

18.4.1 Time domain RF error detection and compensation

Figure 18.25 shows a WLAN frame format. At the short preamble period, there are 10 repeating $1/(4f_0)$ length signals and at the long preamble, $T1$ and $T2$ are $1/f_0$ length repeating signals. At the data burst, GI is a cyclic prefix, that is, a copy of the tail part of the OFDM symbol. By making use of these repeating or copy signals, RF error can be computed as follows.

If the time interval of identical signals $g(t)$ and $g(t+T)$ is T , RF error Δf causes the rotation:

$$g(t+T) = g(t)e^{j2\pi\Delta f T}. \quad (18.36)$$

From Equation (18.36), Δf can be solved as follows:

$$\Delta f = \frac{1}{2\pi T} \tan^{-1}[g(t+T)g^*(t)] \quad (18.37)$$

$$-\frac{1}{4T} < \Delta f < \frac{1}{4T}. \quad (18.38)$$

At the short preamble, $T = 1/(4f_0)$. Then, RF error Δf in the region $[-f_0, f_0]$ can be detected. Similarly at the long preamble region, Δf in the region of $[-f_0/4, f_0/4]$ can be estimated. Since $T1, T2$ has 64 points whose average reduces the Δf detection error, a high accuracy Δf can be obtained. At the data burst region, Δf should be very small because of the preceding Δf detection and compensation. Then Equation (18.37) can be approximated as:

$$\Delta f \cong \frac{1}{2\pi T} \text{Im}[g(t+T)g^*(t)]. \quad (18.39)$$

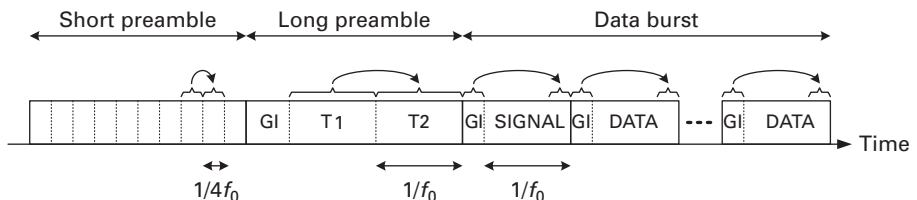


Figure 18.25 Time domain RF error detection.

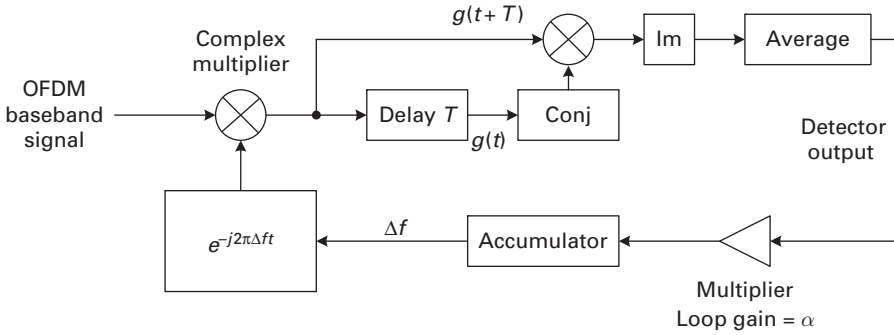


Figure 18.26 Time domain RF error compensation feedback.

During the data burst region, RF error can be compensated by feedback loop as shown in Figure 18.26. In the loop, loop gain α and accumulator are used. When the RF error detector output = 0, the accumulator holds the same value since the input = 0. The accumulator output corresponds to RF error Δf .

18.4.2 Time domain FFT window position detection and RF error detection

The FFT window position can be found by cyclic prefix GI. According to Reference [17], head and tail GI correlation can be calculated as follows:

$$As(k) = \sum_{i=0}^{M-1} \left\{ |r(k+i)|^2 + |r(k+i+N_{PD})|^2 \right\} - 2 \left| \sum_{i=0}^{M-1} r(k+i+N_{PD}) \cdot r^*(k+i) \right| \quad (18.40)$$

where $r(k)$ is the baseband OFDM signal input. Index i is used to average M points. If the time index k is not in the GI period, $r(k+N_{PD})$ and $r(k)$ are independent. Then the second term will be close to zero. Instead, if the time index k is in the GI period, $r(k+N_{PD})$ and $r(k)$ should be close. Then the output of Equation (18.40) will be close to zero. Consequently, the correlation $As(k)$ waveform is negative pulse at the GI period as shown in Figure 18.27(a). Here, N_{PD} is the OFDM symbol length.

The negative pulse will be repeatedly generated. By detecting the pulse rising edge, the OFDM starting position can be detected and then the FFT window position is obtained. Since the negative pulse period corresponds to the GI period, phase rotation φ in one OFDM symbol can be obtained as:

$$\varphi = \text{angle} \left\{ \sum_{i=0}^{M-1} r(k+i+N_{PD}) \cdot r^*(k+i) \right\}. \quad (18.41)$$

Since the phase rotation φ happens in OFDM symbol period ($1/f_0$), the RF error Δf can be calculated as follows

$$\Delta f = \frac{\varphi}{2\pi} f_0. \quad (18.42)$$

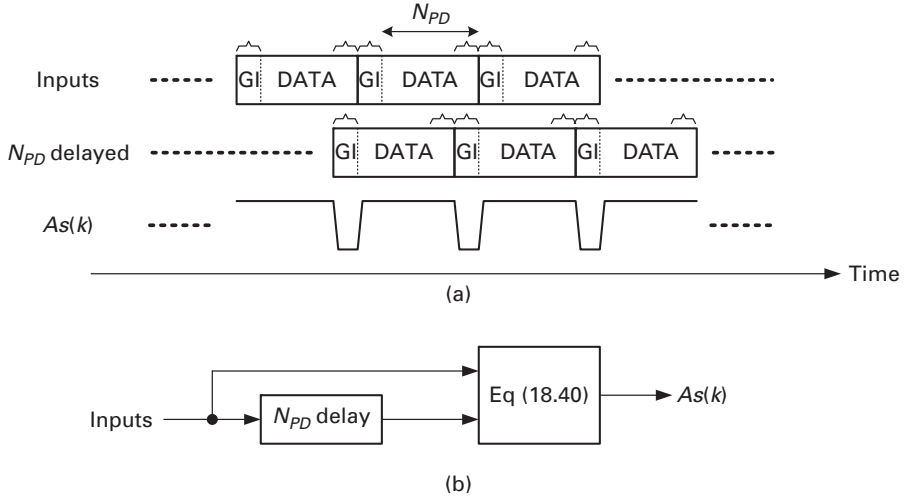


Figure 18.27 Cyclic prefix GI time domain correlation.

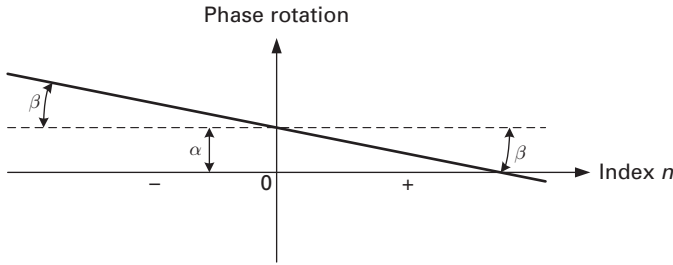


Figure 18.28 Phase rotation with RF error α and sampling clock error β .

18.4.3 Time domain RF error, sampling error, and symbol timing detection and compensation

One subcarrier signal with index of n can be expressed as

$$x_n(t) = d_n \cdot e^{j2\pi n f_0 t} \quad (18.43)$$

where d_n is the n th digital modulation symbol and f_0 is the carrier spacing. Then $t=0$ for $1/f_0$ rotation with RF error Δf and the sampling clock error β is can be expressed as follows:

$$\begin{aligned} \frac{x_n\left(\frac{1}{f_0}\right)}{x_n(0)} &= e^{\frac{j2\pi(n f_0 + \Delta f)(1-\beta)}{f_0}} \\ &\cong e^{j2\pi(\alpha - n\beta)}. \end{aligned} \quad (18.44)$$

Here α is the normalized RF error of $\Delta f/f_0$. According to Equation (18.44), the RF error causes common phase rotation but the sampling clock error β causes the phase error in proportion to the subcarrier index number n as shown in Figure 18.28.

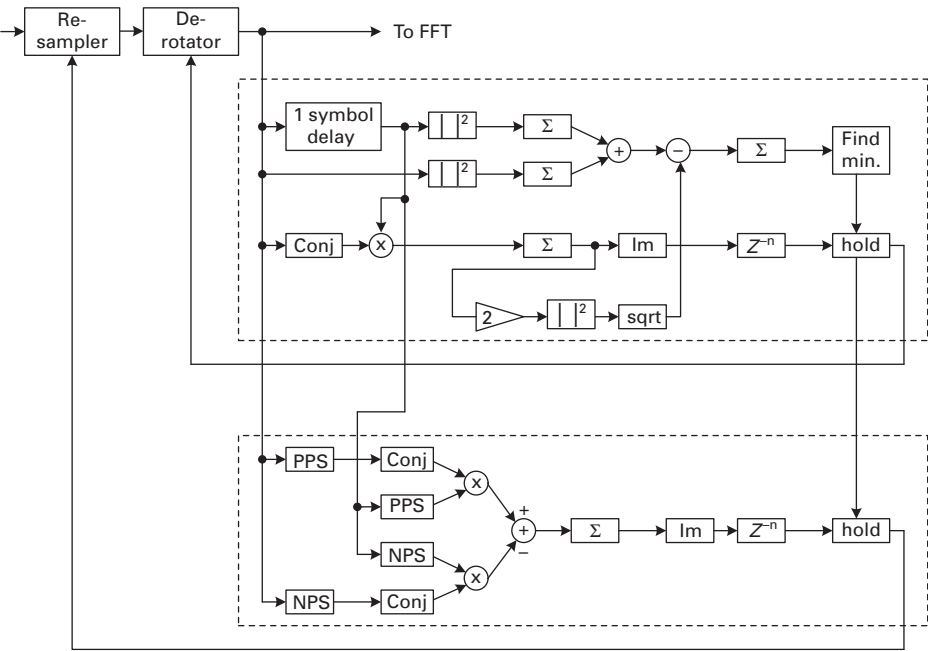


Figure 18.29 Time domain RF error and sampling clock error compensation.

Then by averaging all phase rotation for all subcarriers, the RF error α can be obtained. By calculating the phase error in positive frequency ($n > 0$) and phase error in negative frequency ($n < 0$), the sampling clock error β can be calculated.

Figure 18.29 shows one example of time domain GI correlation, an RF error detection and sampling clock error detection circuit diagram. The “Find min.” block detects minimum points of GI correlation given in Equation (18.40). The upper dashed block outputs RF error and the lower dashed block outputs sampling clock error. PPS and NPS represent positive and negative pass FIR filters, respectively. Since they form a feedback loop and the detection error should be close to zero in converged stable state, the imaginary components of the phase errors are used to feedback outputs. The simplest implementation of PPS and NPS FIR filters’ tap coefficients are given as $[-j, 1, j]$ and $[j, 1, -j]$ respectively. Here, j represents the square root of (-1) .

18.4.4 Frequency domain RF error and sampling error detection

Figure 18.30 shows one Physical Resource Unit (PRU) of an OFDMA-based XGP communication system [4]. It corresponds to 19 OFDM symbols and 24 subcarriers. In one PRU, there are 24 pilots, which transmit pre-determined value. For the same subcarrier number such as F3, F7 and so on, the pilot values are identical and then the phase rotation shown in solid or dashed arrows can be detected. In one PRU, there are 18 phase rotations. The averages of the phase rotation in the lower frequency side (shown in dashed arrows) and in the higher frequency side (shown in solid arrows) can be expressed as follows:

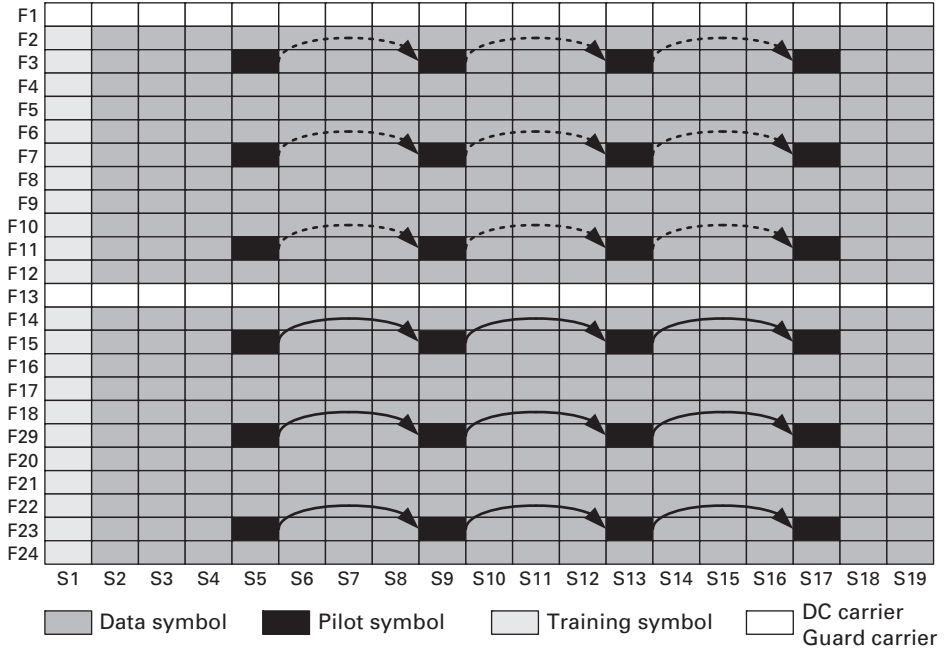


Figure 18.30 XGP PRU structure.

$$\text{LowRotMean} = \frac{p(F3, S5)^* \times p(F3, S9) + \cdots + p(F11, S13)^* \times p(F11, S17)}{9} \quad (18.45)$$

$$\text{HighRotMean} = \frac{p(F15, S5)^* \times p(F15, S9) + \cdots + p(F23, S13)^* \times p(F23, S17)}{9} \quad (18.46)$$

where $p(F, S)$ means the pilot subcarrier value at subcarrier of F and symbol number of S .

If we define LV and HV as

$$\text{LV} = \frac{\text{imag}(\text{LowRotMean})}{\text{real}(\text{LowRotMean})} \quad (18.47)$$

$$\text{HV} = \frac{\text{imag}(\text{HighRotMean})}{\text{real}(\text{HighRotMean})}, \quad (18.48)$$

the RF error FRF in Equation (18.49) and sampling clock error CLKD in Equation (18.50) can be calculated:

$$\text{FRF} = \frac{m \cdot \text{LV} - n \cdot \text{HV}}{\frac{9}{4}\pi(m - n)} \quad (18.49)$$

$$\text{CLKD} = \frac{LV - HV}{\frac{9}{4}\pi(m - n)} \quad (18.50)$$

where n and m are the center carrier index numbers at the lower frequency side and higher frequency side, respectively. The idea of the calculation is based on the property that RF error rotates the same phase for all subcarriers and that sampling clock error rotates each subcarrier phase in proportion to the carrier number index as shown in Figure 18.28.

18.4.5 Delay profile detection and frequency domain FFT position detection

In the frequency domain, from series of pilot or training symbols, the delay profile can be easily obtained. The S1 symbols of PRU shown in Figure 18.30 are training symbols and their values are all pre-determined. Then by IFFT calculation, delay profile can be obtained as follows:

$$\text{Delay Profile} = \text{IFFT} \left[0, \frac{X(F2)}{T(F2)}, \frac{X(F3)}{T(F3)}, \dots, \frac{X(F12)}{T(F12)}, 0, \frac{X(F14)}{T(F14)}, \frac{X(F15)}{T(F15)}, \dots, \frac{X(F24)}{T(F24)}, 0, \dots, 0 \right] \quad (18.51)$$

where $X(F)$ is the FFT output of subcarrier F and symbols S1, and $T(F)$ is the pre-determined training value at subcarrier F and symbols S1. Then the power profile can be obtained as shown in Figure 18.31. In this profile, the power of the main wave is at index = 0 and then the FFT window position is placed at the main wave position. Because of the cyclic property of IFFT and FFT, the tail profile is connecting to the head profile as shown in Figure 18.31.

If the FFT window position shifts ahead, the delay profile shifts backward as shown in Figure 18.32. According to the index number of the peak point of delay profile, the FFT window shift can be detected. Detectable maximum delay is the OFDM symbol length divided by the training subcarrier interval. Since the S1 symbol includes all training symbols and its interval equals unity, the detectable maximum delay is the same as the

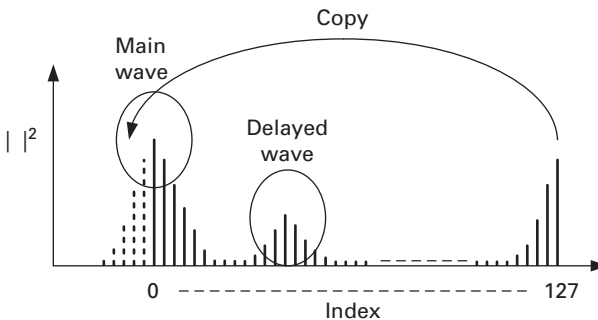


Figure 18.31 Power delay profile.

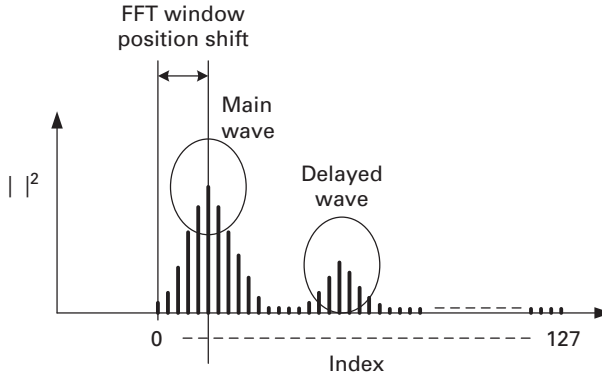


Figure 18.32 FFT window position shift ahead.

OFDM symbol length. If pilots in S5 are used to calculate the delay profile, the detectable maximum delay is 1/4 of the OFDM symbol because the pilot interval = 4.

18.4.6 FFT window shift detection by constellation rotation and compensation

A subcarrier signal of index = n can be expressed as Equation (18.52) in one OFDM symbol $[0 \ 1/f_0]$. The FFT output of the OFDM symbol with index = n is d_n .

$$x_n(t) = d_n \cdot e^{j2\pi n f_0 t} \quad (18.52)$$

When the FFT window shifts k points ($-0.5 < k < 0.5$) forward, the equivalent time shift Δt is

$$\Delta t = \frac{k}{N f_0}. \quad (18.53)$$

Then the FFT output of the OFDM symbol with index = n will become

$$d_n \cdot e^{j2\pi \frac{nk}{N}} \quad (18.54)$$

where N is the size of the FFT. This means that the FFT window shift causes a constellation rotation proportional to the index number n . Then by detecting the equalizer output's rotation, the FFT window shift can be detected. Figure 18.33 shows a block diagram of equalizer output phase rotation detection and compensation. Since the equalizer output should be pre-determined constellation based, phase rotation can be detected with the average of many samples. If averaging all the points is performed, the common phase error CPH in the figure can be detected. If the difference between positive and negative index averages is calculated, the phase rotation which is proportional to the index will be obtained and this corresponds to the FFT window shift WSH in the figure. By using CPH and WSH, the constellation can be compensated as shown in Figure 18.33.

We have discussed various OFDM based transceivers and their design, various diversity techniques, and related error detection and compensation techniques. In the final section, the hardware implementation and related chip performance will be addressed.

Table 18.4 Features and Data of LSI.

Die size	25 mm ²
Process technology	0.18 μm CMOS 6-metal
Supply voltage	1.8 V for ADC and Logic 3.3 V for IO buffer
Number of gates	700 K gates
ADC	Four ADCs 10bits 33 Msps
Power consumption	600 mW

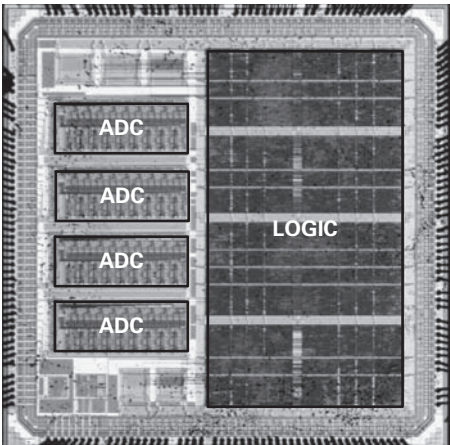


Figure 18.35 Chip photomicrograph of adaptive combiner LSI.

with 0.18 μm CMOS technology. The LSI needs 1.8V voltage supply for internal circuits and 3.3V voltage for I/O buffers.

In this implementation, four ADCs need approximately 25 percent of die area with 0.18 μm CMOS technology. A chip photomicrograph is shown in [Figure 18.35](#).

18.5.2 A 2/4/8 Antennas Configurable Diversity OFDM Receiver LSI [16]

The second implementation result is a 2/4/8 Antennas Configurable Diversity OFDM Receiver LSI for Japan digital TV specification of ISDB-T. The LSI is composed of time domain adaptive array and frequency domain carrier diversity as shown in [Figure 18.22](#). Using one or two LSIs, 2/4/8 antennas diversity configurations are possible as shown in [Figure 18.23](#). In addition to OFDM demodulation, the LSI integrates Viterbi and Reed-Solomon Error correction circuits.

The features of the diversity receiver LSI is shown in [Table 18.5](#). The LSI integrates 4 ADCs, 1.8 M gates and 18.4 Mbit SRAM memory in 49 mm² with 90 nm CMOS technology. The largest memory is time de-interleave RAMs because of the ISDB-T requirement. The core logic supply voltage is 1.2V and I/O circuit needs 3.3 V supply voltage. A chip photomicrograph is shown in [Figure 18.36](#).

Table 18.5 Features and data of Diversity OFDM Receiver LSI.

Process technology	90 nm 7M1P CMOS
Logic	1.8 M gates
Memory and ADC	18.4 M bit/4 ADC
Supply voltage	1.2 V core, 3.3 V I/O
Active power	310 mW typical
Die size	7.0 mm × 7.0 mm
	12 mm × 12 mm
Package	144FBGA

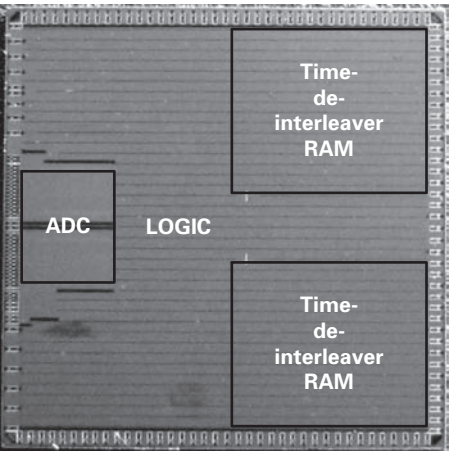


Figure 18.36 Chip photomicrograph of Diversity OFDM receiver LSI.

18.5.3 Joint hardware-software implementation of adaptive array antenna OFDM receiver by FPGA [9]

Figure 18.37 illustrates a block diagram of the array antenna concatenating with a OFDM demodulator. A digital beamforming (DBF) algorithm to generate the coefficients of the array antenna is executed prior to FFT. The output of the array antenna is fed to the OFDM demodulator to retrieve the transmitted data. A joint hardware-software (HW-SW) solution is chosen to provide an easily debugged and flexible platform for adaptive array antenna processing. The supported adaptive array antenna algorithms are OFDM selective Maximum Ration Combining (MRC) as in Section 18.3.2.2 and Sample Matrix Inversion (SMI) as in Section 18.3.2.3.

Since RF error and sampling rate error are inherently unavoidable, the compensation feedback signals of RF error and sampling error are provided from OFDM demodulator to array antenna signal processing. Figure 18.38(a) shows an impact of sampling rate error on a sampled TailGI. Due to the drifting of the sampling portion of TailGI, the samples of TailGI do not remain precisely. Therefore, it obviously deteriorates the

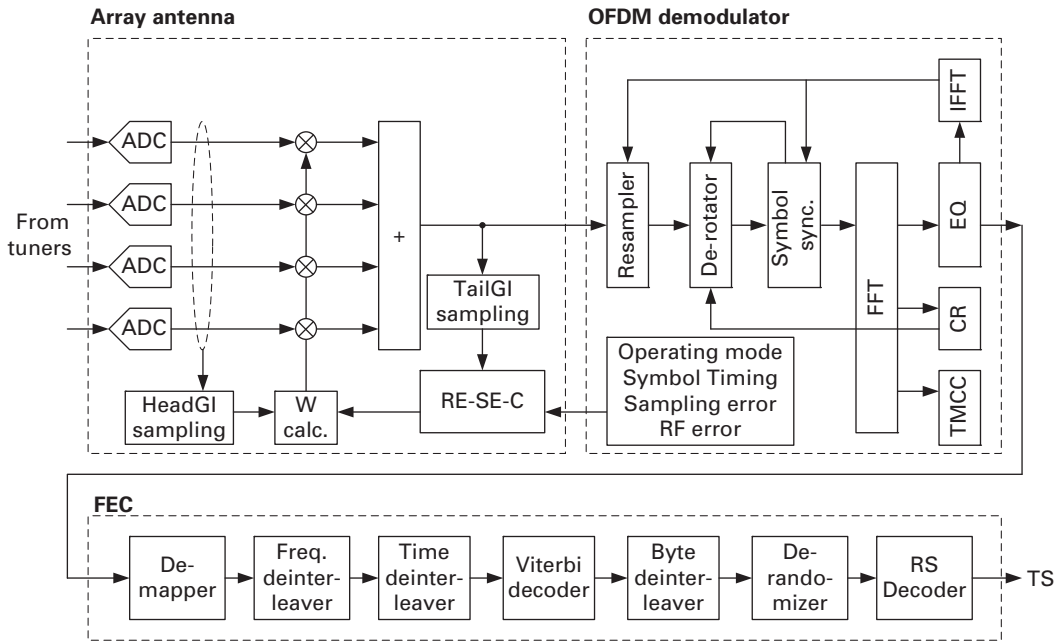


Figure 18.37 Block diagram of adaptive array antenna OFDM receiver (where “RE-SE-C” stands for “RF error and Sampling error compensator”)

cross-correlation between HeadGI of the inputs and TailGI of the output. The block diagram of the compensator of sampling rate error is shown in Figure 18.38(b). The error is estimated in the OFDM demodulator and accumulated on each OFDM symbol to measure the exact duration of effective symbol length T_e . TailGI is then fed to a 13-tap poly-phase FIR filter to eliminate the impact of sampling rate error. The coefficients of the 13-tap FIR filter are determined by the value of the T_e using an LUT method (Look-Up Table). Not only RF error and sampling rate error, but also the operation mode of ISDB-T and the synchronization of the OFDM symbol are provided from OFDM demodulation and they are fed back to the adaptive array antenna.

Figure 18.39 illustrates the functional block diagram of the array antenna, which is composed of HW and SW components. The HW side consists of prefixed functions of the array antenna including four complex multiplications and a complex addition of four-element array antenna. It also takes responsibility for accessing interface memory, timing in the SW side, I/O and so on. Meanwhile, the SW side delivers the flexible functions corresponding to DBF algorithms in C programming language. In detail, signals coming from the four-element array antenna are first converted from analog to digital. Using the information of the starting point of the OFDM symbol provided by OFDM demodulation, samples of HeadGI of the inputs and TailGI of the outputs are extracted and are stored into the interface memory. The compensation of RF error and sampling rate error are conducted in the HW side. The HW side also generates timing control signals to the

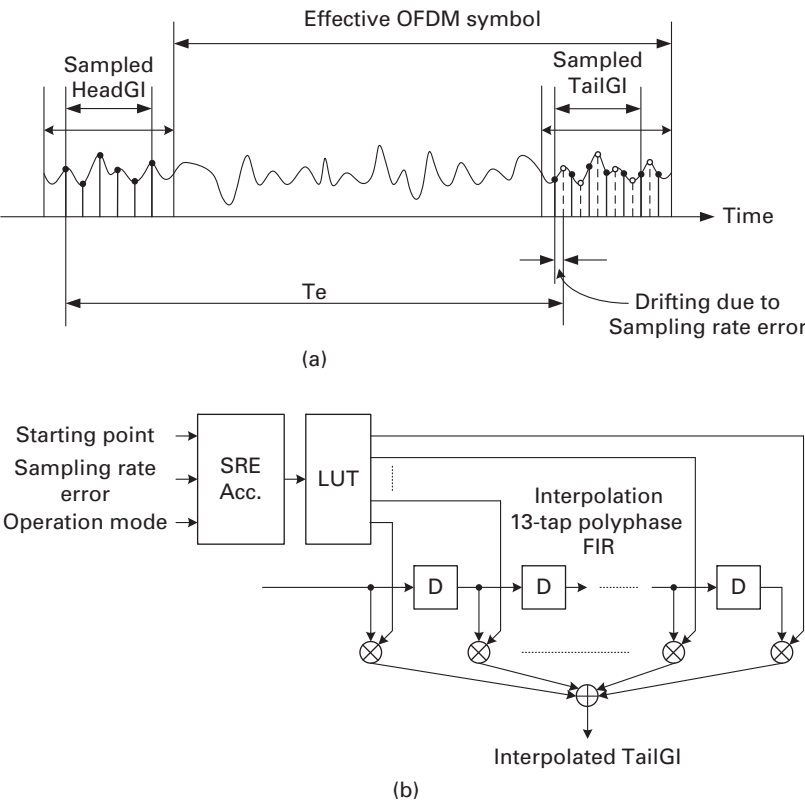


Figure 18.38 (a) Impact of sampling error rate and (b) block diagram of sampling rate error compensator (where “SRE Acc.” Stands for “Sampling Rate error Accumulator”).

SW side to trigger CPU. Then, the coefficients generated by the SW side are used to combine four branches of the array antenna. In the SW side, two banks of block memory storing samples of two successive OFDM symbols directly link to the CPU. Using these samples, the CPU performs a DBF algorithm to generate the coefficients. They are stored back into the corresponding bank of memory and are delivered to the HW side.

Figure 18.40 shows a photograph of the prototype and Table 18.6 lists its hardware features, respectively. Four canned-type tuner devices are used and down-convert the RF signal to IF signals. The four ADCs convert the analog IF signals to digital signals. The lower left boards carries two FPGAs and a Power PC microprocessor embedded in one of the FPGAs is used for the array antenna system. The ISDB-T OFDM demodulation is performed in the right board which carries three FPGAs. Two VirtexII V4000 FPGAs are used to implement OFDM demodulation and one VirtexII V3000 FPGA is used for FEC processing.

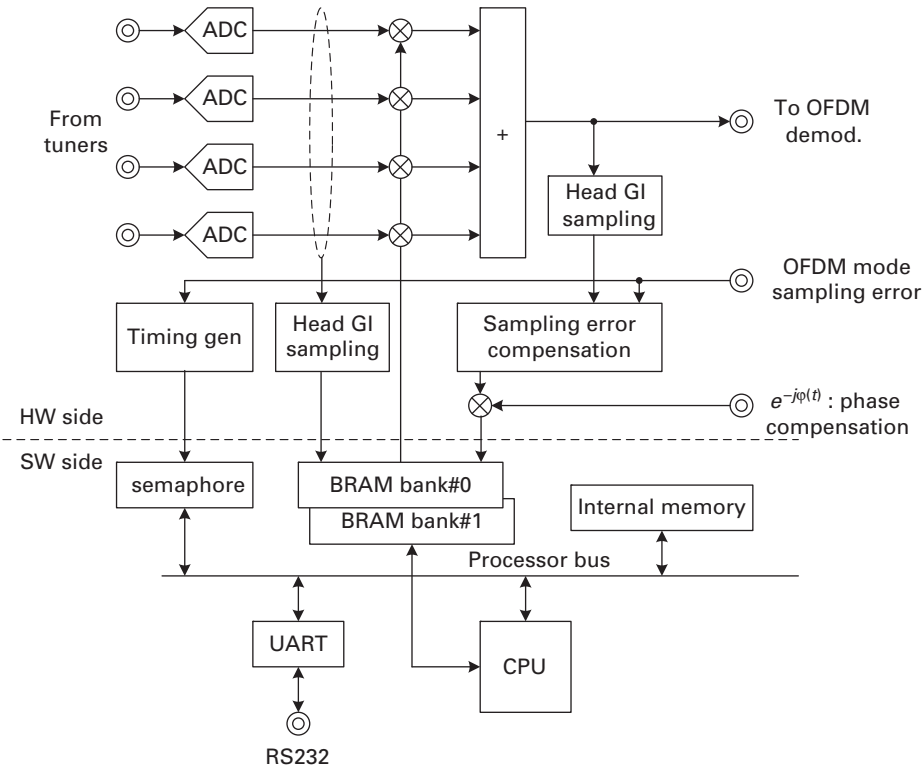


Figure 18.39 Functional block diagram of joint HW-SW solution for array antenna.

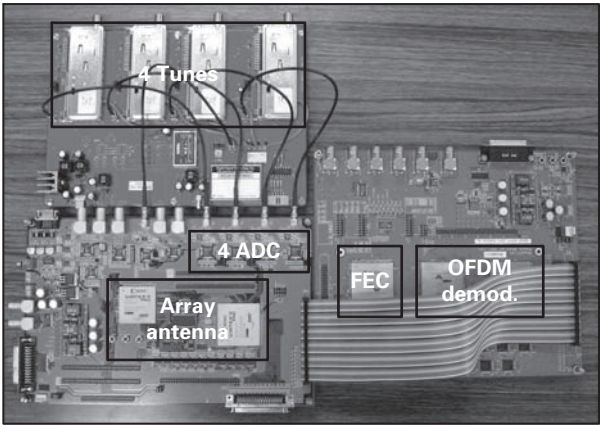


Figure 18.40 Prototype of adaptive array antenna OFDM receiver.

Table 18.6 Features and data of prototype implementation.

OFDM	Array antenna	ADC	4 channel, 10bit
			32 MHz sampling rate
		FDGA	Xilinx Virtex-II Pro VP70 ×1 Xilinx Virtex-II Pro VP20 ×1
		CPU	PowerPc405 260 MHz in FPGA
		Demod.	Xilinx Virtex-II V4000 ×2
		FEC	Xilinx Virtex-II V3000 ×1

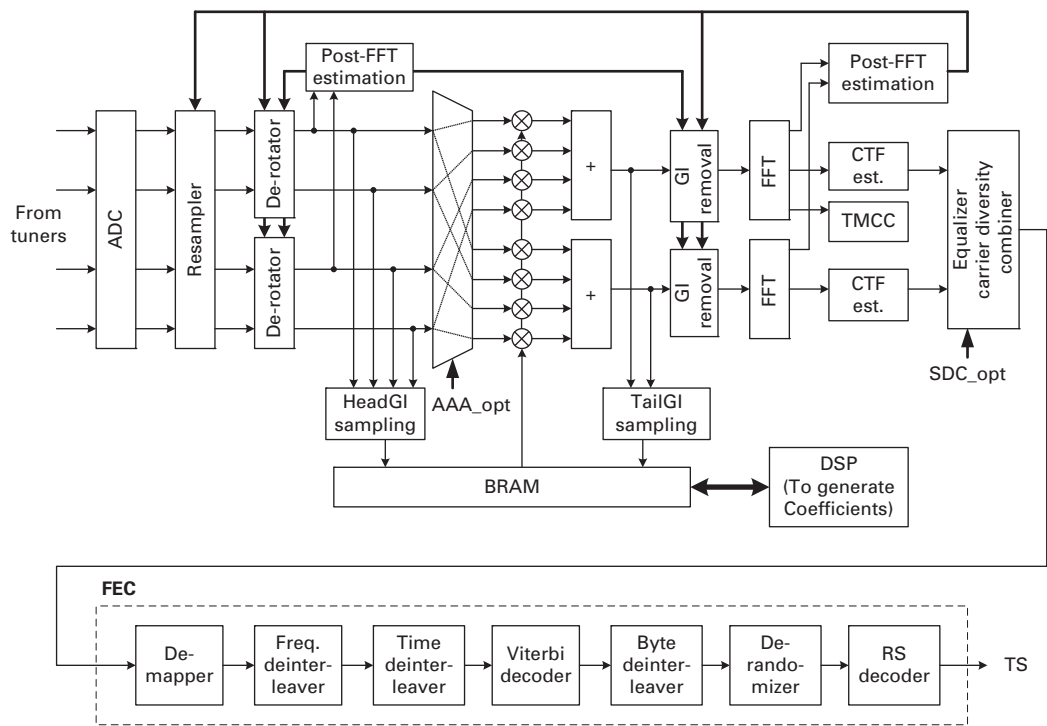
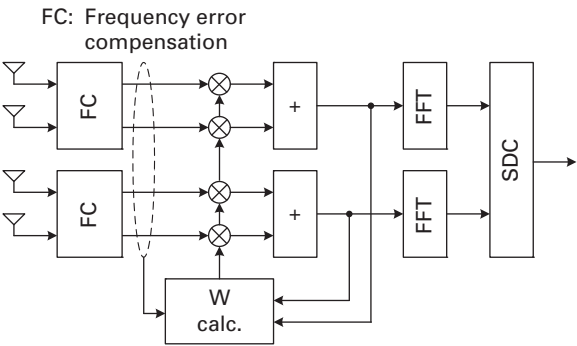


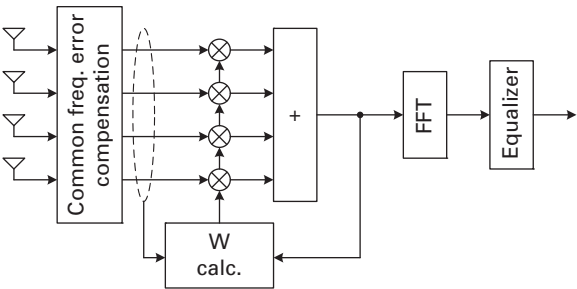
Figure 18.41 Block diagram of AAA-SDC OFDM receiver.

18.5.4 Hybrid pre-FFT adaptive array antenna and post-FFT space diversity combining for Mobile ISDB-T Receiver by FPGA [19]

Figure 18.41 shows a block diagram of a hybrid pre-FFT adaptive array antenna (AAA) and post-FFT space diversity combiner (SDC) system. In this implementation, prefixed processes such as FFTs, CTF estimation, pre-FFT and post-FFT estimations are realized by gates in FPGA. In contrast, DBF algorithms are realized in a digital signal processor (DSP) such as TI C6713 floating point DSP running at 225MHz. The supported adaptive array antenna algorithms are OFDM selective MRC and SMI. Sampled matrix inversion needs 4×4 matrix inversion and then the C6713 floating point calculation can support



(a) System configuration A



(b) System configuration B

Figure 18.42 System configurations A and B.

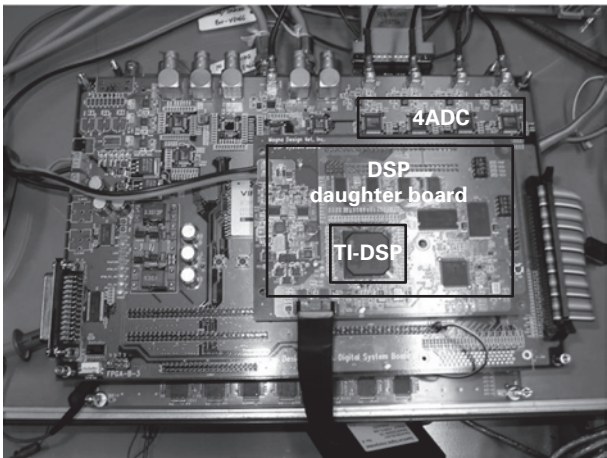


Figure 18.43 Prototype of AAA-SDC OFDM receiver.

Table 18.7 Features of AAA-SDC system.

Array antenna	ADC	4 channel, 10bit 32 MHz sampling rate
	FPGA	Xilinx Virtex-II Pro VP70 $\times$ 1 Xilinx Virtex-II Pro VP20 $\times$ 1
	DSP	TI DSP C6713 TMS320C6712 225 MHz
OFDM	Demod.	Xilinx Virtex-II V4000 $\times$ 2
Carrier diversity	FEC	Xilinx Virtex-II V3000 $\times$ 1

complicated DBF algorithms well. Figure 18.42 shows supported configurations. System configuration A needs two independent 2-element antenna DBF calculation by DSP and System B needs 4-element antenna DBF calculation by DSP. In order to change these configurations, control signals AAA_opt and SDC_opt have to be controlled while DBF software in DSP has to be controlled.

Figure 18.43 shows the photograph of the prototype, which is composed of two large FPGA boards as in Figure 18.40 and a small DSP daughter board on the top. The DSP daughter board carries TI C6713 floating point DSP running at 225 MHz and DRAMs for main memory for DSP. The external 32bit memory bus is connected to the other FPGA board by flat cable as shown in the photograph. Table 18.7 summarizes the system features.

18.6 Conclusions

In this chapter we have discussed various OFDM-based transceiver architecture and their designs such as DTV receiver, OFDM WLAN, and OFDMA transceivers. Since the OFDM-based applications are still expanding, many more OFDM-based communication systems will be developed in future. However, the two key techniques such as Error Detection and Compensation, and Diversity covered in this chapter are fundamental techniques to achieve OFDM-based successful communication system designs. Then those fundamental signal processing techniques should also be further researched in future. Since semiconductor technology progress pushes the system integration level higher and higher, further performance increases will be achieved by heavy signal processing enhancements for both single-input single-output (SISO) and multiple-input multiple-output (MIMO) applications. Then we look forward to digital signal processing researches for future communication systems.

References

- [1] R. W. Chang, "Synthesis of Band-Limited Orthogonal Signals for Multichannel Data Transmission," *Bell Syst. Tech. J.*, vol. **45**, pp. 1775–1796, (1966).
- [2] S. B. Weinstein and P. W. Ebert, "Data Transmission by Frequency-Division Multiplexing using the Discrete Fourier Transform," *IEEE Trans. Comm.*, vol. **19**, pp. 628–634 (1971).

- [3] B. Muquet, Z. Wang, G. B. Giannakis, M. de Courville, and P. Duhamel, "Cyclic Prefixing or Zero Padding for Wireless Multicarrier Transmissions?" *IEEE Trans. Comm.*, vol. **50**, no. 12, pp. 2136–2147 (2002).
- [4] www.xgpforum.com/document/, "Next Generation PHS Specification (Revision 4)," PHS MoU Document, A-GN4.00–01-TS, (2009).
- [5] 3GPP, "Evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN) Release 8," TS36.201 (V8.3.0), Feb. (2009).
- [6] "IEEE Standard for Local and Metropolitan area networks Part 16," Institute of Electrical and Electronics Engineering, Inc. Std. IEEE 802.16E-2005, (2005).
- [7] S. Sakaguchi, M. Hori, H. Asato, *et al.*, "An Adaptive array direction control LSI for mobile digital HDTV receiver," Session: 7.4 Acquisition Systems 7.4–3, IEEE International Conference on Consumer Electronics 2005.
- [8] H. Asato, T. Tabata, D. H. Pham, *et al.*, "A Software-configurable Adaptive Array Antenna Systems for ISDB-T Reception," Session: 11.2 Signal Processing for Multiple Ant., DVB-T and DAPSK/DMT 11.2–1, IEEE International Conference on Consumer Electronics ICCE2006, USA, Jan. 2006.
- [9] D. H. Pham, T. Tabata, H. Asato, S. Hori, T. Wada, "Joint Hardware-Software Implementation of Adaptive Array Antenna for ISDB-T Reception," *IEICE Trans. of Communication*, vol. **E89-B**, No.12, pp. 3215–3224, Dec. 2006.
- [10] S. Nakahara, H. Hamazumi, K. Shibuya, M. Sasaki, "An Application of Diversity Combination Techniques to Broadcasting Wave Relay Station for ISDB-T," *ITE Tech. Report*, vol. **25**, No.31, pp. 7–12, March, 2001.
- [11] N. Kikuma and M. Fujimoto, "Adaptive Antennas," *IEICE Trans. Of Communications.*, vol. **E86-B**, no.3, pp. 968–979, March 2003.
- [12] L. C. Godara, "Application of Antenna Arrays to Mobile Communication, Part II: Beam-forming and Direction-of-arrival Consideration," *Proc. IEEE*, vol. **85**, no.8, pp. 1195–1245, Aug. 1997.
- [13] J. Imai, M. Fujimoto, T. Shibata, *et al.*, "Experimental Results of Diversity Reception for Terrestrial Digital Broadcasting," *IEICE Trans. Commun.*, vol. **E85-B**, no.11, pp. 2527–2530, Nov. 2002.
- [14] S. Sakaguchi, M. Hori, T. Wada, and N. Itoh, "An Adaptive Array Direction Control LSI for Mobile Digital HDTV Receivers," *IEEE ICCE 2005*, pp. 299–300, Jan. 2005.
- [15] S. Hori, N. Kikuma, and N. Inagaki, "MMSE Adaptive Array Utilizing Guard Interval in OFDM Systems," *Electron. Commun. Jpn. 1 Commun.*, vol. **86**, no.10, pp. 1–9, 2003.
- [16] J. Gao, T. Iida, H. Mizutani, *et al.*, "The Implementation of a 2/4/8 Antennas Configurable Diversity OFDM Receiver for Mobile HDTV Application," *Int. J. Digital Multimedia Broadcasting*, Article ID 303627, Dec. 2009.
- [17] M. Speth, F. Classen and H. Meyr, "Frame Synchronization of OFDM Systems in Frequency Selective Fading Channels," *Proc. IEEE 47th Vehicular Technology Conference*, Phoenix, USA, pp. 1807–1811, May 1997.
- [18] K. Ishikawa and T. Wada, "Radio and Sampling Frequency Error Detection Algorithm for Extended Global Platform (XGP)," *Proc. ITC-CSCC2009*, Jeju Korea, S33-D6: Mobile & Wireless Communication (2), pp. 1092–1095, July 7th, 2009.
- [19] D. H. Pham, J. Gao, T. Tabata, *et al.*, "Implementation of Joint Pre-FFT Adaptive Array Antenna and Post-FFT Space Diversity Combining for Mobile ISDB-T Receiver," *IEICE Trans. of Communication*, Vol. **E91-B**, No.1, pp. 127–138, Jan. 2008.

19 Front-end architectures and impairment corrections in multimode and multi-antenna systems

Guillaume Villemaud, Jacques Verdier, Matthieu Gautier, Ioan Burciu, and Pierre-François Morlat

19.1 Introduction

Nowadays the market continues to show a real interest in the development of telecommunication networks based on radiofrequency (RF) systems. Particularly, we can note a proliferation of radiofrequency standards such as WiFi, WiMax, DVB-T, or the 3G or 4G standards and a strong desire to merge those standards in a single terminal. Along with the existing standards, all new standards allow the operators to offer new and better services in terms of speed, quality, and availability. Consequently, in order to handle this important diversity of telecommunication techniques, there is a growing interest in developing new front-end architectures capable of processing several standards.

Of course, one first point is to enhance the digital resources of the terminal to offer Software Defined Radio (SDR) potential so that the structure can be easily reconfigured and evolved. But along with this evolution towards high numerical capacities, terminals are also required to propose an RF part enabling multi-standard, multi-frequency and often multi-antenna operations (referred to as multi-* terminals).

For the multi-standards research domain we can distinguish two different categories of receivers: non-simultaneous receivers using switching techniques [2][3][4][5] and simultaneous receivers [1][16]. To be able to offer full cognitive or SDR potential, a simultaneous case must be privileged. The state of the art of multi-standard simultaneous reception architectures uses the front-end stack-up technique – each chain being dedicated to the reception of only one standard. Nonetheless, this architecture is characterized by some drawbacks such as a bad complexity–performance trade-off, but also the price and the physical size. Thus a new trend is to think directly of the whole terminal as an intrinsically multi-* one, and then try to design the global RF and digital system from this point of view. Several ideas concerning multi-standard single front-end receivers, capable of simultaneously processing two standards, have been recently patented [6]. Using only one front-end reduces the complexity; nevertheless additional problems linked to the analog-to-digital conversion appear.

The same goes for multi-antenna systems, as nearly all new wireless standards propose MIMO (multiple-input multiple-output, i.e. multiple antennas on both parts of the radio

link) modes. Thus, multiplying the number of antennas also suggests to increase the cost of the RF part. Then this additional constraint must also be integrated in the global design of an optimal transceiver [14]. In each particular case, the capability of the digital part must moreover be considered to find new innovative solutions.

Organized into the following four sections, this chapter is devoted to radio front-end architectures and related impairment corrections for multiband, multi-antenna, and multi-mode receivers. In Section 19.2, we will first depict the state of the art of multi-* receivers, starting from classical structures used for single-antenna and single-band operation. Then the different ways of multiplexing signals with shared components are described.

Section 19.3 describes in detail one possible original solution for designing a multi-band architecture, the patented so-called double IQ (in-phase and in-quadrature) receiver [33]. This receiver combine two standards (or two frequency bands) in a unique front-end, and we will show that this kind of structure could take large advantage of the combination of RF and digital operations, particularly to decrease the impact of reducing the constraints on analog components.

For multi-antenna processing, also known to create increased costs, some ways of using numerical resources to decrease analog cost are described in Section 19.4. Then two different approaches to decrease the cost of some multi-antenna receivers will be described in this section. The first one uses the intrinsic ability of a multi-antenna system to digitally process the different copies of the incoming signal in order to relax constraints on certain duplicated components. The second one goes further in combining the different branches of a multi-antenna receiver, using a code multiplexing technique.

Furthermore, Section 19.5 of this chapter will present two examples of combined multiband and multi-antenna structures. The first part of this section describes a structure able to digitize several overlapping channels to be able to jointly manage two interfering signals and thus increase the frequency reuse in a congested frequency band. Then, the second part of this section presents the combination of two previously depicted structures: the double IQ and the code multiplexing ones. In the final section of the chapter, we will give a summary and there will be further discussion.

19.2 State of the art of multi-* receivers

Today, transceivers' designers often continue to see the development of a multi-* system as the addition of separate subsystems, even if integrated in the same package (or chip). We will here emphasize that some more efficient ways of thinking could greatly enhance the global system performance or significantly reduce the overall cost. In order to be able to appreciate this potential gain, this section first provides the details of classical structures used to realize mono-* systems and then presents the different ways of multiplexing different branches.

19.2.1 Front-end stack up for multiband or multi-antenna systems

The reception chain of a radiofrequency transmission can be categorized as follows by depending on the number of antennas and on the number of useful frequency bands that

are simultaneously processed: one-antenna reception chain used to process a useful signal composed of a single frequency band, multiple-antennas reception chain used to process a useful signal composed of a single frequency band, one-antenna reception chain used to process a useful signal composed of several frequency bands and finally multiple-antennas reception chain used to process a useful signal composed of several frequency bands.

Concerning the front-end architectures used by a basic one-antenna receiver processing a useful signal composed of a single frequency band, three technical solutions can be mentioned as state of the art: the heterodyne architecture [8], the homodyne architecture [9] and the sub-sampling architecture [11]. These three architectures are presented by the schematics in Figure 19.1.

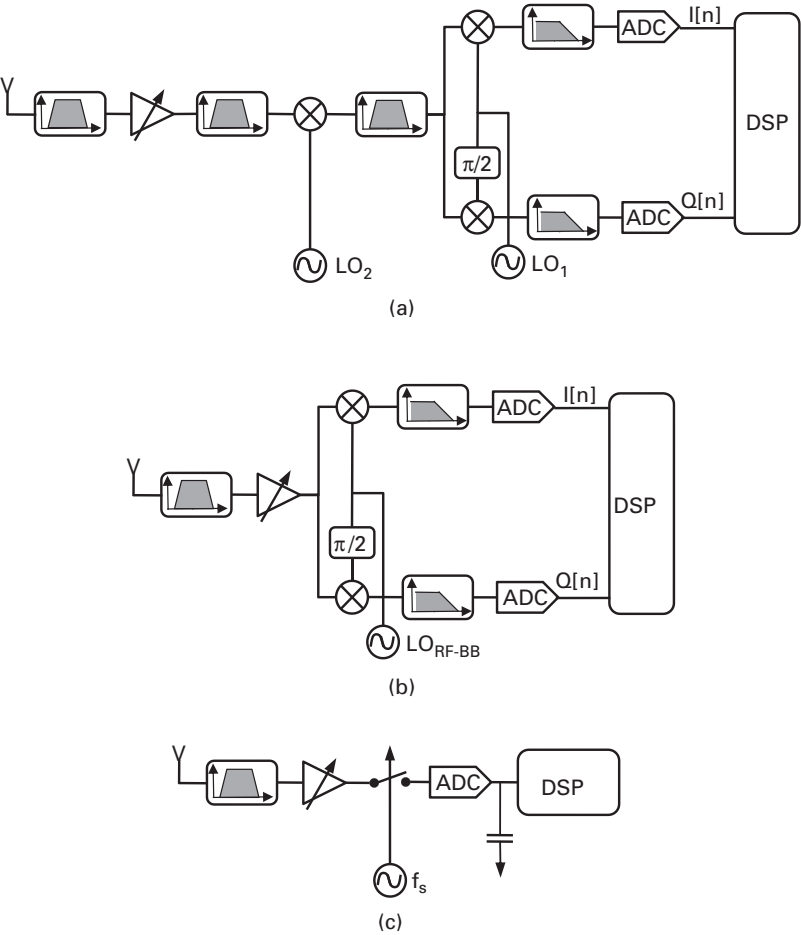


Figure 19.1 Front-end architectures used to realize a one antenna receiver processing a single frequency band useful signal: (a) heterodyne architecture; (b) homodyne architecture; (c) sub-sampling architecture.

The input of the reception chain using one of these three architectures is composed of an antenna that represents the interface between the hertzian transmission channel and the electronic processing chain. The signal obtained at the output of the antenna is filtered by a band filter in order to reject the parasitic signals that have their spectrum at the exterior of the frequency band dedicated to the standard. The state of the art of electronic components allows the on-chip integration of these components [10], even though they are generally not on-chip integrated components. Afterwards, the filtered signal is amplified by a low-noise amplifier (LNA). The feature that makes the difference between the three front-end architectures is the way how the filtered and amplified RF signal is translated in the baseband domain. The heterodyne and the homodyne architectures are based on a mixed signal processing method where the useful signal is multiplied with a sinusoidal signal in order to translate it around a smaller frequency. The sub-sampling architecture is based on a sub-sampling signal processing method where the useful RF signal becomes a periodical spectrum signal and therefore a baseband “copy” of the useful band can be recovered.

Depending on the desired features of the mono-antenna mono-band receiver, each of the three architectures offers certain advantages. Generally, the heterodyne receiver is dedicated to high-speed radiofrequency transmissions due to the performance advantages offered in terms of filtering and gain distribution along the processing chain. In contrast, the homodyne architecture is usually dedicated to low-performance receivers having low cost and low power consumption. Concerning the sub-sampling front-end architecture, its main advantage resides in its reconfiguration flexibility. In fact, this type of sub-sampling receiver can be easily reconfigured to process a different RF frequency band signal. This is a very interesting feature for the SDR domain where the receiver must be able to process different communication standards.

If there are several architecture solutions for a mono-antenna receiver processing a single frequency band RF signal, the state of the art of multi-antenna or multiband simultaneous processing receivers is mainly based on a single architecture: the stack-up of several front-ends [1]. Each of these parallel RF chains is dedicated to only one antenna and one useful frequency band. The choice of the individual architecture of the different stacked up front-ends is made between the three “classical” mono-antenna and mono-band front-end architectures. This choice depends on different constraints imposed by the radiofrequency transmission such as processing performance, equipment power consumption and cost or reconfiguration flexibility. The stack-up architecture is widely used in the radio-communication applications using multi-antenna techniques such as MIMO transmissions or multi-frequency band simultaneous processing due to its design and implementation simplicity. Nevertheless, the use of the parallelized RF signal processing architecture implies a high level of both power consumption and cost of the mobile terminal.

In order to decrease the power consumption and cost of the mobile terminals implementing simultaneous multiple antennas and/or multiple frequency band processing, a natural solution consists in mutualizing some of the electronic components. Several technical solutions were proposed and analyzed in order to multiplex the different useful RF antenna and/or spectral contributions in the RF domain. This allows thereafter using a

single frequency translation chain in order to translate in the baseband domain the resulting signal of the multiplexing stage. In the baseband domain, the different useful contribution can be obtained by implementing a demultiplexing signal processing technique. The next subsection is completely dedicated to the presentation and to the study of these multiplexing and demultiplexing signal processing techniques.

19.2.2 Multiplexing techniques

The RF stack-up architecture is an obvious choice: the fact of having N separated dedicated chains allows the demodulation of each branch with a significant quality (high signal-to-noise ratio, SNR, at the ADC, analog-to-digital converter, input). However, this choice enforces a high complexity of the analog front-end.

In this subsection, the general concept of RF front-ends based on multiplexing techniques is introduced. The aim is to reduce this complexity of the architecture without decreasing the SNR quality after branches demodulation. The approach is to reuse or share the functional elements of the radio architecture by exploiting an additional dimension. The signal of one branch must be separated from those of the other branches in order to facilitate the sharing of a single analog reception chain. This separation could be performed by the multiplexing techniques.

Figure 19.2 illustrates the principle of the multiplexing applied to multi-* receivers having N antennas. Notice that the multiplexing operation (MUX) is implemented analogically, while the demultiplexing operation (DEMUX) is implemented digitally.

In the following, we consider how a radio can be multiplexed between several branches by noting that the multiplexing can be done either in time or in frequency, or by using codes, analogous to the commonly used multiple access schemes in communications. Figure 19.3 illustrates the multiplexing techniques for the three dimensions.

With the application of these multiplexing techniques to multi-* receivers, we can make the following analyses:

Time multiplexing:

In the time-multiplexed approach, the outputs from the antenna array are multiplexed together by using an RF switch. A theoretical structure has been introduced in [13]. The RF switch has to operate at a speed of N times the sampling rate. The weakness of this approach is an SNR loss because only a

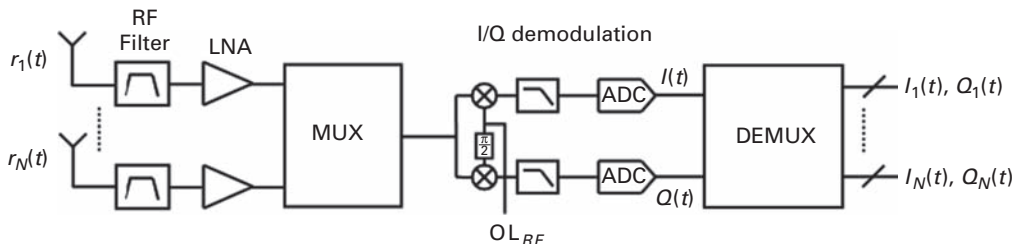


Figure 19.2 Analog and digital front-end of the multiplexing architecture.

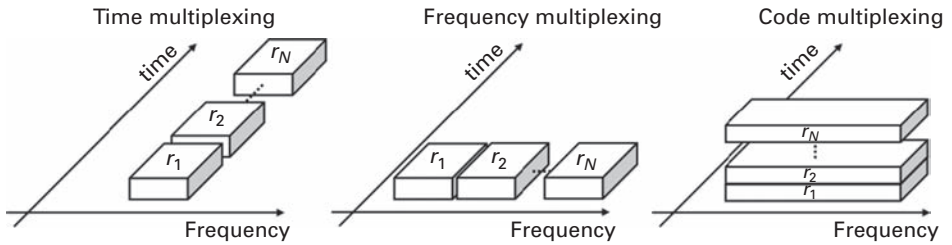


Figure 19.3 Time, frequency, and code multiplexing of an N branches multi-* system.

fraction of the received power has been used by the demodulator. Moreover, implementation constraints of the RF switch are an important barrier for the development of this kind of architecture. Using time multiplexing technique for RF signals that are received during time periods that do not overlap could reduce these constraints.

Frequency multiplexing: Several signals can be multiplexed together by applying a different frequency offset to each of them. Once frequency separated, this would allow these signals to be combined and then passed through a common receiver. In [14], Evans *et al.* propose a multi-antenna receiver based on frequency multiplexing. Several weaknesses of this approach have been observed: the stringent filtering requirements and the need for additional RF local oscillators. Frequency multiplexing could also be applied to multiband RF signals whose frequency bands are widely separated.

Code multiplexing: The third dimension uses a code-division technique in order to combine several individual signals simultaneously on the same frequency band. By using orthogonal codes, the correlation between the different signals becomes very low so that information from each antenna can be extracted. This method has just been recently approached in [15][12]. The codes domain appears to be the most suited for multi-antenna multiplexing.

To illustrate the multiplexing architecture, Figure 19.4 shows the power spectrum of the signal after frequency and code multiplexing for a multi-antenna receiver. We use an $N=4$ antennas system that receives an IEEE 802.11g type signal having a 20 MHz bandwidth and a 2412 MHz RF frequency.

For the code multiplexing architecture, the codes have a rate N times higher than the symbol one, so the resulting signal has variations that are N times faster than the information signal and the bandwidth of the signal frequency spectrum increases N times. For the frequency multiplexing architecture, a frequency offset is applied to each signal in order that they be frequency separated. Figure 19.4 shows that the

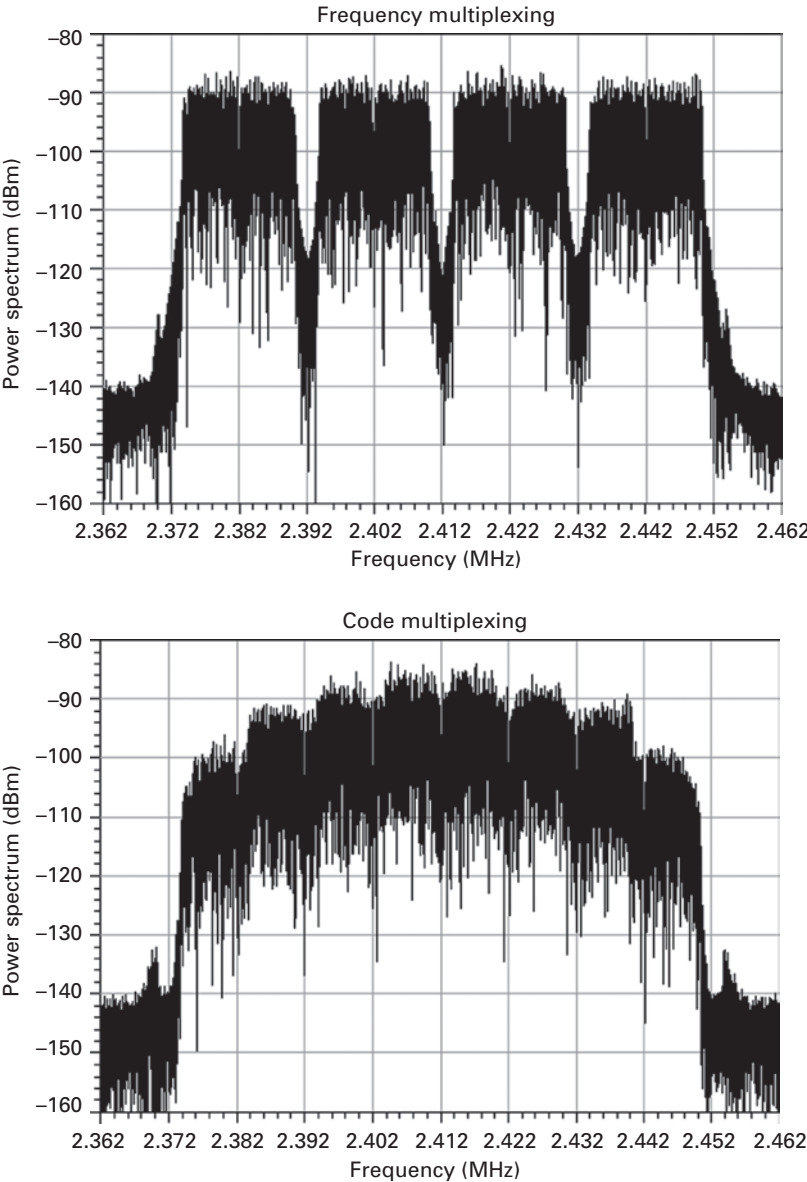


Figure 19.4 Power spectrum of an 802.11g signal after frequency and code multiplexing.

frequency bandwidth required for both structures is 80 MHz. Indeed, the multiplex operation generates an increase of the required bandwidth by a factor of $N=4$.

For the three dimensions, many issues have to be considered in order to validate the architectures in terms of performance, RF defaults sensitivities, implementation, complexity (number of components), power consumption, etc. These metrics vary considerably depending on these three approaches and also really depending on the application and the size of the bands.

However, it would appear that only the full stack-up architecture offers a high level of performance. The multiplexing-based solutions could achieve an important cost saving. In [Section 19.3](#), low complexity implementations are proposed for both frequency and code multiplexing: an original frequency multiplexing architecture based on the double IQ structure is proposed for the reception of two distinguished bands, and the code multiplexing technique based on spread spectrum is introduced.

19.3 Mixed analog–digital front-end architecture for multiband simultaneous reception

Conceiving a multi-* system must take into account that the digital part can easily perform some operations which would require numerous components if done in the analog part. The main idea of the double IQ structure given here as an innovative example of a mixed analog–digital multiband front-end is to reduce the global bandwidth to digitize, while the two mixed frequency bands can be separated in the digital domain with very simple operations. Moreover, after some examples of an UMTS-WiFi receiver’s performance, the third subsection will show that a specific digital processing can help to decrease the constraints imposed on analog components to design such a receiver.

19.3.1 Double orthogonal frequency front-end architecture dedicated to bi-band reception

Double orthogonal translation techniques are generally used in image frequency rejection front-ends [18] [17]. The rejection of the signal having its spectrum in the image frequency band of the useful signal is a key point of the heterodyne front-end architectures. The double orthogonal translation technique uses a signal processing method that realizes a multiplexing of the useful signal and the image frequency band signal during the translation from RF to baseband and is then demultiplexing these signals in the baseband domain. Starting from this structure, [16] proposed and developed a novel front-end architecture dedicated to the simultaneous reception of two separate frequency band signals. The proposed dual-band simultaneous reception architecture is shown in [Figure 19.5](#). The input blocks – antenna, RF filter, and LNA – are parallelized in order to have good low noise amplification and filtering of the two useful signals $S_1(t)$ and $S_2(t)$. Once filtered and amplified, the useful signals are translated in the baseband domain by using a double orthogonal frequency translation structure. An important aspect of this technique that has to be underlined is the choice of the first local oscillator frequency. This frequency is chosen in such a manner that each of the useful signals has a spectrum situated in the image frequency band of the other. In the baseband domain, the four obtained signals are digitized. The baseband component of the two useful signals is obtained by using two dedicated basic operations processing similar to that used by the image frequency rejection structures. The main advantage of this single front-end receiver architecture is the power consumption and cost decrease compared to those of a two-heterodyne front-end stack-up architecture. This is mainly due to the use of less

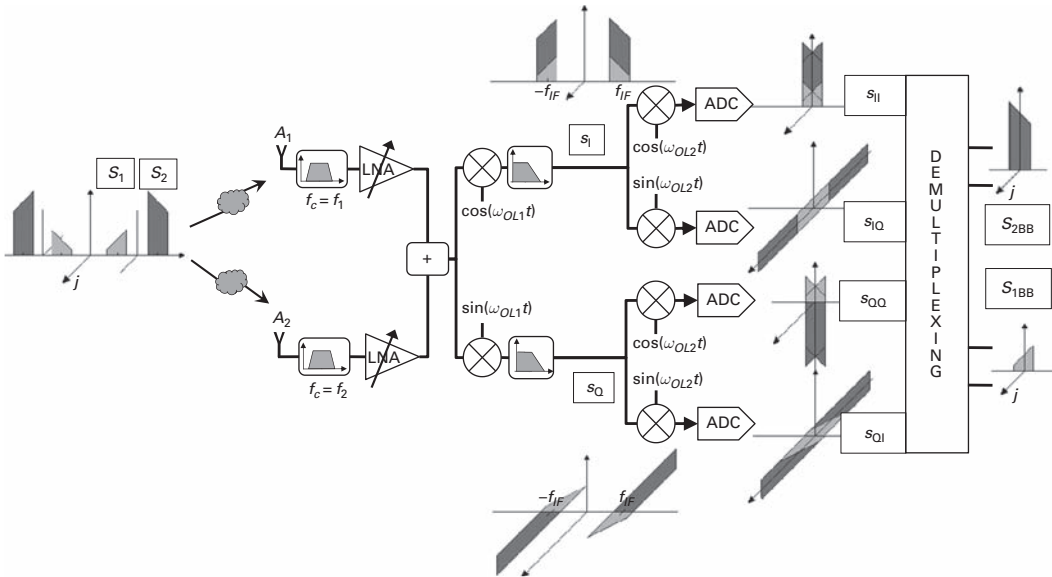


Figure 19.5 Single front-end receiver architecture based on the double orthogonal frequency translation technique.

electronic components such as frequency synthesizers and bandpass RF filter used for the rejection of the image band signal. The main drawback relies on the performance's sensitivity to the level of phase and gain mismatches between the two branches of the orthogonal frequency translation blocks.

19.3.2 UMTS and WiFi simultaneous reception study case

The theoretical analyses [16] of the single double IQ front-end receiver reveal the potential advantages of this type of receiver compared to the two heterodyne stack-up receiver when the transmission scenario involves the processing of two independent frequency band signals. In order to validate these theoretical analyses, the study scenario chosen here is that of a simultaneous UMTS and WiFi reception. A first power consumption study presented in [19] is realizing a comparison between the power consumption of a heterodyne front-end stack-up receiver and the proposed single front-end receiver. The hypothesis chosen for this comparative power consumption study is that two receivers have the same performance in terms of bit error rate (BER) when simultaneously processing the UMTS and the WiFi signals. The updated conclusions of this study show that the single double IQ receiver has a power consumption of 199 mW while the front-end stack-up receiver consumes 249 mW. As illustrated by the chart of Figure 19.6, this 20 percent gain in terms of power consumption is mainly due to the use of only two instead of four frequency synthesizers.

The second study concerning the double IQ receiver assesses the processing performance of this type of receiver and is presented in [16]. In the beginning, two realistic models of a simultaneous UMTS and WiFi transmission are realized using a system-level

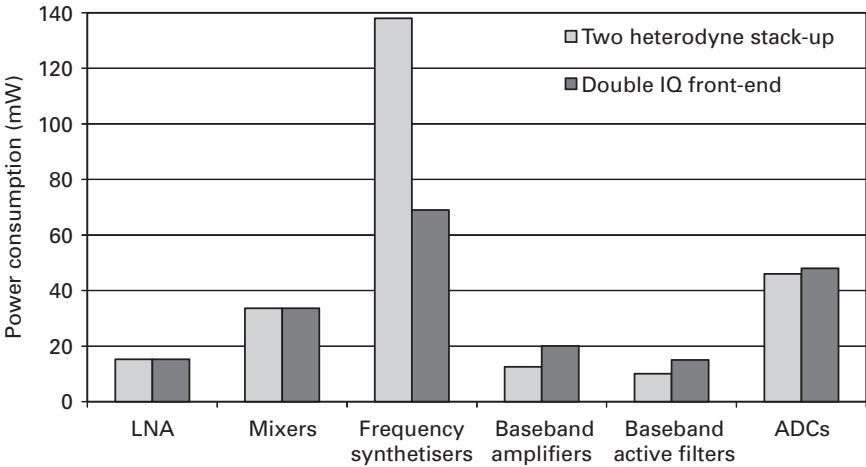


Figure 19.6 Power consumption of the different block types used by two receivers that implement the heterodyne front-end stack-up and the double IQ architecture in order to have the same performance of simultaneous UMTS and WiFi processing.

simulation software (here Agilent ADS) in order to evaluate the BER of the two radio-frequency links. The difference between these two models is the receiver architectures: one is using the single double IQ front-end while the other is using the heterodyne front-end stack-up. The results of the first series of simulations show that the two receivers offer identical performance in terms of BER when the gain and phase imbalance between the two branches of the orthogonal frequency translation blocks are not taken into account. Meanwhile, as shown in Figure 19.7, when these IQ impairments defaults are integrated in the two radiofrequency models, the processing quality of the single double IQ front-end receiver is degraded up to six times for realistic levels of the phase and gain mismatches. A second conclusion that can be drawn from the results presented here is that one of the solutions to reduce the performance sensitivity to the IQ impairments is the increase of the gain dynamics of the power control stage that is implemented in the RF domain. Nevertheless, the main drawback of this solution is the power consumption rise that is inherent in the increase of the gain dynamics of the power control stage.

19.3.3 Digital method dedicated to the mitigation of the orthogonal defaults impact on the double IQ receiver performance

The processing quality of the single double IQ front-end during a two-independent frequency band signal reception depends on the level of gain and phase mismatch between the two branches of the orthogonal frequency translation blocks. In fact, if theoretically each of the two RF useful signal bands is completely reconstructed by the digital demultiplexing while the component corresponding to the other useful band is completely rejected, due to the presence of the IQ mismatches, the complementary component is not completely rejected and becomes therefore a parasitic signal. In other words, the processing quality of a single double IQ signal is decreased by the joint

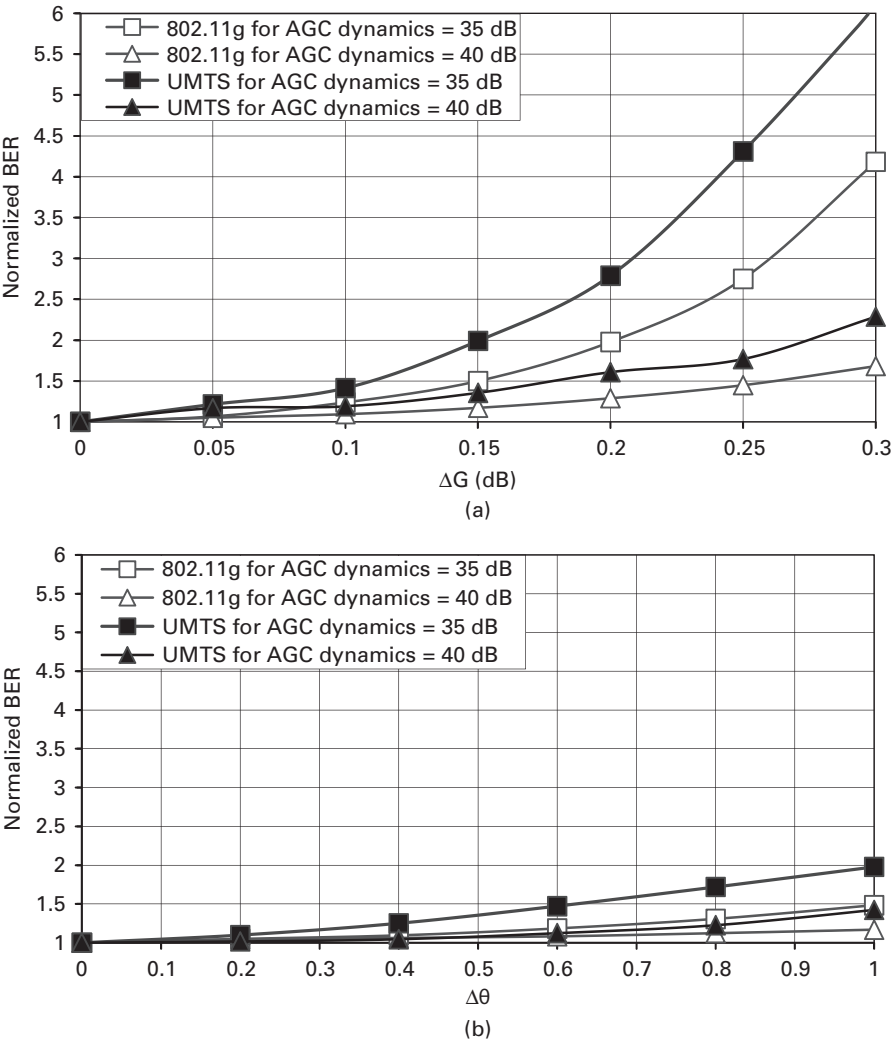


Figure 19.7 Normalized UMTS and 802.11 BER evolutions depending on the (a) gain and (b) phase imbalance characterizing the first orthogonal frequency translation of a double IQ receiver.

presence of the IQ mismatch and the RF complementary frequency band. Two solutions can be performed in order to limit the sensitivity of the processing quality: the increase of the gain dynamics of the power control stage and the use of a digital post-processing method. Knowing that the increase of the gain dynamics is increasing the power consumption of the power control block, the second solution represents a suitable choice for a low power consumption receiver. Concerning the single double IQ front-end receiver, a digital method was designed and implemented for the case of a simultaneous UMTS and 802.11g transmission [27]. The schematic of Figure 19.8 presents the implementation of this digital method along with the double IQ front-end. The left part of the structure corresponds to the double IQ architecture already detailed in Figure 19.5.

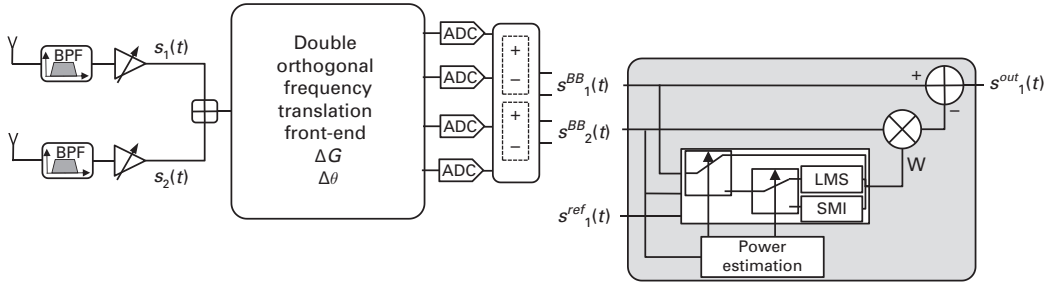


Figure 19.8 Digital context-aware method used to mitigate the influence of IQ mismatches in a double orthogonal translation receiver.

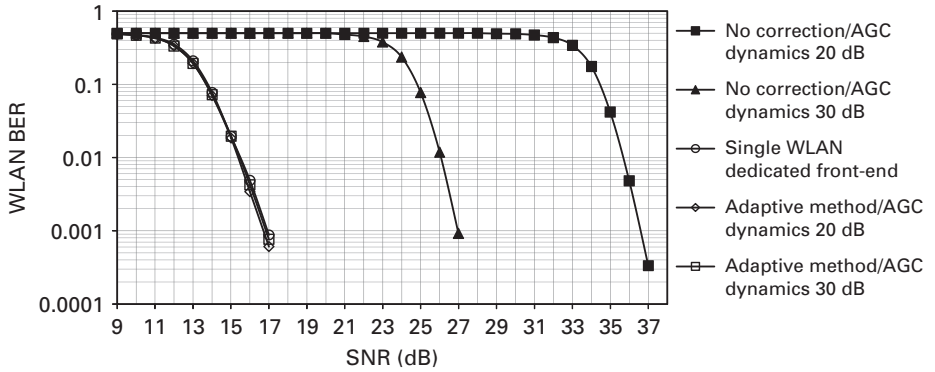


Figure 19.9 WLAN BER evolution during a UMTS and WiFi simultaneous reception for different values of the gain control and for two configurations depending on the implantation of the adaptive IQ mismatches correction method.

The right part represents the processing blocks used to mitigate the IQ impairments. This signal processing uses a training sequence s_1^{ref} in order to estimate through a minimum mean square error (MMSE) algorithm [23] the presence of the parasitic complementary component in s_1^{BB} . By realizing a simple subtraction, this parasitic signal is rejected from the useful baseband component. The rejection level depends on the precision of the MMSE method. In order to limit the complexity and therefore the power consumption of the digital method two MMSE algorithms are integrated. One is a very simple LMS (least mean square), which is iterative with a low computational cost but slow convergence. The second one is a SMI (sample matrix inversion) which is more complex but gives a direct solution. Each of these algorithms is activated depending on the power variation of the RF complementary frequency band.

This context-adaptive method was implemented using Matlab software. In order to have more meaningful results, this was incorporated in a co-simulation platform which also includes the UMTS and WLAN transmission sources, as well as the ADS model of the analog double orthogonal front-end. Figure 19.9 presents several BER evolutions as functions of the WLAN signal's SNR. These results reveal that, for different levels of the

gain control and therefore of the interfering signal, the influence of the IQ mismatches is mitigated by the digital adaptive method. Furthermore, the results obtained when integrating the proposed architecture are compared to those of a dedicated front-end stack-up. When the MMSE method is used along with the proposed structure, the simulation shows that the performance is identical with that of the front-end stack-up.

Thus, combined with a specific digital processing, we can see that such a structure is not only reducing the number of components and the energy consumption, but also could greatly reduce the constraints on the AGCs. In this example, one can note that only 20 dB of dynamic could be sufficient to propose equivalent performance compared to classical structures.

19.4 Some methods for simplifying analog parts of multi-antenna front-ends

In the previous part, one enhanced multiband architecture has been shown. But is the same approach valid for a multi-antenna system? [Section 19.4.1](#) details a deeper study of the impact of RF impairments on a classical multi-antenna receiver. Obviously, a multi-antenna system performing antenna processing in the digital domain already requires numerical resource. Then minimizing the error on a received signal degraded by a radio channel could be viewed as the same problem as mitigating the impact of RF impairments. We will show in [Section 19.4.1](#) that this consideration goes for only certain kinds of impairments, but that indicates the characteristics that could be relaxed to reduce the global cost, tending to a dirty-RF point-of-view [34]. [Subsection 19.4.2](#) presents an innovative structure to combine multi-antenna contributions while preserving the diversity of the signals and therefore the performance of the system, by using code multiplexing in the RF part.

19.4.1 RF impairments mitigation via antenna processing

Due to the increase of receiver branches, the cost of the RF stage of multi-antenna systems (MIMO, but also SIMO, single-input multiple-output, or MISO, multiple-input single-output) grows and also leads to taking into account dirty-RF principles [34]. The idea is the use of poor quality RF stage components, introducing digital processing to compensate the received signal degradation due to a dirty RF stage. In this context, the aim of this subsection is to describe the effect of multi-antenna processing on the reduction of RF impairments effects, in the special case of a 802.11g OFDM 1x2 SIMO receiver using a zero-IF down converter, that is a very sensitive architecture to RF impairments due to its definition. Furthermore, OFDM technique, based on precise frequency modulation is also a critical case for the study of the impact RF defaults.

The impact of three different RF impairments is presented:

- phase noise, due to local oscillator errors
- frequency offset, due to error between receiver and emitters LO
- IQ mismatches (gain and phase).

The impact of SIMO processing in order to mitigate RF impairments is given in terms of relative BER, taking into account a reference BER in the case of no RF impairments

introduced in simulation tools. For considered RF impairments, simulated and measured performance of SISO and SIMO structures are compared and explained on the basis of theoretical considerations of each case. Additive white Gaussian noise and also multipath fading channels are taken into account.

19.4.1.1 SIMO processing

The chosen antenna processing algorithm is the traditional Sample Matrix Inversion based on MMSE criterion [23], taking advantage of the knowledge of the 802.11g training sequence and ensuring a good trade-off between complexity computation and estimation of optimal complex weights. The main assumption of this algorithm is to consider that the propagation channel is constant during an 802.11g frame.

19.4.1.2 RF impairments mitigation

Phase noise

The Theoretical impact of phase noise in an 802.11g receiver is given by [20] for each subcarrier of a considered OFDM symbol m of emitted signal x , through the propagation channel denoted h . Here n denotes the white Gaussian noise applied on subcarrier k

$$Z_m(k) = H_m(k)X_m(k) \underbrace{S_{m,0}}_{CPE} + \underbrace{\sum_{l=0, l \neq k}^{M-1} X(l)H(l)S_{m,l-k}}_{ICI} + N_m(k). \quad (19.1)$$

The term $S_{m,i}$ corresponds to DFT of one realization of the phase noise $e^{j\varphi(n)}$ during the m th OFDM symbol:

$$S_{m,i} = \frac{1}{M} \sum_{n=0}^{M-1} e^{-j(2\pi n i / M + \varphi(n))}. \quad (19.2)$$

The received samples in the frequency domain exhibit two terms due to phase noise. All phase noise correction schemes are based on Common Phase Error (CPE) estimation and mitigation, and are relatively efficient in WLAN OFDM systems for small phase noise working conditions. The correction is possible due to the invariance of CPE within an OFDM symbol using the four pilot subcarriers. However, under high phase noise, the Inter-Carrier Interference (ICI) term dominates over the CPE. In this case, the phase noise suppression becomes really difficult and the performance requirement may not be guaranteed due to the loss of orthogonality between subcarriers. This is especially true under selective channel propagation when $H(l)$ coefficients can take large values and have an important impact on ICI values.

SIMO mitigation phase noise

Figure 19.10 gives the achievable performance in an SIMO configuration for different propagation channels including phase noise impairment with PSD spectrum described as follows:

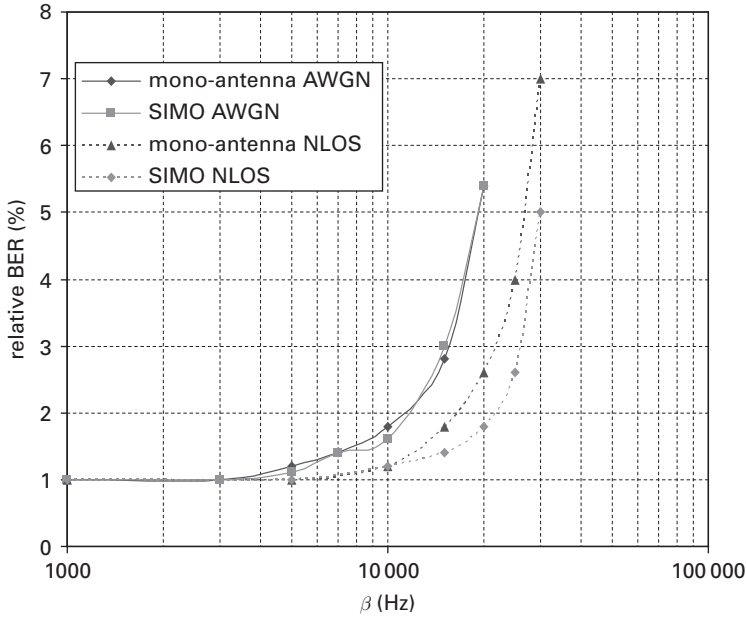


Figure 19.10 Phase noise mitigation in AWGN and NLOS simulated conditions. Results are given in relative BER versus local oscillator phase noise characteristics (β).

$$\begin{array}{ll}
 -60 \text{ dBc/Hz} & \text{at } \beta \text{ Hz} \\
 -80 \text{ dBc/Hz} & \text{at } \beta \times 10 \text{ Hz} \\
 -100 \text{ dBc/Hz} & \text{at } \beta \times 100 \text{ Hz}
 \end{array}$$

This corresponds to a Wiener phase noise exhibiting a -20dB/dec slope with β being the PSD spectrum bandwidth of the phase noise. With β increasing, SNR degrades and hardware components get cheaper.

As described in Figure 19.10, SIMO processing does not mitigate phase noise impairments significantly, especially under the AWGN case. This result can be explained by the fact that the optimal complex coefficients applied on each branch of the SIMO receiver are estimated on the basis of the assumption that they are constant during the whole 802.11g frame, which is not the case when ICI error increases. This observation could be mitigated if the transmission were under a frequency selective channel. In this case, the multi-antenna receiver is able to take advantage of spatial diversity and $H(k)S_{k-l}$ can take small values on each received branch, ensuring a decreasing of the ICI term.

Frequency offset

Reference [21] gives the expression of the sampled received signal for the subcarrier k ($k=0, \dots, M-1$) affected by frequency offset error in the frequency domain

$$Z(k) = H(k)X(k) \underbrace{S_0}_{CPE} + \underbrace{\sum_{l=0, l \neq k}^{M-1} S_{l-k}H(l)X(l)}_{ICI} + N(k) \quad (19.3)$$

where the sequence S_{l-k} is given by:

$$S_{l-k} = \frac{\sin[\pi(l-k+\varepsilon)]}{M \sin\left[\frac{\pi}{M}(l-k+\varepsilon)\right]} e^{j\pi\left(1-\frac{1}{N}\right)(k-l+\varepsilon)}. \quad (19.4)$$

The received signal expression is very close to the received signal after phase noise error, supposing that the ICI term is constant and does not depend on the index of considered OFDM symbol m .

SIMO mitigation for frequency offset

Figure 19.11 presents the BER relative performance for different frequency offset values, in the range of 0 to 55 kHz. Due to the fact that the ICI term is constant in the case of a transmission impaired by frequency offset, a constant error is taken into account by the SIMO processing and complex weights computed by the MMSE algorithm are still optimal during the entire frame. That is why 1×2 SIMO processing ensures a really good frequency offset impact mitigation.

IQ mismatch

Reference [22] gives the theoretical impact of IQ mismatch on the received signal

$$Z(k) = \mu H(k)X(k) + \lambda H(-k)^* X(-k)^* + \mu N(k) + \lambda N(-k)^* \quad (19.5)$$

where μ and λ depend on the I-Q imbalance (θ phase and α gain):

$$\begin{cases} \mu = \cos\left(\frac{\theta}{2}\right) + j\alpha \sin\left(\frac{\theta}{2}\right) \\ \lambda = \alpha \cos\left(\frac{\theta}{2}\right) - j \sin\left(\frac{\theta}{2}\right) \end{cases} \quad (19.6)$$

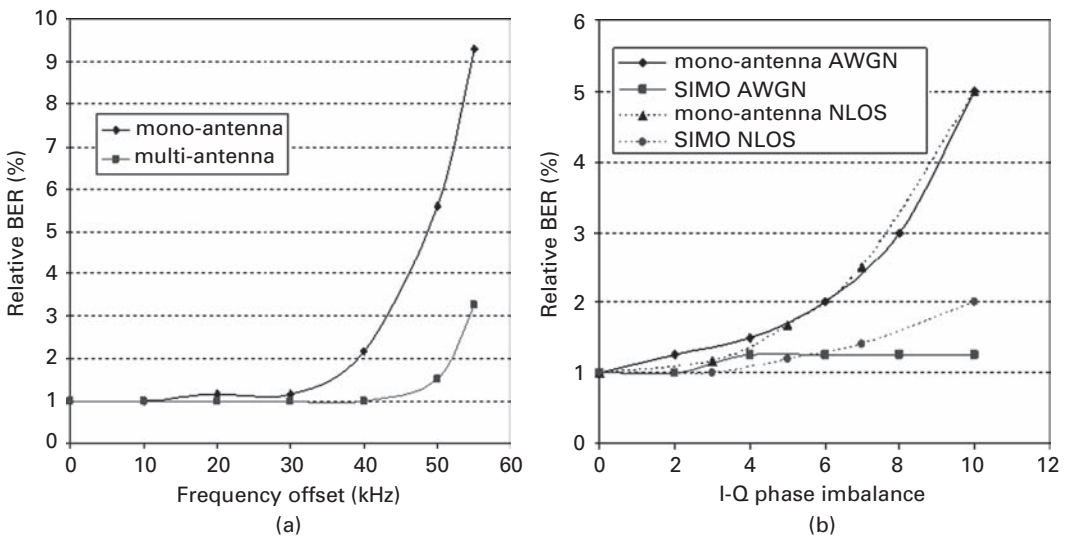


Figure 19.11 Relative BER versus frequency offset (a) and IQ phase imbalance (b) in NLOS working conditions. SISO and 1×2 SIMO performance is given.

SIMO mitigation for IQ mismatch

Figure 19.11(b) gives the performance of SIMO processing in the presence of IQ phase imbalance in the AWGN measured channel and also NLOS with fading channels. It clearly appears that multi-antenna processing ensures a really good IQ phase mismatch mitigation. The same results can be observed taking into account gain imbalance.

We prove that the conventional multi-antenna treatment can significantly reduce damage caused by the most common RF impairments: IQ imbalance, local oscillator phase noise, and frequency offset. The multiplication of RF branches can be made cheaply if we consider this natural impairments compensation, without ever having to resort to specific digital processing. However, we have to remember that using SIMO processing for naturally reducing RF impairments leads to decreasing SIMO processing capabilities in order to reduce BER degradation due to multipath propagation.

19.4.2 Multi-antenna structure with code multiplexing

As introduced in Subsection 19.2.2, the use of the multiplexing technique could simplify the structure of multi-* receiver architecture. In the case of a multi-antenna system, code multiplexing seems to be a suitable technique. This section is dedicated to the implementation of a multi-antenna front-end architecture based on code multiplexing. It is composed of two parts: first, the theoretical aspects are presented, and then implementation issues are given including a complexity evaluation.

19.4.2.1 Structure of the code multiplexing architecture

The multi-antenna signals are received simultaneously on the same frequency band. In order to achieve the spectrum overlapping, decorrelation can be done by the spread spectrum technique. The Direct Sequence Spread Spectrum (DSSS) technique is the basis of Code Division Multiple Access (CDMA) technology [28]. It consists in allocating a spreading code to each branch with all these codes being orthogonal two by two.

The received signal that carries information is multiplied by the code which is a pseudo-random sequence of N binary entities having a rate N times higher than the symbol. Thus, the resulting signal has variations that are N times faster than the information signal and the bandwidth of the signal frequency spectrum increases N times.

The novel antenna diversity receiver using code multiplexing is depicted in Figure 19.12 and consists of two parts: the analog multiplexing and the digital demultiplexing.

In this figure, the different signals could be in baseband or in radiofrequency. A signal X could be represented in RF using the notation X^{RF} or in baseband using the notation X^{BB} . The RF and the baseband notation are related by:

$$X^{RF}(t) = \text{Re}[X^{BB}(t)e^{j2\pi f_0 t}] \quad (19.7)$$

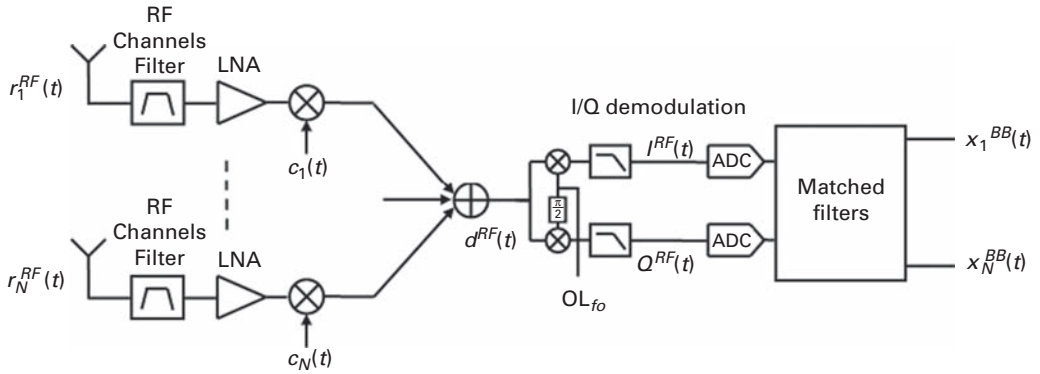


Figure 19.12 Analog and digital front-end of the code multiplexing architecture.

with f_0 being the RF modulating frequency.

In an ideal transmission through a non-dispersive channel, the signal received at the k th antenna is expressed by:

$$r_k^{RF}(t) = h_k s^{RF}(t) + n_k^{RF}(t) \quad (19.8)$$

with h_k the path loss attenuation and $n_k^{RF}(t)$ the total RF noise at the input of the k th antenna; $s^{RF}(t)$ is the transmitted signal.

In order to perform the code multiplexing, each received signal is multiplied using a dedicated code $c_k(t)$. The codes being orthogonal, signals can be clearly overlapped in time and frequency, as they are separable by their code. So, the adding operation between the different contributions can be done:

$$d^{RF}(t) = I^{BB}(t) + Q^{BB}(t) = \sum_{k=1}^N c_k(t) \cdot r_k^{RF}(t) = \sum_{k=1}^N c_k(t) \{h_k s^{RF}(t) + n_k^{RF}(t)\}. \quad (19.9)$$

After the code multiplexing step, the signal $d^{RF}(t)$ is transposed to the baseband frequency by an IQ demodulator, leading to its baseband version $d^{BB}(t)$:

$$d^{BB}(t) = \sum_{k=1}^N c_k(t) \{h_k s^{BB}(t) + n_k^{BB}(t)\}. \quad (19.10)$$

So far, the different steps will be implemented in the analog part of the receiver. The decoding step will be performed digitally by successive correlations (denoted by $\langle \cdot, \cdot \rangle$) between the multiplex signal and the same coding sequence:

$$x_l^{BB}(t) = \langle c_l(t), d^{BB}(t) \rangle = \sum_{k=1}^N \langle c_l(t), c_k(t) \rangle \{h_k s^{BB}(t) + n_k^{BB}(t)\}. \quad (19.11)$$

If the codes are orthogonal, we have $\langle c_l(t), c_k(t) \rangle = \delta[k - l]$.

As a final result, we get:

$$x_l^{BB}(t) = h_l s^{BB}(t) + n_l^{BB}(t). \quad (19.12)$$

The digital baseband symbols received on each antenna $\{x_l^{BB}(t), l = 1, \dots, N\}$ are recovered.

19.4.2.2 Implementation issues

The proposed architecture works for any multi-antenna scheme, for every antenna array, and every digital algorithm. However, a limitation of the concept of using orthogonal spreading codes is that only an even number (2, 4, 8 . . .) of antennas can be received.

The synchronization between coding and decoding is not such an important issue as during a UMTS transmission. Indeed, the propagation delay of the spread signal through the circuit path is well-predicted through accurate circuit analysis and simulation. Therefore, attaining synchronization between the spreading and despreading codes is a trivial matter.

For the purpose of comparison of the stack-up architecture (multiplication of classical structures as in Figure 19.1(a)) with the code multiplexing architecture (Figure 19.12), the complexity of the two structures has to be computed. Table 19.1 reveals the number of components required for each structure and some constraints associated with these components. In this table N is the number of antennas, B_c is the channel bandwidth, f_0 is the RF standard center frequency, and BW stands for bandwidth.

Regardless of technological constraints, the major difference between the two structures is the number of mixers and ADCs. The proposed structure reduces by a factor N the number of ADCs compared to the stack-up structure. This complexity gain should induce an important reduction of the chip area.

The digital complexity of the matched filters (N filters using N coefficients) could be neglected. The reason is that the additional complexity of this processing is not important compared to the initial complexity of algorithms needed to perform the multi-antenna processing.

More details on this structure could be found for example in [12]. A performance study has shown that some equivalent results could be expected with this approach compared to classical ones. This structure will be reused in Subsection 19.5.2 with more interesting results.

19.5 Front-ends combining multiband and multi-antenna capabilities

Going further in the analysis of multi-* front-ends brings us naturally to investigate the problem of combining multiband and multi-antenna systems. Here we take as examples the case of dual channel operation with multi-antenna receivers. The first case considers

Table 19.1 Complexity of the front-ends using a stack-up and a code multiplexing architecture.

	Stack-up structure		Code multiplexing structure	
	Number	Specification	Number	Specification
Antenna	N	$BW = B_c$	N	$BW = B_c$
Filter	N	$BW = B_c$	N	$BW = B_c$
LNA	N	$BW = B_c$	N	$BW = B_c$
Mixer	$2N$	f_0	$N + 2$	f_0
Adder	0		1	N to 1
ADC	$2N$	$BW = B_c$	2	$BW = NB_c$

some overlapping channels only separated by antenna processing, while keeping a classical multi-antenna structure combined with a wideband sampling. The second case here combines the previously exposed double IQ and code multiplexing structures. This structure prefigures what could be a modern view of a multi-* receiver for future standards, such as LTE-Advanced.

19.5.1 Dealing with overlapping channels with multi-antenna processing

Context.

The 2.45 GHz ISM band is well known for hosting lots of competitive radio standards. Even if we only consider IEEE 802.11 standards, channels are defined with only 5 MHz separation while occupying at least a 20 MHz bandwidth. Therefore adjacent channels are strongly overlapping and interfering with one another. Then it is commonly admitted for WiFi system deployment that a frequency allocation panel (FAP) may use only three non-interfering channels in the whole band, otherwise it leads to complex optimization problems [29]. Moreover, additional technologies such as Bluetooth or ZigBee are sharing this band, producing a large amount of possible interference.

Study case

The study case we depict here is based on a terminal able to receive a 40 MHz bandwidth signal (i.e. possibly containing up to five overlapped channels). Considering that this receiver has multimode capabilities (by the way of multiple dedicated chips or ideally software defined radio structure), we would like to investigate possible performance while receiving information on two overlapped channels with different standards. Several approaches could be considered to mitigate or suppress interference. One way is to use interference cancellation techniques like multi-user detection as presented in [30, 31]. This way becomes harder (i.e. resource-consuming) in the case of interference from various standards. Another way is to use multiple-antenna techniques, enabling spatial rejection of the interfering signal. It is this way that we focus on here, considering that more and more terminals tend to be equipped with two or more antennas in order to be compatible with new MIMO standards. Multiple-input multiple-output systems require multiple antennas, hence multiple RF chains and digital processing capabilities, but not all devices always use MIMO modes, depending on the network architecture and the radio channel conditions. However, multiple-antenna resources could be used to perform effective interference mitigation for the simpler radio modes, at the same time operating a real multimode interference mitigation which is not dependent on the chosen standards.

For the sake of better understanding, we will here present only results with non-MIMO standards, i.e. without space-time coding, but results are not limited to mono-antenna standards. Hence, the considered communication standards will be IEEE 802.11b and IEEE 802.11g.

Obviously these standards are not new, but offer the advantage of a large background of theoretical study and complementarities of used waveforms. Standard 802.11b uses a spread spectrum technique (DSSS) while 802.11g uses orthogonal frequency division multiplexing (OFDM) signals. These two kinds of waveforms do not have the same

sensitivity to interference, and we of course suppose here that there is no coordination between concurrent signals, that is no common MAC layer. Moreover, DSSS and OFDM are the bases of lots of standards intended to be used for wireless personal area networks (WPAN), wireless sensor networks (WSN), or even body area networks (BAN).

Subsequently results will be presented on two concurrent communications each on 20 MHz bandwidth, possibly overlapping in the whole 40 MHz bandwidth of the receiver.

From this point it is important to remember that here we only consider the problem of the receiver. One reason is that, as stated before, we do not want to analyze MIMO modes. The point is that our approach does not necessarily require that all nodes have multi-* ability and no pre-coding at the emitter is needed. Another reason is that our aim is to mutualize information from different independent networks, thus received signals are never synchronized (MIMO modes requiring synchronized sources). So we can argue that all results presented here could be applied for SISO but also MIMO systems, MIMO bringing an additional coding gain to the spatial diversity gain considered here.

Multi-* performance.

In multi-* operating mode, we would like to evaluate performance when two overlapped channels with different modes are received. The first scenario presented corresponds to two emitters, one on 802.11b mode, the other one on 802.11g mode. We analyze here the results in a multipath environment, hence non-line-of-sight (NLOS) conditions. Nevertheless received SNR for both channels are kept constant, corresponding in each case to a $5 \cdot 10^{-3}$ BER without any concurrent communication in a SISO configuration (only one received antenna). This supposes that the radio link would be quite good with only one of those two links. But in our configuration, both links are maintained at the same time and moreover on overlapped channels (see Figure 19.13). The separation between the center frequency of each channel, also called channel spacing Δ , is then varied from 0 (co-channel interference) to 15 MHz.

Table 19.2 summarizes mean measured values obtained with a radio platform. We compare BER obtained with or without SISO processing at the receiver. Obviously, with co-channel interference in the SISO configuration, radio links are unusable and BER reaches 0.5. In this co-channel configuration, SISO processing does not bring a sufficient performance enhancement to expect channel reuse. But what can be observed is a quick decrease of BER from 0 to a 5 MHz channel spacing with the use of multi-antenna processing. This suggests that multi-* devices allow a better frequency reuse and can support simultaneous connections with different modes on strongly overlapped channels.

Of course, the obtained BER values depend directly on exact operating conditions (emitters' positions, received power, channel type, etc.). But in every tested condition, the global behavior is that SISO processing allows the maintenance of reliable links even with channel spacing down to 10 MHz, hence strongly overlapping channels.

Another test is presented in Figure 19.14. In this case, channel spacing is fixed at 10 MHz. The 802.11b link (line with *) is maintained as previously at the SNR level corresponding to a $5 \cdot 10^{-3}$ BER without any concurrent communication in an SISO. Then the power of the 802.11g channel is progressively increased. Bit error rate curves are shown in Figure 19.15.

Table 19.2 Obtained BER with measured channels depending on channel spacing and antenna configuration.

Δ	802.11b SISO	802.11b SIMO	802.11g SISO	802.11g SIMO
0 MHz	0.5	0.4	0.5	0.4
5 MHz	0.3	0.08	0.4	0.07
10 MHz	0.2	0.01	0.35	0.02
15 MHz	0.01	0	0.01	1.2×10^{-4}

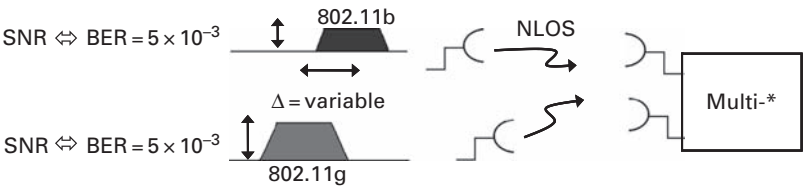


Figure 19.13 Definition of study case with overlapping channels.

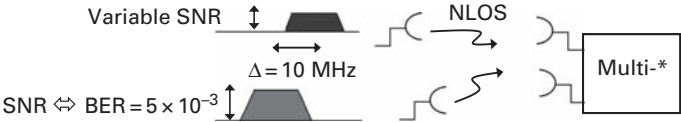


Figure 19.14 Definition of study case with overlapping channels and variable power of 802.11g signal.

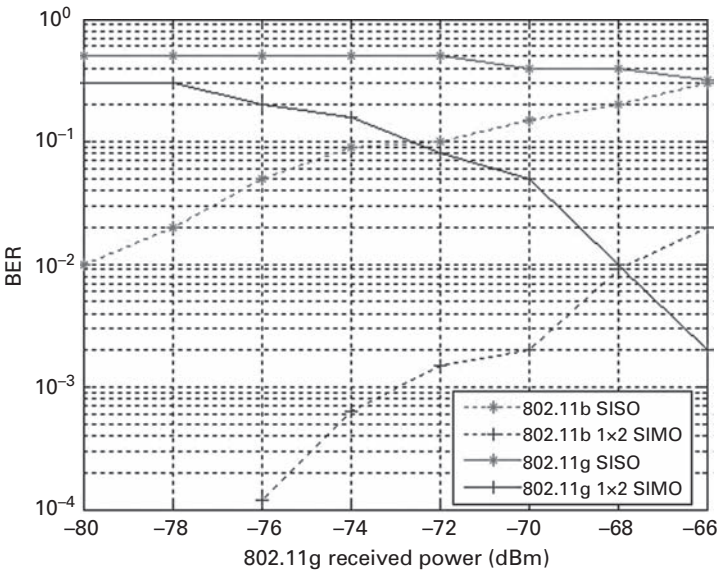


Figure 19.15 BER depending on 802.11g received power for a channel spacing of 10 MHz in a multipath environment.

We can observe in [Figure 19.15](#) that in an SISO case, the 802.11b link is rapidly degraded when the 802.11g interferer increases in power, thus none of those two links appear exploitable. But with antenna processing (line with +), we can see that an interesting trade-off could be made so as to ensure two parallel available radio links.

Finally, this study shows that a smart combination of analog characteristics (here wideband sampling) and digital processing (simple antenna processing) could help to increase notably the frequency reuse, i.e. the global spectrum efficiency. Of course, this study is voluntarily not considering some MIMO processing or some evolved interference cancellation techniques like SIC or PIC [31]. It is obvious that adding more specific digital processing will increase the ability to reuse the spectrum, but here the fact is to first show that required processing (here SMI) could already be integrated in the global optimization of multi-* receivers.

19.5.2 Combining code multiplexing and double IQ structure

The forthcoming challenges of radio-frequency standards involve higher throughputs and low power consumption of compact terminals. In order to obtain higher throughputs, the proposed solutions converge towards the same techniques: OFDM (Orthogonal Frequency Division Multiplexing), adaptive coding and modulation, MIMO (multiple-input multiple-output), as well as scalable bandwidths [25]. Nevertheless, the impact of these techniques on the power consumption of the mobile terminals is important [24]. For example, the future LTE-Advanced standard takes into consideration a discontinuous spectrum usage, as well as multi-antenna techniques for the downlink transmission. In order to address these specifications the actual state of the art of radiofrequency receivers imposes the use of dedicated front-end stack-up architectures. For example, if we consider a two-frequency band discontinuous spectrum transmission scenario where two antennas are used for the reception, the receiver will have to integrate a four front-end stack-up. Each of these front-ends is dedicated to the processing of one of the contributions obtained from the combination of two antennas and two non-adjacent frequency bands. Due to the use of the stack-up of dedicated front-ends, it becomes obvious that this method imposes high complexity, but especially high power consumption [1]. Based on the merging of the dual band simultaneous architecture and the RF code multiplexing structure, a single front-end architecture dedicated to the LTE-Advanced receivers was proposed and studied [26]. This section is devoted to this architecture of single front-end receiver dedicated to the simultaneous reception on two antennas of two independent frequency bands. Meanwhile, the structure can be generalized for several antennas by increasing the length of the orthogonal codes used during the RF multiplexing stage.

The proposed architecture of the single front-end receiver is shown in [Figure 19.16](#). It consists of four main parts: the RF dedicated chains, the RF code multiplexing, the double IQ structure and the digital part that realizes the code demultiplexing, and the multi-antenna processing. In the following, it is considered that the two input signals S and S' are the result of the passage of a bi-band signal through two different transmission channels. Let $\text{Band}_1\text{-Band}_2$ and $\text{Band}_1'\text{-Band}_2'$ be the two pairs of contributions

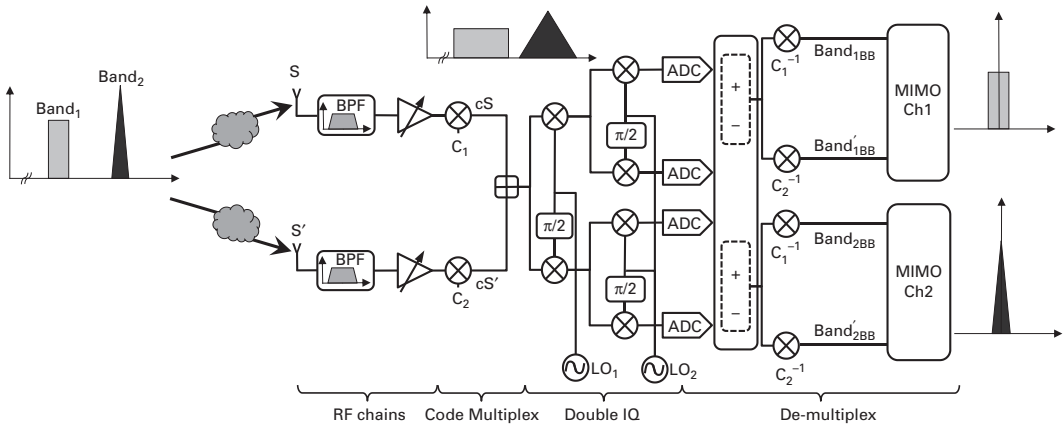


Figure 19.16 Receiver capable of realizing a two antennas processing of a discontinuous spectrum signal.

composing S and respectively S' . Once received by two independent antennas, the two input signals are separately filtered and amplified by two RF filters and two LNA respectively. The multiplexing of the four contributions is realized by a two-step method. To begin, the orthogonal code spreading technique is used in order to multiplex the two input signals once filtered and amplified. Let cS , cS' , $cBand1$ - $cBand2$ and $cBand1'$ - $cBand2'$ be the signals resulting from the use of the coding technique. The orthogonal codes used here have a length of two and a chip time two times smaller than the symbol time of each of the two signals. When multiplied with the orthogonal codes, each of the S_1 and S_2 signals will see each of the components of the pairs $Band1$ - $Band2$ and respectively $Band1'$ - $Band2'$ spread in the same manner around their own central frequency. This multiplexing step is concluded by the addition of the cS_1 and cS_2 signals. As a result of this operation a dual band spectrum signal is obtained. Each of its two frequency bands is composed of the addition between $cBand1$ and $cBand1'$ and respectively $cBand2$ and $cBand2'$. The second block of the architecture assessed in this section is to implement the double IQ technique similar to that used by the dual band simultaneous reception structure. The processing realized by this block consists of a frequency translation of the RF orthogonal coding block output into the baseband domain. The two digital outputs are the baseband translation of the two frequency bands composing the signal at the output of the RF coding block. Therefore, each of them is the addition between the baseband translated signals $cBand1$, $cBand1'$ and respectively $cBand2$, $cBand2'$. In order to de-multiplex each of these two pairs of signals, two matched filters are applied. Once the two pairs of signals corresponding to the two antennas' reception of a dual band signal are separately obtained, two antennas SIMO processing methods are used. Each of them has one of the pairs of decoded signals as input.

The physical layer requirements concerning LTE-Advanced are not yet finalized; a particular RF signal model was chosen for the realized simulations in order to evaluate the performance of such a receiver. In order to obtain an OFDM signal having a discontinuous spectrum, a signal composed of the addition of two 802.11g non-

overlapping channels was used. For the transmission channel the chosen model is the AWGN (additive white Gaussian noise). The results presented in Figure 19.17 show the BER evolution for different E_b/N_0 of the input signal. The receivers are using either the proposed unique front-end architecture or the dedicated front-end stack-up architecture. These results are mainly concerning the SIMO receivers using the two different architectures when receiving the dual-channel signal (expressed as Stack-up-SIMO, DoubleStructure-SIMO-Channel1 and DoubleStructure-SIMO-Channel2, respectively in Figure 19.17). The BER evolution when the receivers are realizing an SISO reception is also shown (see Stack-up-SISO and DoubleStructure-SISO in the figure). Both in the SIMO and SISO case, the reception performance with using the front-end stack-up architecture is slightly better compared to those obtained using the unique front-end structure. This is due to the fact that the codes used by the Double Structure are not completely orthogonal. We observe that the SIMO reception offers the same performance gain when using the two architectures. Meanwhile, the theoretical gain of a SIMO system in an AWGN channel is 3 dB, but the measured channel manages to offer only quasi AWGN conditions. In fact, the two noises added to the initial signal are not completely uncorrelated. This is why the SIMO gain obtained during ADS simulation does not reach this 3 dB level.

Parallel to this performance study, a comparative power consumption study was made. The conclusions of this study show a 33 percent power consumption gain in favor of the single front-end architecture compared to a four heterodyne stack-up receiver (211 mW instead of 315 mW). If the complexity decrease due to the use of fewer components is also taken into account there, it becomes obvious that the proposed structure offers an excellent performance–consumption–complexity trade-off.

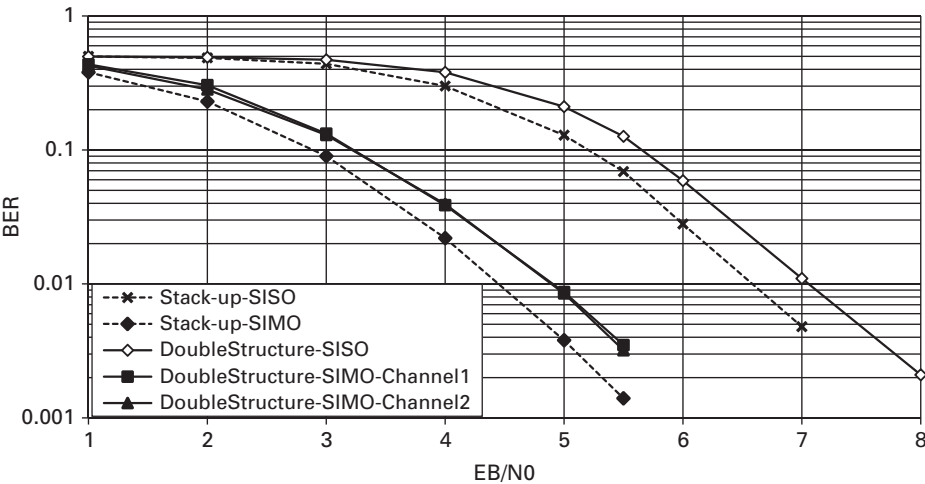


Figure 19.17 BER evolution during the simultaneous SIMO reception of a signal composed of two non-overlapping 802.11g channels.

19.6 Conclusions

After introducing the global context of multi-* terminals, this chapter has detailed in [Section 19.2](#) the state of the art of this kind of architecture, which could be based on stacking up several single-band single-antenna front-ends, or on a multiplexing strategy.

[Section 19.3](#) has shown that some new architectures could be defined to multiplex two bands in for instance a double IQ receiver. This structure has proven to be an interesting trade-off between performance and cost, but requires being associated to some dedicated digital processing to offer full potential.

Then, two multi-antenna structures are discussed in [Section 19.4](#). The first one is a simple example to underline that digital processing capabilities could help to reduce the constraints on analog components, hence decreasing the cost of multiplying the branches. The second one has shown that a code-multiplexing technique is efficient to reduce the number of required IQ blocks in a multi-antenna system, if digital resources are available to demultiplex the signals.

Finally, [Section 19.5](#) has also proposed two different structures, combining multi-antenna and multi-frequency operation. The first one is able to simultaneously receive two strongly overlapping channels using spatial diversity. The second one combines the double IQ structure with the code-multiplexing one to propose an LTE-Advanced receiver with a lower consumption than classical stack-ups.

Therefore, a designer of a multi-* front-end must consider the global problem of both the analog and digital part, but also the possibility of combining or sharing components in the analog domain if simple digital functions could help to retrieve a complete information. This chapter has shown various examples of simple or more complex multi-* structures, each time taking into account the performance due to the digital part.

Double IQ structure is a very smart way of designing a multiband receiver, and its combination with a code multiplexing multi-antenna approach leads to an important decrease of the energy consumption of the system. Certainly this kind of multi-* system will play a very important role in future wireless systems.

More generally, these examples have shown that a complete optimization of complex systems with numerous degrees of freedom can no longer rely on separate studies for the analog and the digital part. With this, the main challenges to be addressed include enhancing the use of the radio spectrum, reducing the energy consumption or the component cost, and enabling a large flexibility of the transceiver. One interesting way to reduce the energy consumption in flexible terminals is to try not only to share the components, but also to find some architecture enabling some low-energy sleeping modes. Reducing the number of components to keep continuously on, or operating with lower characteristics (low gain or narrow band modes for instance), can result in more efficient and flexible terminals.

References

- [1] D Evans, D Raynes and A Payne, "Development and Simulation of a Multi-standard MIMO Transceiver," IST-2001-32125 FLOWS Deliverable no. D20, November 2004.
- [2] S-G Lee, N.-S. Kim, S.-M. Oh, J.-K. Choi, and S.-C. Kim, "A Dual-Band Receiver Architecture For PCS and IMT-2000," Proc. 2nd IEEE Asia-Pacific Conference on ASICs, pp. 235–238, Aug. 2000.
- [3] B. McFarland, A. Shor and A. Tabatabaei, "A 2.4 & 5 GHz Dual Band 802.11 WLAN Supporting Data Rates to 108 MB/s," IEEE Gallium Arsenide Integrate Circuit Symposium, pp. 1–14, Oct. 2002.
- [4] D. Belot, B. Bonhoure, D. Saias, and N. Bertholet, "A DCS1800/GSM900 RF to Digital fully Integrated Receiver in SiGe 0.35 μ m BiCMOS", Proc. 2001 BIPOLAR/BiCMOS Circuits and Technology Meeting, pp. 86–89, Oct. 2001.
- [5] K. Rampmeier, B. Agarwal, P. Mudge, D. Yates, and T. Robinson, "A Versatile Receiver IC Supporting WCDMA, CDMA and AMPS Cellular Handset Applications", Proc. IEEE Radio Frequency Integrated Circuits Symposium, pp. 21–24, May 2001.
- [6] C. Van Der Burgt, "Multi-band Receiver and Method Associated Therewith," U.S. Patent 7120406, June 11, 2004
- [7] H. Tsurumi and Y. Suzuki, "Broadband RF stage architecture for software-defined radio inhandheld terminal applications," *IEEE Communications Magazine*, vol. **37**, no. 2, pp. 90–95, February 1999.
- [8] E. H. Armstrong, "Super-regeneration", *QST*, p. 11, December 1922.
- [9] B. Razavi, "Design Considerations for Direct-Conversion Receivers," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal processing*, vol. **44**, pp. 428–435, June 1998.
- [10] J. W. M. Rogers, J. Macedo, and C. Plett, "A Completely Integrated 1.9 GHz Receiver Front-End with Monolithic Image Reject Filter and VCO," *IEEE Journal of Microwave Theory and Techniques*, vol. **50**, pp. 210–215, Jan. 2002.
- [11] R. G. Vaughan, N. L. Scott, and D. R. White, "The Theory of Bandpass Sampling", *IEEE Transactions on Signal Processing*, vol. **39**, pp. 1973–1984, Sept. 1991.
- [12] M. Gautier and G. Villemaud, "Low complexity antenna diversity front-end: Use of code multiplexing," IEEE Wireless Communication and Networking Conference (WCNC09), April 2009.
- [13] G. Krishnamurthy and K. G. Gard, "Integrated Time Division Multiplexing Front-end Circuit for Multi-antenna RF Receivers," IEEE Radio Frequency Integrated Circuits Symposium (RFIC 2008), June 2008.
- [14] T. Kaiser, A. Bourdoux, H. Boche, *et al.*, *Smart Antennas: State of the Art*, Euraspis Book Series on Signal Processing & Communications, 2005.
- [15] F. Tzeng, A. Jahanian, and P. Heydari, "A Universal Code-Modulated Path-Sharing Multi-Antenna Receiver," IEEE Wireless Communication and Networking Conference (WCNC08), March 2008.
- [16] I. Burciu, G. Villemaud J. Verdier, and M. Gautier, "Low Power Front-End Architecture dedicated to the Multistandard Simultaneous Reception," *International Journal of Microwaves and Wireless Technologies*, August 2010.
- [17] R. Hartley, "Modulation System," U.S. Patent 1666206, 1925.

- [18] J. C. Rudell, J.-J. Ou, T. B. Cho, *et al.*, "A 1.9GHz Wide-Band IF Double Conversion CMOS Integrated Receiver for Cordless Telephone Applications," *IEEE Journal of Solid-State Circuits*, vol. **32**, no. 12, pp. 2071–2088, December 1997.
- [19] I. Burciu, J. Verdier, and G. Villemaud, "Low Power Multistandard Simultaneous Reception Architecture," European Wireless Technology Conference, September 2009.
- [20] A. G. Armanda, "Understanding the effects of phase noise in orthogonal frequency division multiplexing," *IEEE Transactions on Broadcasting*, vol. **47**, no. 2, pp. 153–159, 2001.
- [21] K. Sathananthan, C. Tellambura, "Probability of Error Calculation of OFDM Systems with Frequency Offset," *IEEE Transactions on Communications*, vol. **49**, no. 11, pp. 1884–1888, Nov. 2001.
- [22] M. Windish and G. Fettweis, "Performance Degradation due to IQ Imbalance in Multi-carrier Direct Conversion Receiver: A Theoretical Analysis," IEEE International Conference on Communications, 2006.
- [23] I. J. Gupta, "SMI Adaptative Antenna Arrays for Weak Interference Signals," *IEEE Transactions on Antennas and Propagation*, vol. **10**, no. 1, pp. 1237–1242, 1986.
- [24] E. Seidel, "Progress on LTE Advanced" – the New 4G Standard". White Paper, Nomor Research GmbH., July 2008.
- [25] S. Parkvall *et al.*, "LTE-Advanced–Evolving LTE towards IMT-Advanced," IEEE Vehicular Technology Conference, September 2008.
- [26] I. Burciu, M. Gautier, G. Villemaud, and J. Verdier, "Candidate Architecture for MIMO LTE-Advanced Receivers with Multiple Channels Capabilities and Reduced Complexity and Cost," IEEE Radio and Wireless Symposium, Phoenix, US, Jan 2011.
- [27] I. Burciu, G. Villemaud, and J. Verdier, "Multiband Simultaneous Reception Front-End with Adaptive Mismatches Correction Algorithm," IEEE Personal, Indoor and Mobile Radio Communications Symposium, September 2009.
- [28] R. Kohno, R. Meidan, and L. B. Milstein, "Spread Spectrum Access Methods for Wireless Communications," *IEEE Communications Magazine*, vol. **33**, no. 1, pp. 58–67, Jan. 1995.
- [29] G. de la Roche, R. Rebeyrotte, K. Jaffres-Runser, and J. M. Gorce, "A QoS-based FAP Criterion for Indoor 802.11 Wireless LAN Optimization," IEEE International Conference on Communications (ICC 2006), Istanbul, Turkey, pp. 5676–5681, June 2006.
- [30] P. Mary, J. M. Gorce, G. Villemaud, M. Dohler, and M. Arndt, "Performance Analysis of Mitigated Asynchronous Spectrally-Overlapping WLAN Interference," WCNC 2007, Hong Kong.
- [31] P. Mary, J. M. Gorce, M. Dohler, G. Villemaud, and M. Arndt, "Reduced Complexity MUD-MLSE Receiver for Partially-Overlapping WLAN-Like Interference in Heterogeneous Multipath Channels," VTC Spring 2007, Dublin, Ireland, 2007.
- [32] G. Villemaud, P. F. Morlat, J. Verdier, J. M. Gorce, and M. Arndt, "Coupled Simulation-Measurements Platform for the Evaluation of Frequency-Reuse in the 2.45 GHz ISM band for Multi-mode Nodes with Multiple Antennas," *EURASIP Journal on Wireless Communications and Networking*, Dec. 2009.
- [33] I. Burciu, M. Gautier, G. Villemaud, and J. Verdier, "Method for Processing Two Signals Received by a Single Terminal," International Patent, WO/2010/031944, March 2010.
- [34] G. Fettweis, M. Lohning, D. Petrovic, M. Windish, P. Zillman, and W. Rave, "Dirty RF, a New Paradigm", *International Journal of Wireless Information Networks*, vol. **14**, No. 2, 2007.

Part V

Circuits and system integration in digital front-end

20 Integration and interface of digital front-end with analog front-end

Steven Loh

20.1 Introduction

This chapter describes the integration and interface between the digital front-end and analog front-end with a focus on wireless terminals' ASICs applications. This has been most dynamic over the past decade with rapid progress from a mostly discrete implementation platform to a single system-on-chip (SoC) ASIC that integrates the RF blocks with the Modem [1]. The pressure to perform with size, costs, and power constraints has fueled creativity in system and architecture designs and implementations in submicron CMOS processes. Such development efforts have revolutionized the types and means of digital-to-analog interfaces, which in turn have redefined ASIC partitions and popularized today's mixed signal integration on a common substrate.

Traditionally, the digital front-end resides within the Modem baseband processor ASIC and seeks a partition with an analog baseband IQ interface to the RFIC transceiver. Such direct conversion transceiver architecture offers simplicity in analog design, and provides analog channel filtering and full dynamic gain range in the analog domain. The digital front-end can perform calibration functions in addition to pre-decoding and post-encoding functions. Such a device platform continues to see wide commercial applications in 3G ASICs implementation, mWiMax wireless terminal ASICs, as well as LTE datocard ASICs design.

In narrowband cellular systems like GSM/GPRS/EDGE, transceiver designs too had adopted a similar traditional platform in the mid to late 1990s, but made the transition to more efficient mixed signal RFIC designs thereafter, motivated by the more efficient implementation of baseband analog functions in the digital domain. The advent of improved dynamic range and more power efficient data converters helped push the change. By 2002, the industry saw GSM/GPRS RFIC that integrated the ADC and part of the receiver digital front-end with the RF and analog circuits on the same substrate. The design promptly moves the channel filtering functions after the ADC and performs gain control digitally to undertake a dominant portion of the receiver dynamic range. The RFIC transmitter digital front-end on the other hand continued to reside within the Modem ASIC, and so did the RF power amplifier control interfaces like the ramping function as well as transmit/receive (T/R) switching logic controls.

As RF CMOS devices move towards lower linewidth processes, the related design is allowed to perform more analog functions in the digital domain for industrial cellular

wireless terminals ASICs. The GSM/GPRS/EDGE transmitter architectures began to see more elegant design transformation with the entire digital front-end including the QPSK and 8PSK modulator finding a place within the RFIC to ensure a more efficient interface with the analog blocks. This need was ushered in by the onset of the polar loop design, first ambitiously embracing the power amplifier (PA) in the hope of achieving significant power savings in the EDGE transmitter [2]. However, PA efficiency improvements were restricted by limitations in power management designs. As a result, the polar loop technique began to find applications within the RFIC, while reverting to just the use of a linear PA. The most common internal polar loop architecture seen in commercial GSM/GPRS/EDGE RFICs today closes the loop at the mixer while phase modulation occurs at the voltage control oscillator (VCO) that provides the local oscillator (LO) to the mixer [3].

Up until recently, the semiconductor industry providing RFIC transceivers for 3G mobile terminals was still deeply entrenched in the all analog direct conversion architecture adopting a conventional analog quadrature baseband interface with the Modem DSP. This traditional design offered straightforward implementation with the fewest number of blocks. In the relatively relaxed 3G test requirements, it is no surprise that the all analog approach has become a lucrative and conservative preference. This was especially true given that power and Silicon efficient analog-to-digital converters (ADC) and digital-to-analog converters (DAC) did not appear until recent years. Also, the wireless industry remained in higher linewidths CMOS of 0.18 μm and 0.13 μm for quite a while, unconvinced that deep submicron CMOS is capable of effectively providing the linearity performances required for 3G in Radio Frequencies (RF).

In the imminent progression towards Broadband Wireless Access technologies for mobile terminals, radio transceiver architectures find challenges that force a departure from their traditional design. The makeup of Radio Frequency blocks to Baseband analog blocks that stacks up the entire variable gain function and the essential channel filtering function in an all analog direct down-conversion receiver is no longer capable of efficiently providing front-end processing needs to meet performance requirements. New Standards in Long Term Evolution (LTE) [4] expose the limitations of analog domain processing, which demands reference sensitivity performance over a 30 dB dynamic range as well as close-in rejection capabilities when supporting desired signals with Baseband bandwidth of up to 10MHz. Similarly, an all analog direct conversion transmitter finds the tight distortion limits in support of high-order modulation demands complex compensation mechanisms.

The following section presents several transceiver mixed signal architectures with various analog-to-digital interfaces that have been popularized in today's wireless terminals ASICs design. [Section 20.3](#) will address the system aspects of integration and interfaces in detail. [Section 20.4](#) examines the value of the designs and development work currently in progress, and explores the future directions of these designs.

20.2 Mixed signal transceiver architectures

There are several variants of mixed signal transceiver architectures that are in use commercially today. This section will present them and address the particular

characteristics of each. Since there are distinct differences in design between the receiver and the transmitter, the two shall be presented here separately, first the receiver and then the transmitter.

20.2.1 Receiver design for TDD narrowband GSM/GPRS/EDGE

There are two novel approaches for designing GSM/GPRS/EDGE (GGE) receiver architectures; a direct down-conversion approach, and a Low or Near IF approach [1], [5]. The earlier architecture is passed down from the long-adopted legacy design, except that the digital front-end has taken up some of the analog functions, like channel filtering and dynamic gain range. Figure 20.1 shows a block flow example of such a receiver. The analog-to-digital converter (ADC) dynamic range has been increased to accommodate this system.

Such architecture is a plus in simplicity, but has to contend with DC calibration. It is a narrowband system that adds to its limitations. The system aspects and DC calibration challenges will be addressed in more detail in Section 20.3.

Another common implementation for GGE receivers is the near zero intermediate frequency (NZIF), where instead of direct conversion to baseband, the signal is down-converted to a pre-determined low intermediate frequency. In the GSM/GPRS only receiver, 100 kHz intermediate frequency (IF) is adopted [5]. In order to accommodate EDGE, IF frequency needs to be increased to at least 110 kHz or higher but upper bounded by the challenge of alternate channel interferer rejection [11]. Figure 20.2 illustrates the block flow diagram of a NZIF GGE receiver [13]. The ADC too needs to have an increased dynamic range to accommodate the system requirements of such a receiver.

A NZIF receiver effectively avoids the DC issue by placing the signals where they no longer overlap. Section 20.3 will describe how the DC component can be conveniently

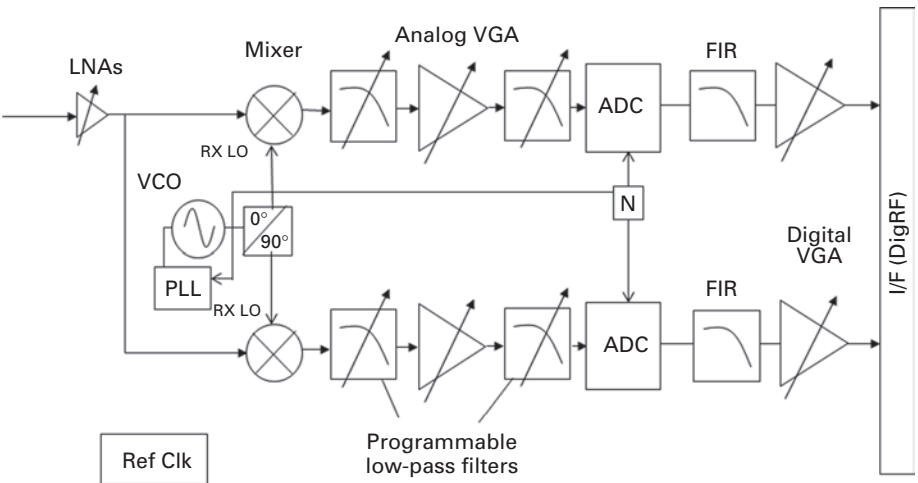


Figure 20.1 GGE direct down-conversion receiver.

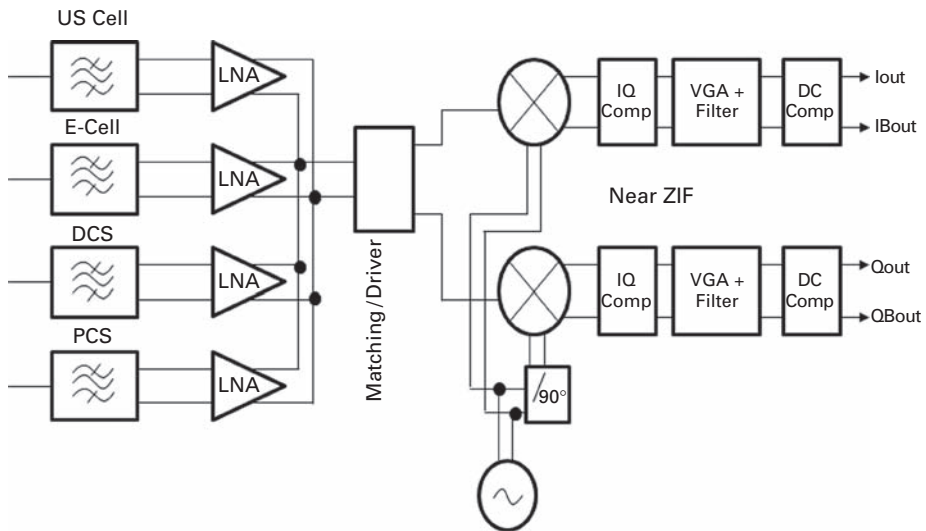


Figure 20.2 Block diagram of GGE NZIF receiver adapted from [13].

removed in the digital front-end. However, such architecture invites another set of challenges in close-in interferer suppression. Unlike direct down-conversion, adjacent and alternate channel interferers will overlap with the desired signal at IF. The system will rely on image rejection to achieve the required signal-to-noise-distortion ratio (SNDR), which will be addressed in [Section 20.3](#). A digital down-converter (DDC) is required to complete the frequency conversion to baseband.

Both GGE receiver architectures are seeing implementation in single Baseband-RF SoC ASICs today in 90 nm or lower linewidth CMOS processes.

20.2.2 Receiver design for FDD LTE/HSPA/WCDMA

Unlike narrowband systems, wideband or broadband receivers can only be effectively implemented in direct conversion architecture. Furthermore, as signal bandwidth varies and expands like in LTE, the incentive is to move channel filtering and dynamic gain range functions into the digital domain to yield an efficient architecture. The general architectural block flow of the digital baseband is shown in [Figure 20.3](#) [11].

The CIC filter block in [Figure 20.3](#) is commonly used to decimate the higher sample rates from the ADC to lower sample rates to reduce digital baseband power consumption. The DC offset compensation corrects for the DC component from the ADC output. Droop compensation can easily be implemented in the digital baseband to equalize amplitude variations introduced by the analog baseband filter and the CIC filter. IQ compensation is especially important for NZIF, which maximizes image rejection on images of adjacent and alternate interferers overlapping into the desired GGE signal. The digital mixer or digital down-converter (DDC) converts the signal from NZIF to baseband; FIR1 provides the digital channel filtering that will also further reject the DC residue that is upconverted to NZIF by the DDC. The digital VGA provides the

Table 20.1 LTE signal Characteristics.

Modulation RF Bandwidths	QPSK	16QAM	64QAM
1.4 MHz	X	X	X
3 MHz	X	X	X
5 MHz	X	X	X
10 MHz	X	X	X
15 MHz	X	X	X
20 MHz	X	X	X

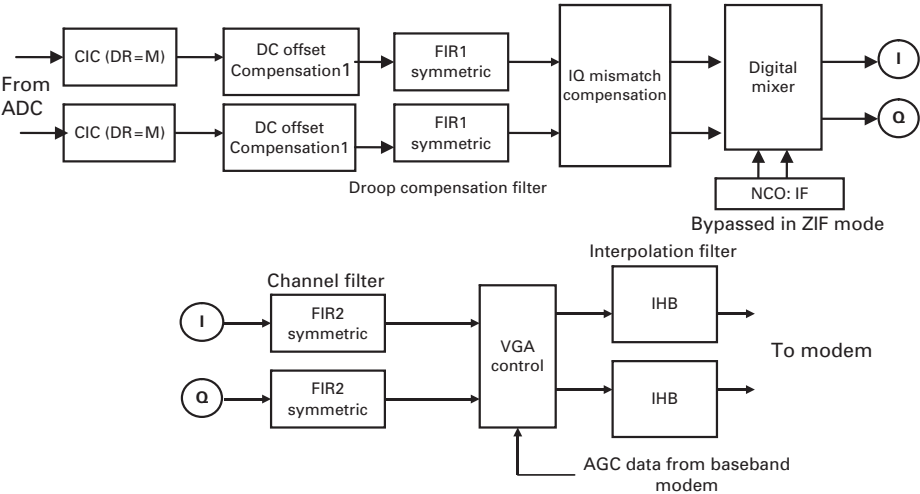


Figure 20.3 Digital baseband architecture for ZIF and NZIF adapted from [11].

remaining bulk of the receiver gain dynamic range. Interpolation may be required to interface with the Modem, which can be achieved using an IHB filter.

The variants in such system implementations find compromises between analog baseband design and ADC design requirements based on area and power targets. Other architectural variations include clocking structure implementations. While there are advantages in using a fixed clock for the ADC and digital front-end, a mixed signal receiver brings about challenges in keeping the clock spurious out of the assigned channel that may favor a variable clock. Details of advantages and disadvantages for the various implementations will be provided in Section 20.3.

There are clear motivations for a common receiver to support HSPA/WCDMA on top of LTE requirements to operate with the various bandwidths and orders of modulation as tabulated in Table 20.1.

A higher order of modulation and bandwidth imposes tighter constraints on receiver Error Vector Magnitude (EVM) requirements, creating more need for calibration to counter analog variations over process, voltages, and temperature (PVT). Section 20.3 will address the various calibrations commonly implemented.

20.2.3 Transmitter design for TDD narrowband GSM/GPRS/EDGE

The GSM/GPRS/EDGE (GGE) transmitters have rapidly gone through several phases of evolution, from the direct conversion architecture to a polar loop architecture that is popularly adopted in commercial ASICs today. Figure 20.4 [3] shows the block flow example of polar loop architecture. It is architecture with the internal polar loop at the up-converter. There have been several variations of internal polar loop architectures developed commercially, and Figure 20.5 [6] shows one developed by Texas Instruments. This architecture has the polar loop at the digital pre-PA driver. Both of these architectures interface with linear power amplifiers (PA).

The modulated signal from the Modem is split into its phase and amplitude content prior to up-conversion, and recombined at RF frequency. The motivation is to exploit the accuracy of phase modulation in the digital domain, and to maintain as much

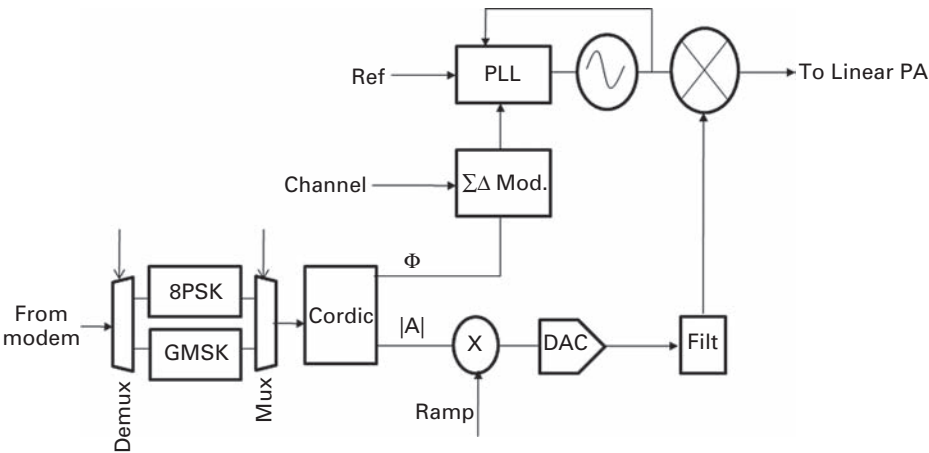


Figure 20.4 Block diagram of polar modulation transmitter adapted from [3].

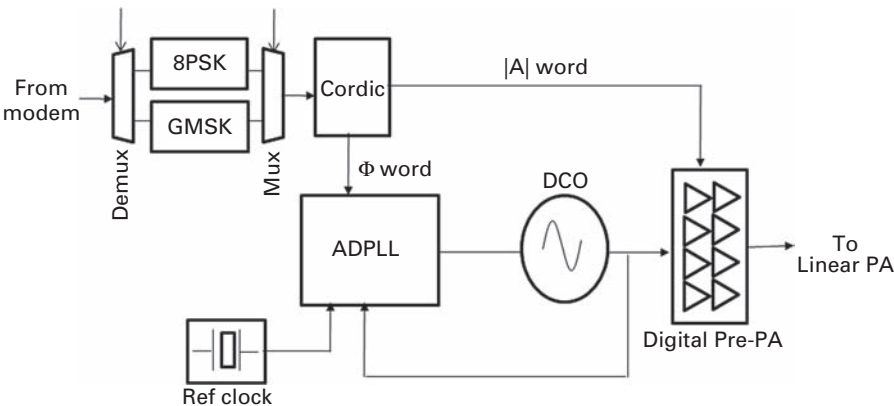


Figure 20.5 All-digital polar transmitter adapted from [6].

commonality as possible between GSM and EDGE. During initial implementations, when the GGE platform was still a two-chip solution, the transmitter digital front-end was moved into the RFIC to better interface with the RF blocks while adopting a standardized DigRF serial data interface with the Modem. Then in 2006, Texas Instruments announced the industry’s first single System-on-Chip (SoC) ASIC for its GSM/GPRS solution. Ever since there was no looking back, almost all commercial GGE terminal ASICs today are single SoC devices to leverage size and cost implications.

The narrowband nature of GGE and the more relaxed EVM requirement of EDGE have afforded the polar loop approach, which carries other challenges that will be described in [Section 20.3](#). The time division duplex mode of operation reduced isolation challenges between digital noise and RF, and facilitated single SoC ASIC implementation. Unlike the GGE receiver, the transmitter architecture is unique and mature, and designers lack the incentive to have a common transmitter to support GGE and wider bandwidth signals that would dictate the use of direct conversion architectures. This explains why the GGE transmitter commonly finds itself in parallel with a direct conversion transmitter in an ASIC that supports wideband standards as well.

20.2.4 Transmitter design for FDD LTE/HSPA/WCDMA

The most simplistic and proven architecture for the 3G/3.5G terminal transmitter is the direct conversion architecture [7]. In the 4G transmitter, there are continued incentives to maintain such architecture due to its ease in design implementation. [Figure 20.6](#) shows a block flow diagram of an analog direct conversion transmitter.

While the architecture easily suffices to support 3G/3.5G signal modulation requirements, it becomes more challenged in broadband and higher-order modulation requirements that is commonly determined by a performance index called the error vector magnitude (EVM). The issues will be detailed in [Section 20.3](#).

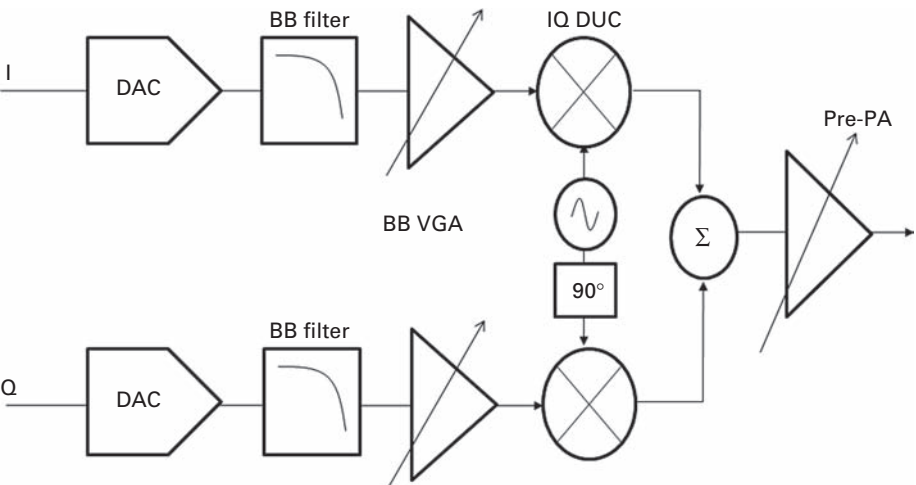


Figure 20.6 Conventional analog direct conversion Transmitter.

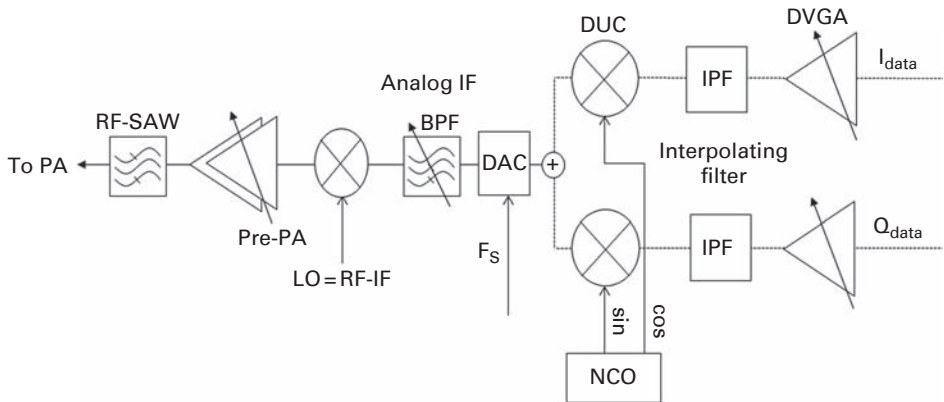


Figure 20.7 Heterodyne transmitter with digital IF modulator adapted from [12].

Another architecture implementation that is geared towards signal modulation quality performance is the digital-IF architecture without the need for an external IF filter, which is particularly useful for broadband, high-order modulation operations. There are several variants to such architecture. Figure 20.7 [12] depicts the block flow of a conventional digital-IF, where the use of an external inter-stage filter is assumed when supporting FDD modes.

The key to this architecture is to offset the desired spectrum from the transmitter RF local oscillator (LO). This creates two desirable properties: any image created due to IQ imbalance now falls outside the desired signal, and any LO feed through also falls outside the signal spectrum, and as such no longer contributes to EVM degradation. There are also challenges with this architecture design, which will be described in Section 20.3.

20.3 Digital–analog system integration and interfaces

An efficient system integration and interface between the digital front-end and the RF-analog is one of the keys in architecture selection and design. This section describes the distinctive digital–analog system features of each of the architectures addressed in the previous section. An analysis is performed on the advantages and disadvantages of each type of system integration and interface related to the architecture that adopted it.

20.3.1 GSM/GPRS/EDGE receiver analog–digital systems

20.3.1.1 Direct down-conversion

A direct down-conversion receiver offers simplicity in design, frequency plan, and has the least challenges among other architectures for baseband analog filter design requirements. The downconverted baseband signal allows the baseband low pass filter corner to be designed for the minimum half bandwidth, while providing required rejections to interferers at pre-determined offsets. As such, it offers the best ratio of rejection offsets

to passband bandwidth. The filtering of interferers is necessary to avoid saturation of the ADC, and at the same time to provide anti-aliasing.

The advent of power-efficient, high dynamic range ADCs can offer the headroom to deal with close-in interferers for digital processing. This does two things: it relaxes the corner of the analog baseband filter by relegating its functions to only suppressing further out interferers, and as an anti-alias filter it facilitates digital channel filtering. Effectively, the improvement in ADC technology is able to move the burden of baseband filtering from the analog domain to the digital domain. In addition, the increased ADC dynamic range has allowed part of the dynamic gain range to be undertaken in the digital domain as well. The lower analog gain range reduces receiver linearity requirements, and reduces residual DC amplification seen at the ADC input.

The key downside of a direct down-conversion narrowband receiver is its limited ability to remove the DC component at baseband. The DC at the receiver baseband is created predominantly by LO self mixing due to LO leakage into the mixer input, and to a smaller degree by second-order product contributions due to strong blockers. Due to the narrowband nature of the desired signal as in GSM/GPRS/EDGE, it is impractical to implement a capacitive DC block on the DC component at baseband without severely degrading the signal. An active DC compensation technique is commonly adopted to resolve this issue. Since all DC compensation algorithms carry high pass response characteristics and will affect desired signal quality, DC correction is traditionally activated prior to the “receiver on” slot and shut off when the receiver is fully turned on. This DC preset is only possible for a TDD receiver that is not in continuous time operation. However, such a receiver would still need to build in a design budget to accommodate dynamic DC contributions during receiver operation.

Besides DC, the direct down-conversion receiver also has the signal image superimposed upon itself at baseband contributing to SNR degradation. This is however a lesser problem as the quadrature down-conversion architecture provides an inherent image rejection property that is easily sufficient for narrowband system requirements like those of GGE. Some degree of IQ mismatch compensation is common to further improve performance.

20.3.1.2 Near ZIF down-conversion

As late as the mid to late 1990s, super heterodyne receivers using Intermediate Frequency (IF) of 225 MHz still saw commercial implementations in GSM wireless terminals worldwide. When the industry moved overwhelmingly to direct conversion, little thought was given to non-zero IF designs. As the ASIC design community begins to confront the issue of effective receiver DC removal as well as the impact of $1/f$ noise, studies were made into Near ZIF systems to avoid the use of external IF filters as with earlier receivers using higher IF.

In Near ZIF systems, the desired signal is commonly downconverted to 100 kHz for a GSM/GPRS receiver and at least 110 kHz or higher for a receiver that supports EDGE. Further down-conversion to baseband is necessary prior to the Modem, and it is accomplished using a digital down-converter (DDC) after the ADC. The key to this architecture is for the occupied bandwidth of the desired signal to be away from the DC, so that when the

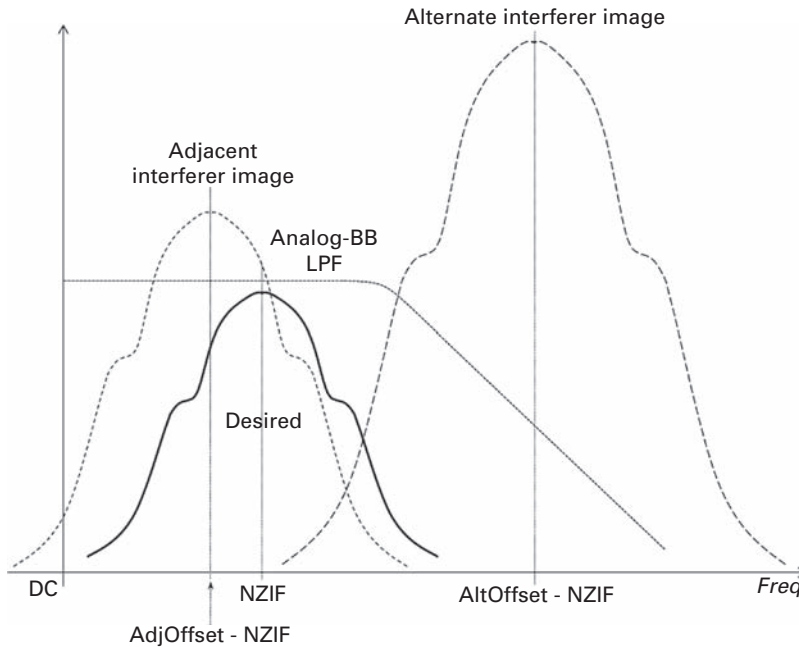


Figure 20.8 Quadrature digital-IF $\Sigma\Delta$ RF modulator.

signal is down converted to baseband in the digital domain the DC component will in turn be up converted to Near ZIF frequency to allow for digital filtering. Such a system only requires the ADC to account for the headroom to accommodate the analog DC component.

The price to pay for Near ZIF architecture is to deal with the images of adjacent or alternate channel interferers that would overlap within the desired occupied bandwidth. Quadrature image rejection is necessary in the Near ZIF system, and can be accomplished by adopting either the Hartley or Weaver architectures [14]. In Near-ZIF, IQ mismatch compensation becomes critical, and the digital front-end will provide IQ mismatch compensation to meet image rejection requirements of at least 35dB. The IQ amplitude mismatch originates primarily from the analog I and Q gain variations. The IQ phase mismatch originates primarily from I and Q LO phase variations. Since most of the mismatches are static, the digital front-end can be pre-calibrated to offset these mismatches. Figure 20.8 depicts the issues described in this subsection.

20.3.2 LTE/HSPA/WCDMA receiver Analog–Digital systems

The goal is to achieve a common receiver that is capable of supporting the performance requirements of LTE in Full-Duplex mode, while effectively operating in HSPA/WCDMA modes. The figure of merit is to attain the required performance margin while minimizing silicon area usage and power consumption. The challenge comes from LTE with a variety of signal characteristics, where the receiver sensitivity will vary according to the signal bandwidth and the order of modulation.

A QPSK modulated signal with Rate Coding at 1/3 only needs a Signal-to-Noise Ratio (SNR) of 0 dB at the Modem to process. If the receiver chain has a typical Noise Figure (NF) of 6 dB at maximum gain, then a 1.4 MHz bandwidth similarly modulated signal will need to be at a level of -106 dBm at the receiver input to be processed correctly. In reality, the effective bandwidth for LTE is smaller than the channel bandwidth and therefore will provide some additional margin. A static 64QAM modulated signal with Rate Coding of 5/6 will require an SNR performance of over 20 dB at the Modem. The same receiver processing this signal with 20 MHz bandwidth will require a minimum level of -75 dBm at its input carrying a similar margin due to the smaller effective bandwidth. This represents a dynamic range of over 30 dB, where the receiver is required to maintain the minimum typical Noise Figure performance of 6 dB.

This will be challenging to achieve using the traditional legacy systems where the full analog gain dynamic range is needed to maintain a single optimum level at the ADC input of the Modem Baseband device. Figure 20.9 shows a block diagram of a legacy receiver system of a RFIC. As the analog variable gain block is lowered to accommodate the higher input level of a 64QAM 20 MHz signal, its Noise Figure (NF) will escalate resulting in an increase in the overall receiver NF exacerbated by a lesser LNA-Mixer gain due to linearity considerations without interstage filter. This will reduce the receiver performance when operating at 64QAM 20 MHz bandwidth compares with lower bandwidths.

A bigger challenge for such legacy systems is the need to suppress close-in blockers using analog filters while supporting wide bandwidth signals. The LTE Narrowband Blocking requirement maintains the need to process a signal up to 10 MHz baseband bandwidth with a tone 200 kHz away from its edge. It has become a common understanding that the tone cannot be suppressed enough in the analog domain. A quick fix

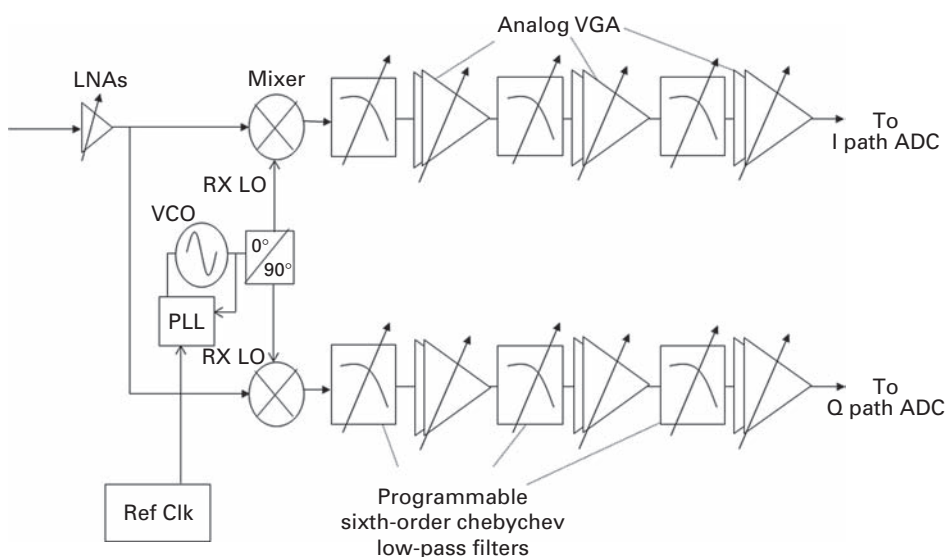


Figure 20.9 Legacy direct conversion receiver.

adopted in the analog legacy system to support LTE uses an interferer detector in the digital front-end to trigger a lower analog gain at the expense of SNR to avoid saturating the ADC when encountering such blockers. Otherwise, an unacceptable amount of power would have to be expended at the ADC for the headroom needed to avoid saturation. This quick fix method however, while sufficing in a laboratory or pilot type testing environment, can potentially become problematic in commercial deployments. The tone in this test case is to simulate a narrowband interferer like that of a GSM signal, which is TDD in nature. Likely to be asynchronous with the LTE system, such a time slotted interferer imposes severe complexity on the detector system in order to be robust. For example, there is no certainty that after detecting the tone in the timeslot prior, the tone will still be present during the current timeslot with the analog gain lowered, and vice versa. The detector system needs to determine when to switch states.

Therefore, while the legacy receiver system supported HSPA/WCDMA modes well with smaller and fixed bandwidth, they are clearly not the most efficient to operate in LTE. The receiver system needs to be redesigned in order to accommodate LTE. A modified receiver system to circumvent the issues addressed also adopts a Zero-IF but splits the variable gain function between the analog and digital domain with most of the dynamic range undertaken by the digital variable gain amplifier as shown in Figure 20.10. System wise, it allows the signal arriving at the ADC to be at an arbitrary level instead of keeping it constant as long as the SNR is sufficient to process the order of modulation. The digital variable gain block will be responsible for feeding a constant level needed for Modem processing. Using this technique, the analog gain can be kept at maximum setting and thereby maintain the optimum NF so as to maximize receiver

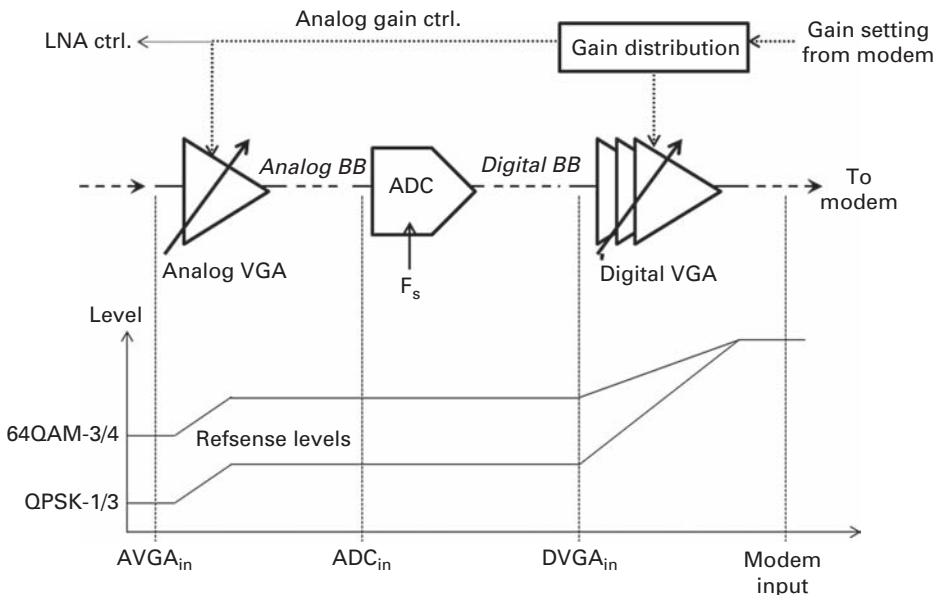


Figure 20.10 Gain partitioning and setting in mixed-signal receiver.

sensitivity for all signal characteristics in Table 20.1. The key is that in the digital domain, variable gain setting has no impact on noise unlike its analog counterpart.

It is also at this arbitrary ADC input level that headroom is designed into the system stack up to accommodate close-in blockers like the Narrowband Blocking tone. The motivation is to let close-in interferers through for rejection by digital filters rather than analog filters. Such a design not only alleviates the need for high-order analog filters, but also allows a comparable ADC to accommodate interferers without expending additional power. Figure 20.11 explains the concept of such a system design that makes full use of the ADC dynamic range to meet all LTE Standards test requirements. Functions performed in the digital domain are not susceptible to process, voltage, and temperature (PVT) variations. This can also be accomplished with much less power and lower Silicon area, and more so as lower linewidth processes are adopted.

The ADC clock may be derived from the RF LO. It not only has low jitter characteristics, but clock harmonics that fall within the assigned bandwidth will coincide with the LO frequency therefore down converted to DC, which is then calibrated out. The DC compensation in a mixed signal receiver is commonly achieved in both the analog and digital domain. Static DC contributions are predominantly from LO self mixing, and may be easily calibrated during pre-start. Dynamic DC transients are mainly from baseband gain changes during operation, and to a smaller degree by blockers in a high IP2 receiver. Since gain change is an instructed function from the Modem, a priori correction of the effected dynamic DC transient is possible.

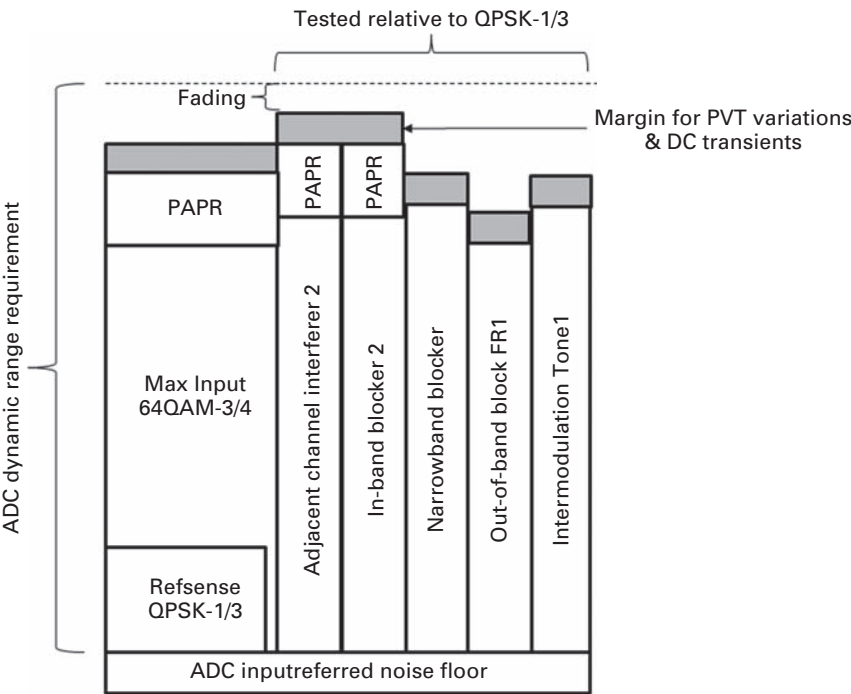


Figure 20.11 Optimization of ADC dynamic range.

In addition, the programmability of a digital baseband affords equalization functionality for the baseband analog blocks in terms of amplitude ripples that is a common characteristic of analog filters. A digital domain FIR filter can be easily designed to compensate for the droops. What in fact is happening is that the digital baseband block is taking the signal from the analog baseband and improving its EVM.

20.3.3 GSM/GPRS/EDGE transmitter analog–digital systems

Most GGE RF today is implemented with the Modem baseband on a single SoC substrate. A slightly earlier solution used a two-chip solution with a DigRF interface between the RFIC and Modem ASIC. Regardless, the GMSK and 8PSK modulator has long resided with the transmitter RF/Analog circuits on a single SoC. Such an implementation is favorable for the polar loop where the modulators now feed directly into the Cordic to separate the phase and amplitude component of the signal. The phase component then feeds into the digital PLL to phase modulate the TX RF VCO, where the amplitude components in the case for EDGE can either be routed into the mixer input port for the architecture shown in [Figure 20.4](#) [3], or to a digital Pre-PA for the transmitter shown in [Figure 20.5](#) [6].

The architecture described offers the most efficient transmitter design in terms of minimizing the number of analog blocks as well as power consumption. It is designed to exploit the advantages of low linewidth CMOS, and in the case of the architecture in [Figure 20.5](#) having claim to be all digital. The long obstacle of the polar loop transmitter architecture is the difficulty in time aligning the phase and amplitude path until they recombine at an RF block. Clearly, routing the path off chip always posts a bigger challenge in time alignment as in the case of a Polar PA transmitter. Fortunately, the narrowband signal with a relaxed EVM requirement affords the time alignment tolerance for an open polar loop PA transmitter. When the advantage in power added efficiency (PAE) is not realized, the de facto course of action is for an internal polar loop design. On-chip designs implemented mostly in digital further eases polar loop implementations, boosting enough confidence for some designers to venture with wider bandwidth signals like WCDMA.

Polar loop time alignment increases in difficulty as signal bandwidth grows, explaining why there has been little success to date for WCDMA applications or beyond. As EVM requirements tighten for higher-order modulation schemes, the time alignment tolerance shrinks making it impractical to implement. In the case of GGE systems adopting the internal polar loop architecture, a linear PA with the analog voltage ramp interface is still required.

20.3.4 LTE/HSPA+/WCDMA transmitter analog–digital systems

20.3.4.1 Direct conversion

Direct conversion architectures continue to be most popularly adopted even for broadband transmitters. Recent designs moved some of the baseband analog functions into the digital domain resulting in a mixed signal RFIC transmitter and thereby prompting a

digital interface with the Modem. One of the drivers for LTE systems is for higher resolution variable gain control for the transmitter, motivating designers to adopt a digital variable gain amplifier instead. This arrangement moves the DAC onto the RFIC, which for wideband modes still comes in two-chip platform solutions. As such, gain partitioning is between the RF variable gain block commonly known as the Pre-PA and the digital variable gain block. The dynamic range of the digital variable gain block is determined by the DAC dynamic range. The transmitter gain loop setting for the required output level needs to take this into account. Gain control is traditionally provided by the Modem via the Serial Port Interface (SPI) when the DAC still resides in the Modem. In the case of a mixed signal transmitter, the DigRF v4 data interface is adopted which also acts as a control interface including gain settings.

A mixed signal transmitter is described above with direct conversion architecture. Although direct up-conversion offers simplicity in frequency planning and analog design, as well as power savings, there are some distinct disadvantages. The pulling effect is inherent for all direct conversion architecture having the LO frequency coinciding with the assigned RF output frequency of the transmitter. RFIC designers have long struggled to contain this issue through various isolation techniques. Synthesizer pulling causes frequency error and contributes to EVM degradation. Another LO-related anomaly is its own feedthrough as an unwanted spurious within the desired spectrum having the same assigned frequency by virtue of direct conversion, degrading the EVM. Feedthrough is commonly caused by residual DC component at the up-converter inputs, and to a lesser degree with the LO to RF isolation. This anomaly can be calibrated out by appropriately setting the offsets at the I and Q input ports of the up-converter. However, at lower output level conditions where the feedthrough needs to be reduced toward the noise floor, closed loop calibration is required over PVT variations. The calibration algorithm commonly resides in the Modem, where the RF output is fed back and processed for the level of feedthrough.

Mismatch distortions can occur at I and Q paths of the transmitter. Quadrature mismatch is commonly characterized by the differences in amplitude and phase between the two paths. Gain differences between I and Q paths contribute to amplitude mismatch, while phase mismatch between I and Q usually comes from LO quadrature errors. The distortion shows up as the level of the spectrum's own image that occupies the same bandwidth by virtue of direct up-conversion. The greater the mismatch, the higher the image that degrades the EVM. Again close loop calibration is required to optimize performance and may share the same feedback path for LO feedthrough calibration.

20.3.4.2 Digital IF system

The shortcomings of direct conversion addressed in the previous subsection becomes the motivation for pursuing digital IF architecture designs. An intermediate frequency moves the RF LO frequency away from the assigned RF channel. The immediate findings are obvious, synthesizer pulling, LO feedthrough, and IQ mismatches become non issues. The operating condition for the first phenomena is no longer valid, while LO feedthrough and IQ mismatches are minimized by digitally implementing IF up-conversion. By design, digital circuits do not yield mismatches and DC spurious. That said, any residual

product arising from the DAC and analog IF blocks no longer coincide with the desired spectrum, and readily filters out. As a result, digital IF transmitters are able to offer superior EVM performance without any close loop calibration.

Like all designs, digital IF architectures have disadvantages. The main disadvantages come from design challenges of the DAC and IF analog filter. Similar to the super heterodyne architecture, digital IF also requires complex frequency planning, particularly to avoid desensitizing the receiver in FDD-type systems. Although the digital frequency shifter within the digital IF block offers programmability and therefore flexibility in IF selection, the larger bandwidth signal add challenges to this problem. Figure 20.7 [12] shows the use of an RF inter-stage SAW filter, which helps alleviate the issue. However, current industrial design trends seek to remove the use of such filters, and in that the digital IF design has proven deficient compared with direct conversion. Also, it should be well understood that a high rate DAC with an IF output will consume significantly more power than its counterpart processing at baseband. Additionally, an analog baseband low-pass filter is far easier to achieve than an analog IF bandpass filter.

20.4 Conclusions and discussions

This chapter has presented the conception and reasons for a mixed signal cellular ASIC transceiver architecture that redefined integration and interface methodologies. It shows how the aggressive cellular ASICs' business needs of minimizing costs, size, and optimizing power efficiency challenged the new design direction of such devices. Development moves progressively toward lower linewidths, higher integration, more digital functions, dictating mixed signal design techniques with new partitioning and interfaces. At the same time, the advent of broadband access technologies to improve spectrum efficiency yielded new transceiver architectures to deliver performance and to do so efficiently.

Mixed signal transceiver architectures are the current trend in cellular wireless terminal ASICs. In that, the traditional RF-analog functions are now partially undertaken by the digital domain. This design direction is consistent with progressively lower linewidth interests for the Modem blocks and digital front-end. The benefits are smaller Silicon area and lower power consumption. In addition, digital implementations are immune to PVT variations facilitating tighter design requirements.

Current development work furthers the direction of mixed signal transceiver implementations, where it moves more analog functions into digital processing like the "Quadrature digital IF" transmitter proposed in [8]. As the data converter improves in performance and power efficiency, more transceiver functions will move into the digital domain. Ultimately, ASIC implementations will follow the path of current narrowband systems, where the transceiver becomes all or mostly digital [9], and pursues integration with the Modem as a single-chip SoC [10].

In the meantime, ASICs are not quite there as single-chip SoCs to support wider bandwidth signals. Current Modem interfaces favor DigRF standards, while some may choose to stay with the IQ analog interface. There is the likelihood that ASIC partition

may move towards RF to remove the high data rate interface of DigRF to support broadband applications. This will further reduce transceiver power requirements.

Other techniques that will ultimately be adopted are currently in research like interference detection, Digital Pre-Distortion (DPD), and RF digital processing. Such methodologies are implemented using digital processing to either alleviate RF anomalies or to reduce the power consumption of the analog blocks. These being shorter-term solutions where digital implementations facilitate extensive reconfigurability and leads eventually towards a universal ASIC solution.

References

- [1] Seippel, D., Hammes, M., Kissing, J., Nicola, P., Vannier, C. (2007). Single Chip Cellular Radios for GSM, GPRS, EDGE, IEEE RFIC Symposium, 9–12.
- [2] Sowlati, T., Rozenblit, D., Pallela R., *et al.* (2005). Polar Loop Transmitter for GSM/GPRS/EDGE, IEEE RFIC Symposium, 13–16.
- [3] Muenker, C., Klepser, B-U, Neurauter, B., and Mayer, C. (2005). Digital RF CMOS Transceivers for GPRS and EDGE, IEEE RFIC Symposium, 265–268.
- [4] 3GPP™ (2010). TSG RAN: E-UTRA User Equipment (UE) Radio Transmission and Reception (Release 9), 45–53.
- [5] Yanduru, N., Griffith, D., Bhagavatheeswaran, S., *et al.* (2006). A WCDMA, GSM/GPRS/EDGE Receiver Front End without Interstage SAW Filter, IEEE RFIC Symposium, 19–22.
- [6] Staszewski, R. B., Wallberg J. L., Hung C. M., *et al.* (2005). All-Digital PLL and Transmitter for Mobile Phones, *IEEE Journal of Solid-State Circuits*, vol. **40**, no. 12, 2469–2482.
- [7] Chiu, C. S., Heng, B. S., Khoo, E. S., *et al.* (2009). A Multi-band High Performance Single-Chip Transceiver for WCDMA/HSDPA, IEEE RFIC Symposium, 201–204.
- [8] Jerng, A. and Sodini, C. G. (2007). A Wideband SD Digital-RF Modulator for High Data Rate Transmitter, *IEEE Journal of Solid-State Circuits*, vol. **42**, no. 8, 1710–1722.
- [9] Parikh, V. K., Balsara, P. T., and Eliezer, O. E. (2008). A Fully Digital Architecture for Wideband Wireless Transmitters, IEEE RWS, 147–150.
- [10] Seippel, D., Hammes, M., Hanke, A., and Kissing, J. (2006). GSM/GPRS Single Chip in 130nm CMOS: Challenges on RF for SoC Integration., IEEE RFIC Symposium, 203–206.
- [11] Samsung Electronics, Inc. (2009). Versatile System for Multimode, Wireless Communication Receiver with ZIF and Near-ZIF Operations, USPTO issued patent 7,558,550. 2009-07-07.
- [12] Leung, V. W., Larson, L. E., and Gudem, P. (2003). An Improved Digital-IF Transmitter Architecture for Highly-Integrated W-CDMA Mobile Terminals, *IEEE Vehicular Tech. Conf.*, vol. **2**, 1335–1339
- [13] Hietala, A. W., Humphreys, S. R., Arkizewski, R., and Morris, S. (2008) Self-shielded Quad-band EGPRS Transceiver, IEEE RFIC Symposium, 9–12.
- [14] Der, L. and Razavi, B. (2003) A 2 GHz CMOS Image-Reject Receiver With LMS Calibration, *IEEE Journal of Solid-State Circuits*, vol. **38**, no. 2, 167–175.

21 Circuits and systems for digital front-ends to support multiple wireless standards

Imtinan Elahi and Khurram Muhammad

21.1 Introduction

While the migration of RF circuits running at multi-gigahertz frequencies to low-cost deep-submicron CMOS process has made single-chip wireless transceivers a reality, the proliferation of multiple wireless standards has imposed new challenges of integrating multiple wireless standards on the same hardware. One of the challenges in multi-standard radios is the need to support multiple frequency bands ranging from less than 1 GHz to more than 5 GHz. In this chapter, we will discuss the design of digital front-end (DFE) circuits that are applicable to at least two wireless standards: these include GSM/GPRS/EDGE (GGE) and WCDMA. We will also show how to extend the design to support LTE. Restricting the coverage to only three wireless standards is primarily for the amount of analysis that needs to be performed for each standard.

In the context of a receiver, we will use digital front-end (DFE) to describe digital circuits that bridge the output of an A/D converter (ADC) to a digital base-band (DBB) processor. We will also restrict discussion in this chapter to the case of implementation-friendly direct-conversion receiver (DCR) architecture. This will assume zero intermediate frequency (ZIF) for WCDMA and LTE signals but will allow for very low IF (VLIF) for GGE.

Figure 21.1 shows a very simplified, unidirectional, block diagram of a receiver, where ARX stands for analog receiver and it includes all RF/analog circuits, including the ADC, between antenna and the DFE. This simple figure clearly demonstrates that DFE design needs to meet certain requirements on both of its ends. On the ARX side, it receives N_1 -bit data stream from an ADC at a sampling rate of fs_{in} , while at the DBB side, it has to comply to the requirements of N_2 -bit data at a sampling rate of fs_{out} ; N_1 is the number of ADC bits, which depends on the type of ADC and it can be as low as 1-bit for sigma-delta ADCs; fs_{in} is the ADC sampling rate, which could either be fixed or derived from the receiver *local oscillator* (LO) clock. As will be explained later, the latter mode is preferred. At the DBB side, both N_2 and fs_{out} will depend on the type of wireless standard supported. In general, fs_{out} is an integer multiple of the supported modulation's baud-rate or chip-rate.

We can infer from Figure 21.1 that the design of DFE also depends on the filtering which takes in ARX to make sure that the combined filtering of ARX and DFE is

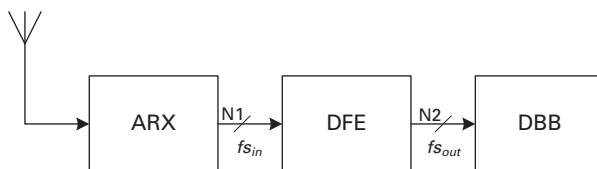


Figure 21.1 Simplified receiver block diagram showing location of the DFE.

sufficient so that the signal which enters the DBB has been cleaned from both close-in and far-out interfering signals.

The organization of the rest of this chapter is as follows: [Section 21.2](#) highlights three major functions of DFE. In [Section 21.3](#), we use the example of a WCDMA ARX receiver to determine specifications for DFE. Similar principles can be applied to determine specifications for other wireless standards. [Section 21.4](#) describes in detail the design of DFE for both GGE and WCDMA cases, highlighting the digital blocks that can be shared between WCDMA and GGE. In [Section 21.5](#), we briefly discuss how the concepts presented in earlier sections of this chapter can be used to design a DFE that supports LTE, in addition to WCDMA and GGE. Finally, some conclusions are drawn in [Section 21.6](#).

21.2 Major Functions of the DFE

Broadly speaking, these are the three major functions of the DFE

- sample rate conversion
- channel selection (or filtering)
- compensation of analog/RF impairments.

The ADC clock rate is usually much higher than the DBB input clock rate, especially if the ADC is an oversampled sigma-delta ADC. The digital base-band is generally implemented in a programmable DSP processor and hence it normally expects data rate at its input that has been reduced to $4\times$ or $2\times$ (or even $1\times$ in some cases) of the supported modulation's baud (or chip) rate. Hence sample rate conversion is one of the major functions of DFE.

Filtering for channel selection is the second major function of DFE. The RF signal that hits the antenna is generally composed of a wanted signal, occupying a bandwidth depending on the supported wireless standard, and a number of blocking/interfering unwanted signals. The function of channel selection is to make sure that all unwanted signals have been *sufficiently* filtered out and a relatively *clean* signal enters the DBB. The quality of a clean signal is generally measured in terms of signal-to-noise-and-distortion ratio (SNDR), where distortion includes, among other terms, the residual unwanted signals. If one could design ideal analog/RF components, including the ADC, or just have an ideal ADC as envisioned by Mitola [1], all filtering would then be done in the DFE to take maximum benefit of scaling of digital circuits in deep submicron technology. Since the technology is still quite far from ideal, all receiver chains designed to date have a filtering function that is split between ARX and DFE. The

primary objective of ARX filtering is to make sure that analog/RF circuits and ADC do not saturate in the presence of an unwanted signal. The unwanted blocking signals that are outside the band of interest, called out-of-band blockers, are generally filtered to such a level that the DFE will need to perform very little, if any, filtering for the out-of-band blockers. It is primarily for the in-band blocking and close-in interfering signals, which happen to be relatively close in frequency to the wanted signal, that impose stricter requirements on DFE filtering.

The third major function of the DFE is to compensate for analog/RF impairments. More and more new designs are relying on digital techniques to compensate for analog/RF impairments, which include DC offset, I/Q imbalance, amplitude/phase distortion, etc. We will only briefly cover this function of DFE in this chapter because more detailed coverage of compensation is covered in other chapters of this book.

Finally, as pointed out in [2], the synchronization (of timing, frequency, and phase errors) belongs to the DBB and will not be covered in the context of DFE.

21.3 Determination of DFE specifications

In this section, we will describe how to determine DFE specifications for the example case of WCDMA to support multiple bands. Similar procedures can be used to determine specifications for other wireless standards. We will restrict discussion for this example to UTRA FDD bands I–X, excluding band-VII [3]. We will also assume a recently published receiver [4] that supports both WCDMA and GGE modes, in cascade with an RF filter, as ARX for this example. For the DBB, we will assume that the expected data rate at DBB output is $2\times$ the WCDMA chip rate, i.e., 7.68 Msps. Other assumptions for the DBB will be mentioned later in this chapter.

ARX specifications

Here are some of the important ARX specifications that impact the DFE specifications:

- overall baseband filtering
- ADC Specifications
- output sampling rate (F_s) and its relationship to the local oscillator clock
- dynamic range (DR)
- number of bits at the output.

For the receiver in [4], these specifications are listed in [Table 21.1](#) for both WCDMA and GGE modes. The filtering listed in this table is only the baseband filtering. The receivers designed so far also rely on some RF filtering, which is primarily achieved using an external surface acoustic wave (SAW) filter. The SAW filter is placed between the antenna and low-noise amplifier (LNA) of the receiver in [5]. Note that the receiver in [5], in combination with the SAW filter, forms ARX for our example. The SAW filter provides > 50 dB of rejection for the transmitter signal coupling to the receiver in the case

Table 21.1 ARX specification [4].

Mode	Baseband filtering	ADC Fs	ADC DR (dB)	No. of bits
WCDMA	Two poles at 2.5 MHz and 3.5 MHz, respectively	FLO/4 for high-bands and FLO/2 for low bands	73	1 (second-order sigma-delta ADC)
GGE	Two poles at 250kHz and 350kHz, respectively	FLO/4 for high-bands and FLO/2 for low bands	80	1 (second-order sigma-delta ADC)

of full-duplex WCDMA. For other out-of-band blocking signals, the SAW filter generally guarantees > 30 dB of rejection.

The amplitude response of a first-order filter is given by

$$|H(f)| = -10 \log_{10} \left[1 + (f/f_c)^2 \right],$$

where f_c is the 3-dB corner frequency of the filter and f is the frequency at which the amplitude response is to be determined. With two first-order filter poles at 2.5 MHz and 3.5 MHz, the combined ARX filtering is 33.4 dB at 20 MHz and 61.2 dB at 100 MHz. Assuming > 30 dB of filtering from the SAW filter, the overall ARX filtering is > 90 dB for blockers at 100 MHz offset or above.

There are two ways of generating the ADC clock, both having certain pros and cons. The first way is to use a separate PLL for the ADC that generates a fixed clock, independent of the LO frequency. In this case, the ADC clock can be chosen to be an integer multiple of the supported modulation's baud/chip rate. In our example of WCDMA, this would be a multiple of 3.84 MHz. To support multiple wireless standards, this PLL will need programmability. The PLL will normally need to meet tough specifications for high-performance ADC with dynamic range greater than 80 dB. These tough specifications will translate to more area and power consumption.

Another disadvantage of having a fixed clock ADC is that this clock, along with its harmonics, can couple to the received signal of interest acting as an interferer signal. As an example, if the ADC is running at 491.52 Msps (which is 128×3.84), the fourth harmonic will be at 1966.08 MHz, which is almost in the middle of the UTRA band-II (1930–1990 MHz). A very careful chip layout would be required to minimize the impact of ADC clock coupling. The obvious advantage of a fixed clock ADC is that relatively simple decimation filters are required to reduce the data rate to DBB input sampling rate. For the DFE designer, another huge benefit is that he/she needs to perform analysis of the design only for one case per supported modulation bandwidth.

The other way of generating the ADC clock is to derive it directly from the receiver LO clock, which is a high-quality clock since it needs to meet tight phase noise performance specifications. It is extremely beneficial to use integer division of LO frequency to derive clocks for both analog (ADC, e.g.) and digital blocks. This way we do not incur any

penalty on the analog performance due to clock jitter in the clock provided to the analog/digital blocks. The mixer is the only block that incurs penalty of LO jitter on the down-converted signal. Moreover, we avoid a second PLL to generate a fixed rate clock for the A/D converter (ADC) saving area and current.

One of the major challenges of an LO driven ADC clock is that the DFE will need to perform sample rate conversion from a variable, LO driven, clock domain to a fixed, baud/chip rate driven, clock domain. As will be shown later, this can be done by digital circuits called resamplers. Another challenge is that the DFE designer needs to look at all the supported bands and perform a very detailed analysis to make sure that overall receiver specifications are met in all cases.

As shown in Table 21.1, the ADC for our example ARX is a 1-bit output second-order sigma-delta ADC, running from an LO driven clock. For UTRA bands with RF frequencies less than 1 GHz, called low-bands in Table 21.1, the ADC clock is derived by dividing the LO clock by 2, while for UTRA bands with RF frequencies more than 1 GHz, called high-bands in Table 21.1, the ADC clock is derived by dividing the LO clock by 4. This means that for UTRA bands I–X, excluding band-VII, the ADC sampling frequency will vary from approximately 434.5 Msps (for lower edge of band-V) to 542.5 Msps (for upper edge of band-I).

Figure 21.2 shows an example of frequency response of second-order sigma-delta ADC output with an input *continuous wave* (CW) signal of 100 kHz. Note the second-order noise shaping characteristics of this ADC at higher input frequencies. The DFE will not only have to filter out unwanted blocking/interfering signals left unfiltered by the ARX, it will also have to make sure that the ADC noise shaping gets filtered.

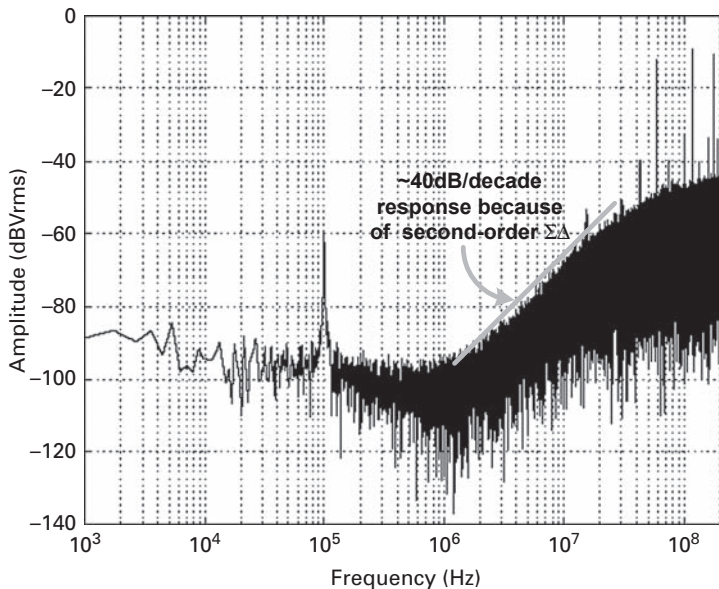


Figure 21.2 FFT measured at ADC output with a tone at 100kHz [5] (© 2005 IEEE).

Close-in filtering requirements

It was mentioned previously, and it will be shown again later, that it is the filtering of close-in unwanted blocking/interfering signals that imposes one of the challenging requirements for the design of the DFE. In this subsection, we will determine these filtering requirements for up to 20 MHz offset from the wanted signal of interest.

Table 21.2 lists the worst-case requirements derived from WCDMA specifications in Chapter 7 of [3], which will help us determine the overall DFE filtering requirements. This table also lists the bands that dictate the worst-case requirement for each test case. Note that the first three interferers, i.e. at 2.7 MHz, 2.8 MHz and 3.5 MHz are either narrow-band (GMSK modulated) or continuous wave (CW) interferers. The interferer at 5 MHz offset, however, is WCDMA modulated and hence it spans from approximately 3–7 MHz. We have listed both low-power and high-power ACS test cases for the sake of completeness, although they impose the same filtering requirement.

In Table 21.3 we have added a 3 dB margin on the blocker level. This will make the filtering requirement 3 dB tighter. We have also added a column that shows ARX base-band filtering with two poles at 2.5 MHz and 3.5 MHz, respectively. Finally, the last

Table 21.2 Summary of 3GPP specs for close-in interferers.

Frequency offset (MHz)	Wanted signal level (dBm)	Blocker level (dBm)	Difference (dB)	Blocker modulation	Limiting bands	Test case
2.7	−96.7	−57	39.7	GMSK Mod	IV, X	Narrow-band blocking
2.8	−93.7	−56	37.7	GMSK Mod	III, VIII	Narrow-band blocking
3.5	−96.7	−44	52.7	CW	IV, X	Narrow-band IMD
5	−92.7	−52	40.7	WCDMA	I, IV, VI, X	ACS – low power
5	−65.7	−25	40.7	WCDMA	I, IV, VI, X	ACS – high power
5.9	−96.7	−44	52.7	GMSK	IV, X	Narrow-band IMD
10	−103.7	−46	57.7	CW	I, IV, VI, X	Intermodulation distortion
20	−103.7	−46	57.7	WCDMA	I, IV, VI, X	Intermodulation distortion

Table 21.3 Determination of DFE filtering requirements.

Frequency offset (MHz)	Difference with 3dB margin (dB) – A	ARX baseband filtering (dB) – B	DFE filtering requirement with 10dB blocker rejection (dB) C = A−B+10
2.7	42.7	5.4	47.3
2.8	38.7	5.7	43
3.5	55.7	7.7	58
5	43.7	6.3 ¹	47.4
5.9	55.7	14.0	51.7
10	60.7	21.9	48.8
20	60.7	33.4	37.3

¹The worst-case filtering is at the edge of the WCDMA modulated interferer, i.e., at 3MHz offset.

column lists the required filtering to be performed in the DFE, with the requirement that the interferer will be filtered 10 dB more than the wanted signal. In this table, we have kept only one entry for 5 MHz case.

It appears that the most stringent requirement comes from the 3.5 MHz blocker. However, this is a CW blocker and, after de-spreading in the DBB, this will become negligibly small compared with the wanted signal even if it is only filtered to make its level the same as the wanted signal (without 10 dB of additional filtering). A similar argument is applicable for the GMSK modulated interferer, since it occupies approximately 200 kHz bandwidth and will de-spread to a negligible level even without 10 dB of additional filtering. In general, we can conclude that the adjacent channel imposes the toughest requirement of approximately 50 dB of filtering at offset of 3–7 MHz.

To summarize the discussion in this section, the DFE in this example should perform the following tasks:

- Convert the data rate from a very high (in this case 434.5–542.5 Msps), LO driven, rate to a fixed 7.68 Msps rate.
- Perform residual filtering of all unwanted blocking/interfering signals left unfiltered by ARX. The close-in filtering requirements are given in [Table 21.3](#).
- Filter out ADC noise shaping and make sure that it does not alias back in the signal of interest.

21.4 Design of DFE to support WCDMA and GGE

In this section, we will present the design of a DFE chain that meets both WCDMA and GGE specifications for the example ARX assumed in the [previous section](#). In the following section, we will discuss considerations that are important in the design of a DFE that supports the LTE standard as well.

The data rate at DFE output should be a multiple of baud/chip rate of the supported modulation. For WCDMA, we assumed DFE output rate to be $2\times$ the WCDMA chip rate (3.84 Mcps), i.e., 7.68 Msps. For GGE, we will assume that the DFE output rate should be $4\times$ GSM baud rate (270.833 kbaud), i.e., 1.0833 Msps. Note that the data rate at DFE output for WCDMA needs to be >7 times more than the data rate for GGE. If we want to share the same circuits for both WCDMA and GGE, they will be designed for the more challenging case of WCDMA, at least towards the end of the DFE chain. These circuits will need to be highly programmable, both in terms of their transfer functions, as well as their clock rates. Because of the scaling advantage of digital circuits in deep submicron processes, we may also have separate circuits for both WCDMA and GGE. It is the right mix of programmability and judicious use of separate circuits that will result in close to the optimum DFE designs.

Since the assumed ADC is a 1-bit output with a sampling rate of at least 434.5 Msps, we need to decimate it by ≥ 56 for WCDMA and ≥ 400 for GGE. We obviously need a multi-rate design in order to achieve an area and power efficient implementation [6]. We will also need a non-integer sample rate converter somewhere in the DFE chain to go

from LO driven, variable, clock rate to a baud/chip rate dependent, fixed, clock rate. We will use a polynomial resampler (RES) [7], implemented using a Farrow structure [8], to achieve this function.

Two important questions that one needs to answer while designing a polynomial resampler are the order of the polynomial for interpolation and the location of the resampler. When using a Lagrange polynomial, we are basically limited to first-, second- or third-order polynomials because after that the advantage of more filtering from a higher-order polynomial is overshadowed by its implementation complexity. Of course, it is highly desirable to use the lowest-order polynomial resampling, if possible, for lower implementation complexity. Before proceeding further with details of DFE, we think it is important to understand the principle of polynomial resampler in some detail.

We would also like to mention here that, for reduced implementation complexity, it is highly desirable to have fixed-coefficient filters, whenever possible. We will, therefore, try to make use of multiplier-less digital filters in the design of the DFE.

Principle of polynomial resampling

In this subsection, we will briefly describe the principle of resampling based on Lagrange polynomial interpolation. The details can be found in [7] and [9].

Figure 21.3 shows an example of input and output sampling times. The objective of a polynomial resampler is to calculate output at kT_I , call it $y(kT_I)$, from the knowledge of N input samples equally spaced by T_S , call it $x(mT_S)$, by fitting a polynomial on N input samples and finding its value at kT_I , where T_S is the input clock period and T_I is the output clock period. One requirement of interpolation is the determination of offset, δ , between the output clock edge and its closest input clock edge. The normalized value of δ with respect to T_S is called μ .

Figure 21.4 shows the Farrow structure for implementing cubic Lagrange interpolation [9]. It is composed of four FIR filter sections that are combined in an output stage. If the outputs of four filter sections are f_0, f_1, f_2 and f_3 , the output, $y(kT_I)$, can be written as

$$y(kT_I) = \mu(\mu(f_3 + f_2) + f_1) + f_0. \quad (21.1)$$

Note the iterative nature of the above equation; instead of performing three multiplications as suggested by Figure 21.4, it is possible to implement the operations in Equation (21.1) as a recursive Multiply-Add-ACcumulate (MAAC) operation that requires only

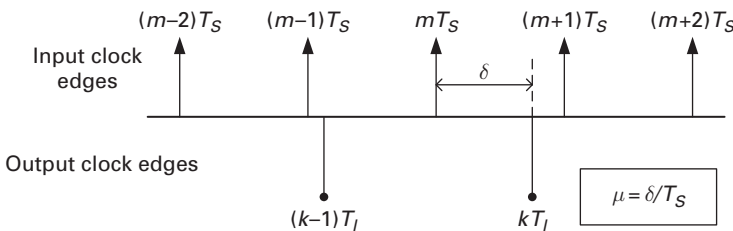


Figure 21.3 Relationship between input and output clocks and definition of μ .

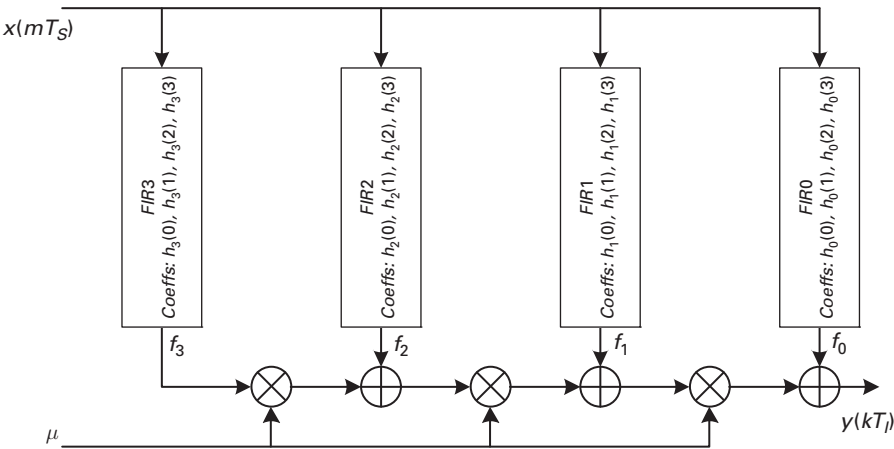


Figure 21.4 Farrow Structure for cubic resampler (adapted from [9]).

Table 21.4 Farrow structure coefficients for Lagrange interpolation.

Order	Filter branch	Filter coefficients			
		$h_i(0)$	$h_i(1)$	$h_i(2)$	$h_i(3)$
Cubic	f_3	$-1/6$	$1/2$	$-1/2$	$1/6$
	f_2	$1/2$	-1	$1/2$	0
	f_1	$-1/3$	$-1/2$	1	$-1/6$
	f_0	0	1	0	0
Quadratic	f_2	$1/2$	-1	$1/2$	0
	f_1	$-1/2$	0	$1/2$	0
	f_0	0	1	0	0
Linear	f_1	-1	1	0	0
	f_0	0	1	0	0

one hardware multiplier and one adder that runs with at least $3\times$ oversampled clock. Table 21.4 lists coefficients of filter branches in the Farrow structure for cubic, quadratic, and linear resamplers. Note that coefficients of the f_2 filter branch are the same for both cubic and quadratic resamplers. Also, f_0 performs the same filtering operation for all resampler orders. In fact, f_1 is the only filter branch that has different coefficients for different resampler orders. These features can be exploited in the implementation of a programmable resampler [10].

One of the challenges in the implementation of resampler is the generation of μ . As was shown in Figure 21.3, μ is generated on the basis of knowledge of the relationship between input and output clocks. If the two clocks are completely arbitrary, then it will be essential to use FIFOs with overflow and underflow logic for appropriate hand-off. It is possible to avoid the use of an FIFO by generating both the “pre-resampler” and “post-resampler” clocks from a very high speed LO driven master clock. An example of such a design is given in [10], where all the variable rate clocks are sub-multiples of LO, while

the fixed rate clock is “on average” equal to the desired rate and μ is generated with the assumption of an ideal fixed-rate clock.

Another important factor to keep in mind when using a resampler is that, similar to an ADC, it may need anti-aliasing filtering at its input. As described in [9], the simplest way to define the functionality of a resampler is that it is composed of a D/A converter (DAC), followed by a filter whose response depends on the type and order of the interpolation function, followed by an ADC. This is shown in Figure 21.5.

Figure 21.6 shows the frequency response for first-, second-, and third-order resampler filters based on Lagrange polynomials. Note that the x -axis is labeled as “Normalized Frequency.” A value of unity on the x -axis corresponds to the sampling frequency at the input of the resampler. Referring back to the simple model in Figure 21.5, the DAC output will have images of the input signal at multiples of the input sampling frequency. The notches in the frequency response of Figure 21.6 will help filter out those images. It is obvious from Figure 21.6 that higher-order interpolation filtering has wider notches and also a flatter response close to the DC, where the wanted signal is located. We can

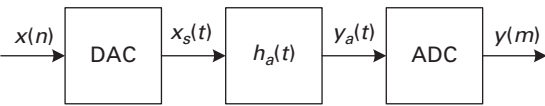


Figure 21.5 The hybrid analog/digital model of the interpolation filter (resampler).

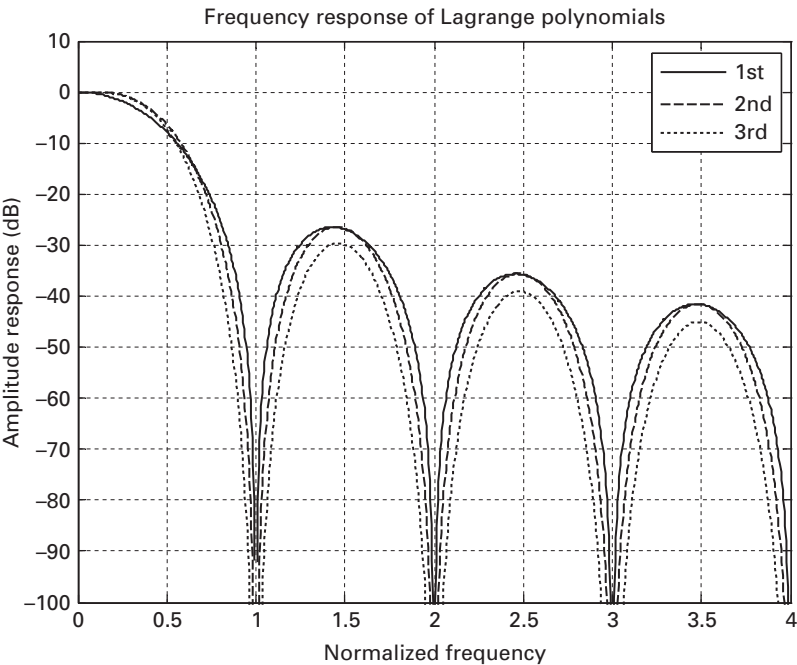


Figure 21.6 Frequency response of first-, second- and third-order Lagrange polynomial based resamplers.

achieve a performance similar to the third-order response by using a first-order interpolation if we make one or both of the following changes in our data path:

- Increase the sampling frequency at the input of the resampler. This is equivalent to moving the resampler upstream, or closer to the ADC.
- Use an anti-aliasing pre-filter (PREF) prior to the resampler. This filter will basically filter out some of the close-in interferer signals that can be significantly stronger than the wanted signal of interest. Having PREF before the resampler will make sure that the images of the unwanted interferer signals around multiples of the resampler input F_s are sufficiently filtered out and do not cause any aliasing problems. We will explain this feature later with the help of examples.

The first rate change filter

After having explained the principle of resampler and its functionality, we are now ready to design the DFE starting at the ADC output. The design of the first stage is critical in order to have low power dissipation in the digital front-end. Since we are using an oversampled sigma-delta ADC, the most recommended decimation filter at its output is a $SINC^N$ filter. A general rule of thumb is to have N at least one more than the order of sigma-delta noise shaping. Our example ARX uses a second-order sigma-delta ADC; therefore N should ≥ 3 . In our example, we will use $N = 4$ [5].

A cascaded integrator-comb (CIC) decimation filter is a good choice to achieve $SINC^N$ response and to provide the necessary rate conversion [12]. Reference [11] shows a non-recursive architecture for first decimation filter. Reference [14] compares CIC architecture with an FIR-based implementation and concludes that FIR-based implementation is comparable in gate count to CIC-based implementation as long as $N \leq 3$ for a $SINC^N$ filter. We noticed that the 1-bit property of the ADC can also be exploited to come up with a look-up table (LUT) based polyphase implementation that is much smaller in gate count than the previously published CIC implementations. The details of our design can be found in [5]. Note, however, that the LUT-based implementation is only applicable for the case of a 1-bit output sigma-delta ADC. For higher number of ADC bits, similar to [2], we recommend using CIC filters. As will be pointed out later in this chapter, CIC filters are also very useful if we need to have programmable decimation ratio to address various types of wireless standards or signal bandwidths. One disadvantage of CIC filter is that it causes increase in word length early in the design chain.

The z -domain transfer function of a $SINC^4$ filter, with each stage having 16-taps is given as $H(z) = \left(\sum_{i=0}^{15} z^{-i} \right)^4$. The total number of coefficients in this filter is 61. Hence

we can re-write $H(z) = \sum_{i=0}^{60} c_i z^{-i}$, where c_i represents the i -th coefficient of the 61-tap

filter. In Reference [5], decimation ratios of 8, 16, and 32 are compared and a decimation ratio of 16 is chosen for area and output clock frequency reasons. To reduce the area further, I/Q interleaving is also employed [13]. Finally, to relieve the design of the high-

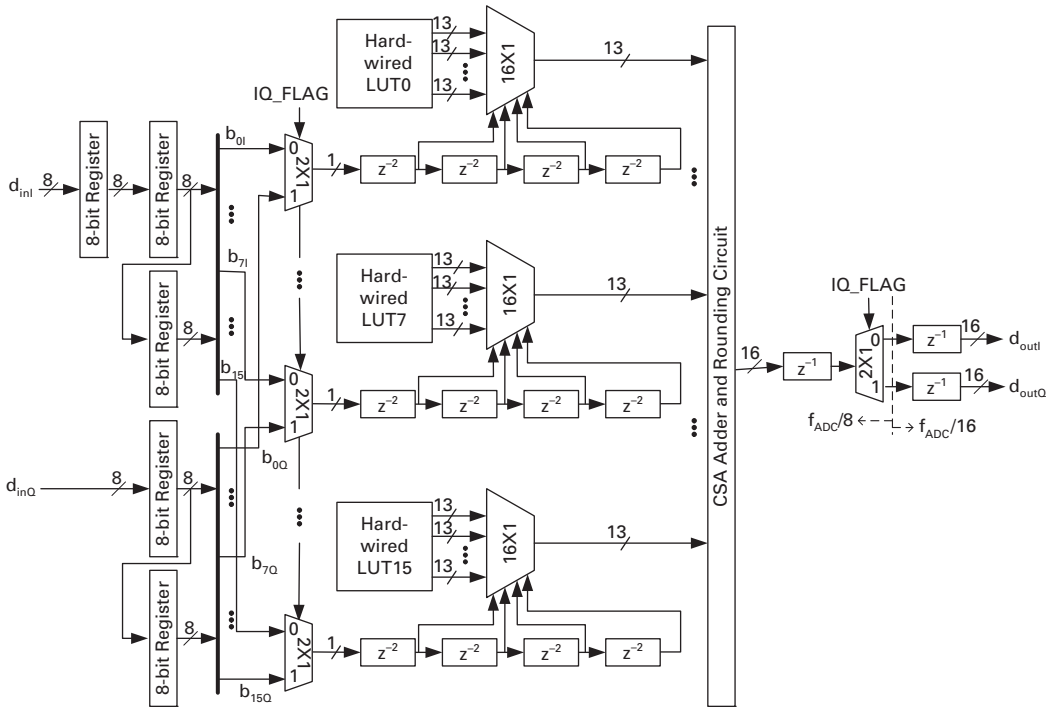


Figure 21.7 I/Q interleaved polyphase implementation of $SINC^4$ decimation by 16 filter [5] (© 2005 IEEE).

speed interface between the ADC and the first rate change filter (RCF1), the ADC output is packed to 8-bit words for transportation across the interface. Figure 21.7 shows the implementation details of the LUT-based RCF1 design [5].

As mentioned earlier, this design is highly efficient only if the ADC output is 1-bit and also if we need fixed decimation ratio from RCF1. It is also worth mentioning here that while I/Q interleaving is an excellent idea for lower area design, it is not a good idea for low-power design. This is because of higher switching activity of all the bits due to I and Q channel data being generally uncorrelated. This, however, is not a disadvantage for a 1-bit sigma-delta ADC output. Because of noise shaping, there is very little, if any, correlation between consecutive bits even on the I (or Q) channel. Hence, the interleaved architecture presented in [5] is close to optimal for both area and power reasons.

Now, with RCF1 performing decimation by 16, the data rate at RCF1 output is reduced from 435.5–542.5 Msps to a much more implementation-amenable rate of 27.16–33.91 Msps.

The RCF1 not only filters out noise-shaping of delta-sigma ADC, it also protects 5 MHz bands around multiples of $f_{SADC}/16$ that alias back to baseband after decimation by 16, where f_{SADC} is the sampling frequency of ADC. In addition, RCF1 provides filtering for far-out blockers.

Figure 21.8 shows the filter response of ARX baseband (ARXB), RCF1, and combined ARXB+RCF1 from DC to 100 MHz. Note that more than 100 dB of attenuation is

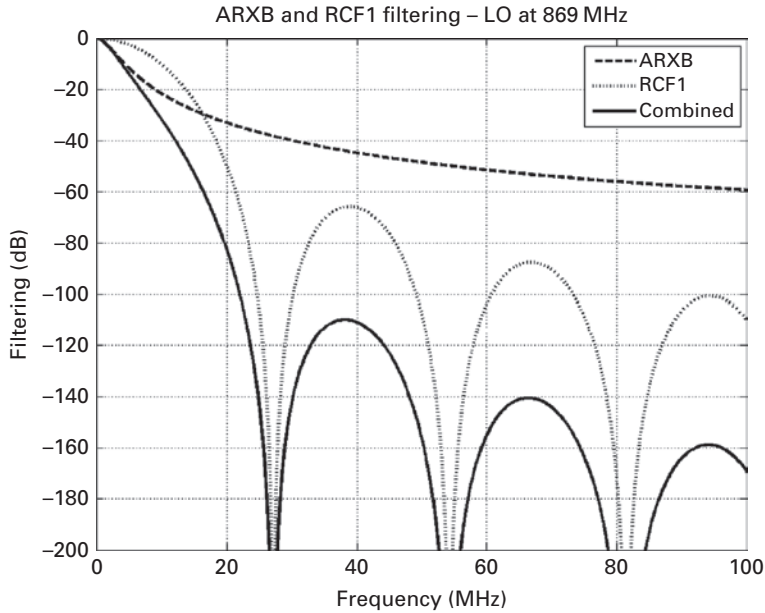


Figure 21.8 ARXB + RCF1 combined filtering at 869 MHz.

achieved for all frequency offsets more than approximately 25 MHz. [Figure 21.9](#) is a zoomed version of [Figure 21.8](#) around $f_{SADC}/16$. This shows the 5 MHz band that will alias back to baseband after decimation by 16. Note that RCF1 alone provides approximately 100 dB of protection. With ARXB, the protection is more than 130 dB.

[Figure 21.10](#) and [Figure 21.11](#) show a frequency response for the highest channel in band-I, corresponding to LO of 2170 MHz. Again, filtering and alias protection are more than required. All that remains to be filtered now by the rest of the DFE chain are the close-in blocking/interfering unwanted signals as listed in [Table 21.3](#). Note that at 20 MHz offset, we have already achieved more than required filtering.

One could argue that we could simplify the design of RCF1 since we do not need so much filtering for WCDMA. We should, however, keep in mind that one of the major functions of RCF1 is to filter out ADC noise shaping, which requires roughly $SINC^4$ filtering.

Prefilter and resampler

The receiver presented in [16] is an example of a GSM/GPRS receiver that has successfully used RCF1 presented in the [previous section](#). A block diagram of the digital data path for this receiver is shown in [Figure 21.12](#). In this figure the block labeled ‘Rate change filter 1’ is composed of a cascade of decimation by 16 RCF1 described in the last subsection, followed by two decimation-by-2 half-band filters. Half-band filters are implementation-friendly decimation (or interpolation) by 2 filters, which are generally used with fixed-coefficients. These fixed-coefficients can be implemented using canonic

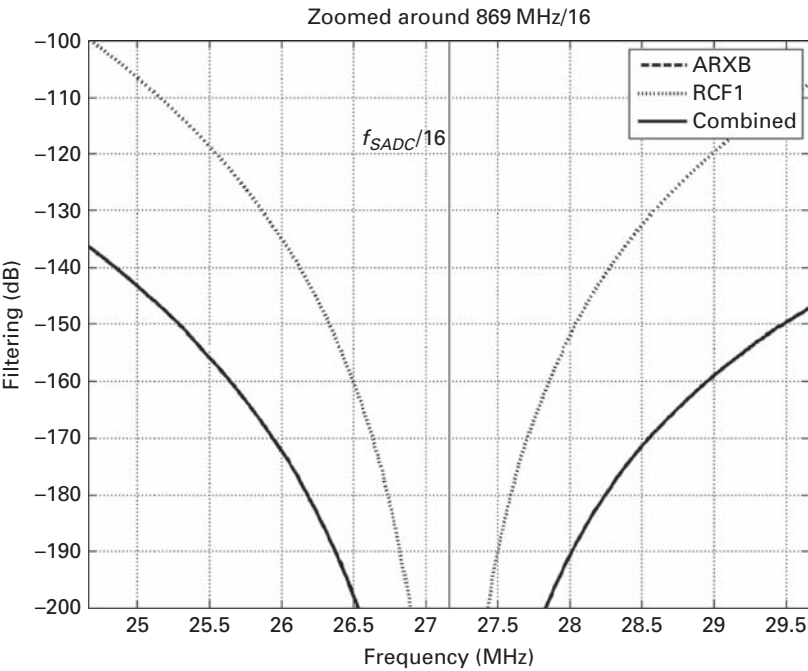


Figure 21.9 Zoomed version of Figure 21.8 around $f_{SADC}/16$.

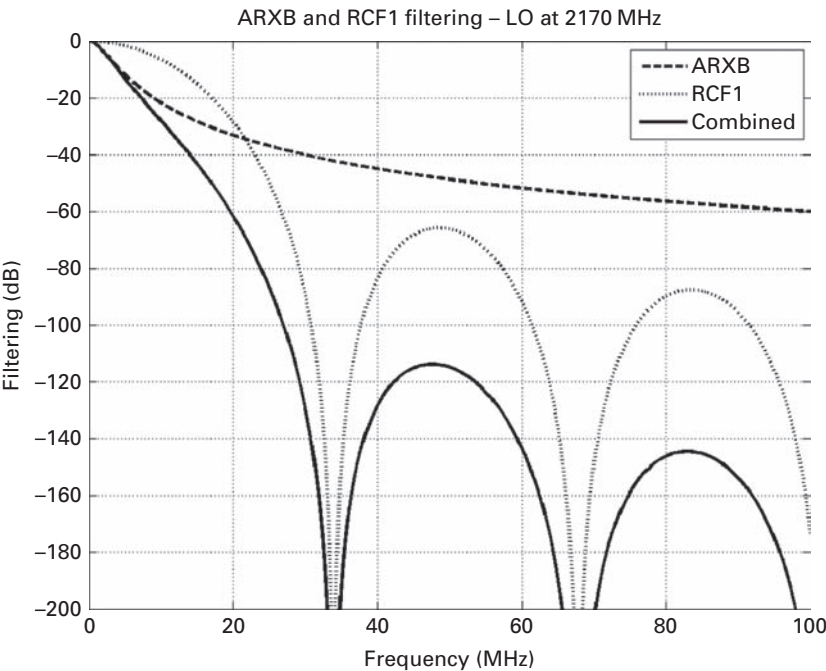


Figure 21.10 ARXB + RCF1 combined filtering at 2170 MHz.

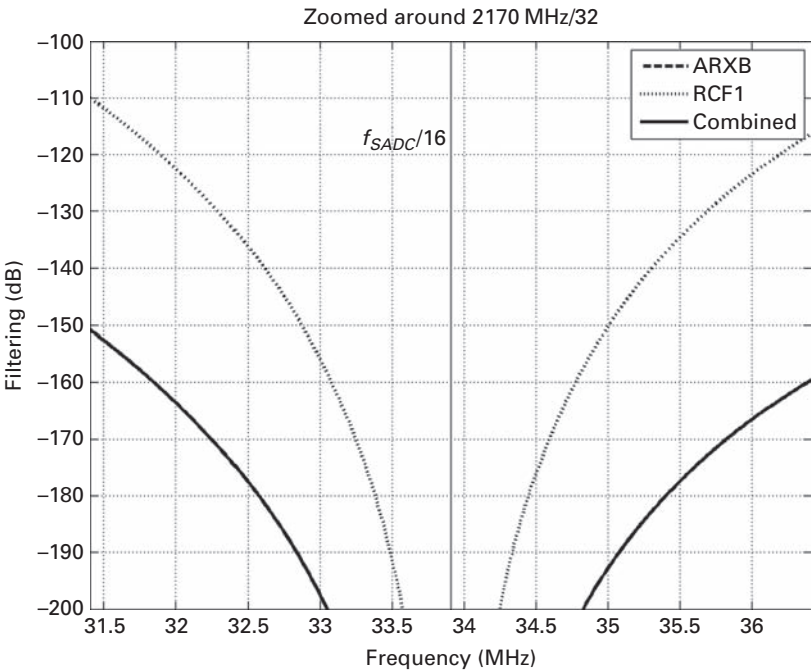


Figure 21.11 Zoomed version of Figure 21.10 round $f_{SADC}/16$.

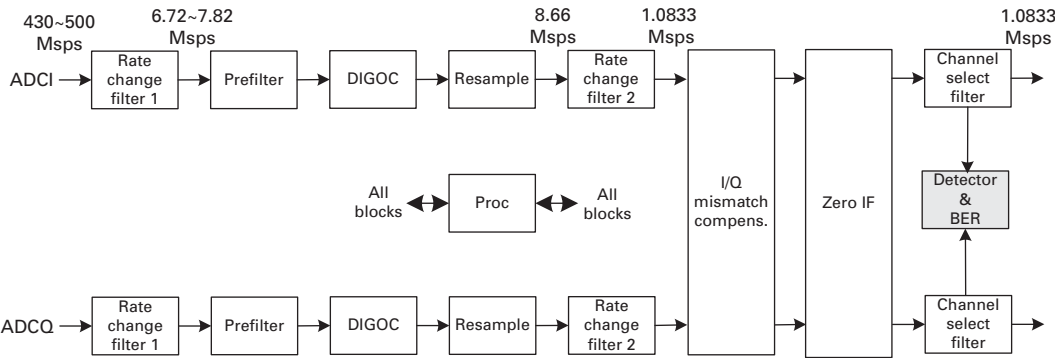


Figure 21.12 DFE for GSM/GPRS receiver using RCF1 as the first filter after the ADC [16] (© 2006 IEEE).

signed digit (CSD) format [13], which requires fewer adders than a simple binary representation. A polyphase (in this case 2-phase) implementation of half-band filters results in a lower power design since all the computations are done at the output data rate for a decimation-by-2 half-band filter [6].

Since the receiver in [16] supports only a 200 kHz wide GSM/GPRS signal (or Bluetooth in a test mode), the RCF1 output is decimated by another factor of 4 to reduce the data rate to 6.72–7.82 Msps, before entering the *prefilter* (PREF) and resampler. Note that we can ignore the block labeled DIGOC in Figure 21.12 for this discussion as it is

used to perform residual DC offset compensation. In this implementation, the PREF is a fixed-coefficients FIR filter, while the resampler is a programmable order, supporting linear, quadratic and cubic interpolation, polynomial resampler [10].

Recall from our discussion on the polynomial resampler that to support a wider band signal and to avoid aliasing caused by images of the resampler input signal, the resampler input sampling rate can be increased. Since we want to support a wider bandwidth WCDMA signal (approx. 5 MHz wide channel), in addition to relatively narrower band GGE signal, we can use a prefilter followed by a resampler directly at RCF1 output. An example of such a design using a sixth-order IIR filter as a prefilter is given in [15].

Infinite impulse response IIR filters are generally avoided in sensitive digital data paths primarily because of stability concerns and the issue of limit cycles. Their advantage compared with FIR filters is that the same filtering response can be achieved using much smaller area. By making sure that none of the IIR poles falls outside the unit circle and by performing detailed analysis for limit cycles, one can take advantage of the low-area IIR filters. Another draw-back of IIR filters is that their phase is nonlinear and one needs to make sure that group-delay distortion caused by nonlinear phase does not degrade the signal-to-noise-and-distortion ratio (SNDR).

Similar to the prefilter in [15], we will use a sixth order IIR filter comprising three independent second order IIR filter stages. Each of these stages has partially programmable filter coefficients. In addition to providing the requisite low-pass filtering prior to the resampler block, PREF can also be used to partially correct for amplitude droop and group delay distortion caused by the ARX filters. The transfer functions of three filter sections for GGE mode are given by [15]

$$H_{PREFA1}(z) = \frac{0.5 + z^{-1} + 0.5z^{-2}}{256 - (498 - 2^{7-\text{shift}A1})z^{-1} + (244 - 2^{7-\text{shift}A1})z^{-2}}, \quad (21.2)$$

$$H_{PREFA2}(z) = \frac{0.5 + z^{-1} + 0.5z^{-2}}{256 - (498 - 2^{7-\text{shift}A2})z^{-1} + (244 - 2^{7-\text{shift}A2})z^{-2}}, \quad (21.3)$$

$$H_{PREFB}(z) = \frac{0.5 + z^{-1} + 0.5z^{-2}}{256 - (486 - 2^{7-\text{shift}B})z^{-1} + (232 - 2^{7-\text{shift}B})z^{-2}}. \quad (21.4)$$

For WCDMA, the filter transfer functions are

$$H_{PREFA1}(z) = \frac{16 \cdot (1 + 2z^{-1} + z^{-2})}{256 - (384 - 2^{7-\text{shift}A1})z^{-1} + (192 - 2^{7-\text{shift}A1})z^{-2}}, \quad (21.5)$$

$$H_{PREFA2}(z) = \frac{16 \cdot (1 + 2z^{-1} + z^{-2})}{256 - (384 - 2^{7-\text{shift}A2})z^{-1} + (192 - 2^{7-\text{shift}A2})z^{-2}}, \quad (21.6)$$

$$H_{PREFB}(z) = \frac{16 \cdot (1 + 2z^{-1} + z^{-2})}{256 - (320 - 2^{7-\text{shift}B})z^{-1} + (128 - 2^{7-\text{shift}B})z^{-2}}. \quad (21.7)$$

In the above equations, $shiftA1, shiftA2, shiftB \in \{0, 1, \dots, 7\}$. Note the similarity of the transfer functions for GGE and WCDMA. By adding a negligibly small area overhead, it is possible to come up with a programmable PREF that supports both GGE and WCDMA modes.

Figure 21.13 shows PREF frequency response in GGE mode for a few coefficient settings. Note that with $shiftA1, shiftA2, shiftB = 7$, there can be a gain of approximately 10 dB around 350 kHz. This can be useful in correcting the droop of analog low-pass filters. Figure 21.14 shows PREF frequency response in WCDMA mode. The three

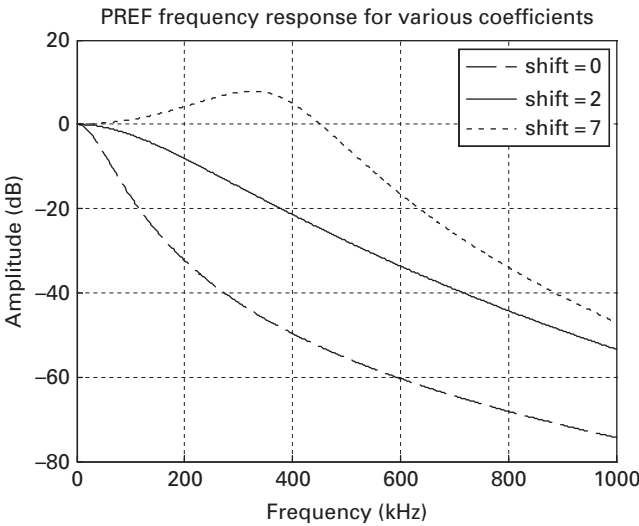


Figure 21.13 PREF frequency response in GGE mode for various coefficient settings.

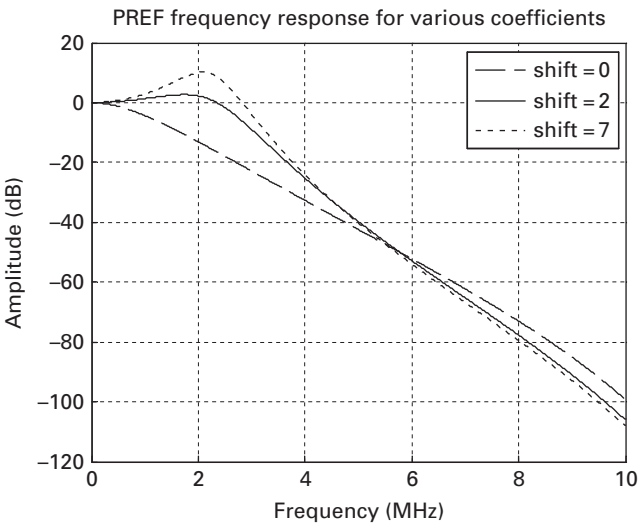


Figure 21.14 PREF frequency response in WCDMA mode for various coefficient settings.

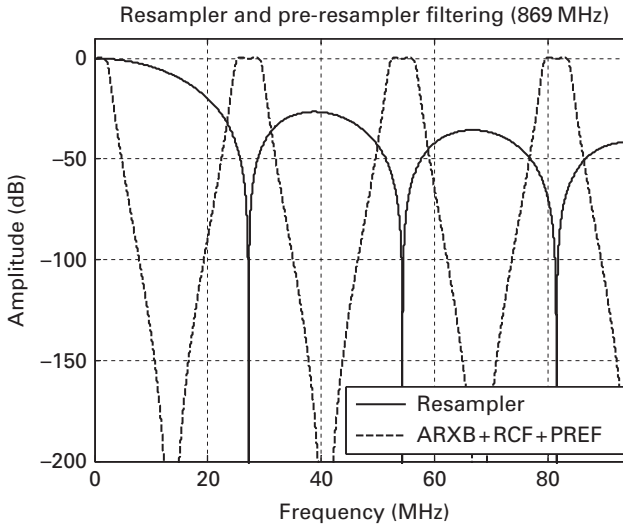


Figure 21.15 Resampler and ‘pre-resampler’ filtering for LO of 869 MHz.

filter sections in these two figures are shown to have the same shift value only for the sake of convenience of these plots; in reality each section can have a different shift value.

In the following analysis, we will look at PREF and resampler response for WCDMA for two extreme cases: first channel of band-V with LO at approximately 869 MHz and the last channel of band-I with LO at approximately 2170 MHz. PREF coefficients will be selected to compensate for the amplitude droop caused by ARXB for both cases.

Figure 21.15 shows the amplitude response of a first-order resampler (solid line) and the combined ARXB+RCF1+PREF filtering, repeated at multiples of the sampling frequency of PREF, that equals $869/32 = 27.1562$ MHz for this case. PREF coefficients were set to 1, 7 and 2, respectively, for *shiftA1*, *shiftA2*, *shiftB*.

Figure 21.16 shows the combined amplitude response of ARXB+RCF1+PREF+RES from 0–3.5 MHz and ARXB+PREF combined group-delay distortion from 0–2.5 MHz. Note that a very flat amplitude response is obtained. The group-delay distortion caused by PREF is also less than 0.2 μ sec.

In Figure 21.17, the solid black curve shows the combined ARXB+RCF1+PREF+RES response, where RES represents the filtering of Farrow structure in the resampler. In this case, we used a first-order resampler. The solid black curve can be considered as the frequency response of a signal that needs to be sampled at the resampler output clock. We chose it to be $4 \times$ WCDMA chip rate, i.e., 15.36 MHz. The solid gray lines are at multiples of 15.36 MHz, showing the regions that will be aliased back to the baseband after resampling. The dotted gray line is at $2 \times$ chip rate and it shows that with resampler output rate of 15.36 Msps, there will be negligible aliasing in-band. With resampler output rate of $2 \times$ chip rate, however, there will be significant aliasing from an unfiltered adjacent channel, as will be shown later. The dotted rectangle shows the largest 5 MHz wide band signal that will alias back to the baseband. It spans from 28.22 MHz to 33.22 MHz. With LO of 869 MHz, the sampling frequency at PREF output is

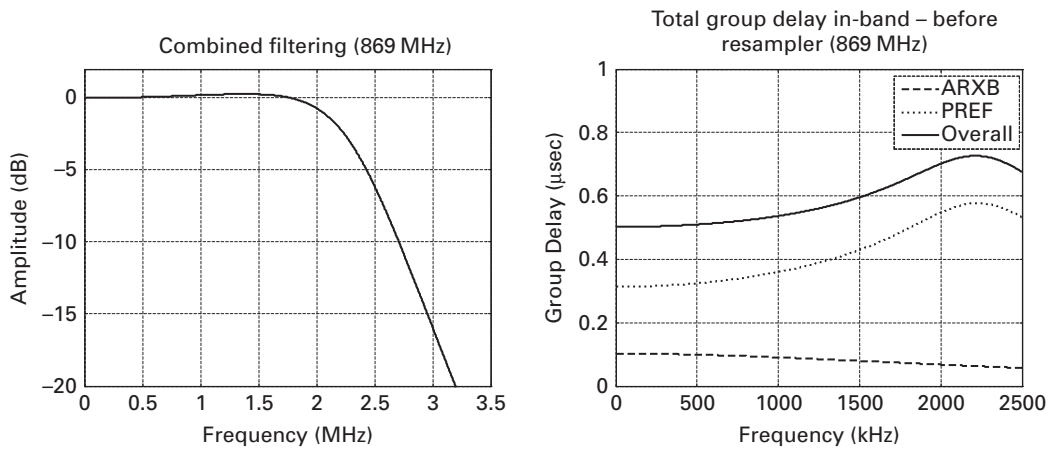


Figure 21.16 Overall amplitude response and group-delay distortion at resampler (869 MHz).

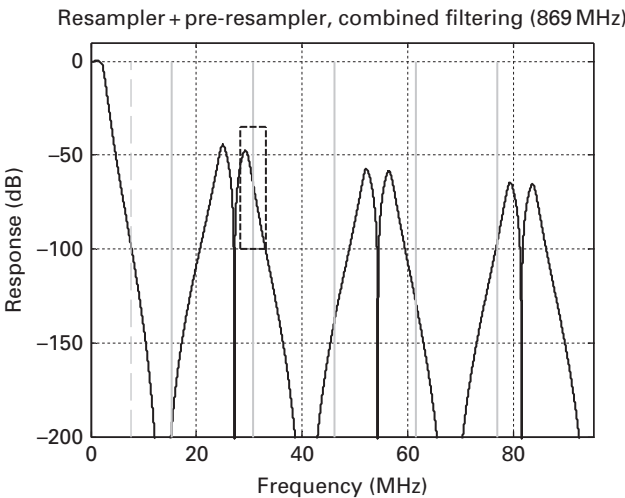


Figure 21.17 Combined filtering response – vertical lines at multiples of 15.36 MHz.

27.1562 MHz, therefore, the first image of the adjacent channel will span from 29.6562 MHz to 34.6562 MHz. The minimum filtering in this rectangle is 47dB and this should still be enough considering that in the 5 MHz band corresponding to the adjacent interferer, the average filtering is more than 80 dB and this will eliminate the adjacent interferer after despreading. Note, however, that with increasing LO frequency, the two “humps” around the second gray solid line at 30.72 MHz will move farther in the right direction and aliasing will get worse.

Figure 21.18 shows the response for LO of 2170 MHz. For this case, PREF sampling frequency is $2170/64 = 33.9062$ MHz and the image of the first adjacent channel spans 26.4062 MHz to 31.4062 MHz. The minimum filtering in this window is approximately

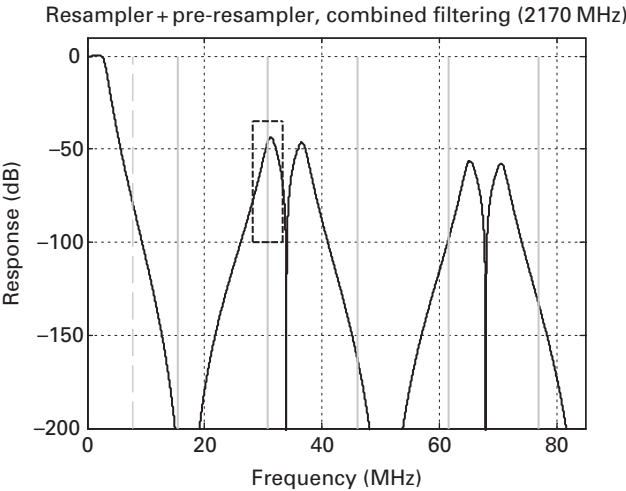


Figure 21.18 Combined filtering response at LO of 2170 MHz.

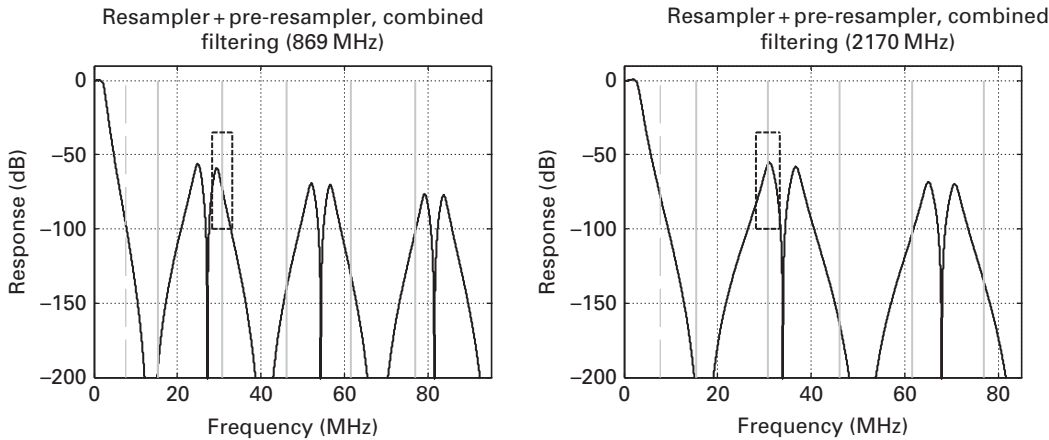


Figure 21.19 Combined filtering response with second-order resampler.

43 dB. Although, this might still be enough since the average filtering is still more than 60 dB, we can use a second-order resampler without adding too much complexity to design. Figure 21.19 shows the response with second-order resampler for two channel extremes and clearly minimum filtering of more than 55 dB is achieved. Finally, Figure 21.20 shows the overall amplitude response and group-delay distortion for LO of 2170 MHz. For this LO, PREF coefficients of 2, 7, and 2 were used, respectively, for *shiftA1*, *shiftA2*, *shiftB*.

Decimation by 2 at resampler output

In this subsection, we will explain that instead of taking the resampler output directly to $2\times$ chip rate, it is better to use $4\times$ chip rate at the resampler output and then follow the

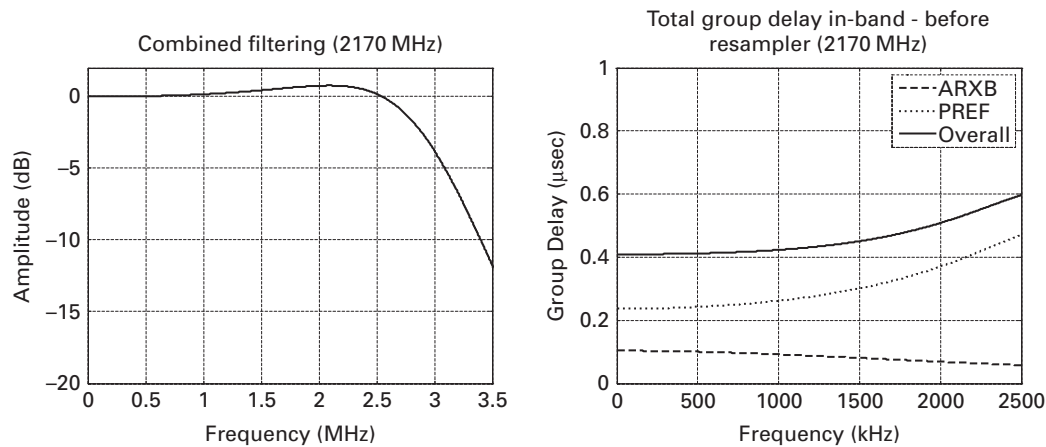


Figure 21.20 Overall amplitude response and group-delay distortion at resampler (2170 MHz).

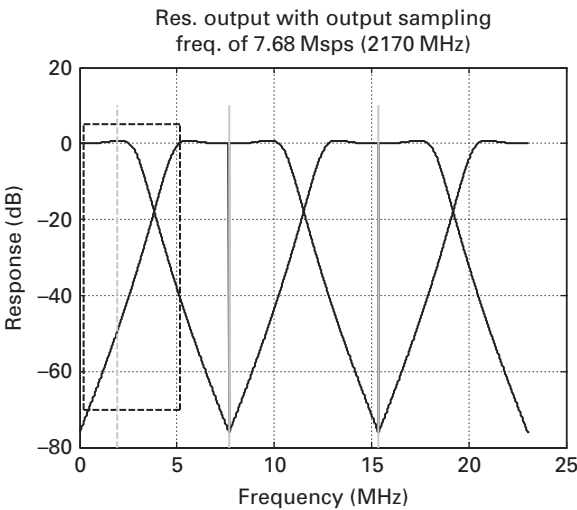


Figure 21.21 Resampler output with output sampling frequency of $2\times$ chip rate.

resampler by a simple decimation-by-2 filter to reduce the data rate to $2\times$ chip rate. The analysis will be done for LO of 2170 MHz without loss of generality and we will also assume a second-order resampler.

Figure 21.21 shows the response at the resampler output with the assumption that it is directly resampled to $2\times$ chip rate. The vertical gray lines are at multiples of 7.68 MHz and the vertical dotted gray line is at 1.92 MHz. The dotted rectangle shows the adjacent interferer with respect to the first image at 7.68 MHz. The minimum in-band filtering of 50 dB is achieved at 1.92 MHz. It might just be enough to reduce the impact of the first adjacent interferer. However, in order to achieve more margin with adjacent channel, we can resample the output to $4\times$ chip rate and then use a decimation-by-2 filter at the

resampler output. Figure 21.22 shows the resampler output response with output resampled to $4\times$ chip rate and it is obvious that in this case the adjacent channel is far from the channel of interest.

We use a simple FIR filter with coefficients of $[-1\ 0\ 3\ 6\ 6\ 3\ 0\ -1]/16$, called second rate change filter or RCF2, at the resampler output before performing decimation by 2. Figure 21.23(a) shows the frequency response of RCF2 and the combined response at RCF2 output. Figure 21.23(b) shows that the adjacent channel is filtered more than 80 dB in-band causing negligible aliasing. Note in Figure 21.23(a) that there is only approximately 10 dB of attenuation at an offset of 3 MHz and 40 dB of attenuation at 4 MHz. Both

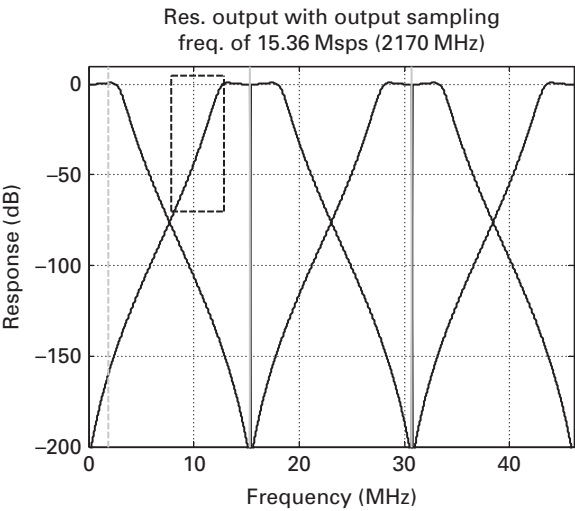


Figure 21.22 Resampler output with output sampling frequency of $4\times$ chip rate.

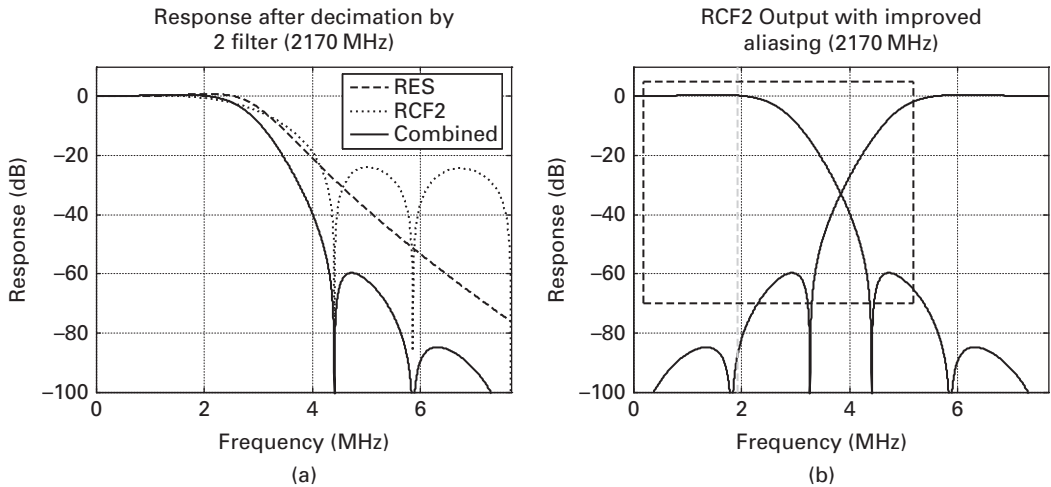


Figure 21.23 RCF2 response and improved in-band aliasing.

of these frequencies correspond to the adjacent interferer. We need a final channel selection filter to achieve more filtering for the adjacent channel and also at 2.7 MHz, corresponding to the 3GPP narrow-band blocking case in [Table 21.3](#).

The same resampler and RCF2 circuits can also be used for GGE. The clocks at the resampler output as well for RCF2 will, however, be different for GGE. For WCDMA, we went from a variable, LO driven, clock rate to a fixed 15.36 Msps at the resampler output and 7.68 Msps at the RCF2 output. For GGE, in order to meet all the specs, the resampler output clock needs to be 2.1667 Msps, which is $8\times$ the GSM baud rate, and 1.0833 Msps at the RCF2 output.

So far, our DFE chain consists of RCF1, followed by PREF, resampler and RCF2 for both WCDMA and GGE and the same circuits are used for both the modes with very limited programmability. Since the data rates are significantly reduced now, it is possible to either use a programmable DSP or fully programmable filters/circuits to achieve the rest of the signal processing. We can also use separate circuits for GGE and WCDMA to make the overall DFE design more power efficient. In the next two subsections, we will give some details of post RCF2 circuits separately for GGE and WCDMA.

GGE data path

Most of the competitive designs use down-conversion to VLIF in the ARX mixer. Generally IF values of 100–150 kHz are used. Higher values of IF make the receiver less vulnerable to the effects of DC offset and second-order intermodulation distortion, but more vulnerable to the gain and phase imbalance between I and Q channels [15]. One needs to make a trade-off among the achievable, compensated, *image rejection ratio* (IRR) and the performance of DC offset compensation circuits, along with IIP2 of the receiver, to decide a certain value of IF. As an example, > 32 dB of IRR is needed for an IF of 100 kHz and > 50 dB of IRR is needed for an IF of 200 kHz [18]. Let us assume that IRR is sufficient to allow an IF frequency of approximately 120 kHz. In this case, a *high-pass filter* (HPF) can be used at the RCF2 output to filter the residual DC offset left uncorrected by other analog/digital circuits. A lower corner frequency of HPF is desirable to make sure that the signal of interest is not filtered out. At the same time, the corner frequency should not be too low, otherwise it will take longer for the filter to settle and that will also result in reduced SNDR.

A digital I/Q mismatch compensation (IQMC) circuit is needed in a GGE receiver to improve the overall IRR. Without digital compensation, the IRR of the analog/RF receiver can be as low as 25 dB. An example I/Q mismatch compensation circuit is described in [18]. More details on receiver compensations can also be found in [Chapter 16](#).

After I/Q mismatch compensation, the VLIF signal can finally be down-converted to zero-IF. This is achieved in the digital domain by employing a complex mixer. As mentioned in [16], the IF frequency of the receiver can be made programmable for more flexibility. Note also that since the complex mixer in our example implementation runs in a fixed clock domain, a simple *look-up table* (LUT) based VLIF to zero-IF mixer can be used, instead of a CORDIC (COordinate Rotation DIgital Computer) based down-conversion

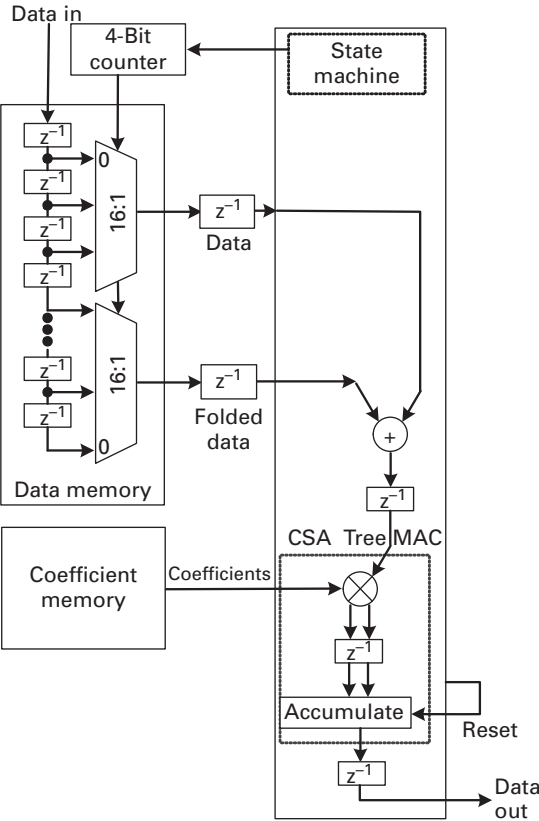


Figure 21.24 A symmetric, fully programmable, 32-tap FIR filter for channel selection [16] (© 2006 IEEE).

mixer. If the resampling were done after the down-conversion mixer, as for example in [2], then a down-conversion mixer using CORDIC would be a better choice. CORDIC is an iterative algorithm that is used for computing trigonometric, logarithmic, and other mathematical functions. An example of CORDIC implementation for Cartesian to polar coordinate conversion is described in [19].

The only filtering operation that now remains is the final channel selection, which primarily cleans the signal from adjacent interfering signals. Channel selection can be performed before VLIF to ZIF down-conversion, but that would entail a complex filter. Hence the best place to perform channel selection is at the end of the DFE chain. Although one may design a fixed-coefficient filter to perform channel selection, our experience shows that the flexibility that we achieve by having a fully programmable final channel selection filter outweighs the overhead in terms of area and power of making it fully programmable. The GSM/GPRS test chip presented in [16], for example, uses a cascade of three symmetric, fully programmable, 32-tap FIR filters to achieve final channel selection. The implementation details of one 32-tap filter are shown in Figure 21.24. In the later version of this chip, presented in [17], only a single, 64-tap, symmetric, fully programmable, channel selection filter was used.

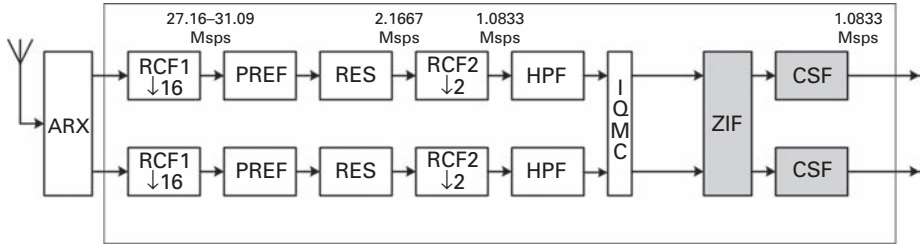


Figure 21.25 DFE in GGE mode.

Figure 21.25 shows a block diagram of the DFE for GGE mode. The block labeled ZIF represents the digital down-converter (composed of a LUT-based sine and cosine generator and complex multiplier [17]) for VLIF to ZIF conversion and the block labeled CSF is the final channel selection filter. Both ZIF and CSF are shaded gray in this figure. As will be shown in the next section, we do not need ZIF for WCDMA and it is better to use a separate CSF for WCDMA. All the other blocks can be shared with appropriate clocks and limited programmability.

WCDMA data path

We can, in principle, use a non-zero IF for WCDMA. One example would be to use an IF of 2.5 MHz, which would simplify DC offset correction circuits. Among other things, this would, however, require the ADC to have higher dynamic range in twice as much bandwidth, making ADC design more complex and power hungry. Similar to GGE, a non-zero IF would also require superior IRR performance. In our example, we will, therefore, assume that a WCDMA has been down-converted to zero-IF by the ARX mixer and hence there is no need for a digital down-conversion operation for WCDMA.

To compensate for residual DC offset left uncorrected by other analog/digital circuits, we can use an HPF at the RCF2 output. It is possible to design an HPF that supports both GGE and WCDMA modes. Both the clocks as well as its corner frequency can be made programmable.

Since WCDMA is a baseband signal, the image of the signal falls on itself and even a 25 dB of IRR might be enough for a WCDMA signal. One may, therefore, live without a digital IQMC circuit to support WCDMA. If the requirement on overall SNDR is higher, as for example for HSDPA signal, any distortion will result in lowering the overall SNDR. Hence it is important to improve IRR beyond 25 dB. We can use a circuit similar to the one for GGE, but operating at WCDMA data rates, to achieve higher IRR.

All that remains now is to perform final channel selection. We can use a channel selection filter (CSF), similar to GGE, to perform this operation. This CSF will, of course be running at 7.68 Msps, instead of 1.0833 Msps, and hence it will be more power hungry for the same number of coefficients. Our analysis shows that we need roughly an 18-tap filter for the CSF to achieve the overall filtering requirements determined in Table 21.3.

If we can achieve approximately 50 dB of attenuation at 2.7 MHz and more than approximately 60 dB of attenuation at 3–3.84 MHz, then all requirements will be met.

Figure 21.26 shows the frequency response of an 18-tap RRC filter (dotted line) and the combined response at the CSF output (solid black line). For comparison, response of a closer to ideal 64-tap RRC is also shown. The vertical gray line is at 2.7 MHz and it can be seen that approximately 50 dB attenuation is achieved at 2.7 MHz and more than 60 dB from 3–3.84 MHz.

Finally, Figure 21.27 shows the overall frequency response with LO of 869 MHz with an 18-tap CSF (same as for Figure 21.26) and it is obvious that all requirements are again met.

Although compensations are discussed in detail in Chapter 15, it is worth mentioning here that we can perform *digital automatic gain correction* (DAGC) prior to the CSF. This will result in a reduced number of bits going to the CSF and hence a lower area design of CSF. Figure 21.28 shows the DFE data path for WCDMA mode.

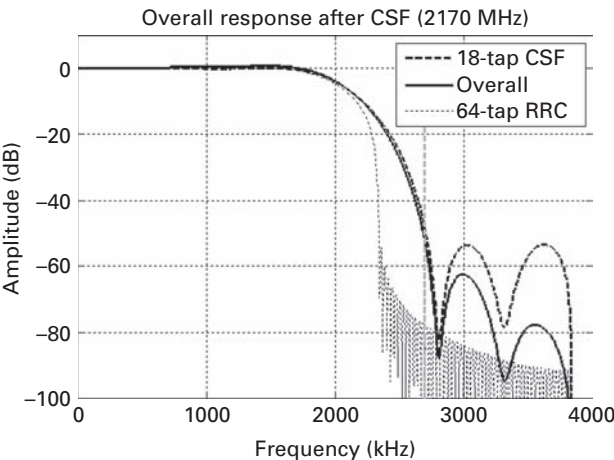


Figure 21.26 Frequency response with 18-tap CSF (LO of 2170 MHz).

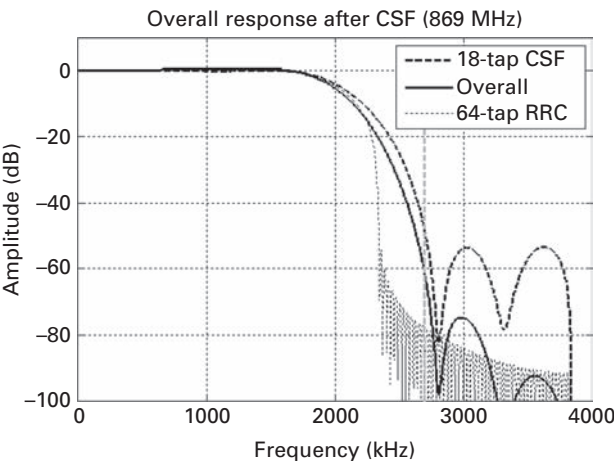


Figure 21.27 Frequency response with 18-tap CSF (LO of 869 MHz).

DFE to support both GGE and WCDMA modes

We can combine the two DFE block diagrams in Figure 21.25 and Figure 21.28 to come up with one DFE that supports both GGE and WCDMA. This is shown in Figure 21.29. There are separate data path blocks at the IQMC output. The blocks shown in gray color have programmable clocks, which means they will be designed to meet the timing requirements of the more stringent WCDMA mode. PREF and HPF can be designed to have limited programmability to support both GGE and WCDMA modes.

21.5 Considerations to support LTE

In the previous section, we demonstrated the design of a DFE data path to support both GGE and WCDMA modes. It is obvious that we need programmability and reconfigurability in order to support multiple wireless standards. This becomes more important to support LTE because, unlike GGE and WCDMA that support only one signal bandwidth, LTE currently supports six different channel bandwidths: 1.4 MHz, 3 MHz, 5 MHz, 10 MHz, 15 MHz, and 20 MHz [20]. While it is relatively straightforward to extend the discussion in the last section to support LTE in 1.4 MHz, 3 MHz, and 5 MHz modes, we need to carefully reconsider the design of the entire data path to support wider

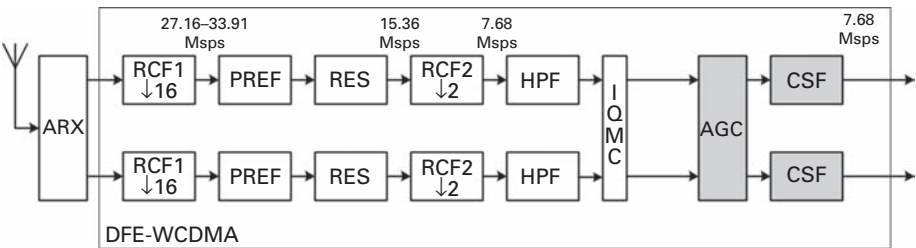


Figure 21.28 DFE in WCDMA mode.

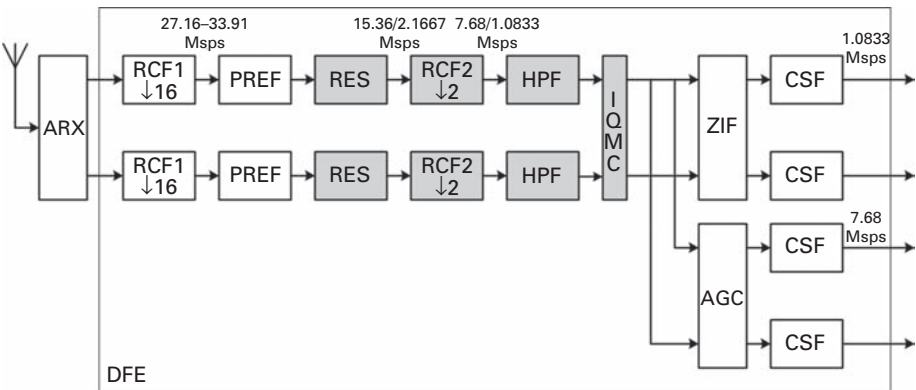


Figure 21.29 DFE supporting both GGE and WCDMA.

bandwidths, in particular the 20 MHz wide signals. In this section, we will mention few guidelines that can be used to design a DFE to support LTE in addition to WCDMA and GGE.

- First, the data rate at DBB will now have to be much higher to support 20 MHz bandwidth signal. Since it will be zero-IF, the signal will actually occupy -10 MHz to $+10\text{ MHz}$ frequency range. This means that the data rate at DBB will be at least 30.72 Msps (to make sure that it is still a good integer multiple of 3.84 Mcps), although 61.44Msps might be more desirable. Since the data rate at the resampler output in our example DFE of the [last section](#) can be as low as 27.16 Msps, we may have to change RCF1 to decimation by 8, instead of 16.
- As shown in [Figure 21.30](#), RCF1 can have an amplitude droop of approximately 10 dB for the case of 20 MHz channel bandwidth, assuming the same RCF1 as presented in the [previous section](#). This is on top of the droop (or amplitude distortion depending on the way ARXB is implemented) of ARXB filters. If left uncompensated, this amplitude droop will result in degraded receiver performance. Since the amplitude droop will vary with channel frequency for an ADC running from an LO derived clock, we may need a programmable filter to compensate for it or use CSF for both channel selection and droop compensation.
- Instead of a fixed decimation filter at ADC output, a filter with programmable decimation ratio is more desirable to support multiple channel bandwidths. Since CIC filters can be easily designed to support a programmable decimation ratio [21], we suggest the use of CIC at the front-end of the DFE chain. As was mentioned earlier,

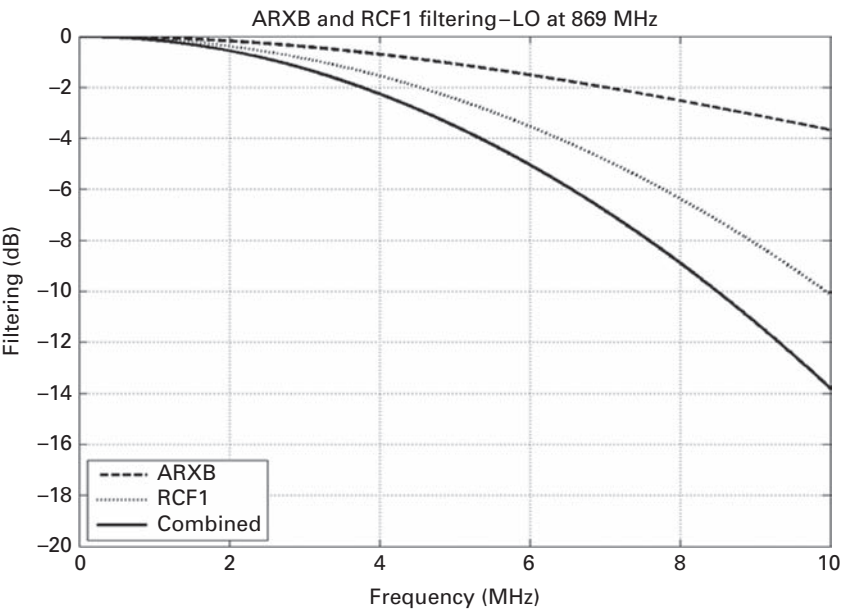


Figure 21.30 RCF1 response in 10 MHz showing approximately 10dB of droop.

CIC filters are also a better choice if the ADC output is more than 1-bit long. One may also consider using a fixed decimation ratio CIC followed by another CIC with programmable decimation ratio.

- One may also consider moving the resampler downstream. One advantage of zero-IF design is that we can, at least in principle, change the sequence of blocks without impacting the overall functionality. (The word lengths will need to be adjusted, however.) As an example, it is possible to perform channel selection filtering prior to performing the resampling operation. The CSF filter in this case will then also act as anti-aliasing filter for the resampler.
- Finally, more and more frequency bands are constantly being added in various wireless standards. As an example, release 9 of LTE supports FDD bands 1–14, 17–21, and TDD bands 33–40, for a total of 27 bands [20]. Although some of the bands share the same frequencies in the down-link path, such a large number of bands may require the ADC sampling frequency to be highly programmable. In our example DFE design for WCDMA and GGE for bands I–X (excluding band VII), we mentioned that the ADC clock is either LO divided by 2 (for low-bands) or LO divided by 4 (for high-bands). We will need more programmability on the LO divider ratio to support all the bands. As an example, it makes more sense to use an LO divider ratio of 5 to support band VII.

21.6 Conclusions

More and more wireless standards are continuously being introduced with drastically varying requirements and the task of designing a single DFE to address all, or even most, of them is quite daunting. Receiver programmability and reconfigurability are the two important requirements to design a DFE to address multiple wireless standards. In this chapter, we presented design details of a receiver DFE to support both WCDMA and GGE wireless standards. We also presented a few ideas to extend this design to cover LTE. To avoid having two PLLs in the receiver chain, one for the ARX mixer and another for the ADC and digital clocks, we presented a DFE running from ARX-LO driven clocks, entailing the use of a resampler to go from variable, LO driven, clock rate to a fixed, baud/chip-rate driven, clock rate. Although the discussion in this chapter focused on design of the DFE using partially programmable hardware blocks, the design of a specialized DSP processor to perform some of the relatively slower speed operations is also worth considering.

References

- [1] J. Mitola, “The software radio architecture,” *IEEE Communication Magazine*, vol. 33, no. 5, pp. 26–38, May 1995.
- [2] T. Hentschel and G. Fettweis, “The digital front-end: Bridge between RF and baseband processing,” in *Software Defined Radio: Enabling Technologies*, W. Tuttlebee, ed. Chichester: Wiley, 2002, pp. 151–198.

- [3] 3GPP Technical Specification Group Radio Access Network; User Equipment (UE) Radio Transmission and Reception (FDD), 3GPP TS 25.101, release 9, Sep. 2009.
- [4] D. Griffith, V. Srinivasan, and S. Pennisi, *et al.*, "A 28mW WCDMA/GSM/GPRS/EDGE transformer-based receiver in 45nm CMOS," IEEE Symposium on Radio Frequency Integrated Circuits (RFIC), pp. 9–12, May 23–25, 2010.
- [5] K. Muhammad, I. Elahi, and T. Jung, "A low-area decimation Filter for ultra-high speed 1-bit $\Sigma\Delta$ A/D Converters," Proceedings of 2005 IEEE Custom Integrated Circuits Conference, session 5.3, Sep. 18–21 2005.
- [6] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Englewood Cliffs, NJ: Prentice-Hall, 1993.
- [7] L. Erup, F. M. Gardner, and R. A. Harris, "Interpolation in digital modems. II. Implementation and performance," *IEEE Trans. Commun.*, vol. **41**, no. 6, pp. 998–1008, Jun. 1993.
- [8] C. W. Farrow, "A continuously variable digital delay element," Proc. IEEE Int. Symp. on Circuits and Systems, pp. 2641–2645, June 1988.
- [9] J. Vesma, "Timing adjustment in digital receivers using interpolation," MSc thesis, Tampere University of Technology, Tampere, Finland, Nov. 1995.
- [10] I. Elahi, and K. Muhammad, "A programmable order polynomial resampler in 90nm CMOS for a multi-standard radio," to be submitted to IEEE TCAS-II.
- [11] Y. Gao, J. Tenhunen and H. Tenhunen, "A fifth-order comb decimation filter for multi-standard transceiver applications", Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS), pp. III-89–III-92, May 2000.
- [12] E. B. Hogenauer, "An economical class of digital filters for decimation and interpolation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. **29**, pp. 155–162, Apr. 1981.
- [13] K. K. Parhi, *VLSI Digital Signal Processing Systems: Design and Implementation*, New York, John Wiley & Sons, 1999.
- [14] M. Shiraishi, "A simultaneous coefficient calculation method for INC^N FIR filters," *IEEE Trans. Circuits and Systems-I: Fundamental Theory and Applications*, vol. **50**, no. 4, pp. 523–529, Apr. 2003.
- [15] I. Elahi, "Robust receiver design using digitally intensive techniques to overcome analog impairments," Ph.D. dissertation, University of Texas at Dallas, December 2005.
- [16] K. Muhammad, Y.-C. Ho, T. Mayhugh, *et al.*, "The first fully integrated quad-band GSM/GPRS receiver in a 90-nm digital CMOS process", *IEEE Journal of Solid-State Circuits*, vol. **41**, no. 8, pp. 1772–1783, Aug. 2006.
- [17] K. Muhammad, C.-M. Hung, D. Leipold, *et al.*, "A low-cost quad-band single-chip GSM/GPRS radio in 90nm digital CMOS", IEEE Symposium on Radio Frequency Integrated Circuits (RFIC), pp. 197–200, Jun. 7–9, 2009.
- [18] I. Elahi, K. Muhammad, and P. T. Balsara, "I/Q mismatch compensation using adaptive decorrelation in a low-IF receiver in 90 nm CMOS process," *IEEE Journal of Solid-State Circuits*, vol. **41**, no. 2, pp. 395–404, Feb. 2006.
- [19] I. Elahi and K. Muhammad, "A low-complexity GSM base-band detector for RF BIST", IEEE Symposium on Radio Frequency Integrated Circuits (RFIC), pp. 599–602, Jun. 7–9, 2009.
- [20] 3GPP Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Transmission and Reception (FDD), 3GPP TS 36.101, release 9, Mar. 2010.
- [21] E. B. Hogenauer, "An economical class of digital filters for decimation and interpolation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. **29**, pp. 155–162, Apr. 1981.

22 Programming models and implementation platforms for software defined radio configuration

Tanguy Risset, Riadh Ben Abdallah, Antoine Fraboulet, and Jérôme Martin

22.1 Introduction

Software means programmable. Hence software defined radio means that the radio should now be *programmable*. We know what computer programming means, and we agree, up to a certain level, on how it should be done. But do we know what programming a radio means? Several questions are still open: what will an SDR platform look like in ten years? Will there exist *software radio code*? What will be the technical challenges and commercial issues behind this code?

Programming is more precise than configuring or tuning, it implies a much greater level of freedom for the programmer. But it also means much cheaper implementations in many cases and in particular a re-use of the same hardware for different protocols (i.e. with different programs). This is, to our point of view, the main difficulty of software radio programming: reconfiguration and in particular dynamic reconfiguration. Dynamic (i.e. very fast) reconfiguration is now mandatory because some protocols, 3GPP-LTE (Third Generation Partnership Program Long Term Evolution) for instance, propose channel adapting for *each frame*, requiring a setting of the channel estimation parameter in a few milliseconds.

In this chapter we will try to give an overview of the technical difficulties of designing a programming environment for software defined radio. Then we will present one particular solution which aims at defining a virtual machine dedicated to the domain of software defined radio.

22.2 Programming environment and tools for SDR

In this section we present existing SDR platforms and give an insight into how they are programmed or configured. This brief review leads to the concept of waveform description language developed in [Subsection 22.2.2](#) and reveals the need of a middleware dedicated to SDR programming ([Subsection 22.2.3](#)). Finally we introduce the Radio Virtual Machine (RVM) concept, which is one possible choice for SDR middleware, in [Subsection 22.2.4](#). Our proposal for such a radio virtual machine will be detailed in [Section 22.3](#).

22.2.1 Hardware platforms for SDR

Unlike for desktop computers, software radio hardware is not yet standardized. The only common point among all SDR hardware platforms is the complexity of their architecture. They are usually composed of many *processing components*, dynamically reconfigurable, and interconnected with fast communication links. The source of this complexity is, of course, the needed processing power and the need to implement possibly *all* protocols. Software radio attempts to implement in software most of the radio processes that are initially hardwired. In practice, digital treatments are implemented by algorithms that require many giga-operations per second (GOPS) to meet protocols' real-time constraints. For instance, a turbo-decoding may require 150 gops which is far from what embedded processors can achieve.

This problem is solved with the use of dedicated processing elements: FFT, turbo decoder, FIR filters, digital predistortion and matrix inverse for instance. These components, also named IP for *intellectual property*, help to achieve the required computing power. Another solution to this problem is to use digital signal processors (DSP), dedicated to stream processing. Note that the complexity of software radio treatment keeps on increasing. For instance, the 3GPP-LTE protocol provides a data throughput which can reach 300 Mbps. Hence it is very likely that the inherent difficulty of building a software radio hardware platform will remain for a while.

As mentioned in [16], most of the existing SDR platforms are prototypes built by public or private research laboratories. Indeed, SDR platform design is a real challenge for hardware architects as well as for software developers: the good trade-off among computing power, power consumption and flexibility (i.e. programmability) is very hard to find.

We present some existing SDR platforms by classifying them in two categories:

1. *DSP centric platforms* that use only software components (DSP, GPP, etc.), and hence are highly flexible but must usually be associated to hardware IPs to meet real-time performance.
2. *Heterogeneous platforms* that try to mix dedicated hardware and software components.

DSP Centric Platforms

Many companies (Sandbridge, picoChip, Fujitsu, Icera, Infineon, NXP, etc.) propose integrated circuits based on DSPs, here are some examples:

- The PicoArray processor [14] from picoChip. This circuit integrates hundreds of small processors. PicoArray can be programmed in ANSI C with a dedicated programming environment. Global computing power can reach 200 GOPS. Associated to some selected IPs (FFT, turbo-decoder, etc.), this circuit is able to implement a complete W-CDMA modem.
- X-GOLD™ SDR-20 from Infineon technologies is a signal processing processor for baseband processing for multi-standard mobile phones. Infineon proposes hardware/software solutions with this platform that support recent protocols (GSM, W-CDMA, LTE, ...).

- EVP (Embedded Vector Processor) [35] from NXP: this architecture is able to support various modes of the LTE protocol.
- In [25], Lin *et al.* present a DSP-based system composed of four SIMD (single instruction multiple data) vector processors. This architecture can realize two different processing chains: IEEE802.11a and W-CDMA.
- Tan *et al.* [34] present an original approach for baseband treatments realized on general purpose processors. They implement IEEE802.11a/b and g physical layers using multi-core architectures.

It is important to realize that although a number of important *software* platforms are mentioned in the literature, many of them have a limited computing power and a bad power consumption, they also need to be associated to dedicated hardware IPs.

Heterogeneous platforms

These platforms integrate dedicated IPs, usually controlled by a processor. Experimental version may contain FPGAs for implementing recent signal processing algorithms.

- Small Form Factor (SFF SDR) Development Platform from Lyrtech: this board embeds a DSP and an FPGA, for baseband processing, connected to a RF board. A development environment is given for programming this machine.
- Universal Software Radio Peripheral (USRP): is a hardware platform conceived for the GNU Radio project [33]. It connects to a computer with a USB interface.
- Kansas University Agile Radio (KUAR) platform [27]: is an experimental SDR platform that includes a Pentium M processor and a Xilinx Virtex2 FPGA. The board connects to a PC either through a gigabit Ethernet interface or through a PCI-express link.

Even from the small platform list presented here, it seems clear that the work of developing a new protocol for each of these platform is a huge task. Expressing this protocol in an existing language like C or Java will not help because of the granularity of the basic operators used in SDR platforms (e.g: FFT operation). Moreover, there is a real challenge in expressing in a high-level language the dynamic reconfiguration which is hardly realized in hardware but will necessarily appear soon. We really need a way to describe protocol physical layers (so-called waveform processing) that can be understood by all the existing SDR platforms. This gave rise to the concept of *waveform description language*.

22.2.2 Waveform description language

A waveform description language (WDL) is simply a programming language dedicated to the expression of the physical layer of a communication protocol: how bits are modulated and transmitted on the antenna for emission and the reverse way for reception. It basically describes the waveform that will be sent in the air for a given packet to be communicated. The design of a waveform description language is of course highly connected to the existence of at least one SDR platform able to implement this physical

layer. Meanwhile, the long-term objective is to have a common WDL for all existing SDR platforms.

Many works have pointed out the difficulties in expressing waveforms [20, 24, 36, 37]. Here is a summary of the properties that a waveform description language should try to respect:

- Be a formal language. Waveform specifications based on large textual documents are often error-prone for developers, the waveform specification should be *compiled* on each SDR platform.
- Implement a clear and extensive Hardware Abstraction Layer (HAL) adapted to most existing SDR platforms. As for the processor, a clear hardware abstraction layer will ease the adaptation to different SDR hardware architectures. It also clarifies what is the assumption on the targeted SDR platform and helps in writing specifications that are independent of any hardware platform.
- Use of a component-based model which is well adapted to SDR. Many protocols can be expressed by connecting components: FFT, Viterbi decoder, scrambler, etc. These common operators should be easily identified into the language.
- Use an object-oriented programming paradigm to ease the mapping between hardware components and software objects semantic.
- Follow the “Write once, Run anywhere” philosophy, which is a slogan created by Sun to describe benefits of the Java virtual machine. This gave rise to the idea of radio virtual machine (RVM).

We now briefly present some attempts that have been made to realize waveform description language. These works are results of academic research; it is worth noting that there is also a military project titled *Advanced Transmission Language and Allocation of New Technology for International Communications and Proliferation of Allied Waveforms (ATLANTIC PAW)* whose goal is to provide a unique standard for expressing waveforms.

Wilink Waveform Description Language

In [36], Wilink presents a waveform description language named WDL, proposed within the *Programmable Digital Radio (PDR)* in UK. It is a behavioral system description: a hierarchical block decomposition using state machines within boxes.

Waveform Description Language (WDL) uses a combination of principles defined in various research domains: graphical interface concepts found in block diagram languages such as Ptolemy, Cossap, SPW. Hierarchical state machines, used for instance in Argos or SDL are used together with synchronization mechanisms inherited from synchronous languages like Esterel. Data types manipulated are defined in object-oriented languages such as Java or C++.

Figure 22.1 gives an insight of how a waveform is expressed in WDL. It describes a part of the FM^3TR waveform which is an international test waveform initially developed by the Air Force Research Labs.

Figure 22.1 is a graphical representation of an entity: a transmission module. This transmission module is itself refined: it is a state machine, each state being itself a block diagram. At the lowest granularity, computations are expressed in Java-like syntax.

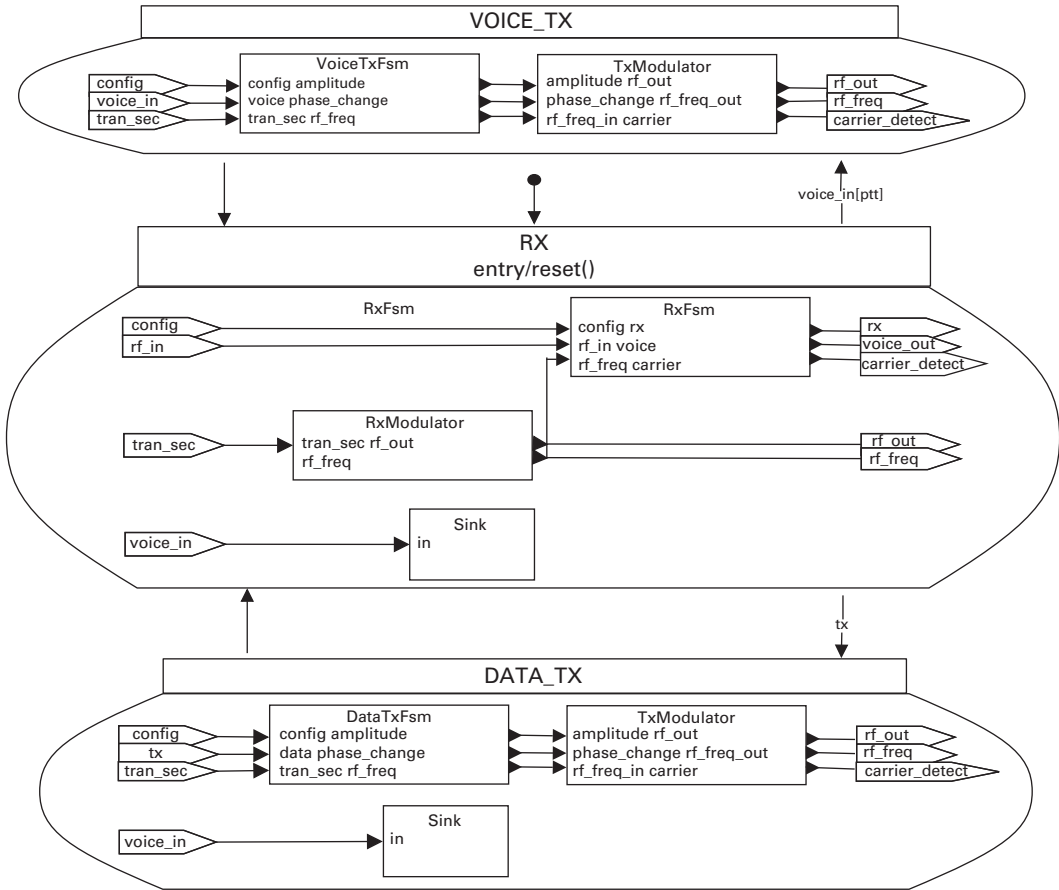


Figure 22.1 Example taken from [36] a WDL description of the FM^3TR physical layer.

Each WDL block gets an internal schedule, hand-shake mechanisms are used to synchronize with input/output data. The refinement occurs up to a granularity level that is suited to the available library provided by the development framework associated with WDL. The WDL framework has to be provided for each target architecture. This language does not permit dynamic reconfiguration of hardware. To our knowledge there is no real implementation of WDL showing performance issues.

VANU RDL (Radio Description Language)

Radio Description Language (RDL) is a waveform description language developed by VANU, Inc. Chapin *et al.* present its foundation in [10]. A radio application is described by an oriented graph where nodes are basic signal processing operators provided in the form of software libraries. An RDL *interpreter* is an execution environment that realizes the required processings, it maps the description graph to the targeted platform by configuring the available hardware. The interpreter can also use software blocks, i.e. software implementation of signal processing primitives.

```

module RxConvDecoder {
    parameter EncoderFormat format;
    dataout GsmFrame output;
    datain GsmFrame input;
}

assembly RxConvDecoderDef
implements RxConvDecoder {
    module ConvDecoder convDecoder;
    module DecoderFrameGenerator frameGen;
    module DecoderStreamGenerator streamGen;

    // data flows
    streamGen.mProtectedOutput -> convDecoder.input;
    streamGen.mUnprotectedOutput -> frameGen.mUnprotectedInput;
    streamGen.mHeaderOutput -> frameGen.mHeaderInput;
    convDecoder.output -> frameGen.mProtectedInput;

    // link ports
    frameGen.mOutput -> output;
    input -> streamGen.mInput;
}

```

Figure 22.2 Example of a RDL graph specification.

RDL defines two basic elements:

- *Modules*: basic signal processing operators defining the nodes of the graph.
- *Assembly*: i.e. graphs, composed of modules and sub-graphs.

Common types used in signal processing languages such as ports, channels, and streams are available in RDL.

An example is illustrated in [Figure 22.2](#). This assembly uses three modules and each of these modules can be itself refined in another assembly or directly implemented by an available library primitive.

An implementation of the GSM protocol has been realized using RDL [8]. The RDL program required 36 modules and 28 assemblies representing approximately 3200 lines of code. Again dynamic reconfiguration cannot be achieved with this language.

E²R FDL (Functional Description Language)

The abbreviation E²R stands for European research project *End to End Reconfigurability* which aims at developing architectures of reconfigurable communicating systems. Burgess *et al.* [6] have studied the possibility of defining a language able to unify waveform specifications.

The E²R project has defined a software architecture for radio equipment. This architecture, referred to as “Configuration and Control Architecture,” is shown in [Figure 22.3](#). It isolates three abstraction levels: (i) hardware abstraction, (ii) system abstraction, and

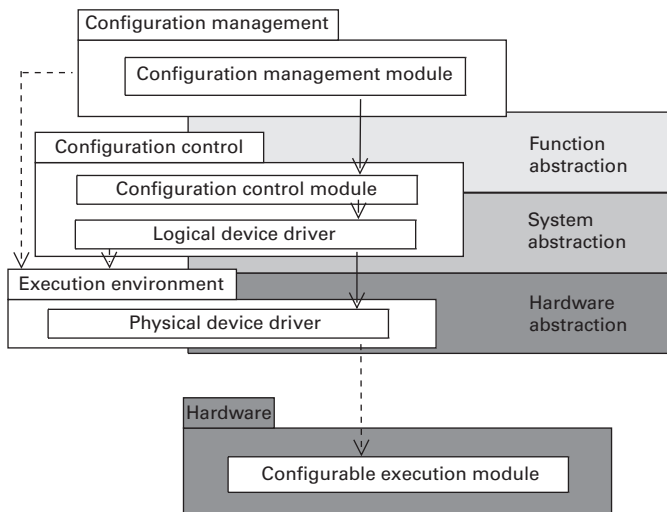


Figure 22.3 Configuration and control architecture of E²R radios.

(iii) function abstraction. This work is interesting because it highlights the difficulty of the waveform description language definition and implementation.

The FDL (Functional Description Language) defined in the E²R project is based on XML and basically proposes a hierarchical composition of components as does RDL. Again, this structural description is useful for describing different protocols but cannot implement dynamic reconfiguration.

In [38], Zhong *et al.* have implemented in software an IEEE802.11a emitter using FDL. They show that the FDL program size is not a big problem but that the execution performance is much worse than with dedicated hardware.

SPEX Language

The SPEX language is a programming language for SDR developed by the Michigan University SDR group. This language is dedicated to DSP-centric platforms and more precisely takes advantage of SIMD DSP. A SPEX description is split in three abstractions: Kernel SPEX, Stream SPEX, and Synchronous SPEX.

- Kernel SPEX is an imperative language supporting native dsp arithmetic operations. It is used to define kernel signal processing algorithms (e.g. FIR, FFT, etc.).
- Stream SPEX describes assembly of kernel programs: connections with channels and sequential scheduling.
- Synchronous SPEX allows parallel construction and synchronization with real-time constraints.

The SPEX language is interesting but is clearly oriented to pure software implementation: code is compiled and mapped on a multi-core architecture. It is not clear yet whether these architectures will succeed for SDR.

22.2.3 **Middleware for SDR programming**

The usage of middleware in networked application is now widely accepted. Middleware requires the definition of standardized interfaces enabling heterogeneous distributed software components to communicate. Middleware is the software that implements an intermediate abstraction between applications and different execution platforms, and usually provides a higher level API than the one provided by the HAL level. Software radio, considered as a particular field of software programming, needs a specific middleware for radio applications. We present here two attempts that have been made to define an SDR middleware.

Software Communication Architecture (SCA)

The SCA architecture [22, 5] was initially conceived in the American military program JTRS (Joint Tactical Radio Systems) for the development of SDR. It is now considered as a middleware dedicated to hardware SDR platforms.

The SCA environment is composed of three main elements: the *Core Framework*, a Corba Object Request Broker, and a real-time *operating system*. This environment is represented in Figure 22.4.

The SCA framework is the most popular software radio software architecture. Many companies developing middleware for SDR provide implementation of SCA on different hardware platforms, such as for instance Prismtech, Zeligsoft, and OIS. However, it is obvious that the use of the Corba specification has an important impact on performances and makes this system not well adapted to actual commercial communication protocols.

UPC radio software framework

Gelonch *et al.* from Polytechnical University of Catalonia (UPC) propose in [19] the P-HAL framework: Platform and Hardware Abstraction Layer. The P-HAL framework abstracts hardware radio platforms by functional services and these services can be categorized in three classes:

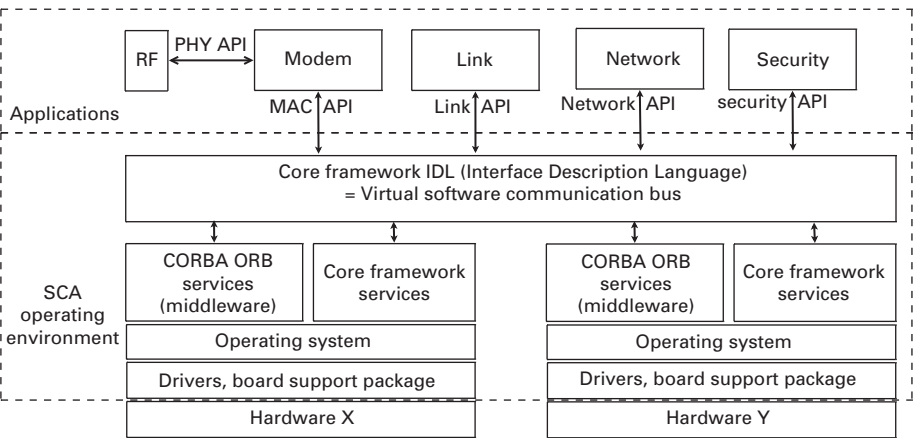


Figure 22.4 Software architecture of SCA.

- Real-time control of radio processes.
- Exchange of data between different processing elements.
- Parameter setting and supervision of functional modules.

By providing a time division in slots, the P-HAL environment tries to introduce a simpler real time radio programming environment. In [18], Gomez *et al.* present a comparison between P-HAL and SCA on a DSP platform and claim much better performance for P-HAL.

22.2.4 Radio virtual machine concept

From what we have seen above, we would like to see a software defined radio system as a software layer offering application programming interfaces (API) enabling the definition of waveform, and their implementation on a hardware platform. In 2000, Gudaitis and Mitola proposed in [17] to apply the virtual machine concept to software defined radio, using the term *radio virtual machine* (RVM).

Java virtual machines have allowed the expansion of Java applications. Performance weaknesses of bytecode execution have led to the proposal of the just-in-time compiler (JIT) that bridged the performance gap. According to Gudaitis and Mitola a radio virtual machine is a particular virtual machine (VM) with its own programming language, which we call *source code*, that can be compiled into *bytecode* for the VM. This VM will, as is the case for Java, provide a common hardware abstraction for software and firmware/hardware developers. This would split the radio application development cycle in two (possibly parallel) phases:

- Software development of the radio application in VM bytecode, common to all platforms.
- Platform specific development to optimize execution of the VM on each particular targeted platform.

Gudaitis and Mitola enumerate a list of important goals that an RVM should achieve:

- Provide a programming language that permits an easy expression of physical layers of most protocols and that can be compiled into an executable form (bytecode).
- Provide an abstraction based on the component model paradigm.
- Avoid Java virtual machine pitfalls: provide mechanism to handle real-time constraints and easy access to hardware.
- Include an arbitrary bit-width arithmetic.

Some patents have been approved or pending [15], [28], [7] but the field is still wide open for innovation.

22.3 An existing radio virtual machine implementation

This section presents a particular prototype RVM implementation that has been realized during R. Ben Abdallah's Ph.d. Thesis [1], within a collaboration between three French

institutes: CEA, Inria, and Insa-Lyon. The goal of this work was to explore the technical viability of using a virtual machine for radio application on a real SDR hardware platform, namely the Magali [9, 11] chip developed at the CEA Leti laboratory.

As we have frequently noticed above, the main difficulty of SDR programming relies on the dynamic reconfiguration: such a system should be able to configure some of its components within a few hundred microseconds (e.g. within the same frame for LTE advanced protocol). The reconfiguration process is generally triggered depending on system status and external events such as radio channel impairments. This issue has to be taken into account in the design of an SDR programming model.

22.3.1 Waveform programming model

The approach that we have proposed is the following: we introduced a two steps programming model called *reconfigurable khan process network* that formalizes the reconfiguration phase and separates it from the computation phase. Then we propose a computation model which should support most of the existing heterogeneous platform. The main restriction of this computation model is that it is not well adapted to massively parallel systems where scheduling is done within each of its processing elements: our model requires a centralized scheduler.

Computation model

Khan process networks (KPN) were introduced in the seventies by Gilles Kahn [23]. The KPN is a distributed computation model with a precise semantic. In KPN, an application is a set of sequential processes communicating through channels which are blocking FIFOs. It is now widely used for modeling signal processing systems. In a KPN, a process cannot test whether a channel is empty or not, it cannot *choose* which channel to read from. This is not an important restriction for modern signal processing programs except during reconfiguration where the channels between processes are changing. The reconfigurable khan process network (RKPN) computation model is composed of sequential processes connected by blocking FIFO (i.e. as for a KPN), and its behavior alternates between computation phases (i.e. standard KPN computation) and configuration phases during which channels between processes are allowed to change. Figure 22.5 shows the configuration phase and the following computation phase of a RKPN.

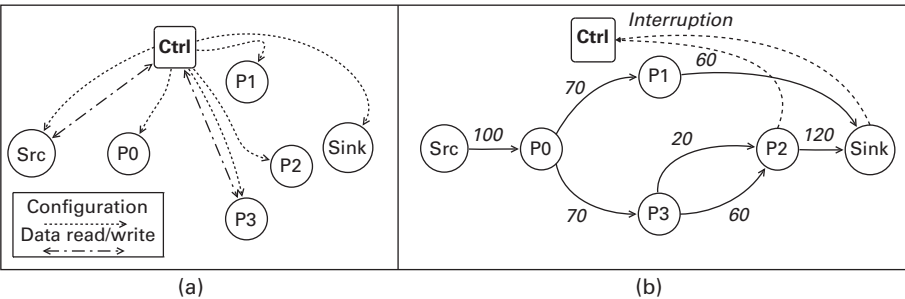


Figure 22.5 Example of a RKPN, during reconfiguration (a) and during computation (b).

We have added the following constraints to this model in order to match with the SDR application and hardware platforms:

- There is a particular node that we call *controller* that controls the configuration of the KPN. This node cannot be the source or sink of an FIFO but can receive interruptions from other nodes. This node can also, during reconfiguration phases, access to other nodes' memory and of course reconfigure other nodes and connections between them.
- During each computation phase, each node knows in advance the *number* of data units that it should process. We call that a *static control program*, it helps in optimizing implementation and is usually not an important restriction for telecommunication protocols (as opposed to multi-media algorithms), as soon as reconfiguration is available.
- For convenience, one usually adds two particular nodes called *source* and *sink* to materialize the world outside the SDR (e.g. RF front-end on one side and higher OSI level protocols on the other side).

Note that this computation model does not mention the use of a virtual machine, it is adapted to other SDR middleware solutions. It takes into account the reconfiguration requirements of SDR applications. Another novelty is that the data of the signal processing stream can have an impact on the configuration and control of the system. Indeed, the controller can access data stored in nodes (Figure 22.5(a)), possibly do some computations with them, and finally configure the network. This is useful for instance when implementing advanced channel adapting algorithms.

Execution model

The Execution model is an abstraction of the hardware platform on which will be executed the SDR program. It is important to emphasize the fact of having a component-based model for all IPs of the platform. We impose that an IP should be a *radio component* and have the following interfaces (a radio component model is shown in Figure 22.6):

- Configuration interface: for tuning functional parameters of the IP.
- Communication interface: for the main input/output data stream.

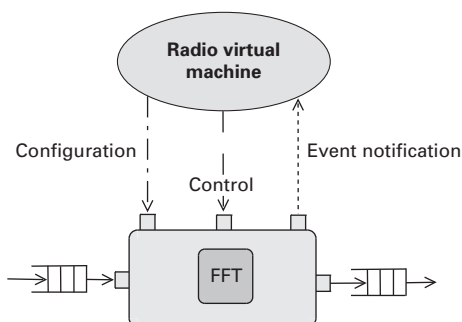


Figure 22.6 Example of radio component: FFT.

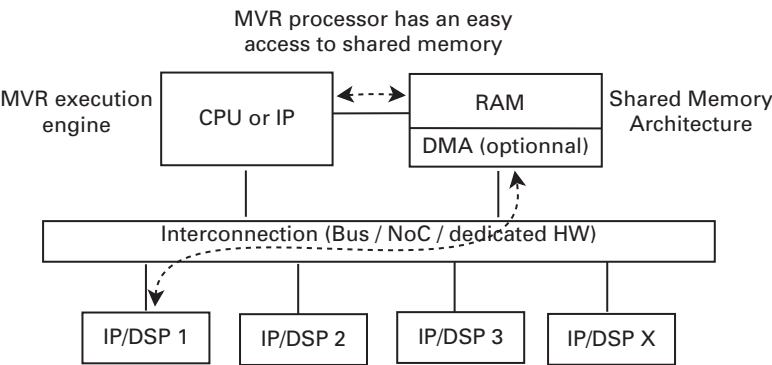


Figure 22.7 Execution model adapted to RKPN.

- Control interface: for being started, stopped, checked, etc. by the controller.
- Notification interface: for notifying the controller of particular events (corresponds to an interrupt mechanism).

Note that such a radio component can be a programmable device (DSP or GPP) or a dedicated IP (FFT, matrix inversion, etc.).

The proposed execution model is simply a set of radio components, such as defined above, interconnected by an efficient communication mechanism and associated to a particular IP that is the controller (GPP or dedicated IP). The execution platform will of course contain some memory elements (RAM). We choose the shared memory to communicate data between IPs and the controller but other communication mechanisms could be envisaged. Experience shows that direct memory access modules (DMA) are necessary to achieve acceptable communication performance. An example of target execution platform architecture is shown in [Figure 22.7](#).

22.3.2 Physical Layer Description Language

The Physical Layer Description Language (PLDL) is the programming language that is proposed in [1] to describe physical layer protocols. Its main component is what we call the RVM API, which is a set of primitives and data structures dedicated to the RVM concept.

The PLDL is adapted to the RKPN programming model and it is executed by the controller which: (i) allocates and frees radio resources, (ii) configures radio components, (iii) controls the execution of components, and (iv) accesses data stored in components. The PLDL program is platform independent; it can be executed on most SDR platforms as well as on a desktop PC. One simply has to provide an implementation of the RVM API on each targeted platform. We use an object-oriented syntax for describing the RVM API; in the following **rvm** basically represents the controller and we describe its possible actions as methods of the **rvm** class.

Here is a brief description of the primitives of the RVM API:

Radio resources allocation/release

- *comp_desc* **rvm.allocate** (*comp_type*): Allocates a hardware component or creates an instance of a software component. This method returns a descriptor of the allocated component: *comp_desc*.
- **rvm.free** (*comp_desc*): Releases the hardware component or frees the software component designed by the descriptor *comp_desc*.

Radio component configuration

- *core_config_desc* **rvm.build_core_config** (*comp_desc*, *core_param_list*): Builds a native configuration for the component *comp_desc* using specified functional parameters. Returns a configuration descriptor.
- *com_config_desc* **rvm.build_com_config** (*comp_desc*, *com_param_list*): Builds a communication configuration for the concerned component (for instance: IP input/output controller configuration). Returns a configuration descriptor.
- **rvm.free_config** (*com_config_desc* || *core_config_desc*): Releases the memory used for the specified configuration.
- **rvm.configure** (*comp_desc*, *core_config_desc*, *com_config_desc*, *event_type*): Configures the component specified by *comp_desc* using configurations stored in RVM memory: *core_config_desc* and *com_config_desc*. The parameter *event_type* specifies if the component must send a notification to the RVM for a particular event.
- **rvm.connect** (*comp_desc_src*, *port_num_src*, *comp_desc_dest*, *port_num_dest*, *data_type*): Interconnects source component output port with destination component input port (i.e. configures a FIFO). The parameter *data_type* may be useful to configure communication channels.

Radio component execution control

- **rvm.start** (*comp_desc*): Activates the behavior of a component previously configured. Note that some hardware components are implicitly triggered by data arrival, but this is not the case for the RVM software components.
- **rvm.stop** (*comp_desc*): Stops the behavior of the specified component.
- **rvm.wait** (*event_type*, [*comp_desc_event_source*]): Blocks the execution of the RVM until an event notified by a component arrives (i.e. an interruption arrives). Optionally the source of the expected event could be specified.

The PLDL also includes methods for manipulating data of the data flow. In this case, it is up to the RVM programmer to check for memory consistency using the synchronization method **rvm.wait**. This also implies some technical choices such as a global memory addressing scheme.

Data flow access methods

- *coarse_data_struct* **rvm.read** (*mem_ptr*, *size*): Copies a data block of size *size* stored at address *mem_ptr* into the RVM local memory. Returns a pointer to the raw data.

- *rvm_data_table* **rvm.convert2rvm** (*coarse_data_struct*, *data_type*): Converts raw data from data flow to a data type understandable by the RVM language (*data_type*).
- *coarse_data_struct* **rvm.convert2raw** (*rvm_data_table*, *data_type*): Converts data from an understandable data type by the RVM language to a raw format specified by *data_type* (inverse of **rvm.convert2rvm**).
- **rvm.write** (*coarse_data_struct*, *mem_ptr*): Copies the data from the local RVM memory to the system memory at address *mem_ptr* (inverse of **rvm.read**).

Figure 22.8 gives an insight of what might be a waveform description using PLDL. For a detailed description of the language, refer to [1].

```
[.....]

print("\n-----\nDATA FIELD DEMODULATION\n-----\n")

-- (0) Initialization --
Ndbps, Ncbps, Nbpsc, coding_rate, modulation = initparam802_11a(rate)

-- (1) Create required instances --
scra1 = scrambler.allocate()

-- (2) Connect the modules --
rvm.connect( vite1, 1, scra1, 1, binary_type )
rvm.connect( scra1, 1, dma2, 1, binary_type )

-- (3) Configure the modules --
param0 = ext_symbol_size
param1 = 16
param2 = symbol_size
param3 = nsym
dma_engine.configure(dma1, "RECEIVE ext_symbol_size; DESTROY 16",
                    NO_IT, param0, param1, param2, param3 )

phase_drift = phase_drift + 64*phase_amount

rotor.configure(rotol, phase_drift, phase_amount,
               nsym*symbol_size, NO_IT)
fft.configure(fft1, mode_fft, fft_size, nsym*symbol_size, NO_IT)
equalizer.configure(equal, coef, nsym*symbol_size, NO_IT)
constellation.configure(cons1, Ncbps, Nbpsc, nsym*symbol_size, NO_IT)
deinterleaver.configure(dein1, Ncbps, Nbpsc, nsym*Ncbps, NO_IT)
depuncturer.configure(depul, coding_rate, nsym*Ncbps, NO_IT)
viterbi.configure(vite1, nsym*Ndbps, NO_IT)

[.....]

-- (4) Launch modules --
rvm.start( dma1 )
rvm.start( roto1 )
rvm.start( fft1 )
rvm.start( equal )

[.....]

-- (5) Wait for result --
rvm.wait( SIGTER )
print("\nProcess terminated. RVM wakes up.\n")

[.....]
```

Figure 22.8 Example of a waveform description using PLDL formalism corresponding to (part of) the data field demodulation phase of the IEEE802.11a protocol.

Table 22.1 Comparison of virtual machine candidate to implement a RVM.

	Lua	Neko	Python	Squawk	Kaffe	LeJOS	TinyVM	NanoVM	Waba
small memory footprint	x	x		x	x	x	x	x	x
performance	x	x		x		x	x	x	x
extensibility	x	x	x						
memory size	x		x	x	x	x	x	x	x
documentation	x		x	x	x	x			

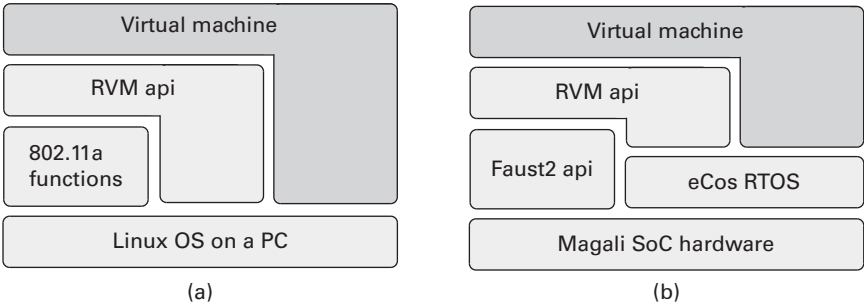


Figure 22.9 Software architecture of the Radio Virtual Machine, on a personal computer (a) and on the Magali Chip (b).

22.3.3 RVM implementation issues

There exist a lot of virtual machines, including light versions of Java VM (e.g. Squawk [32], JVM[26]). Java might be a good choice provided the targeted platform implements the Jazelle processor [Por05] otherwise Java VMs are too complex. Table 22.1 summarizes the pro and cons of the VM we have isolated as potential candidates for being integrated into an embedded system such as a software defined radio system. We have chosen the Lua virtual machine [21] because it has been conceived to be light, embedded, and easily extensible to define domain specific languages.

We have first implemented as a proof of concept, the RVM API on a standard PC, on which a software implementation of the IEEE802.11a protocol was available [2]. In this case, allocating a component would create a thread and call a function corresponding to the computation done by the component. The resulting software architecture is shown in Figure 22.9(a).

An implementation of the RVM was then realized on the Magali Chip. Magali is a system on chip (SoC) developed at the Cea Leti, also called Faust2 as being the second generation of the Faust SoC [13]. Magali is dedicated to physical and MAC layers of 4th-generation telecommunication protocols such as 3GPP-LTE or IEEE802.16e (WiMax). It is composed of an asynchronous network on chip with a 2D-mesh topology, with each router of the network being connected to the components of the system.

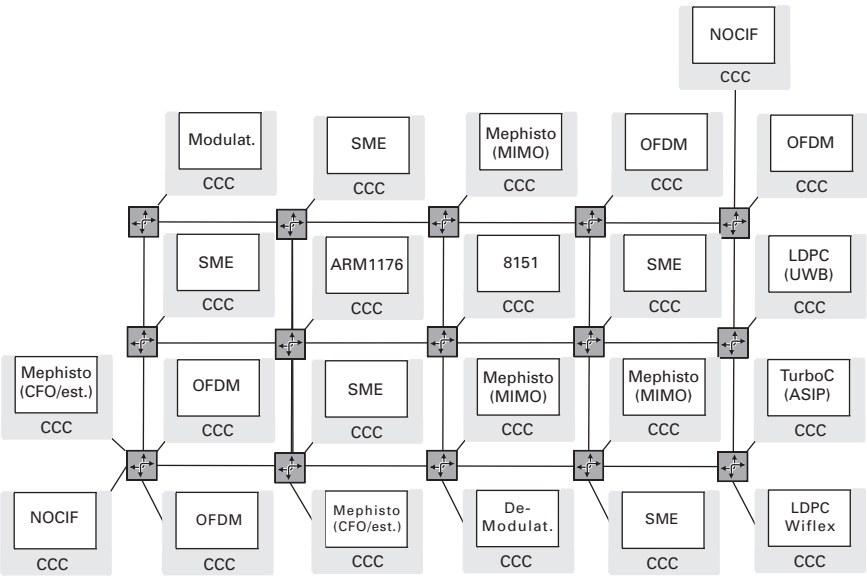


Figure 22.10 The Magali system on chip.

Figure 22.10 presents the different components available on the chip. These components are dedicated IPs necessary to realize OFDMA processing: FFT, LDPC, turbo-decoder, etc. In addition to that, Magali contains the following specific components:

- The *smart memory engine* (SME) is a programmable DMA extensively used to move data between components. All the memory available on the chip is accessed through these SMEs.
- The Mephisto processor is a VLIW DSP used to realize efficiently digital signal processing algorithms: channel estimation, MIMO decoding, digital predistortion, etc.
- The ARM1176 is a general purpose processor used for the global control of the Magali platform as well as for MAC layer processing. In our case this processor will be used as the RVM global controller described previously, i.e. the Lua virtual machine will be executed on this processor. This processor can also access directly to the memory without going through an SME.
- Associated to each router is a dedicated component called CCC (Communication and Configuration Controller) that is used to regulate data exchange between components and their configurations.

As the reader can see the Magali platform is quite complex and it is very difficult to precisely present all the implementation work that occurred for having a radio application running on our radio virtual machine. We simply enumerate the different stages of the implementation and then we will compare this implementation with a native (i.e. handcrafted) implementation of the same application on Magali.

The reader must be aware that an important choice was made: we decided to implement a *soft* virtual machine. The Lua VM has been ported on the ARM processor (with

eCos OS). Another possibility would have to use a *hard* virtual machine, i.e. a dedicated processor for bytecode interpretation. But this would mean building another chip.

The [first part](#) of the work was to adapt the RVM API presented in [Section 22.3.2](#) to the Magali platform. The Magali programming environment provides an API called F2 API, the RVM API encapsulates this native API. Then, we ported the Lua virtual Machine to the ARM processor so that programs such as the one presented in [Figure 22.8](#) can be executed on the Magali platform. Finally we wrote the PLDL program for a 3GPP-LTE receiver configured according to a particular operating mode in order to validate and experiment our RVM prototype.

22.3.4 Performance results for a CFO IEEE802.11a

The Carrier Frequency Offset (CFO) represents the phase shift between emitter and receiver clocks due to hardware imperfections. First, we have implemented the CFO error correction algorithm present in the IEEE802.11a protocol. Then, we have measured the memory footprint and the execution time required by the CFO application for each of the three configurations of implementation on the Magali chip: (i) native implementation (hand-coded, as it was done without RVM), (ii) with RVM API, or (iii) with the virtual machine. [Figure 22.11](#) illustrates these three configurations in the case of an FFT operator.

The performance evaluation results have been obtained on the cycle accurate VHDL simulator of Magali chip. These results are depicted as follows:

The sizes of the executable programs corresponding to the three configurations (native, RVM API and RVM) are presented in [Table 22.2](#).

The execution time is presented in [Figure 22.12](#). It is clear that the overhead of the RVM API is not very important. This proves that our PLDL fits well with the Magali platform which is a real SDR platform, whereas adding a virtual machine to enable portability has an important cost: the memory footprint is doubled and the execution time

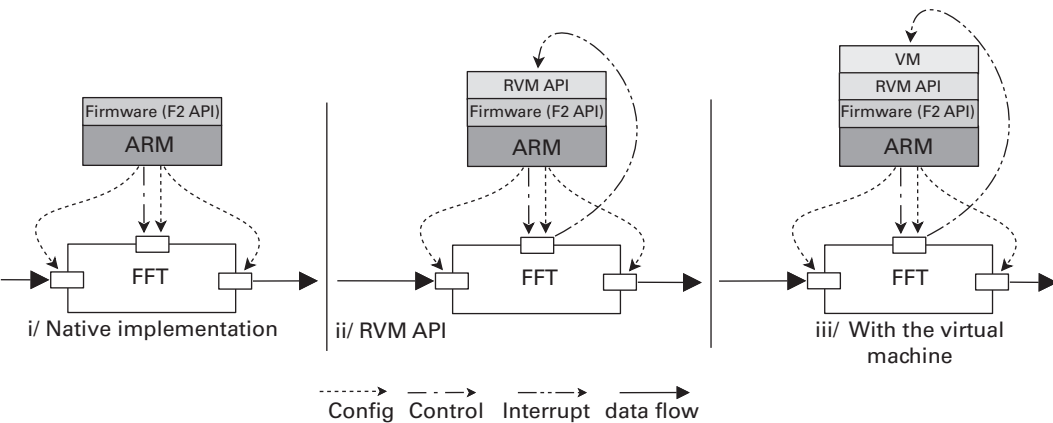


Figure 22.11 Different possibilities for executing a FFT on the Magali platform once the RVM was realized.

Table 22.2 Size of the programs for the three configurations of the CFO of IEEE802.11a.

Configuration type	Memory footprint (kB)
Native	96
RVM API	100
Complete RVM	212

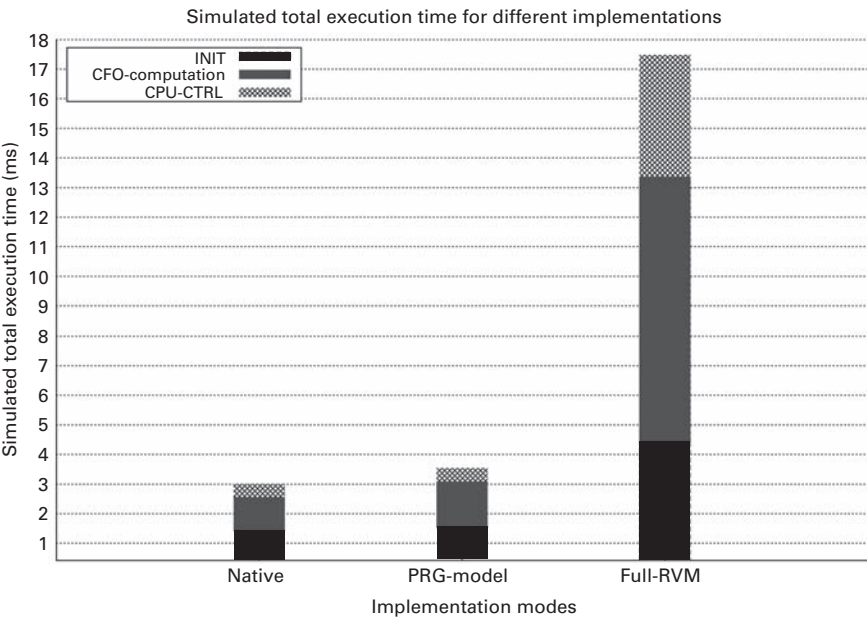


Figure 22.12 Execution time for the three configuration of the CFO of IEEE802.11a.

is almost multiplied by ten. For a more precise presentation of the performance results, see [3]

However, this VM implementation is a prototype and can be optimized much more. Analysis of the execution time shows that most of the time overhead is spent in native function calls, which imply access to a hash function, and data type transformation between Lua types and native types. These performance problems can be reduced with the use of well-known techniques such as runtime compilation or binary translation techniques [4, 12]. Another possibility would be to use hardware accelerators for the chosen virtual machine such as Hard-Int [31], Jazelle [29] or picoJava[30].

Reference [1] presents a complete 3GPP-LTE receiver implemented on top of the RVM. Real-time constraints were not met in this functional demonstrator of the RVM, but as mentioned above, virtual machine optimization techniques can be applied. In any case it is very important to realize that optimization mechanisms present on existing SDR platforms should be taken into account when implementing a waveform description language for a given platform, in order to meet real-time and power consumption constraints.

22.4 Conclusion

In this chapter we have briefly presented the technical problems that occur when trying to program a software defined radio system. The complexity and variety of today's hardware SDR prototypes highlight the need for an abstraction layer dedicated to software define radio systems. We have mentioned some attempts to give a common format for software radio programs, then we have investigated more precisely the concept of virtual machine for software radio which seems to be very promising.

Radio virtual machine is attractive because its goal fulfills the requirements mentioned above: writing one software radio program which executes on every software radio platform. In the [second part](#) of this chapter, we have presented a radio virtual machine prototype developed on the Magali SDR platform at the Leti laboratory. This experiment shows that a radio virtual machine powerful enough to reach real-time performances required by modern telecommunication protocols must be optimized from the very beginning of its realization. We have proposed to use just-in-time compilation and binary translation techniques during RVM design and implementation. More generally, memory management and data representation issues should be carefully studied during RVM design.

It is very likely that, as it has been the case for computers or parallel machines, the software tool-chain available on the SDR system will have a huge technical and economical impact on the future of communicating objects. Depending on the available computing power, the software models and techniques presented here will not only apply to baseband but also to the digital processing part in the front-end such as digital pre-distortion, digital up-conversion and down-conversion for instance.

References

- [1] R. Ben Abdallah. Machine virtuelle pour la radio logicielle. *Thèse de doctorat (Ph.D. Thesis), Laboratoire CITI (Centre d'innovation en télécommunication et intégration de services), INSA de Lyon*, 2010.
- [2] R. Ben Abdallah, T. Risset, A. Fraboulet, and Y. Durand. The radio virtual machine: A solution for SDR portability and platform reconfigurability. *Parallel and Distributed Processing International Symposium*, pages 1–4, 2009.
- [3] R. Ben Abdallah, T. Risset, A. Fraboulet, and J. Martin. Virtual machine for software defined radio: Evaluating the software vm approach. In *International Conference on Computer and Information Technology*, pages 1970–1977, 2010.
- [4] J. Aycock. A brief history of just-in-time. *ACM Computing Surveys (CSUR)*, **35**(2):113, 2003.
- [5] J. Bard and V.J. Kovarik. *Software Defined Radio: The Software Communications Architecture*. Wiley, 2007.
- [6] R. Burgess and S. Mende. Configuration languages-theory and implementation. E2R Project Whitepaper: <http://e2r.motlabs.com/whitepapers>.
- [7] R. Burgess. Configuration method, September 28 2004. US Patent App. 10/950,562.
- [8] J. Chapin and V. Bose. The vanu software radio system. In *2002 Software Defined Radio Technical Conference, San Diego*, 2002.

- [9] F. Clermidy, C. Bernard, R. Lemaire, *et al.* A 477mW NoC-based digital baseband for MIMO 4G SDR. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, pages 278–279. IEEE, 2010.
- [10] J. Chapin, V. Lum, and S. Muir. Experiences implementing GSM in RDL (The Vanu Radio Description Language™). In *IEEE Military Communications Conference, 2001. MILCOM 2001. Communications for Network-Centric Operations: Creating the Information Force*, volume 1, 2001.
- [11] F. Clermidy, R. Lemaire, Y. Thonnart, X. Popon, and D. Knetas. An open and reconfigurable platform for 4g telecommunication: concepts and application. In *12th Euromicro Conference on Digital System Design, Architectures, Methods and Tools (DSD'2009)*, pages 449–456, August 2009.
- [12] C. Cifuentes and V. Malhotra. Binary translation: static, dynamic, retargetable? In *Proceedings of the 1996 International Conference on Software Maintenance*, pages 340–349. Citeseer, 1996.
- [13] Y. Durand, C. Bernard, and D. Lattard. FAUST: On-chip distributed architecture for a 4g baseband modem SoC. In *Design & Reuse IP-SoC, Grenoble, France, Grenoble, France*, December 2005. IEEE Computer Society.
- [14] A. Duller, D. Towner, G. Panesar, A. Gray, and W. Robbins. Picoarray technology: the tool's story. In *Proceedings of Design, Automation and Test in Europe, 2005*, pages 106–111, vol. 3, March 2005.
- [15] G. R. Ferris. Digital wireless basestation, July 24 2002. US Patent App. 10/182,043.
- [16] R. Farrell, M. Sanchez, and G. Corley. Software-defined radio demonstrators: An example and future trends. *International Journal of Digital Multimedia Broadcasting*, article ID 547650, 2009.
- [17] M. Gudaitis and J. Mitola III. The radio virtual machine. *SDR Forum 21st General Meeting. 14–16 November, Mesa, AZ, 2000*, 2000.
- [18] I. Gomez, V. Marojevic, J. Salazar, and A. Gelonch. A lightweight operating environment for next generation cognitive radios. *Euromicro Symposium on Digital Systems Design*, pages 47–52, 2008.
- [19] A. Gelonch, X. Revès, V. Marojevik, and R. Frrús. P-HAL: a middleware for SDR applications. In *SDR Forum Technical Conference*, 2005.
- [20] C. Grassmann, M. Sauermann, H.-M. Bluethgen, and U. Ramacher. System level hardware abstraction for software defined radios. In *Proceeding of the SDR 04 Technical Conference and Product Exposition. SDRForum 2004*, 2004.
- [21] R. Ierusalimsky, L.H. De Figueiredo, and W.C. Filho. Lua-an extensible extension language. *Software Practice and Experience*, **26**(6):635–652, 1996.
- [22] JTRS. Software communications architecture specification. standards joint program executive office (jpeo) joint tactical radio system (jtrs), 2006. version 2.2.2.
- [23] G. Kahn. The semantics of a simple language for parallel programming. *Information processing*, **74**:471–475, 1974.
- [24] T. Kempf, M. Adrat, E.M. Witte, *et al.* A Concept for Waveform Description based SDR Implementation. In *4th Karlsruhe Workshop on Software Radios (WSR'06)*, 2006.
- [25] Y. Lin, H. Lee, M. Woh, *et al.* SODA: A high-performance DSP architecture for software-defined radio. *IEEE Micro*, **27**(1):114–123, 2007.
- [26] T. Lindholm and F. Yellin. *Java Virtual Machine Specification*. Addison-Wesley Longman, 1999.

- [27] G. J. Minden, J. B. Evans, L. Searl, *et al.* Kuar: A flexible software-defined radio development platform. In *2nd IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, 2007. DySPAN 2007*, pages 428–439, April 2007.
- [28] C. Moy, A. Kontouris, and A. Bisiaux. Telecommunication device with software components, May 2007. US Patent 7,212,813.
- [29] C. Porthouse. Jazelle for execution environments. ARM Whitepaper, available online, 2005.
- [30] W. Puffitsch and M. Schoeberl. PicoJava-II in an FPGA. In *Proceedings of the 5th international workshop on Java technologies for real-time and embedded systems*, page 221, ACM, 2007.
- [31] R. Radhakrishnan and L. K. John. Microarchitectural techniques to enable efficient Java execution. Academic Dissertation. University of Texas at Austin, 2000.
- [32] D. Simon, C. Cifuentes, D. Cleal, J. Daniels, and D. White. JavaTM on the bare metal of wireless sensor devices: the squawk Java virtual machine. In *Proceedings of the 2nd international Conference on Virtual Execution Environments*, page 88, ACM, 2006.
- [33] D. C. Tucker and G. A. Tagliarini. Prototyping with gnu radio and the usrp - where to begin. In *Southeastcon, 2009. SOUTHEASTCON '09. IEEE*, pages 50–54, March 2009.
- [34] K. Tan, J. Zhang, J. Fang, *et al.* Sora: High performance software radio using general purpose multi-core processors. *NSDI 2009*, 2009.
- [35] K. van Berkel, F. Heinle, P. P. E. Meuwissen, K. Moerman, and M. Weiss. Vector processing as an enabler for software-defined radio in handheld devices. *EURASIP Journal of Applied Signal Processing*, 2005:2613–2625, 2005.
- [36] E. D. Willink. The waveform description language: moving from implementation to specification. In *IEEE Military Communications Conference, 2001. MILCOM 2001. Communications for Network-Centric Operations: Creating the Information Force*, volume 1, 2001.
- [37] S. Yoo and A. Jerraya. Introduction to hardware abstraction layers for SoC. *Embedded Software for SoC*, pages 179–186.
- [38] S. Zhong, C. Dolwin, K. Strohmenger, and B. Steinke. Performance evaluation of the functional description language in a SDR environment. In *Proceedings SDR Forum Technical Conference 2007*, 2007.

23 Programmable discrete-time integrated receiver front-ends

Norm Filiol, Robert Klett, and Thomas Riley

Traditional continuous-time receivers face issues of high area and power requirements and low selectivity, as outlined in the following [Section 23.1](#), when used in integrated software defined radios. Low selectivity puts increased demands on the analog-to-digital converters used in software defined radio receivers.

Organized into six sections, this chapter is devoted to programmable, sampling-based radio receivers. Early sampling combined with discrete-time filters can reduce the power required for analog-to-digital conversion by increasing selectivity while maintaining easy programmability and reduced power consumption. The required theory to understand such sampling filters is outlined in [Section 23.2](#). In each of [Sections 23.3](#) and [23.4](#), we look at the advantages and disadvantages, associated challenges, and the state of the art of programmable zero intermediate frequency and low intermediate frequency discrete-time receivers, respectively. As an exercise, we review a case study of an integrated AM/FM super-heterodyne receiver in [Section 23.5](#). Finally, in [Section 23.6](#), we present a summary and conclusions.

23.1 Background

For most analog filter design, it is useful to consider the biquadratic filter which has the general transfer function of:

$$H(s) = \frac{n_2 s^2 + n_1 s + n_0}{s^2 + s \frac{\omega_0}{Q} + \omega_0^2} .$$

Any filter can be synthesized with a cascade of biquadratic filters with carefully chosen constants. For the examples used in this section, the general low-pass equation is used with $n_2 = n_1 = 0$, giving:

$$H(s) = \frac{n_0}{s^2 + s \frac{\omega_0}{Q} + \omega_0^2} .$$

23.1.1 Active resistor–capacitor filters

Due to the impracticalities of integrating inductors on silicon at frequencies below 1 GHz, higher-order on-chip filters are usually designed using resistors, capacitors, and active gain elements. One such filtering technique is known as active resistor–capacitor (R – C). This active analog filter typically consists of a series of biquadratic sections made of operational amplifiers, resistors, and capacitors. Figure 23.1 presents an example of a second-order low-pass transfer function implemented using a Tow–Thomas biquadratic active R – C filter.

These filters have a large variance in their frequency response, as the pole positions are set by resistor and capacitor values whose absolute errors are not correlated. Supply voltage variations should have minimal effect on the pole and zero locations as long as amplifier gain remains high enough for the negative feedback path to dominate the frequency response. Temperature variations, on the other hand, compound the effect of process variation and mismatch between the R and C components. This results in the need to include a tuning circuit that calibrates the filter with reference to an accurate time constant.

Linearity of active R – C filters is dominated by the linearity of the resistors and capacitors, as long as the DC gain and bandwidth of the operational amplifiers are sufficiently large.

Noise is dominated by the noise voltage from the resistors at the operational amplifier's inputs and the operational amplifier itself. The noise contributed from the input resistors can be reduced by using smaller resistors, as noise voltage density is proportional to resistance, but this requires larger capacitors to maintain the same time constants. The noise contributed by the operational amplifiers is dependent on the size and current consumption of the devices and larger devices and higher currents are required to reduce noise.

Recent results [1, 2, 3] present multimode low-pass filters which can be programmable with respect to power consumption and performance, transfer function and cut-off frequency. These filters achieve corner frequencies as high as 20 MHz.

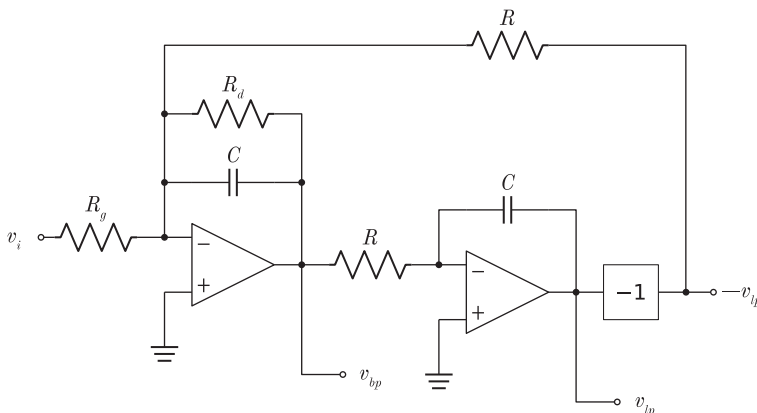


Figure 23.1 Schematic for a Tow–Thomas biquadratic active R – C filter where, $R_d = Q R$, $R_g = \frac{R}{2-Q}$, and $C = \frac{1}{\omega_0 R}$.

23.1.2 Switched-capacitor filters

Switched-capacitor filters overcome the time constant variation of active R – C filters by implementing time constants using capacitor ratios and highly accurate clocks.

Figure 23.2 shows a way to implement a resistor equivalent using a *switched capacitor*. The result is $R \approx T_c/C_R$, where $T_c = 1/f_c$ is the clock period and C_R is the value of the switched capacitor. This changes the time constant calculation from $1/RC$ to $C_R f_c/C$. Figure 23.3 presents the switched capacitor equivalent of the filter shown in Figure 23.1.

Modern semiconductor processes can achieve capacitor matching as accurate as one percent or less [4, 5]. In most applications, this means that pole and zero locations can be set accurately across process, voltage, and temperature variations without the need for calibration circuits.

Switched-capacitor circuits use discrete-time techniques resulting in the need for a sampling clock rate equal to or greater than two times the highest expected frequency, as will be discussed in Section 23.2.

Switched-capacitor filters are typically limited by the slew rate, bandwidth and DC gain of the operational amplifiers. If enough gain and bandwidth are achieved in the amplifiers, linearity is limited by the sampling switches and capacitors in the input and

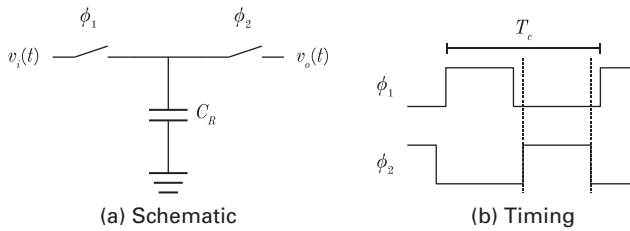


Figure 23.2 Switched-capacitor implementation of a resistor.

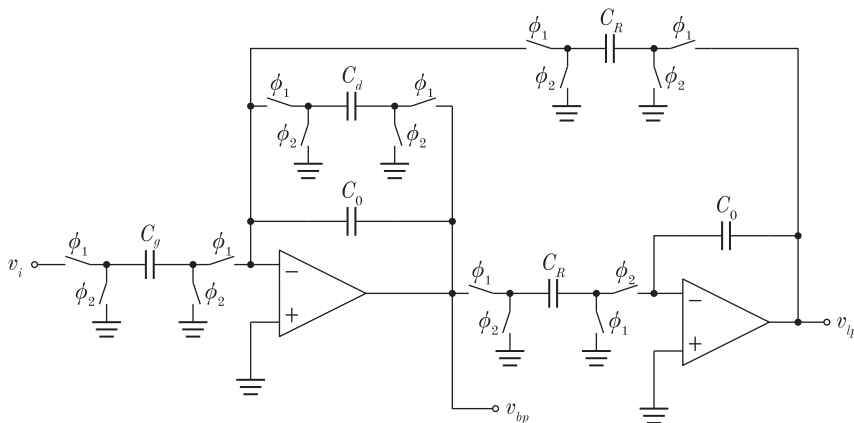


Figure 23.3 Switched-capacitor version of a low-pass Tow–Thomas biquadratic filter, where $C_R = \omega_0 T_c C_0$, $C_d = \frac{\omega_0 T_c C_0}{Q}$, and $C_g = \frac{A_V Q}{\omega_0 T_c C_0}$.

feedback paths. Switched-capacitor filters can have issues with respect to clock feed-through and a continuous-time filter is required prior to the first sampler to prevent aliasing. Noise is dominated by the contributions of components at the input, with the contribution from the capacitors being kT/C . In order to obtain low-noise samples, large capacitors are required. This, combined with the need to reduce noise contributions from the amplifiers, results in high op-amp current consumption and limits the frequency of operation.

A recent publication [6] presents the use of switched-capacitor techniques to implement a 23-tap finite impulse response (FIR) bandpass anti-aliasing filter in a high intermediate frequency, 2.4 GHz WLAN receiver.

23.1.3 Transconductor–capacitor filters

This is a type of analog filter where active transconductance (g_m) elements and capacitors are used to form various filter building blocks such as integrators and active resistors. It is also possible to synthesize gyrator-based inductors whose inductance values depend on the ratio between the transconductance and capacitance values. A gyrator is a circuit that transforms an impedance at its output to an inverted impedance at its input.

Figure 23.4 presents the second-order filter from the previous examples, implemented using g_m – C techniques.

These filters achieve low noise and can achieve low power consumption. However, they face a trade-off between power consumption, bandwidth, and linearity. To achieve high linearity requires increasing current consumption or utilizing multi-transistor, linearizing circuits which introduce parasitic poles. Another problem with these filters is that the transconductance value must be calibrated across process, voltage, and temperature corners to achieve high accuracy. Considering this, g_m – C filters can operate at higher frequencies than other filter types because capacitive loads do not limit the rate of change of an output current and thus do not add a parasitic R – C pole to the frequency response.

A recent publication [7] presents a low-pass filter, which uses a gyrator-based L – C ladder, for a direct conversion multimode receiver, tunable from 100 kHz to 2.7 MHz (GSM, Wideband CDMA) with a tuning accuracy of five percent achieved using a

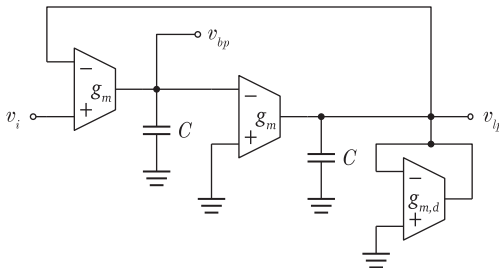


Figure 23.4 Schematic for a g_m – C low-pass Tow–Thomas biquadratic filter with $g_m = \omega_0 C$ and $g_{m,d} = \frac{\omega_0 C}{Q}$.

master–slave approach. A similar paper [8] presents a low-pass filter tunable from 2.5 MHz to 12 MHz used in a dual channel, triple mode WLAN receiver.

23.2 Theory

This section presents the basic theory needed to understand the discrete-time techniques in this chapter. Detailed descriptions of these concepts are beyond the scope of this work and more information can be found in [9, 10].

23.2.1 Sampling theory

Nyquist theory states that in order for an analog signal to be uniquely represented in discrete time, it must be sampled at a rate which is at least two times faster than the highest frequency component of the signal being sampled. Figure 23.5 shows the spectrum of a sampled signal. It can be seen that the sampled data is replicated around integer multiples of the sampling frequency, f_s , and that the signal is only unique in a bandwidth equal to one half of the sampling rate.

This theory is based on ideal impulse sampling, where the sampled output, $y(t)$, is the multiplication of the input signal and an infinite train of impulses, represented by the Dirac delta function [10]:

$$y(t) = \sum_{n=-\infty}^{\infty} x(t)\delta(t - nT_s) \quad (23.1)$$

where $x(t)$ is the input signal, $\delta(t)$ is the Dirac delta function, and T_s is the sampling period.

If $x(t)$ is given as:

$$x(t) = A \sin(2\pi f_{in}t) \quad (23.2)$$

where A and f_{in} are the amplitude and the frequency of the input sinusoid, respectively, then we have (23.3), after substituting back into (23.1).

$$y(t) = A \sum_{n=-\infty}^{\infty} \sin(2\pi f_{in}t)\delta(t - nT_s). \quad (23.3)$$

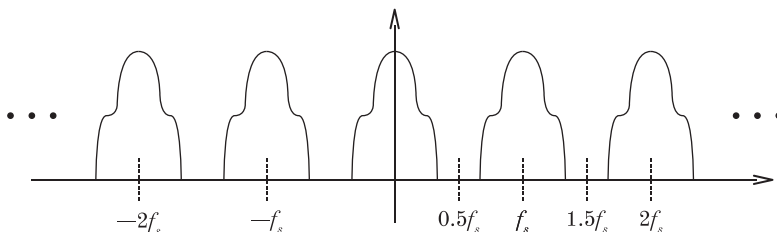


Figure 23.5 Frequency domain representation of a signal sampled at f_s .

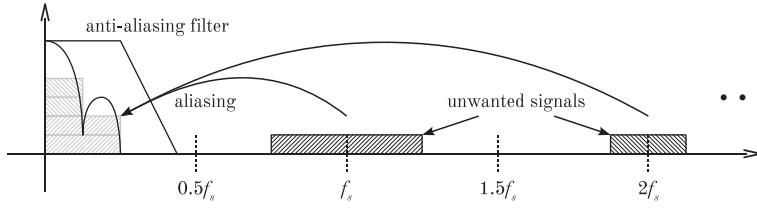


Figure 23.6 Demonstration of anti-aliasing filter.

Taking the Fourier transform of (23.3) we obtain [10]:

$$Y(f) = \frac{A}{2fT_s} \sum_{k=-\infty}^{\infty} [\delta(f - f_{in} - kf_s) - \delta(f + f_{in} - kf_s)]. \quad (23.4)$$

This results shows that impulse sampling in the time domain results in infinite replicas of the input signal in the frequency domain, as in Figure 23.5.

23.2.1.1 Aliasing

Referring to Figure 23.5, it can be seen that any signals that lie between $Nf_s/2$ and Nf_s will be replicated at baseband after sampling. This replication is called aliasing. In order to prevent these undesired signals from replicating at baseband, a continuous-time filter is required before the sampler to eliminate signals above $f_s/2$ or in some bandwidth around multiples of f_s . Such a filter is known as an anti-aliasing filter. This concept is illustrated in Figure 23.6.

In an integrated radio receiver, such a filter is needed before the analog-to-digital converter (ADC) and is usually implemented using g_m -C or active R -C filtering techniques.

23.2.2 $\mathcal{Z}$ -transform

The $\mathcal{Z}$ -transform converts a discrete time signal, $x(n)$, to a complex plane representation given by:

$$\mathcal{Z}\{x(n)\} = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \quad (23.5)$$

where $z = e^{j2\pi f / f_s}$ [9].

The $\mathcal{Z}$ -transform serves a similar function as the Laplace transform, which is used to describe *continuous-time* signals in the frequency domain [11], except the $\mathcal{Z}$ -transform is used to describe *discrete-time* signals in the frequency domain.

For example, for a finite duration discrete-time signal $x(n) = [1 \ 2 \ 4 \ 1 \ 3]$, the $\mathcal{Z}$ -transform is $X(z) = 1 + 2z^{-1} + 4z^{-2} + 1z^{-3} + 3z^{-4}$.

As the $\mathcal{Z}$ -transform can be used to describe any discrete-time signal, it can also be used to describe a discrete-time filter by describing the filter's impulse response as a transfer function. The frequency response of the discrete-time filter will replicate around multiples

of f_s , but we will assume a continuous-time anti-aliasing filter, as in Figure 23.6, makes only the baseband replica of the discrete-time signal filter important. A bandpass sampler or filter can also be achieved, provided that a continuous-time bandpass filter is available to filter the undesired aliases. The order of a discrete-time filter is determined by the value of the largest exponent present in the filter transfer function. For example, if the previously described $X(z)$ were the transfer function of a discrete-time filter, that filter would be fourth order.

A useful tool for evaluating discrete-time filters is the z -plane, which can be used to indicate the positions of poles and zeros of a discrete-time transfer function on the unit circle in the complex plane [10]. The unit circle is shown in Figure 23.7. Figure 23.8 shows the relationship between the z -plane and the s -plane and poles and zeros in the time and frequency domains.

23.2.3 Infinite impulse response filters

Infinite impulse response (IIR) filters have an impulse response of infinite duration. If an impulse is applied to the input of an IIR filter, the output will be non-zero for an infinite amount of time afterwards. An example of a continuous-time IIR filter is an R - C filter. An example of a discrete-time IIR filter is an accumulator-based integrator.

Infinite impulse response filters are less computationally intensive than their finite impulse response counterparts for a given transition band, as IIR filters have the ability to place both poles and zeros and FIR filters can only place zeros. However, as there is some form of feedback or memory in an IIR filter, stability may need to be considered.

23.2.4 Finite impulse response filters

Finite impulse response (FIR) filters have an impulse response that is finite in duration. If an impulse is applied to the input, the output of the filter will change for a finite amount of time after which it will return to zero. The output of the filter is zero at all other times. An example of a discrete-time FIR filter is a comb filter. An example of a continuous-time FIR filter is a *boxcar*, or windowed integrator, filter.

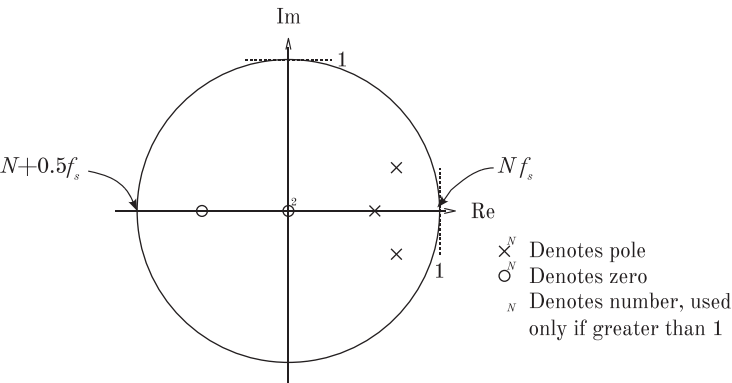


Figure 23.7 The unit circle with the poles and zeros of a low-pass filter depicted on the z -plane. The unit circle is described as $z = e^{-j2\pi f/fs}$.

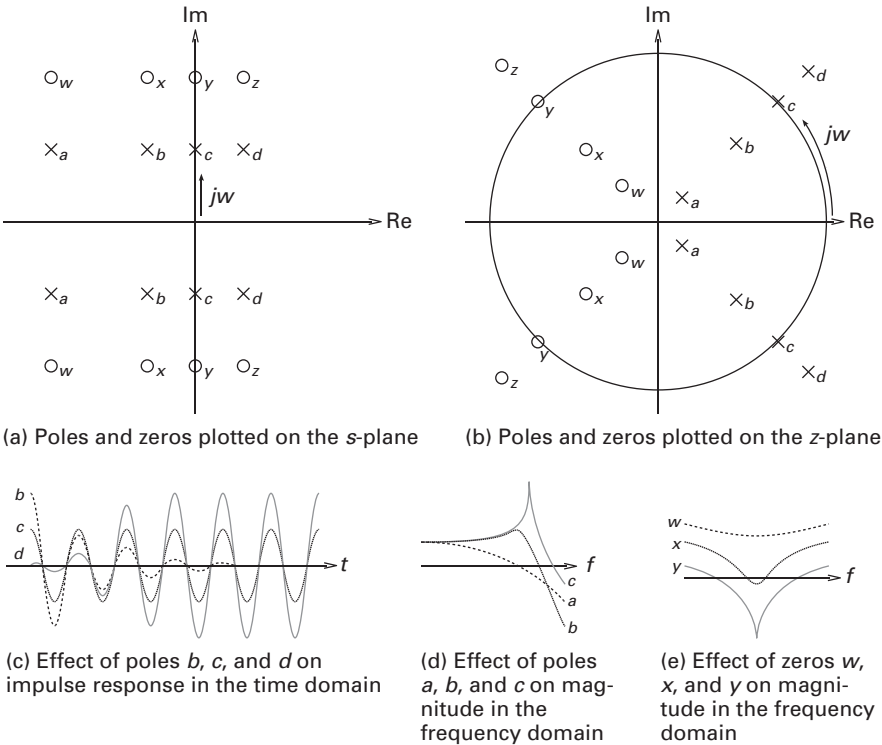


Figure 23.8 Plots showing the relationship of the s -plane and z -plane and the effects of zero and pole placement.

Finite impulse response filters can be designed to have linear group delay and tend to be less sensitive to errors in tap coefficients, compared to IIR filters. However, a large number of taps is required to achieve highly selective responses, when compared to IIR filters. Finite impulse response filters can achieve more rapid transition bands than IIR filters. However, stop-band attenuation is limited to a finite value. A block diagram of an example FIR filter can be found in [Figure 23.9](#) and the transfer function for the filter can be found in (23.6). The coefficients, a_0 to a_4 , are called *tap weights*.

$$H(z) = \sum_{n=0}^{n=4} a_n z^{-n}. \tag{23.6}$$

23.2.5 Sensitivity comparison

When looking at discrete-time FIR and IIR filters, it is useful to perform an intuitive sensitivity analysis.

In an IIR filter, each pole location is controlled by a ratio of two or three capacitors or an R – C time constant. Errors in these ratios or time constants translate directly to errors in

the pole locations. Each of these errors is independent. Pole-zero and transfer function plots for an IIR filter are shown in Figure 23.10, including the effects of capacitor error on pole locations. Getting the average pole locations correct helps somewhat, but it can be seen from the transfer functions that these errors directly affect the signal passband of the filter.

Figure 23.11 presents the generic equation for an N -tap FIR filter and the z -plane summing of the terms, a_i . The transfer function, found in (23.7), has different magnitude

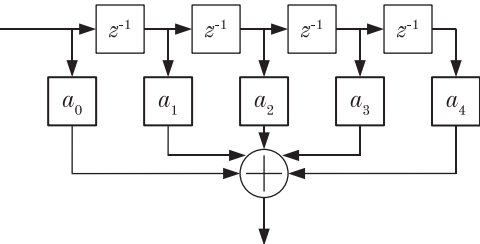


Figure 23.9 An example fourth-order five-tap FIR filter.

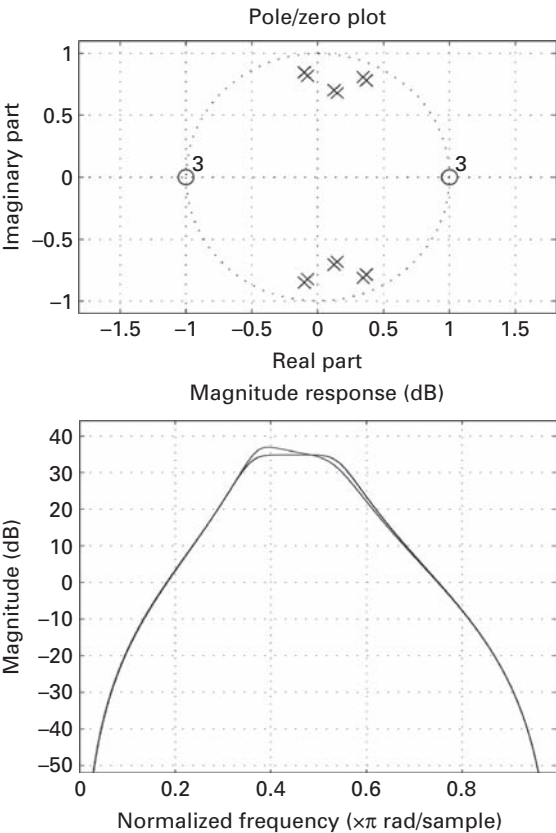


Figure 23.10 Plots of the pole and zero locations and the gain transfer function of an ideal sixth-order IIR filter, including the typical effect of random errors in pole positions.

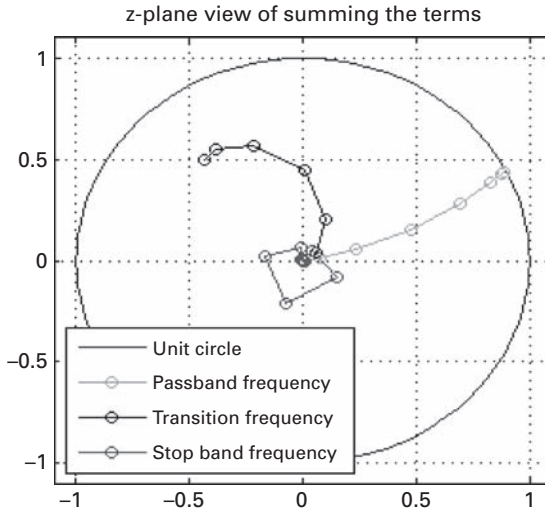


Figure 23.11 Plot of zero locations on the z -plane for an FIR filter implementing the transfer function in (23.7), with typical result from adding random errors to the component values.

and phase for different frequencies. The terms (vector segments) add up in the passband and cancel out in the stop-band. The amplitude of the passband vector is set by the average sum of a_i and the amplitude of the stop-band is set by the average error of a_i . Random tap error results in reduced stop-band attenuation but has only a small affect on the pass-band and transition band. The effect of each ratio error is spread over the stop-band.

$$H(\omega) = \sum_{i=0}^{N-1} a_i z^{-i} \quad \text{where } z = e^{j2\pi\frac{\omega}{\omega_s}}. \quad (23.7)$$

Figure 23.12 shows an FIR filter with ten-percent random error in the taps. It can be seen that this results in stop-band attenuation being limited to 60 dB, which is still more than adequate for many applications. Replicas of the discrete-time transfer function can be seen above 100 MHz which is immaterial, given the anti-aliasing filter. If more attenuation is required, it is possible to cascade multiple filters.

23.2.6 Decimation

Decimation is a form of sampling rate conversion in discrete-time systems known as down-sampling.

In down-sampling, the sampling rate is decreased by an integer factor, M , by picking every M th sample of the signal and throwing away the other $M-1$ samples. In order to avoid the effects of aliasing, this rate conversion is usually preceded by a sinc-type decimation filter with notches placed at f_s/M . In block diagrams, decimation is usually represented by a block labeled $\downarrow M$.

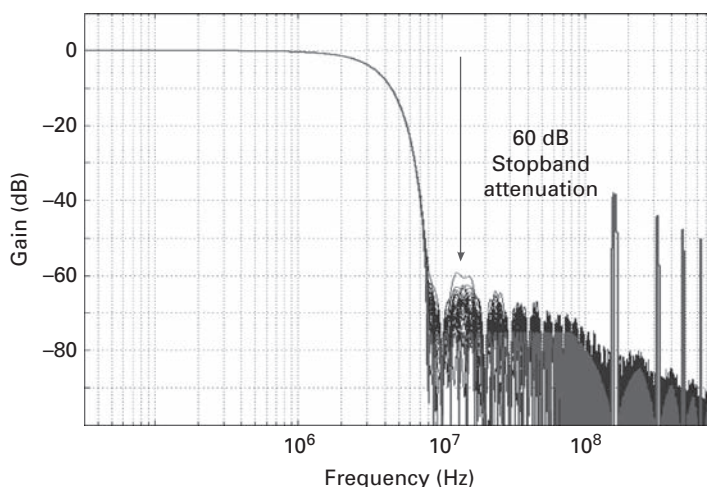


Figure 23.12 Plot of transfer function magnitude for an FIR filter with a typical result from added random error.

Successive stages of filtering and down-sampling are useful in discrete-time receivers to achieve both increased selectivity and lower sampling rates at the analog-to-digital converter (ADC) input.

Decimation can be used to reduce hardware utilization. In Figure 23.13, two FIR filters are presented. Note that the two circuits are equivalent and have the transfer function $H(z) = a_0 + a_1z^{-1} + a_2z^{-2} + a_3z^{-3} + a_4z^{-4}$. The implementation in Figure 23.13(b) utilizes feedback to reduce the amount of hardware when compared to Figure 23.13(a). The implementation in Figure 23.13(a) is performing multiplication and summation on information which is later discarded by decimation, whereas the implementation in Figure 23.13(b) never performs the calculations to be discarded.

23.2.7 Charge domain sampling

A simplified schematic of a sampler is shown in Figure 23.14, along with the associated waveforms. The sampler consists of an ideal switch and capacitor.

The input to the switch is the signal being sampled, $s(t)$, and the voltage on the capacitor is $v(t)$. When the switch is closed at time $T_1 - T_a$, the voltage on the capacitor, $v(t)$, is equal to the input voltage, $s(t)$. A sample of $v(t)$ is taken at the instant the switch is opened (points T_1 and T_2).

In a practical sampler, the switch has finite *on* resistance when it is closed, as shown in Figure 23.15. In this case, $v(t)$ is no longer equal to $s(t)$. It is equal to the convolution of $s(t)$ and the impulse response of the R - C time constant. When the switch is opened, a sample of $v(t)$ is taken. This sample is the integral of $h(T_1 - T_a)s(t)$ over the time the switch was closed, T_a .

Note that the sample is taken when the switch is opened and can still be considered as being ideal. It is the signal being sampled that has been filtered by the R - C time constant.

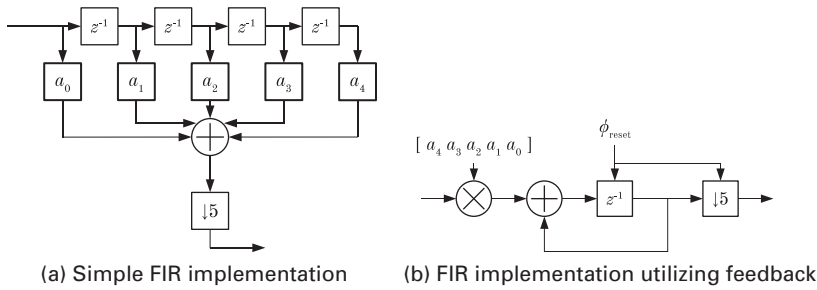


Figure 23.13 Example discrete-time implementations of an FIR filter including a decimate-by-five function.

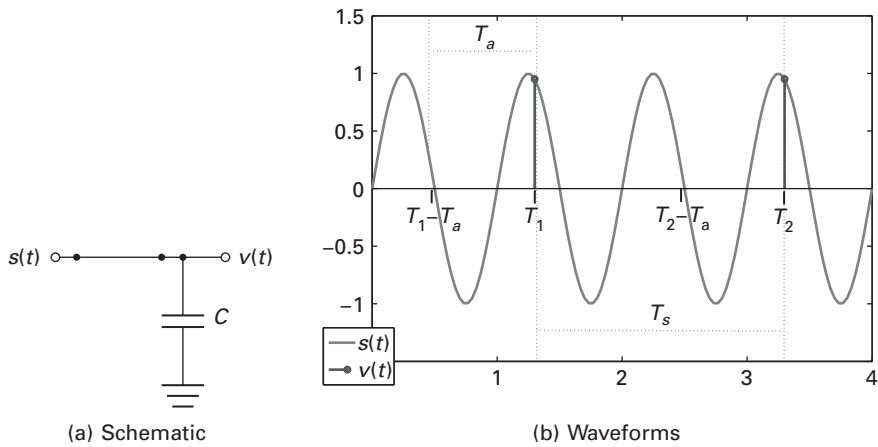


Figure 23.14 Schematic and waveforms of a simple sampler with an ideal switch in the closed position.

This continuous-time filtering occurs before the sample is taken, and results in a decimation from continuous-time to discrete-time.

This R - C is both process and temperature dependent and results in undesired distortion and attenuation of the signal. Typically, R_{on} is minimized to make the time constant as short as possible to minimize this effect. The effect of this time constant becomes more pronounced as sampling rates increase and low R_{on} becomes more difficult to achieve.

If the impulse response of the continuous-time filter in the sampler is accurately controlled, a useful anti-aliasing filter can be built into the sampler. Figure 23.16 shows a simplified schematic of a low-pass sampler that achieves this. It consists of a g_m element, capacitors, and switches. The associated clock waveforms are also shown in Figure 23.16(b). The transfer function of the filter is given as:

$$H(f, z) = \frac{g_m}{C R f_s} \text{sinc}\left(\frac{N \pi f}{f_s}\right). \quad (23.8)$$

This sampler provides a continuous-time anti-aliasing filter with a sinc response that has notches at integer multiples of the sampling frequency, f_s . If the sampling rate is much higher than the Nyquist rate of the signal being sampled, the amount of anti-

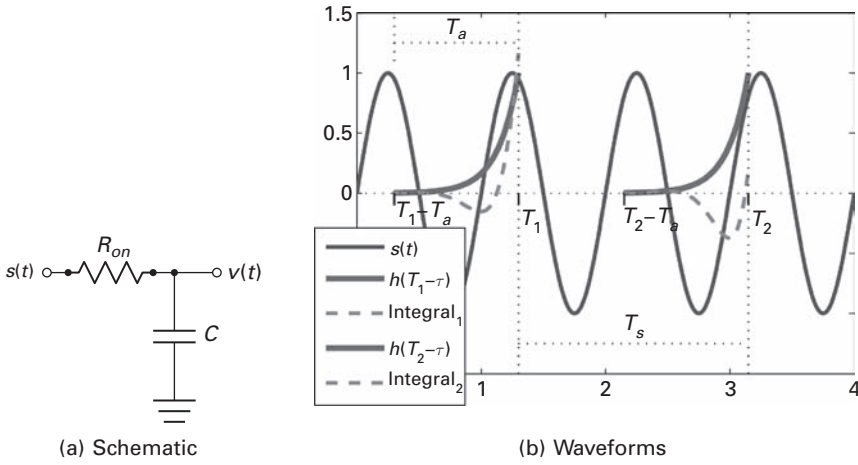


Figure 23.15 Schematic and waveforms of a simple sampler with a switch with *on* resistance in the closed position.

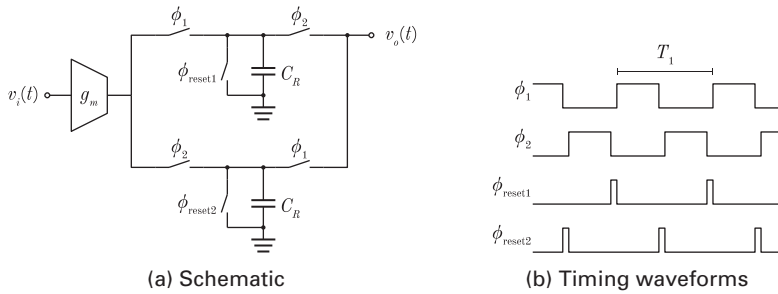


Figure 23.16 Schematic and timing waveforms for charge-domain sampler with built-in continuous-time anti-aliasing sinc filter [12].

aliasing achieved can be substantial. The attenuation provided at multiples of f_s can be calculated as:

$$\text{Att}_{dB} \approx 20 \log \left(\frac{B}{f_s} \right), \quad (23.9)$$

where B is the bandwidth of the signal being sampled and f_s is the sampling rate.

If the transconductance of the g_m element in Figure 23.16 is made to be time varying, a continuous-time FIR filter with an arbitrary impulse response can be synthesized. This is discussed in [13].

The sampler shown uses two paths, each with a sampling rate equal to $f_s/2$. When the outputs are muxed together the overall sampling rate at the output is equal to f_s . This allows for the continuous-time sinc function to have notches at multiples of f_s . The technique of using multiple paths is known as interleaving and can be extended to N -paths, each running at f_s/N . The practical result is the ability to achieve high effective sample rates with low bandwidth components, reducing design difficulty and current consumption. This can be traded-off against area consumption.

The N -path architectures suffer from a phenomenon known as *pattern noise* [14, 15]. Pattern noise results from mismatches between the various parameters of the different paths. For example, a DC offset mismatch on one of the paths will result in an output spur at f_s/N . It has been shown that frequency planning [14, 15] and careful layout can be used to mitigate this problem. Additionally, it has been shown that background digital calibration can be used to significantly reduce the effects of gain mismatch and offset mismatch between paths [16].

23.3 Zero intermediate frequency receivers

In traditional radio receivers that employ an intermediate frequency (IF), image rejection filtering is required before down-conversion, as shown in Figure 23.17. Typically, dual conversion was used to help make this filtering problem more tractable [17].

Alternatively, one can down-convert the input signal directly to DC using a local oscillator (LO) frequency that is identical to the radio frequency (RF), as in Figure 23.18. Note that with $f_{RF} = f_{LO}$, the output in Figure 23.17(b) becomes $\frac{1}{2} A_1 A_2 [\cos \varphi(t) - \cos(4\pi f_{RF} + \varphi(t))]$ and all that is present at DC is the modulation on the RF signal. In this case, it can be argued that there is no image to reject.

In order to solve the problem of potential phase ambiguity, in-phase (I) and quadrature-phase (Q) channels, along with a quadrature LO, are used to ensure that this direct conversion approach works. A block diagram of such a *zero* IF receiver is shown in Figure 23.19. Note that the low-pass filtering blocks that come after the mixers are well suited to discrete-time implementations.

23.3.1 Advantages and disadvantages of zero IF receivers

The zero IF receiver architecture seems to solve the image problem completely, eliminating the need for expensive filters prior to down-conversion and multiple IFs. As the resulting signal is at DC, subsequent signal processing and conversion to the digital domain can be implemented more easily, due to the lower frequencies involved.

However, there are two significant disadvantages to this architecture. The first is that flicker noise ($1/f$ noise) is now an issue, as it resides in the bandwidth of interest. Lowering this noise is difficult and requires large area in the transistors used to implement mixers, amplifiers, and filters.

Another disadvantage is that zero IF receivers are highly susceptible to both static DC offsets in the baseband analog circuits and dynamic DC offsets. Another source of static DC offsets is LO leakage into the RF input, causing self-mixing of the LO. Dynamic DC offsets are caused by large signals encountering even-ordered, and more specifically, second-order, nonlinearity in the front-end components, such as the low-noise amplifier (LNA) and mixers.

Unlike low IF image-reject receivers, I/Q mismatch results in baseband distortion but image leakage is not an issue.

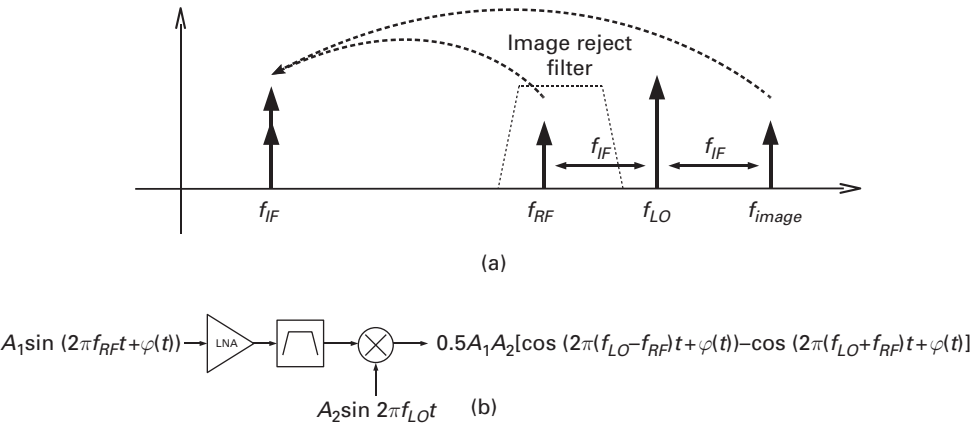


Figure 23.17 A demonstration of mixing products: (a) graphically shows down-conversion, including image frequency and image-reject filter; and, (b) shows the block diagram of a mixer and the math involved.

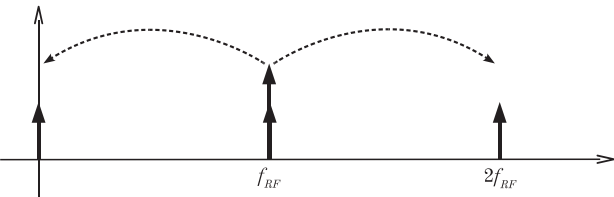


Figure 23.18 Graphical demonstration of a direct-conversion (zero IF) mixer.

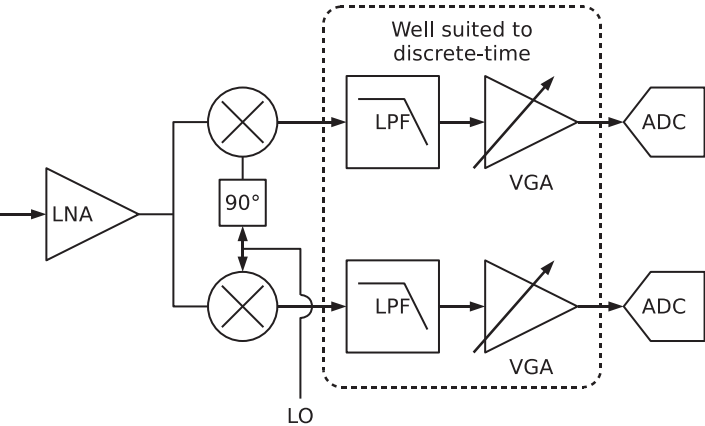


Figure 23.19 Block diagram of a zero-IF receiver.

23.3.2 Specific challenges of zero IF receivers

Static DC offsets in the receive chain components can be calibrated out. Dynamic DC offsets in the front-end are more difficult to deal with as it is not possible to control the amplitude or number of signals present in a wireless channel. This means that high second-order intercept points (IIP2) are required, making fully differential circuit topologies and higher current consumption necessary. A technique that helps to cancel the effects of dynamic DC offsets is presented in [18].

23.3.3 Example zero IF receiver

The low-pass filters that come after the mixers in Figure 23.19 can be implemented using discrete-time techniques. This approach has the advantages of highly accurate low-pass transfer functions and a high degree of programmability. Multiple stages of filtering and decimation can be used to bring the sampling rate down to a frequency that optimizes the current consumption of the ADCs.

A receiver that uses such a discrete-time filter was recently presented [19] and is shown in Figure 23.20. It can be seen that it consists of two LNAs and I/Q mixers followed by a passive R - C filter and discrete-time filtering. The focus of the work was on the discrete-time filtering that follows the mixer. It consists of five filters in two stages, having an overall decimation factor of eight prior to the ADC. The overall transfer function of the discrete-time filters is given in (23.10):

$$H(f, z) = \overbrace{\frac{g_m}{C_R f_s} \operatorname{sinc}\left(\frac{\pi f}{f_s}\right) \frac{1}{1 - \left(\frac{C_R}{C_R + C_H}\right) z^{-1}}}^{\text{First stage}} \overbrace{\left(\underbrace{(1 + z^{-1})^2}_{\text{First filter}} \right)^{2,1} \underbrace{(1 + z^{-1})^2}_{\text{Second filter}} \underbrace{(1 + z^{-1})}_{\text{Third filter}}}^{\text{Second stage}}. \quad (23.10)$$

The first stage consists of a continuous-time FIR filter and a discrete-time IIR filter, as shown in Figure 23.21. The addition of C_H forms a discrete-time pole when combined with the equivalent resistance of C_R . This first stage provides an anti-aliasing filter with notches at multiples of the sampling rate, f_s , as well as a single pole of discrete-time roll-off. The second stage consists of three discrete-time FIR filters implemented using capacitors and switches. Each filter decimates the sampling rate by a factor of two. The first filter in the second stage can be programmed to be first or second order.

By programming the R - C filters and the sampling clock, the receiver can be used for GSM (BW = 134 kHz, f_s = 64 MHz), WCDMA (BW = 1.92 MHz, f_s = 224 MHz) and WLAN (BW = 8.4 MHz, f_s = 320 MHz). The overall transfer function plots for each standard are shown in Figure 23.22.

The discrete-time filter presented is implemented in a 0.13 μm technology, consumes 11 mA from a 1.5 V supply, achieves a noise figure of 23.2 dB and 102 dB of anti-aliasing filtering, while in GSM mode. The area consumed is 1.8 mm².

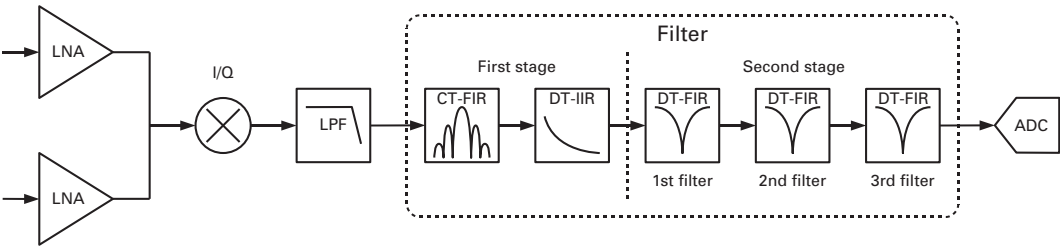


Figure 23.20 Block diagram for example zero IF receiver.

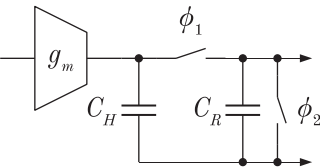


Figure 23.21 Example continuous-time FIR and discrete-time IIR filter (first stage).

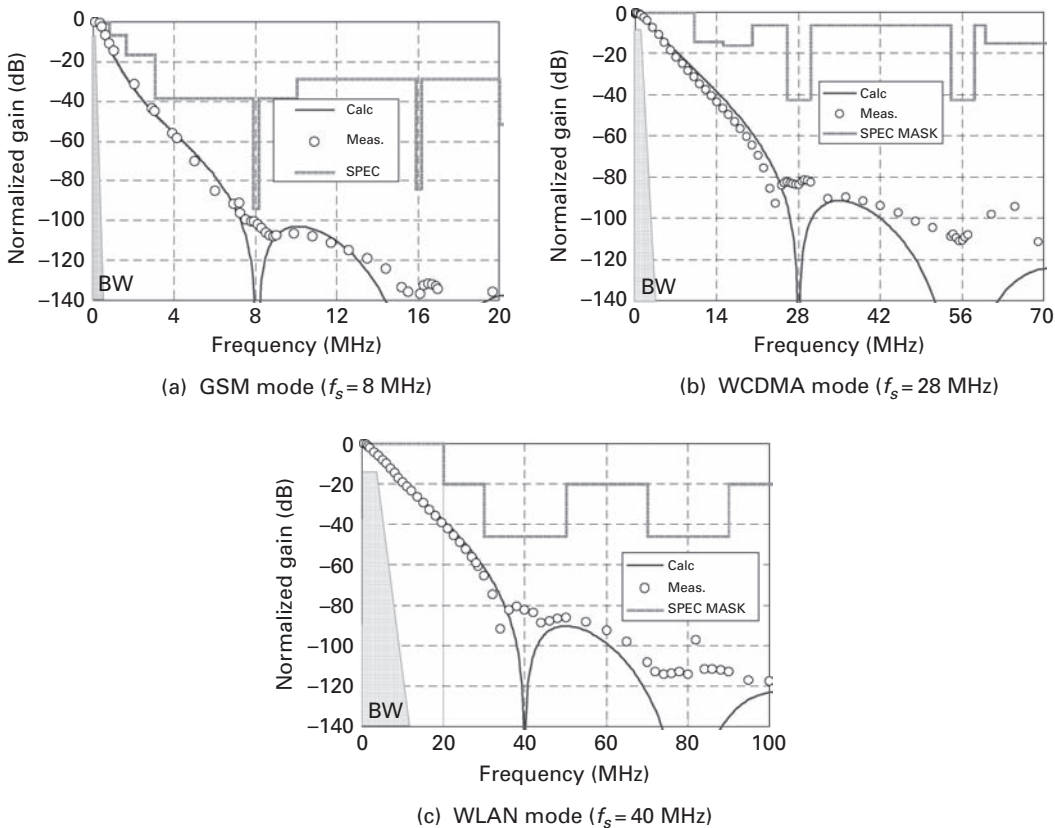


Figure 23.22 Gain transfer function plots for example zero IF receiver [19] © 2007 IEEE.

23.3.4 State-of-the-art comparison

Table 23.1 provides a summary of state-of-the-art programmable discrete-time filters for zero IF receivers, compiled from recent literature. While it is difficult to do a direct comparison, some parameters of interest are presented.

23.4 Low intermediate frequency receivers

Recent work [25, 26, 27] in addressing the disadvantages of the zero IF receiver, as discussed in Section 23.3, has mainly focused on the low intermediate frequency (IF) receiver. A low IF receiver is very similar to the zero IF receiver shown in Figure 23.19 except that “the input RF signal is down-converted to a low IF ([intermediate] frequency being at least half of the channel bandwidth) using quadrature signal paths” [28]. Unlike the zero IF case, there is an image present and mismatch between the I/Q paths limits the amount of image rejection achieved.

23.4.1 Advantages and disadvantages of low IF receivers

As the signal in a low IF receiver is never down-converted to DC in the analog domain, the problems of flicker noise from the transistors, DC offsets and self-mixing in the front-end are greatly reduced. However, as the intermediate frequency is low, the image frequency will fall in-band, which is the primary disadvantage. To account for this, in-phase and quadrature paths are used to differentiate between positive and negative frequencies and filter out the undesirable channels. To accomplish this, the I/Q paths, which contain parallel automatic gain control and baseband filters, must be well matched in phase, gain, amplitude, slope, and ripple.

23.4.2 Specific challenges of low IF receivers

It is possible to use calibration to reduce I/Q mismatch and achieve image rejection appropriate for a particular specification [25]. These techniques are generally known and are not as challenging an issue as dynamic DC offsets caused by front-end nonlinearity in direct conversion receivers. Digital compensation for I/Q mismatch is addressed in related chapters of Parts III and IV of this book.

23.4.3 Example low IF receiver system

The block diagram of a low IF receiver, used as part of a fully integrated quad-band GSM transceiver, is shown in Figure 23.23 [25]. The receiver was implemented in 90 nm CMOS and consumes 60 mA from a 1.4 V supply. The overall transceiver achieves a sensitivity of -110 dBm and supports GSM850, EGSM900, DCS1800 and PCS1900, with a total silicon area of 7.6 mm^2 .

Table 23.1 State of the art in zero IF discrete-time receivers.

Reference	Technology	Supply [V]	P_{DC} [mW]	NF [dB]	Programmability	Area [mm ²]	ADC part of filter	Notes
[19]	0.13 μ m CMOS	1.5	GSM : 16.5	GSM : 23.2	GSM/WCDMA/WLAN modes; input bandwidth 0.8–2.5 GHz; channel bandwidth 135 kHz–8.4 MHz	1.8	No	LNA, I/Q mixer, filter only; re-timing techniques used in FIR to reduce capacitor area
[20]	65 nm CMOS	1.2	77	4.8	Wi-Fi/Wimax modes; input bandwidth 2.4–2.7 GHz; channel bandwidth 5/10/20 MHz	1.2	Yes; no decimation; sampling frequency 675 MHz; successive approximation interleaved	Merged mixer, filter and ADC; feedback loop for parasitic capacitance; IIR pole
[21]	90 nm CMOS	1.2	17.4	6.0	GSM & 802.11g modes input bandwidth 200 MHz–5.5 GHz; channel bandwidth 500 kHz–20 MHz	0.13	No	Rx, LNA, passive mixer; area does not include ADC or synthesizer
[22]	90 nm CMOS	1.0	12	—	second/fourth-order Butterworth, Elliptic or Chebyshev response; channel	0.57	No	input-referred noise: 98 nV/ $\sqrt{\text{Hz}}$; 4-tap FIR g_m and discrete-time g_m ; frequency dependent on g_m , needs tuning; area only includes filter

Table 23.1 (cont.)

Reference	Technology	Supply [V]	P_{DC} [mW]	NF [dB]	Programmability	Area [mm ²]	ADC part of filter	Notes
[23]	90 nm CMOS	1.2	21.84	19	bandwidth 400 kHz–30 MHz None; channel bandwidth 250 MHz	0.27	No	uses <i>zero splitting</i> to get 250 MHz; area includes filter and synthesizer; noise figure only includes filter
[24]	90 nm CMOS	1 and 2.5	GSM: 83	5.5	tested for GSM and 802.11 g; input bandwidth 800 MHz–6 GHz; channel bandwidth 200 kHz–20 MHz; for 802.11g, input sampling frequency 480 MHz; for GSM, input sampling frequency 72 MHz	3.88	Yes; sampling frequency 40 MHz; 9 bit (802.11g, Nyquist)	1 V supply in filter and synth, 2.5 V supply in LNA and mixer; decimate and filter, passive mixer for linearity

The receiver in Figure 23.23 consists of an LNA followed by I/Q paths that convert the desired signal to a low intermediate frequency of 100 kHz where it is converted to the digital domain using passive sigma-delta ADCs. The final down-conversion to DC, channel select filtering, and demodulation are performed in the digital back-end.

The receiver converts the output voltage of the LNA to a current and a pair of direct sampling mixers [29] sample the signal at the Nyquist rate, perform anti-aliasing filtering, IIR filtering and decimate the sampling rate by a factor of two or four, resulting in an output sampling rate from 425 to 475 MSps. A simplified, single-ended schematic, illustrating the basic concept of the direct sampling mixer is shown in Figure 23.24. The approximate transfer function of the direct sampling mixer is presented in (23.11).

$$H(f, z) \approx \frac{g_m}{C_R f_1} \operatorname{sinc}\left(\frac{N\pi f}{f_1}\right) \frac{1}{1 - \left(\frac{C_R}{C_H + C_R}\right) z^{-1}} \quad (23.11)$$

This circuit provides windowed integration, continuous-time anti-aliasing and sampling, combined with a discrete-time pole.

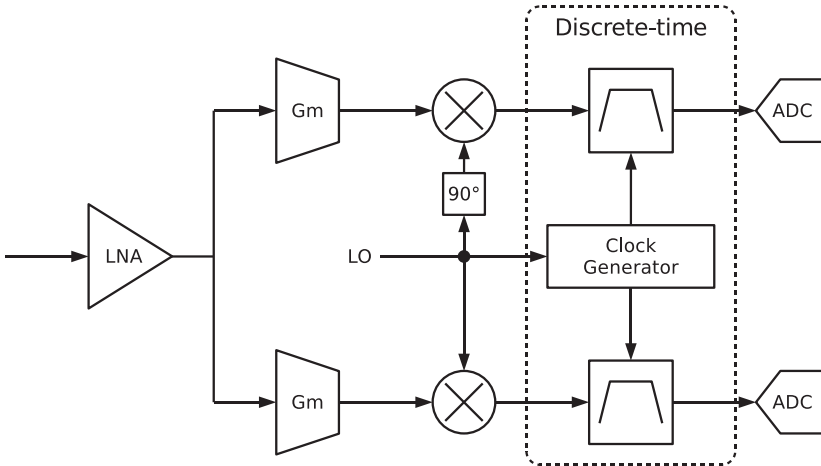


Figure 23.23 Block diagram of example low IF receiver.

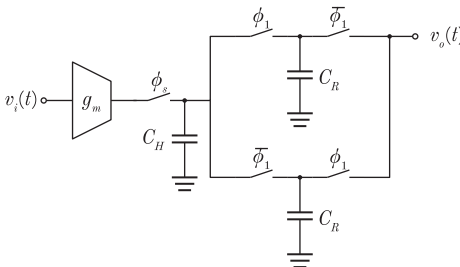


Figure 23.24 Single-ended simplified schematic of example low IF filter in PCS mode where $f_s = 1800\text{MHz}$, which directly samples to DC, and $f_1 = 450\text{MHz}$.

In this particular receiver, in order to achieve more blocker rejection before the final conversion to digital, the first discrete-time pole illustrated in [Figure 23.24](#) is supplemented by series of additional discrete-time poles. Passive, second-order, single-bit, sigma-delta ADCs are used to convert the filtered discrete-time signal to the digital domain. The greater than 425 MHz sampling rate provides an oversampling ratio greater than 700, which gives approximately five bits of resolution from oversampling alone. This high oversampling ratio, combined with the second-order noise shaping of the ADCs, gives a dynamic range equivalent of approximately thirteen bits.

High flexibility is achieved from this architecture due to programmable filtering and decimation. A Bluetooth test mode is included to support 1 MHz signal bandwidths, further demonstrating the flexibility of this approach.

23.4.4 State-of-the-art comparison

There are only a few discrete-time filter implementations for low IF receivers presented in the literature surveyed. The main ones are shown in [Table 23.2](#).

One of the discrete-time receivers [26] in [Table 23.2](#) is interesting because it has some differences compared to other types of discrete-time filters presented so far. It is a windowed-integration type sampling mixer that has a bandpass filtering response centred at a 50 MHz IF frequency, where other filters in the survey were low-pass. This block provides anti-alias filtering, FIR filtering, and down-conversion to a low IF. It is designed to be used as a final IF stage in a super-heterodyne radio receiver architecture.

23.5 Case study of a super-heterodyne AM/FM radio receiver

Super-heterodyne receivers employ multiple stages of filtering and down-conversion before demodulation [30]. Such architectures have the advantage of moving the image frequency away from the desired RF frequency through use of highly separated IF frequencies, relaxing filtering requirements and saving cost and area. This allows for high image-rejection without the I/Q matching issues present in low IF approaches.

Unlike low IF or zero IF receivers, signal gain can be distributed across different frequencies, reducing the risk of instability in high gain amplifiers. The final IF frequency can be low enough to allow channel select filtering before the ADC, but high enough to avoid the DC offset and flicker noise issues that occur in zero IF receivers.

Super-heterodyne receivers fell out of favor when the demand for high on-chip integration and reduced external components grew. No technologies were available to provide highly integrated, highly selective filtering at high frequencies, and the costs associated with multiple mixers was high. This pushed designers to low IF or zero IF quadrature down-conversion followed by baseband filters, eliminating the need for lossy, external surface acoustic wave (SAW) or ceramic filters.

Table 23.2 State of the art in low IF discrete-time receivers.

Reference	Technology	Supply [V]	P_{DC} [mW]	NF [dB]	Programmability	Area [mm ²]	ADC part of filter	Notes
[25]	90 nm CMOS	1.4	91	2	GSM850/EGSM900/DCS1800/PCS1900; programmable low IF, nom. 100 kHz; 500 kHz Bluetooth test mode	TX/RX: 7.36	Yes; passive sigma-delta	110 dB sensitivity; Bluetooth test mode; 50–55 dB IMRR
[26]	0.8 μ m BICMOS	5.0	85	–	clock-programmable center frequency; 3 dB frequency 26.3 kHz	0.7	No	50 MHz bandpass; FIR anti-aliasing; image reject; sub-sampling I/Q; 14.5 kHz low IF; 192-tap FIR
[27]	130 nm CMOS	1.5	61.5	–	No	TX/RX: 10	Yes, sigma-delta	Bluetooth receive at 2.4 GHz; discrete-time direct sampling; 83 dBm sensitivity

As an example, but without loss of generality, we will look at a high performance AM/FM super-heterodyne receiver that employs programmable discrete-time filtering techniques. The receiver is a prototype and was manufactured in a 130 nm CMOS process.

A block diagram of the receiver is presented in [Figure 23.25](#). There are four discrete-time filters (DTF), three in the FM path and one in the AM path. Two of the filters, DTF_IR and DTF_CS, are common to both the AM and FM receive paths.

The FM input consists of 200 kHz channels at output frequencies between 65 MHz and 108 MHz. An external bandpass filter is used to provide some out-of-band filtering prior to the chip. The FM path consists of an LNA, tracking filter (DTF_RF), 10.7 MHz image-reject IF bandpass filter (DTF_IR), 200 kHz channel-select filter (DTF_CS), and a 10-bit ADC.

The AM input ranges from 144–1710 kHz. The RF input is passed through a discrete-time low-pass filter, DTF_AM, and up-converted to an IF frequency of 10.7 MHz. The 10.7 MHz output from DTF_AM is then passed to DTF_IR and goes through the same receive path as FM signals.

Additional filtering, AM/FM demodulation, and audio functions are performed in the digital back-end. A pair of audio sigma-delta digital-to-analog converters (DACs) convert the left and right digital audio streams into analog waveforms.

The different filters will be described in more detail in the following sections. Particular attention will be paid to the RF tracking filter, as the other filters use variations or sub-sets of the same techniques.

23.5.1 RF tracking filter

The first filter in the FM chain is a tracking filter composed of a transconductance amplifier (the LNA shown in [Figure 23.25](#)), and three windowed integration samplers (DTF_RF2). A simplified schematic of this filter is given in [Figure 23.26](#).

The filter is tuned by changing the sampling clock frequency, generated from a programmable frequency synthesizer. The output IF frequency is located at a constant 10.7 MHz regardless of the sampling clock. When receiving 89.3 MHz, for example, the LO is at 100 MHz and the input sampling rate is at 600 MHz. In this case, the output has an overall sampling rate of 300 MHz.

Several discrete-time filtering functions are combined in this block. The resulting transfer function through each path of three paths is shown in [Figure 23.27](#).

The LNA splits the output current into five weighted currents using g_m elements which form the taps in an FIR filter. This filter has notches at 110.7 MHz (the image frequency), 221.4 MHz, and 300 MHz, when tuned to receive 89.3 MHz.

On each cycle of the 600 MHz input clock, up to two of the FIR taps are connected to each of the three windowed integrators (samplers), for a duration of two cycles. The taps applied are then shifted by one on the next clock cycle. In this way, all five taps of the FIR filter are applied to each of the samplers over a single 100 MHz output period. The two-cycle, windowed integration implements a continuous-time sinc filter with notches at multiples of 300 MHz. The combined filtering function has a six-cycle impulse response, formed from the convolution of the 5-tap FIR filter and the two cycle long continuous-

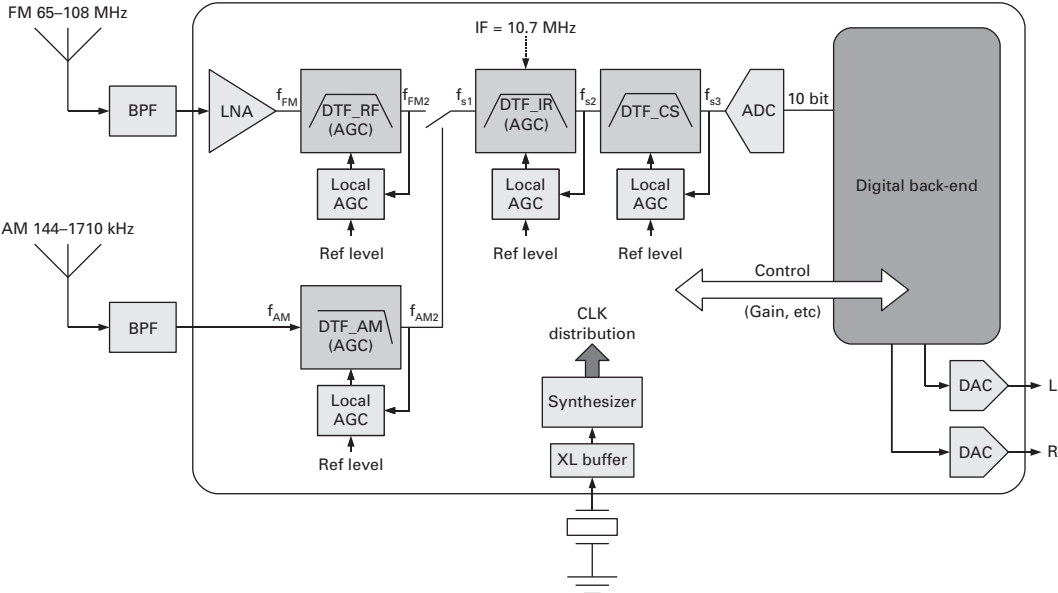


Figure 23.25 Overview of AM/FM tuner using discrete-time front-end.

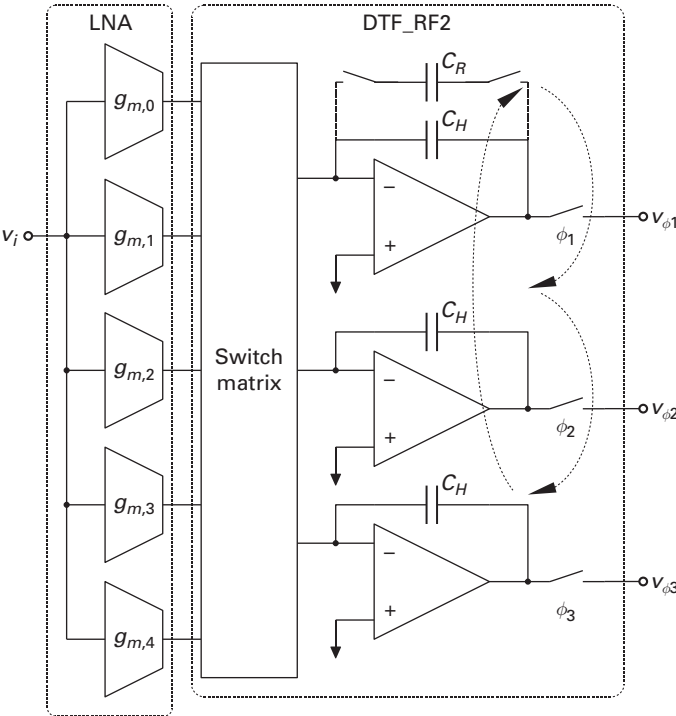


Figure 23.26 Simplified schematic of DTF_RF2.

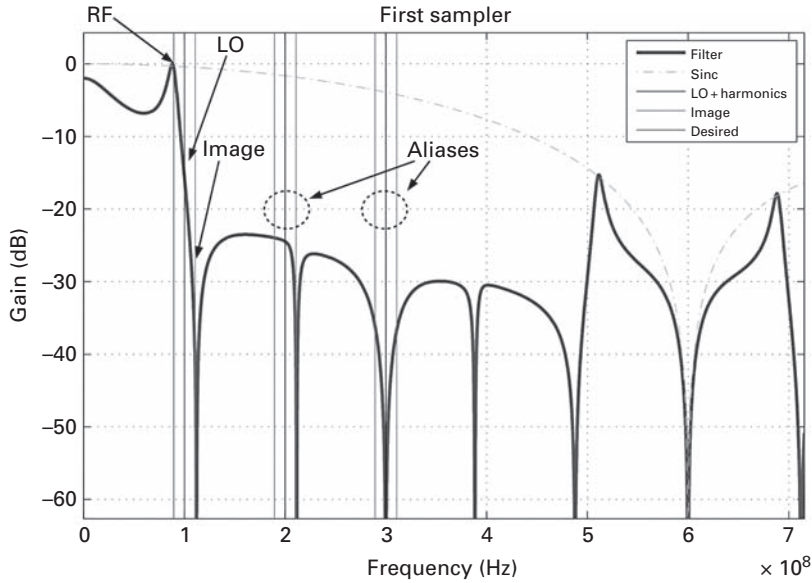


Figure 23.27 Plot of RF tracking filter transfer function.

time box-car anti-aliasing filter. This convolution is shown graphically in Figure 23.28. The output after six cycles can be found in (23.12):

$$\begin{aligned}
 v_o = & a_0 \int_{t_0}^{t_1} v_i(t) dt + a_0 \int_{t_1}^{t_2} v_i(t) dt + a_1 \int_{t_1}^{t_2} v_i(t) dt + a_1 \int_{t_2}^{t_3} v_i(t) dt \\
 & + a_2 \int_{t_2}^{t_3} v_i(t) dt + a_2 \int_{t_3}^{t_4} v_i(t) dt + a_3 \int_{t_3}^{t_4} v_i(t) dt + a_3 \int_{t_4}^{t_5} v_i(t) dt \\
 & + a_4 \int_{t_4}^{t_5} v_i(t) dt + a_4 \int_{t_5}^{t_6} v_i(t) dt.
 \end{aligned} \quad (23.12)$$

The three paths apply identical anti-aliasing and FIR filters to the input signal. The rotating capacitor, C_R , shown in Figure 23.26, is moved from sampler to sampler at a fixed rate. By setting the correct rotation rate and ratio between the sampling capacitor, C_H and the rotating capacitor, an IIR filter is implemented and provides a complex pole at 89.3 MHz, identified as RF in Figure 23.27.

Another feature of the tracking filter is that the three samplers split the LO frequency into three equally spaced phases at 0, 120, and 240 degrees. The outputs of the samplers have relative phase shifts of 0, 120, and 240 degrees for the IF frequencies at -10.7 MHz (89.3 MHz $- 100$ MHz) and at $+10.7$ MHz (110.7 MHz $- 100$ MHz). This is illustrated in Figure 23.29 for -10.7 MHz.

It is important to note that the order of the phases coming out of the RF tracking filter is different for -10.7 MHz and $+10.7$ MHz. This means it is possible to differentiate between the desired signal and the image. If these three phases are time shifted appropriately and added, the result is the desired signal at 10.7 MHz, sampled at 300 MHz.

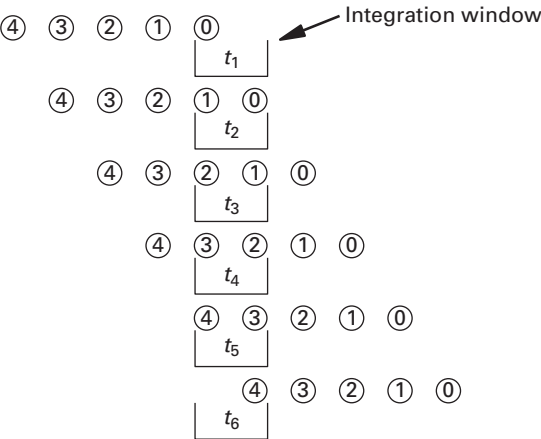


Figure 23.28 Graphical explanation of six-cycle convolution used in DTF_RF2.

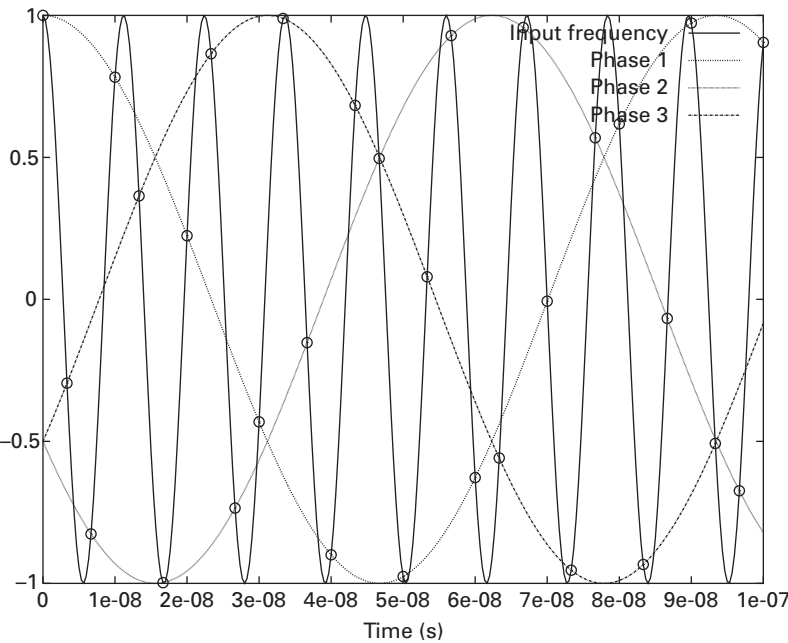


Figure 23.29 Time-domain depiction of down-conversion, showing the three-phase output.

However, because the order of the combiner paths is fixed, relative to the RF tracking filter outputs, only frequencies around -10.7 MHz result in a 10.7 MHz output. Frequencies around $+10.7$ MHz produce no output from the combiner at 10.7 MHz.

A simplified combiner is shown in Figure 23.30(a). Phasor diagrams that illustrate this concept are shown in Figure 23.30(b) and Figure 23.30(c). The combiner filter is implemented as part of the 10.7 MHz IF bandpass filter.

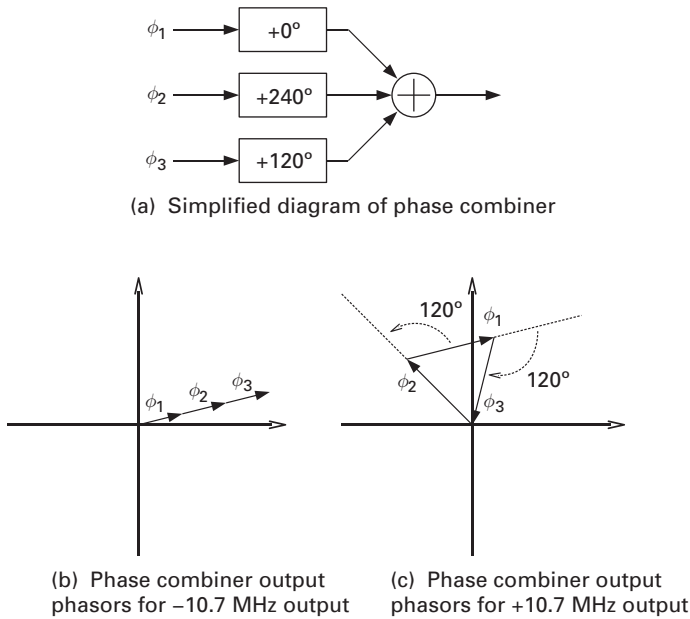


Figure 23.30 Depiction of phase combiner and output phasor diagrams for DTF_RF2.

23.5.2 Intermediate frequency filter

The IF filter performs bandpass image-reject filtering at the 10.7 MHz IF frequency, in both AM and FM modes. The output signal is centered at 1.3 MHz, the second IF in the receiver. This stage has a fixed input sampling rate of 144 MHz and a fixed output sampling rate of 12 MHz. Due to the variable output sampling rate of the RF tracking filter and the AM low-pass filter, a single-pole continuous-time reconstruction filter is placed at the input of the IF filter.

A simplified schematic of the IF filter is shown in Figure 23.31. It uses weighted g_m elements to form an 86-tap FIR filter and ten windowed integrator samplers to convolve a one-cycle (144 MHz) boxcar filter with the 86-tap FIR filter. This one-cycle boxcar results in a continuous-time anti-aliasing sinc filter with notches at multiples of 144 MHz. Because there is no IIR filter embedded in DTF_IR, no rotating capacitor is needed and the sampling capacitors are reset at the output sampling rate of the filter.

The three-phase combiner is embedded in DTF_IR. As can be seen in Figure 23.32, the 86 taps are applied to each of the three paths. Each tap from the three paths is added together before entering the switch matrix. For example, $g_{m0,1} + g_{m0,2} + g_{m0,3}$ forms g_{m0} .

The overall transfer function for the IF filter is shown in Figure 23.33. It can be seen that the theoretical, simulated, and measured results meet expectations. Due to high-frequency parasitics being present in the real circuit, measured results show more attenuation at higher frequencies than expected. This attenuation helps to improve the overall performance of the receiver.

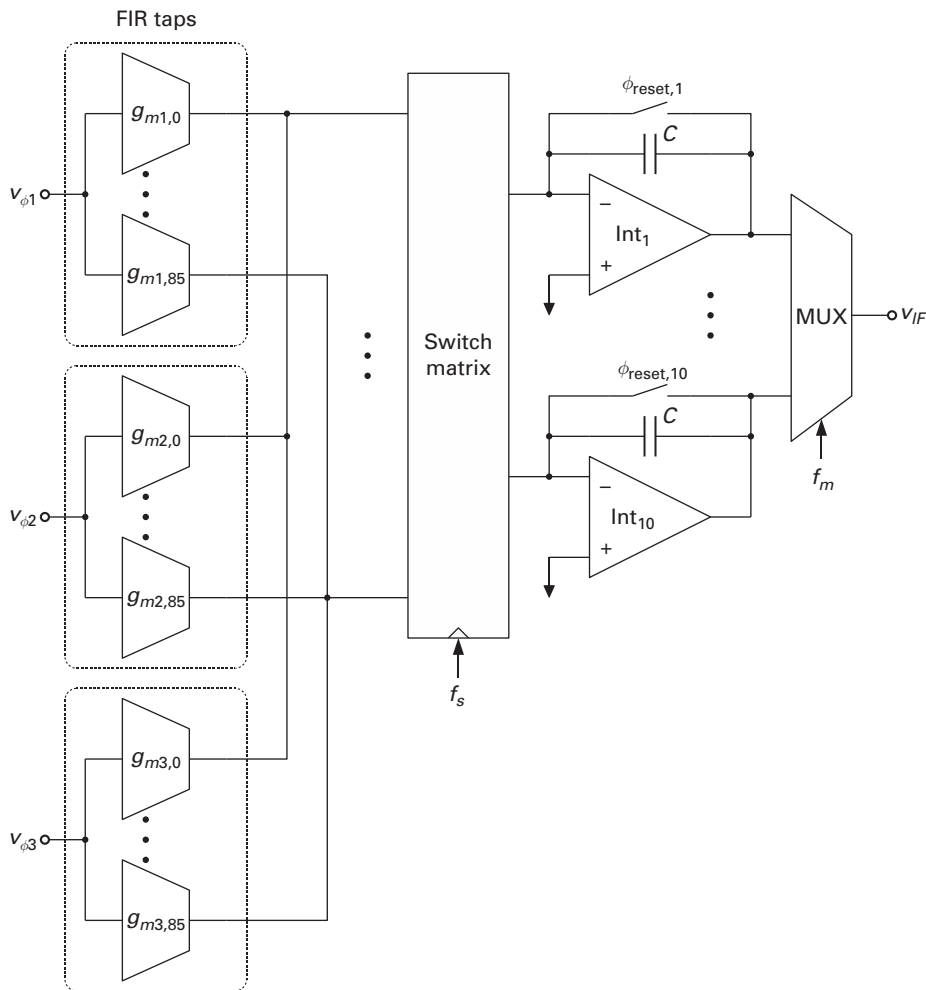


Figure 23.31 Simplified schematic of DTF_IR filter.

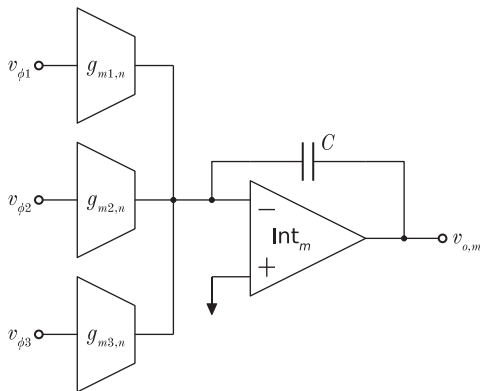


Figure 23.32 A snapshot of DTF_IR during one cycle, showing one tap connected to one integrator.

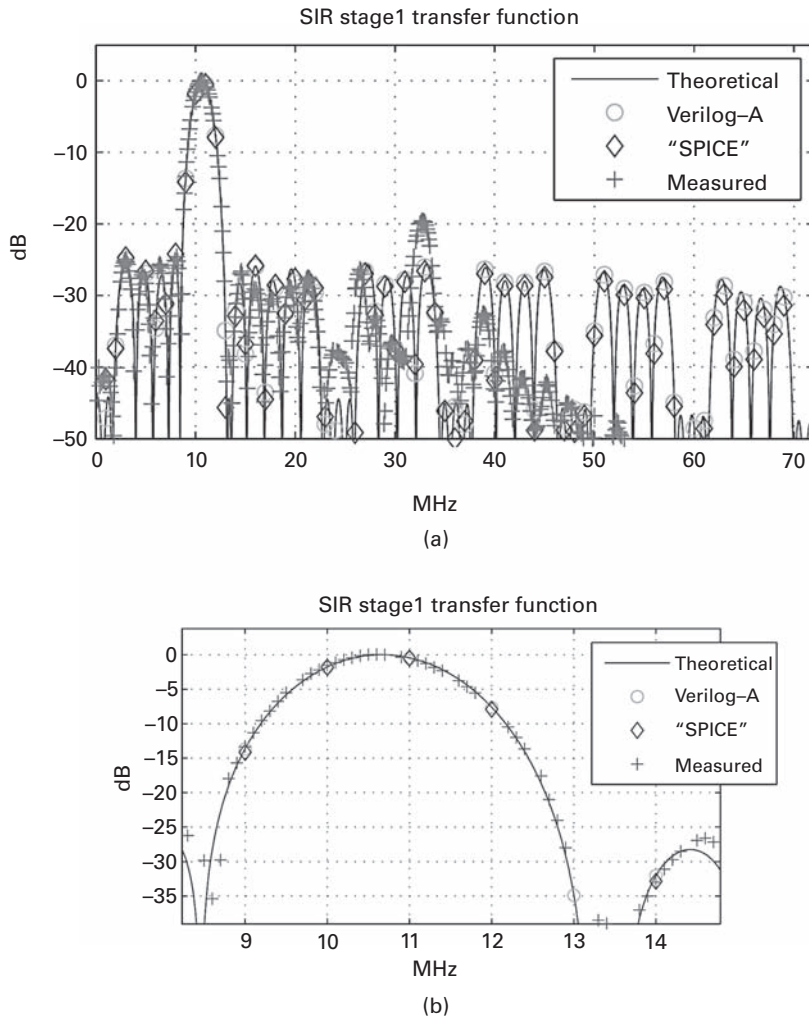


Figure 23.33 Plots of theoretical and measured frequency response of IF filter where (b) has a higher-resolution view of the passband.

23.5.3 Channel-select filter and analog-to-digital converter

The final discrete time stage before the ADC is DTF_CS, which performs 200 kHz channel-select filtering for FM reception. Additional filtering for AM channel selection is performed in the digital back-end.

A simplified block diagram of DTF_CS is shown in [Figure 23.34](#). It can be seen that it consists of three stages, DTF_Stage1, DTF_VGA, and DTF_Stage2. The second block is a variable gain amplifier (VGA) and will not be described here.

DTF_Stage1 has an input and output IF frequency of 1.3 MHz. The input sampling rate is 12 MHz and the output sampling rate is 4 MHz. It consists of a 35-tap FIR filter and

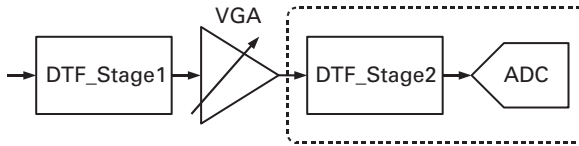


Figure 23.34 Block diagram of DTF_CS and ADC.

14 windowed-integration samplers. The FIR filter is convolved with a three-cycle (12 MHz) boxcar. The three-cycle boxcar filter provides a continuous-time anti-aliasing sinc filter with notches at multiples of 4 MHz. The ideal versus simulated, transistor level, transfer functions are shown in [Figure 23.35](#).

DTF_Stage2 has an input IF frequency of 1.3 MHz and an output IF frequency of 300 kHz. The input sampling rate is 4 MHz and the output sampling rate is 1 MHz. It consists of a 65-tap FIR filter and 26 windowed-integration samplers. The FIR filter is convolved with a four-cycle (4 MHz) boxcar. The four-cycle boxcar filter provides a continuous-time anti-aliasing sinc filter with notches at multiples of 1 MHz. The ideal simulated versus transistor level transfer functions are shown in [Figure 23.36](#).

The ADC is shown in [Figure 23.37](#) and uses a dual-slope architecture. It is integrated as part of the DTF_Stage2 filter, in a similar way to the successive approximation register (SAR) ADC in [20]. Each of the 26 samplers consists of a single ADC, and the integrator shown in [Figure 23.37](#) is also one of the 26 windowed-integration samplers. The 26 ADCs each run at $\frac{1}{26}$ MHz ≈ 38.5 kHz and the 10-bit outputs are muxed together so that the overall sampling rate is 10-bits at 1 MHz.

Each ADC works in the following way. The signal current charges the capacitor, C ; A_Comp compares the resulting voltage, V_x , to the reference voltage, V_{ref} . Depending on the polarity of the comparator output, the charge pump, A_CP, is configured to sink or source current to discharge C . At this time, the A_COUNTER output is sampled. When C is fully discharged, V_x will be above or below V_{ref} and the output of A_Comp will change state again. A_CP is disabled and the A_COUNTER output is sampled a second time. The difference between the start value and the end value in the A_COUNTER is proportional to the charge present due to the signal current.

23.5.4 AM low-pass filter

A simplified schematic of the AM path low-pass filter, DTF_AM, is given in [Figure 23.38](#). It is tuned by changing the sampling clock frequency, generated from a programmable frequency synthesizer. The output IF frequency is located at a constant 10.7 MHz regardless of the sampling clock. When receiving 1.71 MHz, for example, the LO is at 12.41 MHz and the input sampling rate is 99.28 MHz. The output sampling rate is 24.82 MHz, in this case. The three phase output is then sent to the IF filter where it is passed through the same receiver chain as FM signals. This filter uses a similar architecture to the other filters and includes a 4-tap FIR response combined with a sinc anti-aliasing filter formed from three windowed-integration samplers.

The ideal and measured transfer functions for DTF_AM are plotted in [Figure 23.39](#).

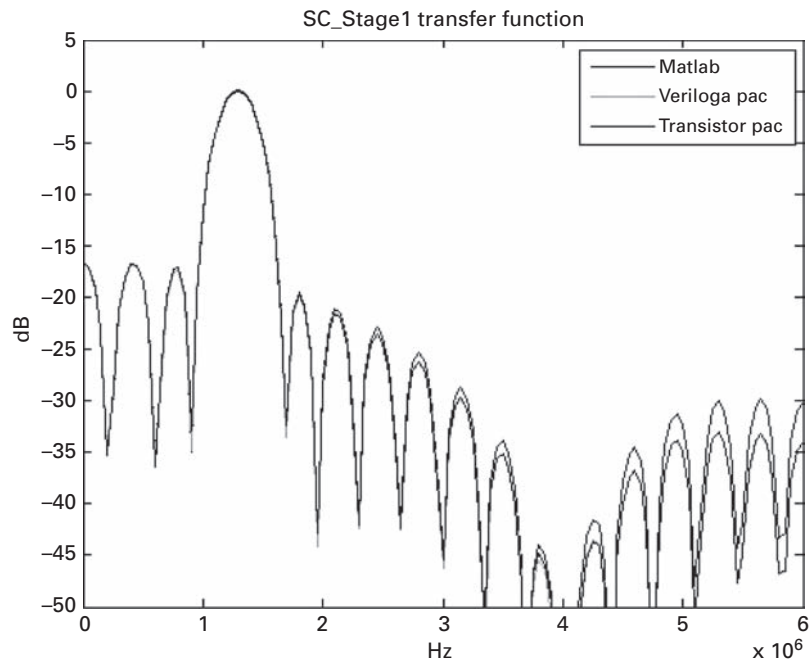


Figure 23.35 Transfer function of DTF_Stage1.

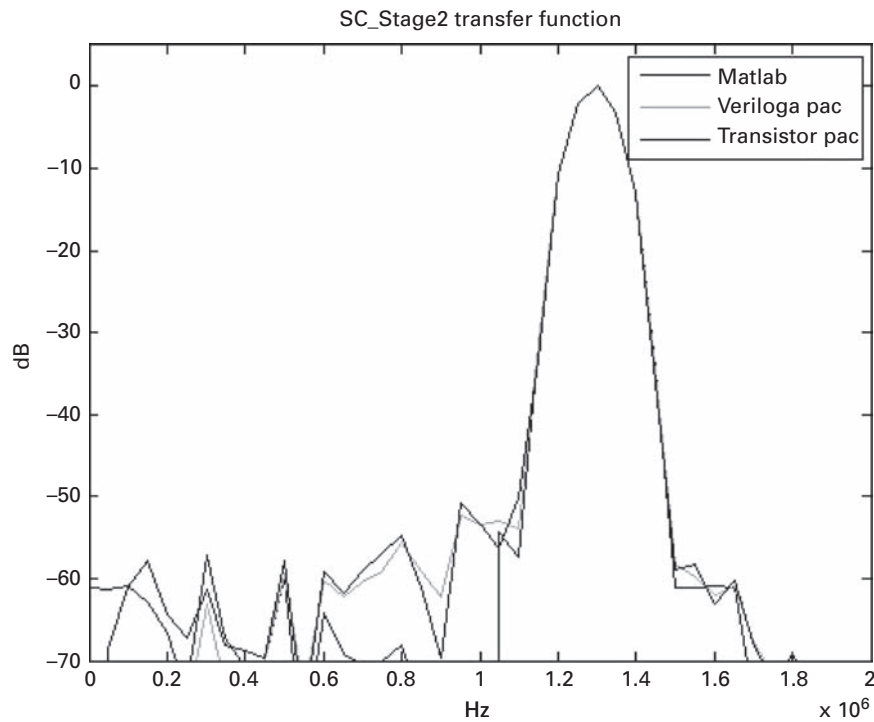


Figure 23.36 Transfer function of DTF_Stage2.

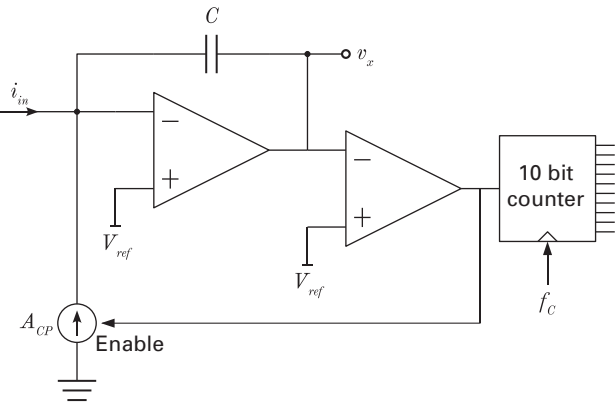


Figure 23.37 Simplified schematic of dual-slope ADC used in DTF_CS.

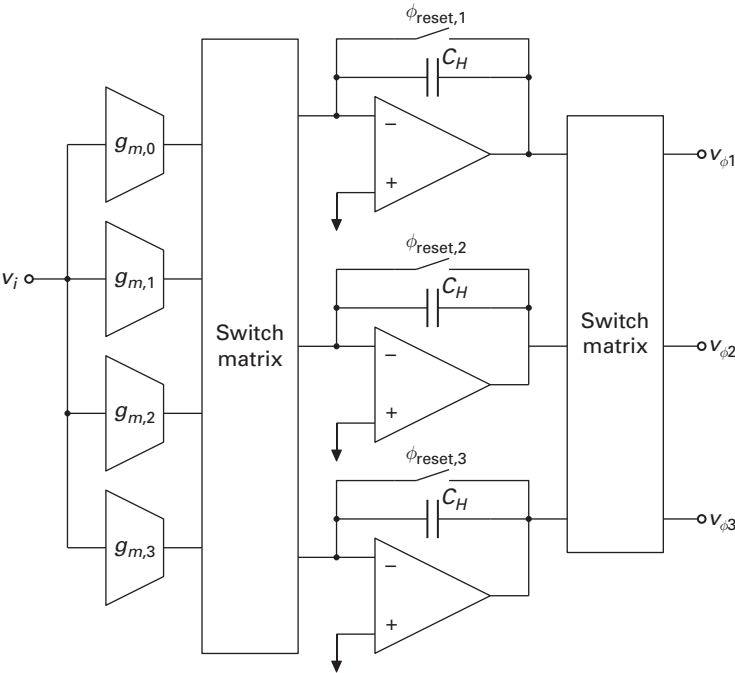


Figure 23.38 Simplified schematic of DTF_AM.

23.6 Summary and conclusions

In this chapter we looked at sampling and discrete-time filtering theory, and how it can be, and has been, applied to analog, integrated, software defined radio receivers. We also briefly looked at traditional methods for obtaining programmable on-chip filtering and the related draw-backs, namely power and area consumption and limited selectivity. It

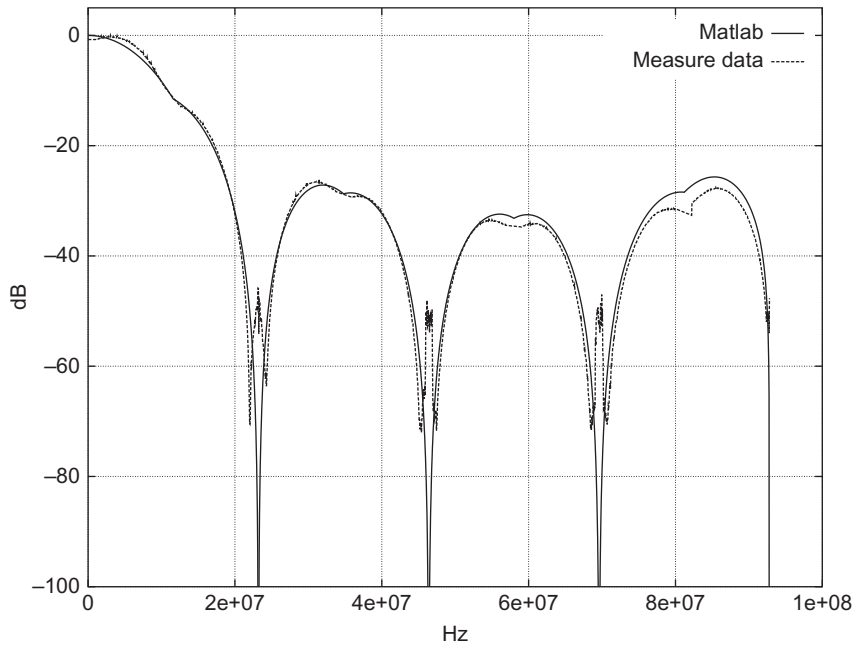


Figure 23.39 Measured and theoretical transfer function results of DTF_AM filter.

should be clear that the sampling techniques discussed address the issues of traditional methods and, while they introduce some new issues, offer an interesting alternative for use in software defined radio receivers.

Discrete-time processing techniques show the promise of implementing multiple bandpass filtering stages and multiple down-conversions on-chip, giving designers the possibility of considering a super-heterodyne type receiver architecture. This adds flexibility to the receiver design space, allowing difficult problems to be mitigated through frequency conversion and staged filtering, when required.

23.7 Acknowledgements

The authors would like to thank Kaben Wireless Silicon Inc. for supporting this research. The authors would also like to extend a special thanks to Professor Miles Copeland of Carleton University and Dr. Neil Birkett of Maplebrook Consulting, for reviewing this manuscript and providing valuable feedback.

References

- [1] S. D'Amico, V. Giannini, and A. Baschirotto, "A 4th-order active G_m -RC reconfigurable (UMTS/WLAN) filter," *IEEE Journal of Solid-State Circuits*, vol. **41**, pp. 1630–1637, July 2006.

- [2] H. Amir-Aslanzadeh, E. J. Pankratz, and E. Sánchez-Sinencio, "A 1-V +31 dBm IIP₃, reconfigurable, continuously tunable, power-adjustable active-RC LPF," *IEEE Journal of Solid-State Circuits*, vol. **44**, pp. 495–508, February 2009.
- [3] A. Vasilopoulos, G. Vitzilaios, G. Theodoratos, and Y. Papananos, "A low-power wideband reconfigurable integrated active-RC filter with 73 dB SFDR," *IEEE Journal of Solid-State Circuits*, vol. **41**, pp. 1997–2008, September 2006.
- [4] M. J. McNutt, S. LeMarquis, and J. L. Dunkley, "Systematic capacitance matching errors and corrective layout procedures," *IEEE Journal of Solid-State Circuits*, vol. **29**, pp. 611–616, May 1994.
- [5] A. Hastings, *The Art of Analog Layout*. Upper Saddle River, New Jersey: Prentice Hall, 2001.
- [6] D. Jakonis, K. Folkesson, J. Dabrowski, P. Eriksson, and C. Svensson, "A 2.4-GHz sampling receiver front-end in 0.18- μ m CMOS," *IEEE Journal of Solid-State Circuits*, vol. **40**, pp. 1265–1277, June 2005.
- [7] D. Chamla, A. Kaiser, A. Cathelin, and D. Belot, "A switchable-order G_m -C baseband filter with wide digital tuning for configurable radio receivers," *IEEE Journal of Solid-State Circuits*, vol. **42**, pp. 1513–1521, July 2007.
- [8] T. Maeda, N. Matsuno, S. Hori, *et al.*, "A low-power dual-band triple-mode WLAN CMOS transceiver," *IEEE Journal of Solid-State Circuits*, vol. **41**, pp. 2481–2490, November 2006.
- [9] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*. Lexington, Massachusetts: MIT Lincoln Laboratory, fourth edn., 2007.
- [10] R. J. Baker, *CMOS Mixed-Signal Circuit Design*. Hoboken, New Jersey: John Wiley and Sons, 2009.
- [11] E. W. Kamen, *Introduction to Signals and Systems*. New York: Macmillan Publishing Company, 1987.
- [12] A. A. Abidi, "The path to the software-defined radio receiver," *IEEE Journal of Solid-State Circuits*, vol. **42**, pp. 954–966, May 2007.
- [13] S. Karvonen, T. A. D. Riley, and J. Kostamovaara, "Charge-domain FIR sampler with programmable filtering coefficients," *IEEE Transactions on Circuits and Systems – Part II: Express Briefs*, vol. **53**, pp. 192–196, March 2006.
- [14] F. Borghetti, C. D. Fiore, P. Malcovati, and F. Maloberti, "Synthesis of the noise transfer function in N-path sigma delta modulators," in *IEE International Conference on ADDA*, pp. 171–176, July 2005.
- [15] A. K. Ong and B. A. Wooley, "A two-path bandpass $\Sigma\Delta$ modulator for digital IF extraction at 20 MHz," *IEEE Journal of Solid-State Circuits*, vol. **32**, pp. 1920–1934, December 1997.
- [16] D. Fu, K. C. Dyer, S. H. Lewis, and P. J. Hurst, "A digital background calibration technique for time-interleaved analog-to-digital converters," *IEEE Journal of Solid-State Circuits*, vol. **33**, pp. 1904–1911, December 1998.
- [17] T. H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*. Cambridge: Cambridge University Press, 1998.
- [18] M. Faulkner, "DC offset and IM2 removal in direct conversion receivers," *IEE Proceedings – Communications*, vol. **149**, pp. 179–184, June 2002.
- [19] T. Sano, T. Maruyama, I. Yasui, H. Sato, and T. Shimizu, "A 1.8 mm², 11 mA, 23.2 dB-NF, discrete-time filter for GSM/WCDMA/WLAN using retiming technique," in *IEEE Custom Integrated Circuits Conference*, pp. 703–706, September 2007.
- [20] F. Montaudon, R. Mina, S. Le Tual, *et al.*, "A scalable 2.4-to-2.7 GHz Wi-Fi/WiMAX discrete-time receiver in 65 nm CMOS," in *IEEE International Solid-State Circuits Conference*, pp. 362–619, February 2008.

- [21] A. Geis, J. Ryckaert, J. Borremans, *et al.*, "A compact low power SDR receiver with 0.5–20 MHz baseband sampled filter," in *IEEE Radio Frequency Integrated Circuits Symposium*, pp. 285–288, June 2009.
- [22] M. Kitsunezuka, S. Hori, and T. Maeda, "A widely-tunable, reconfigurable CMOS analog baseband IC for software-defined radio," *IEEE Journal of Solid-State Circuits*, vol. **44**, pp. 2496–2502, September 2009.
- [23] A. Yoshizawa and S. Iida, "A 250-MHz cutoff charge-domain baseband filter with improved stopband attenuations," in *IEEE Radio Frequency Integrated Circuits Symposium*, pp. 491–494, June 2009.
- [24] R. Bagheri, A. Mirzaei, S. Chehrazi, *et al.*, "An 800-MHz–6-GHz software-defined wireless receiver in 90-nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. **41**, pp. 2860–2876, December 2006. 23.1
- [25] K. Muhammad, Y.-C. Ho, T. L. Mayhugh, Jr., *et al.*, "The first fully integrated quad-band GSM/GPRS receiver in a 90-nm digital CMOS process," *IEEE Journal of Solid-State Circuits*, vol. **41**, pp. 1772–1783, August 2006.
- [26] S. Karvonen, T. A. D. Riley, S. Kurtti, and J. Kostamovaara, "A quadrature charge-domain sampler with embedded FIR and IIR filtering functions," *IEEE Journal of Solid-State Circuits*, vol. **41**, pp. 507–515, February 2006.
- [27] R. B. Staszewski, K. Muhammad, D. Leipold, *et al.*, "All-digital TX frequency synthesizer and discrete-time receiver for Bluetooth radio in 130-nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. **39**, pp. 2278–2291, December 2004.
- [28] J. Laskar, S. Chakraborty, A.-V. Pham, and M. M. Tantzeris, *Advanced Integrated Communication Microsystems*. Hoboken, New Jersey: John Wiley and Sons, 2009.
- [29] K. Muhammad and R. B. Staszewski, "Direct RF sampling mixer with recursive filtering in charge domain," in *IEEE International Symposium on Circuits and Systems*, pp. 23–26, May 2004.
- [30] S. Dell'Aera and T. Riley, "Sampling IF filters and the return of the superheterodyne receiver," *Microwave Journal*, November 2005.

24 Multi-port front-end and transceivers for V-band multi-gigabit/s communication systems

Serioja Ovidiu Tatu, Emilia Moldovan, and Sofiene Affes

24.1 Introduction

The millimeter wave spectrum has been identified as a candidate of choice to support multi-gigabit/s data transmissions. The increasing interest of recent years has pushed the regulatory agencies to provide new opportunities for unlicensed spectrum usage with fewer restrictions on radio parameters. In order to provide more flexibility in spectrum sharing, the FCC introduced an opening of 7 GHz unlicensed spectrum at millimeter wave frequencies around 60 GHz, from 57 to 64 GHz.

As known, in the case of comparable bandwidths and data-rates, an important advantage of using millimeter wave frequencies instead of microwave ones is the reduced ratio between the bandwidth and the central frequency, leading the way to transceiver simplicity. In addition, compared to microwave frequencies, the strong signal attenuation at 60 GHz allows an efficient reuse. This helps to create small indoor cells for hot spot secure wireless communications. This spectrum is suitable for multi-gigabit/s wireless communication systems, which could be home or office high-speed wireless networking and entertainment, such as extremely fast downloading of files via wireless Gigabit Ethernet, and wireless High Definition Multimedia Interface (HDMI).

At the same time, the 60 GHz communication systems are also very attractive for secure line-of-sight multi-gigabit/s Low Earth Orbit inter-satellite links. The communications cannot be received on the Earth surface due to severe oxygen absorption in the atmosphere (around 20 dB/km at sea level). However, for this challenging application, major improvements must be made to today's emerging transceivers. There are still multiple challenges in terms of circuit and module design and fabrication, and system architectures. In the authors' opinion, the use of multi-port interferometric approach in conjunction with innovative fabrication technologies will open the way to low-cost reliable multi-gigabit/s millimeter wave communication systems. We are persuaded that the work presented in this chapter will lead the way to transceiver simplicity together with system miniaturization, ensuring higher data-rates, compared to the conventional microwave Ultra Wideband (UWB) approach. For example, a data-rate of 14 Gb/s is achievable using single carrier 16 symbols Quadrature Amplitude Modulation (16 QAM)

or Orthogonal Frequency Division Multiplexing (OFDM) scheme with seven Quadrature Phase Shift Keying (QPSK) modulated carriers.

The multi-port circuit theory was first developed in the 1970s by scientists for accurate automated measurements of the complex reflection coefficient in microwave network analysis. These multi-port pioneers highlighted its usefulness in microwave low-cost circuit characterizations (S-parameters). Since 1994, the multi-port techniques were further developed by Professors R. G. Bosisio and Ke Wu (Poly-Grames Research Centre of “École Polytechnique de Montreal,” Canada) and collaborators, who proposed its use as a microwave and millimeter wave demodulator in connection with homodyne and heterodyne receivers. Several multi-port architectures for specific applications have been developed and implemented. Basically, this is a passive circuit, composed of several couplers, interconnected by transmission lines and phase shifters. The multi-port acts as an interferometer; its output signals are linear combinations of phase shifted reference and input unknown signals. By using appropriate devices connected to output ports, this circuit can provide quadrature down-conversion or direct modulation.

Section 24.2 presents the multi-port interferometer theory focused on quadrature down-conversion, direct modulation of a millimeter wave signal, and antenna array implementation. This describes and analyzes various multi-port circuit implementations and their advantages versus the conventional approaches.

Section 24.3 deals with practical implementation, modeling, and analysis of a 60 GHz multi-port. Taking advantage of the advancements in fabrication technologies at microwave and millimeter wave frequencies, bulky waveguide circuits are integrated into various substrate integrated circuits (SICs). Multi-port circuit schemes and practical realizations using Miniature Hybrid Microwave Integrated Circuit (MHMIC) technology are presented and discussed. Despite its physical limitations, the conventional Rectangular Waveguide (RWG) technology is still essential to connect the SICs to standard millimeter wave equipment and circuits.

Section 24.4 proposes multi-port interferometric V-band transceiver architectures for wireless local area networks (WLANs), dedicated to high data-rate communications, including digital signal processing considerations. The use of a multi-port interferometer in quadrature down-conversion is an innovative approach, due to its intrinsic properties, such as wide bandwidth, reduced local oscillator (LO) power requirement, excellent isolation between RF input ports, and very good suppression of harmonic and spurious products. In addition, millimeter wave direct modulators and array antennas (implemented using multi-ports, as presented in Section 24.2), are considered in the proposed system architectures.

Advanced system simulation results, based on realistic circuit models, are presented and discussed in detail in Section 24.5, followed by conclusions in Section 24.6.

Computed aided design (CAD) tools, such as Advanced Design Systems (ADS) of Agilent Technologies and High Frequency Structure Simulator (HFSS) of Ansoft have been used for circuit designs and system simulations. Circuit prototype photographs, along with simulation and measurement results reinforce this presentation.

24.2 Multi-port interferometer theory

The multi-port, in fact a “six-port” circuit, was first developed by Cohn and Weinhouse [1] to evaluate the phase of a microwave signal, and extended by Engen and Hoer [2]–[4] for accurate automated measurements of the complex reflection coefficient in microwave network analysis.

Between 1972 and 1994, similar researches were pursued by other laboratories, including Poly-Grames Research Centre of “École Polytechnique de Montreal.” The proposed six-port is a passive circuit, composed of hybrid couplers and/or power dividers, connected by transmission lines. It has two inputs, one for the unknown signal and the other for the reference signal, and four outputs. The output signals represent linear combinations of the input signals. The evaluation of the reflection coefficient is based on the measurement of output signal power levels. Detailed discussion is presented in [5].

In 1994, the idea of a multi-port circuit to be used in direct conversion microwave receivers was proposed for the first time at the Poly-Grames Research Centre of “École Polytechnique de Montreal.” The first reported results were related to the narrow-band single-carrier demodulation of digital data [6]–[8]. The multi-port architecture was very similar to the original approach, and a calibration process was needed to perform measurements. Since 2001, various multi-port interferometer techniques with no need of calibration have been developed [9]–[12].

24.2.1 Multi-port interferometer for quadrature down-conversion

The multi-port circuit presented in Figure 24.1 is composed of three 90° hybrid couplers and a Wilkinson power divider. This architecture is typical for a multi-port microstrip implementation. The instantaneous phase shift and the relative amplitude between RF unknown and reference input signals, a_6 and a_5 , respectively, are obtained by output power measurements, as explained below.

Generally, for a multi-port, the output signals, b_i , can be expressed as a function of the dispersion parameters S_{ij} :

$$b_i = \sum_{j=1}^6 S_{ij} a_j, \quad i = 1, \dots, 6 \quad (24.1)$$

The scattering matrix of the proposed multi-port phase discriminator can be obtained using the diagram presented in Figure 24.1. For this purpose, the following scattering matrices of a 90° hybrid coupler and Wilkinson power divider, as given in Equations (24.2) and (24.3), respectively, are used:

$$[S] = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & j & 1 & 0 \\ j & 0 & 0 & 1 \\ 1 & 0 & 0 & j \\ 0 & 1 & j & 0 \end{bmatrix}, \quad (24.2)$$

In order to obtain the IF output signals, four power detectors are connected to the multi-port outputs. As known, the output voltage of an ideal power detector is proportional to the square magnitude of the RF input signal:

$$v_i = K_i \cdot |b_i|^2 = K_i \cdot b_i \cdot b_i^*, \quad i = 1, \dots, 4. \quad (24.9)$$

Supposing that identical power detectors are used, $K_i = K$, the output voltages become:

$$v_{1,3}(t) = K \frac{a^2}{4} \cdot \{1 + \alpha^2 \mp 2 \cdot \alpha \cdot \cos[\Delta\omega \cdot t + \Delta\varphi(t)]\}, \quad (24.10)$$

$$v_{2,4}(t) = K \frac{a^2}{4} \cdot \{1 + \alpha^2 \mp 2 \cdot \alpha \cdot \sin[\Delta\omega \cdot t + \Delta\varphi(t)]\}. \quad (24.11)$$

As seen, the output voltages at the pairs of ports 1 and 3, and 2 and 4, respectively, are phase opposites. Therefore, the quadrature output signals can be obtained using two differential amplifiers in the IF stage:

$$v_{IF}^I(t) = A_{IF} \cdot [v_3(t) - v_1(t)] = \alpha \cdot K \cdot a^2 \cdot A_{IF} \cdot \cos[\Delta\omega \cdot t + \Delta\varphi(t)], \quad (24.12)$$

$$v_{IF}^O(t) = A_{IF} \cdot [v_4(t) - v_2(t)] = \alpha \cdot K \cdot a^2 \cdot A_{IF} \cdot \sin[\Delta\omega \cdot t + \Delta\varphi(t)]. \quad (24.13)$$

Finally, after the second frequency conversion and low-pass filtering, the baseband quadrature signals are obtained as follows:

$$I(t) = \frac{1}{2} \cdot \alpha \cdot a^2 \cdot K \cdot A_{IF} \cdot A_{BB} \cdot \cos[\Delta\varphi(t)], \quad (24.14)$$

$$Q(t) = \frac{1}{2} \cdot \alpha \cdot a^2 \cdot K \cdot A_{IF} \cdot A_{BB} \cdot \sin[\Delta\varphi(t)]. \quad (24.15)$$

Therefore, a baseband signal $\Gamma(t)$ can be defined in the complex plane:

$$\Gamma(t) = I(t) + jQ(t) = \frac{1}{2} \cdot \alpha \cdot a^2 \cdot K \cdot A_{IF} \cdot A_{BB} \cdot e^{j\Delta\varphi(t)}. \quad (24.16)$$

In order to highlight the differences versus the previous approach, the conventional quadrature mixer and a quasi-conventional multi-port proposal using anti-parallel diodes are presented in [Figure 24.2](#).

The signal flow is calculated considering the scattering matrix of a Wilkinson divider, as presented in Equation (24.3), and the scattering matrix of a 180° hybrid coupler given by:

$$[S] = \frac{j}{\sqrt{2}} \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \end{bmatrix}. \quad (24.17)$$

Therefore, the multi-port scattering matrix becomes:

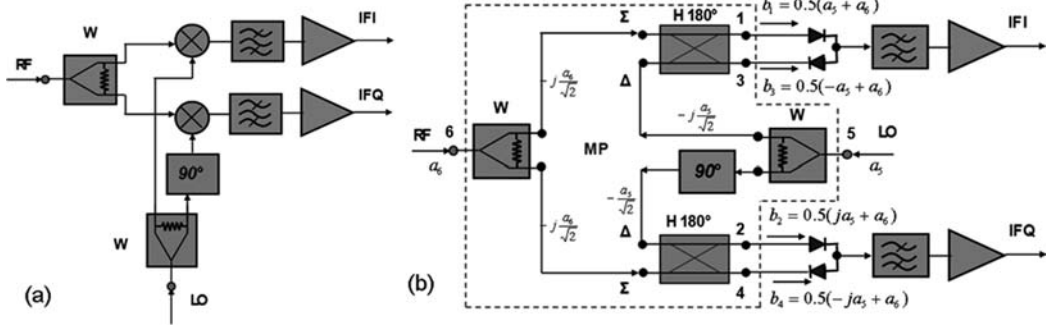


Figure 24.2 Block diagram of a conventional IQ mixer (a) and a quasi-conventional multi-port mixer (b).

$$[S] = \frac{1}{2} \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & j & 1 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & -j & 1 \\ 1 & j & -1 & -j & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \end{bmatrix}. \quad (24.18)$$

The 180° hybrid coupler output signals are injected into the anti-parallel diode inputs. Considering four identical diodes, and their general nonlinear characteristic:

$$I = c \cdot V + d \cdot V^2 + e \cdot V^3 + f \cdot V^4 + \dots, \quad (24.19)$$

the in-phase current, before filtering, can be expressed as:

$$I^I = c \cdot (V_1 + V_3) + d \cdot (V_1^2 - V_3^2) + e \cdot (V_1^3 + V_3^3) + f \cdot (V_1^4 - V_3^4) + \dots \quad (24.20)$$

$$I^I = c \cdot \sqrt{Z_0} \cdot (b_1 + b_3) + d \cdot Z_0 \cdot (b_1^2 - b_3^2) + e \cdot Z_0 \cdot \sqrt{Z_0} \cdot (b_1^3 + b_3^3) + f \cdot Z_0^2 \cdot (b_1^4 - b_3^4) + \dots \quad (24.21)$$

Based on the calculus of b_i (see Equation (24.7)), as specified in Figure 24.2, after the required intermodulation product filtering, the only selected term is:

$$I_{IF}^I = -d \cdot Z_0 \cdot a_5 \cdot a_6. \quad (24.22)$$

In the same way, the related quadratic signal, after filtering, can be computed:

$$I_{IF}^Q = j \cdot d \cdot Z_0 \cdot a_5 \cdot a_6. \quad (24.23)$$

For down-conversion purposes, only the frequency difference term is selected after filtering.

We note that a main difference between the two down-conversion approaches exists. In the first implementation, the diodes are optimized to operate as power detectors. The output spectrum contains basically only low IF and DC signals. In the second one, due to

the use of the whole nonlinear characteristic of diodes, the output spectrum is richer and must be carefully filtered. Therefore, in order to obtain comparable down-conversion results, the LO power must be considerably increased (with more than 20 dB) if the conventional approach is used [13].

24.2.2 Multi-port interferometer for direct modulation

The millimeter wave modulator is an essential element for the transmitter. Depending on the specific modulation scheme, a multi-port modulator or a millimeter wave switching network are proposed for low-cost UWB applications [14].

The multi-port modulator uses the multi-port interferometer presented in Figure 24.1, together with two pairs of mono-ports, having adjustable reflection coefficients. As seen in Figure 24.3, the normalized wave signal passing through the upper or lower hybrid coupler is multiplied by $j\Gamma_i$, $i = 1, 2$. The output signal of the direct modulator combines these signals as follow:

$$b_5 = 0.5 \cdot (\Gamma_1 + j\Gamma_2) \cdot a_6. \quad (24.24)$$

Therefore, the phase and the amplitude of the normalized output signal are related to the mono-port return loss values. The direct modulation of a millimeter wave signal is straightforward; the modulated constellation can be rotated in the IQ plane adding a constant phase shift.

As a first example using short or open circuits, the return loss values ($\Gamma_{1,2}$) are equal to -1 and $+1$, respectively, and conform to the direct modulation chart presented in Table 24.1; consequently, a direct QPSK multi-port modulator is obtained [14].

Different other modulated signals can be obtained using only the real part impedances for both mono-ports, as illustrated in Tables 24.2 to 24.4.

As detailed in the previous analysis, in order to obtain all modulation states, the ratio of the mono-port resistance maximum/minimum values (excepting open and short circuits, that are easily implemented with stubs) is equal to $7^2 = 49$ for Square

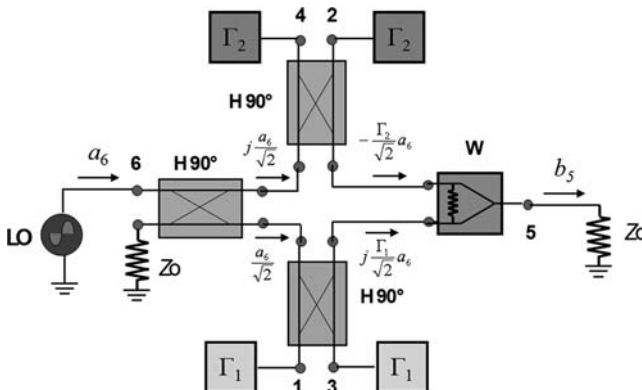


Figure 24.3 Simplified block diagram of multi-port direct modulator.

Table 24.1 QPSK direct modulation chart.

symbols	Gray code	Γ_1	Γ_2	Z_1	Z_2	b_5
1	00	1	1	Open	Open	$0.5 \cdot (1 + j) \cdot a_6$
2	01	-1	1	Short	Open	$0.5 \cdot (-1 + j) \cdot a_6$
3	11	-1	-1	Short	Short	$0.5 \cdot (-1 - j) \cdot a_6$
4	10	1	-1	Open	Short	$0.5 \cdot (1 - j) \cdot a_6$

Table 24.2 8PSK direct modulation chart.

symbols	Gray code	Γ_1	Γ_2	Z_1	Z_2	b_5
1	000	1	0	Open	Z_0	$0.5 \cdot a_6$
2	001	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	$\frac{\sqrt{2}+1}{\sqrt{2}-1} \cdot Z_0$	$\frac{\sqrt{2}+1}{\sqrt{2}-1} \cdot Z_0$	$\frac{0.5}{\sqrt{2}} \cdot (1 + j) \cdot a_6$
3	011	0	1	Z_0	Open	$0.5 \cdot j \cdot a_6$
4	010	$-\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	$\frac{\sqrt{2}-1}{\sqrt{2}+1} \cdot Z_0$	$\frac{\sqrt{2}+1}{\sqrt{2}-1} \cdot Z_0$	$\frac{0.5}{\sqrt{2}} \cdot (-1 + j) \cdot a_6$
5	110	-1	0	Short	Z_0	$-0.5 \cdot a_6$
6	111	$-\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$	$\frac{\sqrt{2}-1}{\sqrt{2}+1} \cdot Z_0$	$\frac{\sqrt{2}-1}{\sqrt{2}+1} \cdot Z_0$	$\frac{0.5}{\sqrt{2}} \cdot (-1 - j) \cdot a_6$
7	101	0	-1	Z_0	Short	$-0.5 \cdot j \cdot a_6$
8	100	$\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$	$\frac{\sqrt{2}+1}{\sqrt{2}-1} \cdot Z_0$	$\frac{\sqrt{2}-1}{\sqrt{2}+1} \cdot Z_0$	$\frac{0.5}{\sqrt{2}} \cdot (1 - j) \cdot a_6$

Table 24.3 Square 16-QAM direct modulation chart.

symbols	Gray code	Γ_1	Γ_2	Z_1	Z_2	b_5
1	1000	0.75	-0.75	$7 \cdot Z_0$	$1/7 \cdot Z_0$	$(0.375 - j \cdot 0.375) \cdot a_6$
2	1001	0.75	-0.25	$7 \cdot Z_0$	$3/5 \cdot Z_0$	$(0.375 - j \cdot 0.125) \cdot a_6$
3	1011	0.75	0.25	$7 \cdot Z_0$	$5/3 \cdot Z_0$	$(0.375 + j \cdot 0.125) \cdot a_6$
4	1010	0.75	0.75	$7 \cdot Z_0$	$7 \cdot Z_0$	$(0.375 + j \cdot 0.375) \cdot a_6$
5	1110	0.25	0.75	$5/3 \cdot Z_0$	$7 \cdot Z_0$	$(0.125 + j \cdot 0.375) \cdot a_6$
6	1111	0.25	0.25	$5/3 \cdot Z_0$	$5/3 \cdot Z_0$	$(0.125 + j \cdot 0.125) \cdot a_6$
7	1101	0.25	-0.25	$5/3 \cdot Z_0$	$3/5 \cdot Z_0$	$(0.125 - j \cdot 0.125) \cdot a_6$
8	1100	0.25	-0.75	$5/3 \cdot Z_0$	$1/7 \cdot Z_0$	$(0.125 - j \cdot 0.375) \cdot a_6$
9	0100	-0.25	-0.75	$3/5 \cdot Z_0$	$1/7 \cdot Z_0$	$(-0.125 - j \cdot 0.375) \cdot a_6$
10	0101	-0.25	-0.25	$3/5 \cdot Z_0$	$3/5 \cdot Z_0$	$(-0.125 - j \cdot 0.125) \cdot a_6$
11	0111	-0.25	0.25	$3/5 \cdot Z_0$	$5/3 \cdot Z_0$	$(-0.125 + j \cdot 0.125) \cdot a_6$
12	0110	-0.25	0.75	$3/5 \cdot Z_0$	$7 \cdot Z_0$	$(-0.125 + j \cdot 0.375) \cdot a_6$
13	0010	-0.75	0.75	$1/7 \cdot Z_0$	$7 \cdot Z_0$	$(-0.375 + j \cdot 0.375) \cdot a_6$
14	0011	-0.75	0.25	$1/7 \cdot Z_0$	$5/3 \cdot Z_0$	$(-0.375 + j \cdot 0.125) \cdot a_6$
15	0001	-0.75	-0.25	$1/7 \cdot Z_0$	$3/5 \cdot Z_0$	$(-0.375 - j \cdot 0.125) \cdot a_6$
16	0000	-0.75	-0.75	$1/7 \cdot Z_0$	$1/7 \cdot Z_0$	$(-0.375 - j \cdot 0.375) \cdot a_6$

16-QAM, and $\left(\frac{\sqrt{2}+1}{\sqrt{2}-1}\right)^2 \approx 34$ for 8PSK and 16-Star QAM. Therefore, QPSK, 8PSK, and Star modulations are easier to implement. It is to be noted that, for demodulation purposes, the same modulations are less sensitive to both amplitude and phase noise.

Table 24.4 16-Star QAM direct modulation chart.

symbols	Gray code	Γ_1	Γ_2	Z_1	Z_2	b_5
1	0000	1	0	Open	Z_0	$0.5 \cdot a_6$
2	0001	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	$\frac{\sqrt{2}+1}{\sqrt{2}-1} \cdot Z_0$	$\frac{\sqrt{2}+1}{\sqrt{2}-1} \cdot Z_0$	$\frac{0.5}{\sqrt{2}} \cdot (1+j) \cdot a_6$
3	0011	0	1	Z_0	Open	$0.5 \cdot j \cdot a_6$
4	0010	$-\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	$\frac{\sqrt{2}-1}{\sqrt{2}+1} \cdot Z_0$	$\frac{\sqrt{2}+1}{\sqrt{2}-1} \cdot Z_0$	$\frac{0.5}{\sqrt{2}} \cdot (-1+j) \cdot a_6$
5	0110	-1	0	Short	Z_0	$-0.5 \cdot a_6$
6	0111	$-\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$	$\frac{\sqrt{2}-1}{\sqrt{2}+1} \cdot Z_0$	$\frac{\sqrt{2}-1}{\sqrt{2}+1} \cdot Z_0$	$\frac{0.5}{\sqrt{2}} \cdot (-1-j) \cdot a_6$
7	0101	0	-1	Z_0	Short	$-0.5 \cdot j \cdot a_6$
8	0100	$\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$	$\frac{\sqrt{2}+1}{\sqrt{2}-1} \cdot Z_0$	$\frac{\sqrt{2}-1}{\sqrt{2}+1} \cdot Z_0$	$\frac{0.5}{\sqrt{2}} \cdot (1-j) \cdot a_6$
9	1000	0.5	0	$3Z_0$	Z_0	$0.25 \cdot a_6$
10	1001	$\frac{1}{2\sqrt{2}}$	$\frac{1}{2\sqrt{2}}$	$\frac{2\sqrt{2}+1}{2\sqrt{2}-1} \cdot Z_0$	$\frac{2\sqrt{2}+1}{2\sqrt{2}-1} \cdot Z_0$	$\frac{0.25}{\sqrt{2}} \cdot (1+j) \cdot a_6$
11	1011	0	0.5	Z_0	$3Z_0$	$0.25 \cdot j \cdot a_6$
12	1010	$-\frac{1}{2\sqrt{2}}$	$\frac{1}{2\sqrt{2}}$	$\frac{2\sqrt{2}-1}{2\sqrt{2}+1} \cdot Z_0$	$\frac{2\sqrt{2}+1}{2\sqrt{2}-1} \cdot Z_0$	$\frac{0.25}{\sqrt{2}} \cdot (-1+j) \cdot a_6$
13	1110	-0.5	0	$1/3 \cdot Z_0$	Z_0	$-0.25 \cdot a_6$
14	1111	$-\frac{1}{2\sqrt{2}}$	$-\frac{1}{2\sqrt{2}}$	$\frac{2\sqrt{2}-1}{2\sqrt{2}+1} \cdot Z_0$	$\frac{2\sqrt{2}-1}{2\sqrt{2}+1} \cdot Z_0$	$\frac{0.25}{\sqrt{2}} \cdot (-1-j) \cdot a_6$
15	1101	0	-0.5	Z_0	$1/3 \cdot Z_0$	$-0.25 \cdot j \cdot a_6$
16	1100	$\frac{1}{2\sqrt{2}}$	$-\frac{1}{2\sqrt{2}}$	$\frac{2\sqrt{2}+1}{2\sqrt{2}-1} \cdot Z_0$	$\frac{2\sqrt{2}-1}{2\sqrt{2}+1} \cdot Z_0$	$\frac{0.25}{\sqrt{2}} \cdot (1-j) \cdot a_6$

24.2.3 Multi-port interferometer for antenna array

An antenna array with adaptive electronically steerable beams will allow mobility, and easy setup compared to a directional one using fixed high gain [15], [16].

The geometrical model of a four-element planar antenna array is presented in Figure 24.4. The patch antennas are spaced by $\lambda_0/2$, where λ_0 is the free-space wavelength. Due to the angle of arrival θ , a difference between two consecutive propagation paths, Δx , will appear.

Consequently, the RF input signals a_i will be phase shifted one versus the other by an angle $\Delta\phi$. As known, this phase shift is directly related to the path difference, expressed by the next equation:

$$\Delta x_i = \lambda_0 \cdot \frac{\Delta\phi_i}{2\pi}, \quad i = 2, 3, 4. \quad (24.25)$$

Using the previous equation, and the geometrical model of Figure 24.4, the angle of arrival can be obtained as follows:

$$\sin \theta = \frac{\Delta x_i}{(i-1) \frac{\lambda_0}{2}} = \frac{\Delta\phi_i}{(i-1)\pi}, \quad i = 2, 3, 4. \quad (24.26)$$

The block diagram of a four element antenna array is shown in Figure 24.5. Its architecture is based on a 4×4 Butler matrix using an original topology, adapted to

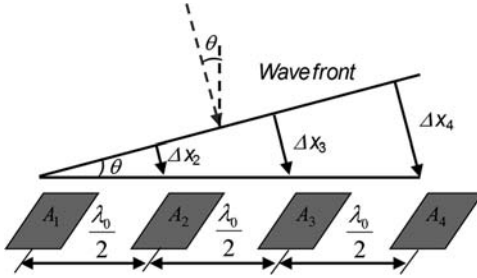


Figure 24.4 Geometrical model of a four-element antenna array.

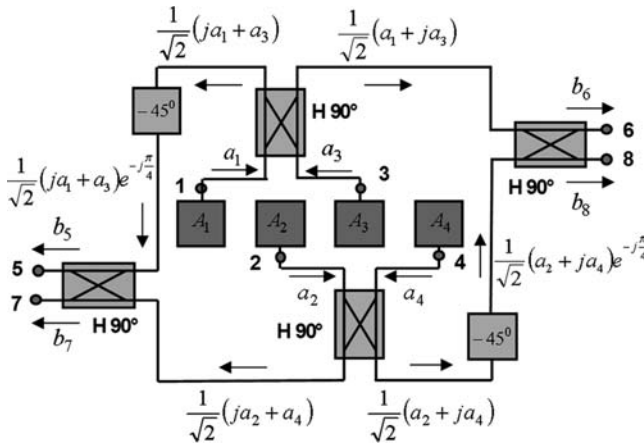


Figure 24.5 Block diagram of four-element antenna array.

millimeter wave frequencies, which avoids any cross line. The four patch antennas are connected to a multi-port circuit, having four inputs and four outputs.

This circuit is composed of four 90° hybrid couplers and two 45° phase shifters, implemented by $\lambda_g/8$ transmission lines. Due to their small dimensions, the patch antennas are integrated on the same substrate.

In order to calculate the multi-port output signals, b_5 to b_8 , the pattern factor of a patch antenna in the H-plane is considered [17]:

$$F_H(\theta) = \cos \theta \cdot \frac{\sin \left[\frac{\beta \cdot W}{2} \cdot \sin \theta \right]}{\frac{\beta \cdot W}{2} \cdot \sin \theta}. \quad (24.27)$$

In this equation, β is the free-space propagation constant and $W = \lambda_g/2 = \lambda_0 / (2\sqrt{\epsilon_{eff}})$ represents the patch width, selected to obtain a proper radiation resistance R to the input ($R = 50 \Omega$). Hence, the previous equation becomes:

$$F_H(\theta) = \cos \theta \cdot \frac{\sin \left[\frac{\pi}{2\sqrt{\epsilon_{eff}}} \cdot \sin \theta \right]}{\frac{\pi}{2\sqrt{\epsilon_{eff}}} \cdot \sin \theta}. \quad (24.28)$$

According to Equations (24.25) and (24.26), and using the geometrical model of Figure 24.4, the four input signals of the multi-port antenna array, a_1 to a_4 , are:

$$a_i = a \cdot e^{-j\Delta\phi_i} \cdot F_H(\theta) = a \cdot e^{-j(i-1)\pi \cdot \sin \theta} \cdot F_H(\theta), \quad i = 1, 2, 3, 4. \quad (24.29)$$

Based on the block diagram of Figure 24.5 and the S matrix of the 90° hybrid coupler given by Equation (24.2), the four multi-port output signals b_5 to b_8 can be expressed as follows:

$$b_5 = \frac{1}{2} (-a_1 \cdot e^{-j\frac{\pi}{4}} + ja_2 + ja_3 \cdot e^{-j\frac{\pi}{4}} + a_4) \quad (24.30)$$

$$b_6 = \frac{1}{2} (ja_1 + ja_2 \cdot e^{-j\frac{\pi}{4}} - a_3 + ja_4 \cdot e^{-j\frac{\pi}{4}}) \quad (24.31)$$

$$b_7 = \frac{1}{2} (ja_1 \cdot e^{-j\frac{\pi}{4}} - a_2 + a_3 \cdot e^{-j\frac{\pi}{4}} + ja_4) \quad (24.32)$$

$$b_8 = \frac{1}{2} (a_1 + ja_2 \cdot e^{-j\frac{\pi}{4}} + ja_3 - a_4 \cdot e^{-j\frac{\pi}{4}}). \quad (24.33)$$

Figure 24.6 shows the computed normalized output signals versus the angle of arrival, using the previous equations. For each output signal, an individual maximum is obtained by shifting θ in a 180° range. The side-lobes are at least 8 dB below the main lobe and the angles of arrival corresponding to the maximum signals are around: -45° , -15° , 15° , and 45° . Therefore, the main lobe of the antenna array can be shifted by 30° multiplies.

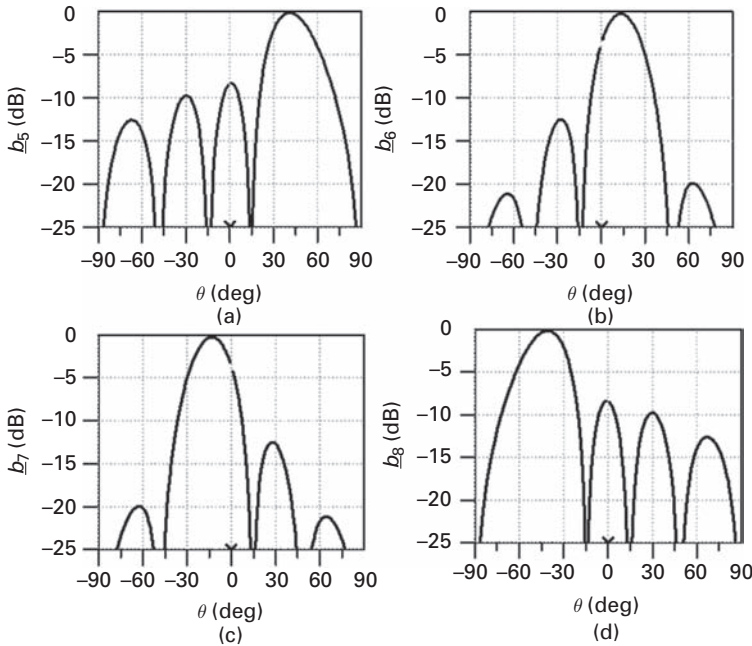


Figure 24.6 Computed normalized output signals versus the angle of arrival.

The antenna array system based on the multi-port interferometer is used in a multiple-input multiple-output (MIMO) architecture as explained in [Section 24.4](#). As seen, this spatial selectivity is achieved by using four fixed received/transmitted beams. A gain improvement, compared with an omnidirectional antenna pattern, is obtained for each angle of arrival. ADS Momentum simulations of this antenna array using the fabrication technology and components presented in the [next section](#) show an average directivity of 11 dBi and an average gain of around 8 dB.

Therefore, the proposed beamforming will improve the overall signal-to-noise ratio of the communication link, decreasing the global bit error rate (BER).

24.3 Multi-port V-band practical implementation, modeling, and analysis

In the last decade the multi-port circuit was implemented in various architectures and technologies. The choice of the fabrication technology depends on the operating frequency and application. Miniature Hybrid Microwave Integrated Circuits (MHMIC) [9] and Monolithic Microwave Integrated Circuits (MMIC) [10] technologies were chosen to implement the prototypes of various communication receivers and radar sensors, in homodyne and heterodyne architectures, at Ka and V band frequencies. The Substrate Integrated Waveguide (SIW) [18] and conventional RWG [19] technologies were used at W band frequencies, to support higher power levels requested in radar sensors and phase noise measurement applications.

The typical topology of the circuit was illustrated in [Figures 24.1 and 24.3](#). This multi-port circuit represents the core of the interferometric I/Q demodulator and direct modulator, respectively. Advanced Design System (ADS) Momentum software of Agilent Technologies was used to perform the circuit design in microstrip implementations. A careful design targets the reflection and insertion loss, and the isolation between the RF and reference signals. It is to be noted that, due to its specific architecture, the insertion loss of an ideal multi-port is 6 dB. Therefore, the phase result of the transmission signals is critical to obtain the desired combination of input signals through the four outputs. The isolation between the millimeter wave inputs is also a very important criterion of the communication systems. Moreover, in order to reduce the DC offset caused by the multiple reflections at the outputs, the return losses at all ports, at the operating frequency, must be excellent.

A V-band multi-port circuit was designed in MHMIC technology using a ceramic substrate, having a 9.9 relative permittivity and 125 μm thickness. The circuit topology is based on the schematic presented in [Figures 24.1 and 24.3](#), as seen in the microphotographs of [Figure 24.7](#). In order to avoid reflections at all unused ports of the multi-port circuit during two port measurements, integrated 50 Ω loads are connected to the open circuit quarter-wave transmission lines (representing virtual RF short-circuits in the operating band). The diameter of the Wilkinson power divider and that of the hybrid coupler are around 1 mm and 0.7 mm, respectively. The 50 Ω line width is practically equal to the substrate thickness.

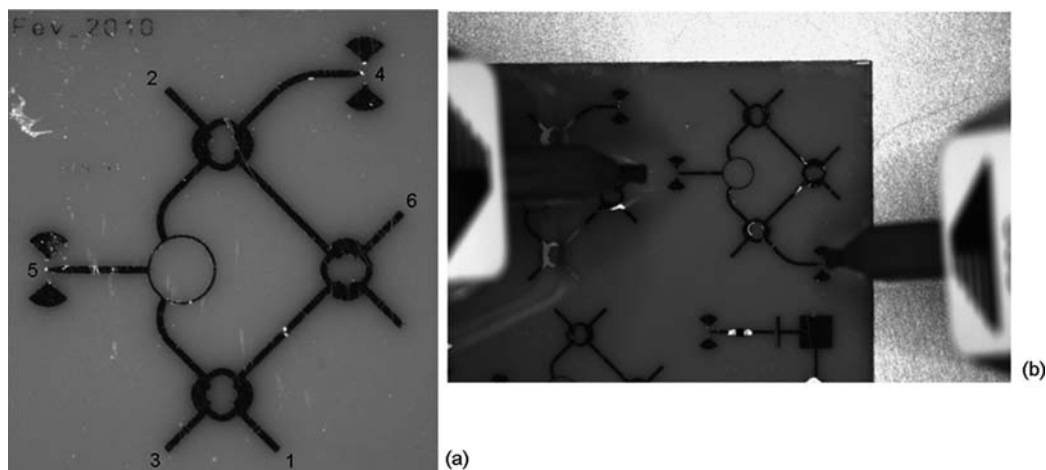


Figure 24.7 Microphotograph of the V-band MHMIC multi-port prepared for S_{54} measurement (a), and a detail of its on-wafer measurements (b).

On-wafer multi-port S parameter measurements are performed using a millimeter wave Precision Network Analyzer (PNA) of Agilent Technologies and a Cascade Microtech probe station equipped with ground signal ground (GSG) 150 μm pico-probes.

Figure 24.8 presents typical S parameter measurement results of this multi-port circuit, over 4 GHz frequency band, between 60 and 64 GHz. The lower measurement limit of 60 GHz is imposed by the PNA frequency range of 60–90 GHz. However, these results can be easily extrapolated over 57–60 GHz band using the acquired measurements, combined with the multi-port ADS Momentum simulations. As seen, good performances in term of return loss at all ports and RF input ports' isolation, are obtained. In addition, the power split between the RF reference input and the four outputs is quasi-constant over the band, and close to the theoretical value of -6 dB, as presented in Figure 24.8(c). It is to be noted that comparable measurement results have been obtained for the other RF input port. As seen in Figure 24.8(d), the use of the V-band hybrid couplers engenders a quasi-constant 90° phase difference over a wide band, suitable for a high-performance quadrature down-converter or direct modulator.

In order to perform advanced system simulations, a multi-port model based on the block diagrams of Figures 24.1 and 24.3 is implemented in ADS, using full S-parameter measurements of the Wilkinson power divider and 90° hybrid coupler. This model has been preferred instead of one based on multi-port full S-parameter measurements, which requires the fabrication of 15 different circuits to cover all two-port measurements. The multi-port model has been validated through comparison with its available measurements, such as presented in Figure 24.8. A very good agreement is obtained between the model simulation and the measurement results.

To demonstrate that the multi-port is a four “ q_i points” circuit [2] having all points spaced by 90° , a harmonic balance simulation is performed at 61 GHz. Both RF input signals have the same frequency (homodyne case). The phase difference between the

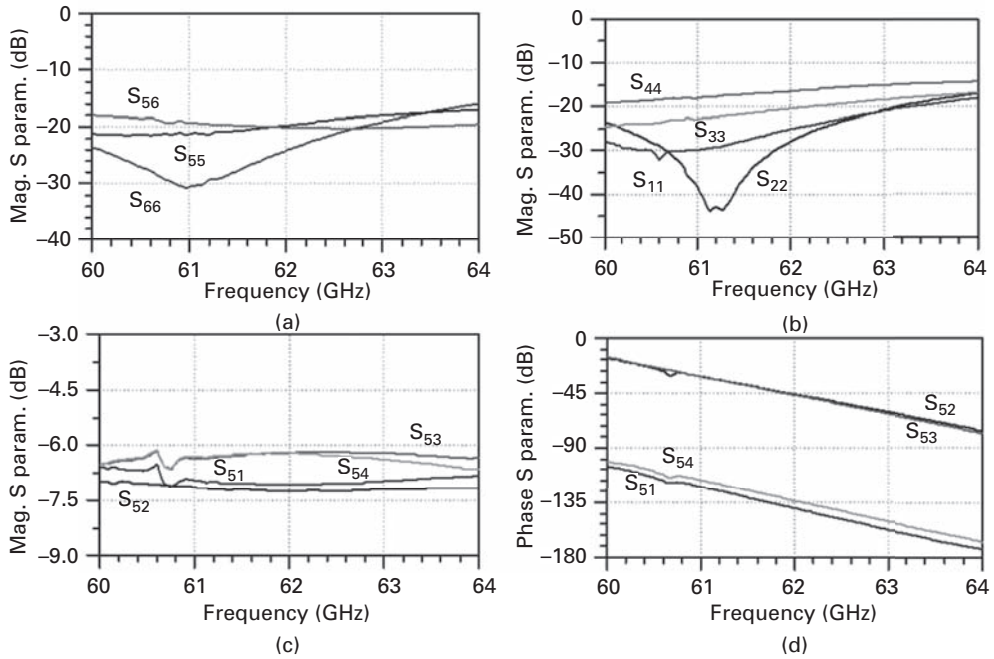


Figure 24.8 Measured typical S-parameters of the V-band multi-port.

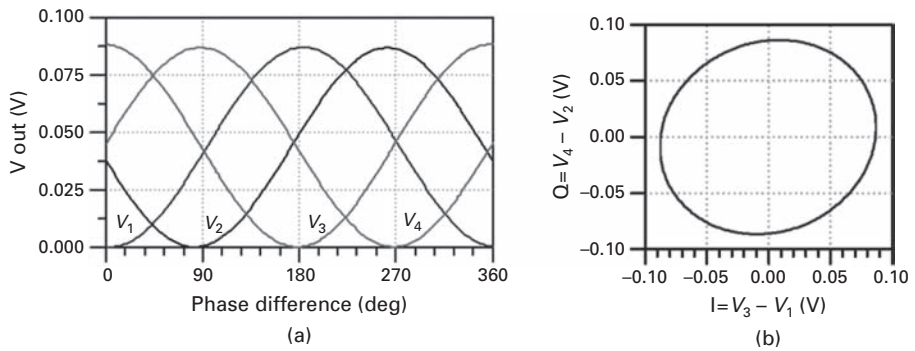


Figure 24.9 Multi-port detected voltages vs. input phase difference (a) and the quadrature signals (b).

millimeter wave inputs is swept in a 360° range and the RF input signal power is set to 0 dBm. Power detectors are connected at the four outputs.

The multi-port detected output voltages versus the RF input signals' phase difference are presented in Figure 24.9(a). As seen, the output voltage minimum values are shifted by 90° multiples, as requested for this architecture. In addition, the output voltages at the pairs of ports 1 and 3, and 2 and 4, respectively, are phase opposites. Therefore, I/Q output signals can be obtained using a differential approach, as sustained by Equations (24.12) and (24.13), and illustrated in Figure 24.9(b). As seen, the shape is quasi-circular

and practically centered in the origin (low DC offset), proving very good capabilities of the proposed multi-port to perform quadrature operation.

In order to demonstrate the multi-port demodulation capabilities, an ADS envelope simulation is performed in homodyne architecture, using a schematic based on Figure 24.1. Various modulation schemes have been implemented and analyzed. Let's see some results. In the case of Amplitude Modulation (AM), the $\mathbf{I}$ vector, defined by Equation (24.16), has a linear variation, while maintaining a constant phase. In the case of frequency modulation (FM), the constant magnitude of $\mathbf{I}$ vector turns on a circle, and rotates clockwise or anticlockwise with the instantaneous frequency difference between the input ports.

As known, in a digital modulation, each combination of magnitude and phase is called a *symbol* and represents a digital bit stream. QAM technique is used to analyze error vectors in the down-conversion process.

In the ADS envelope simulation, the LO power is set at as low as -25 dBm, and the low-noise amplifier (LNA) input signal at -45 dBm. The LNA model is based on the data sheet of the Hittite HMC-ALH382, high gain GaAs HEMT LNA (57–65 GHz, 21 dB of gain, NF = 3.8 dB). The millimeter wave power detector models are based on the Spice model of the zero-bias Schottky diode, model HSCH-9161, of Agilent Technologies. Matching networks are used to improve the return loss over the 60 GHz band. Each differential amplifier module uses two dual wide-band OPA 2658 low-power current feedback operational amplifiers (Burr-Brown). Its model is implemented in ADS based on the manufacturer's data sheet. The total gain of the module is 20 dB, and is validated by measurements on a test bench.

Figure 24.10(a) shows dual-ring 16-Star QAM demodulation results. The modulation scheme looks like a dual 8-PSK having a radius ratio of 2. In addition, Figure 24.10(b) shows the results of the well-known Square 16-QAM demodulation.

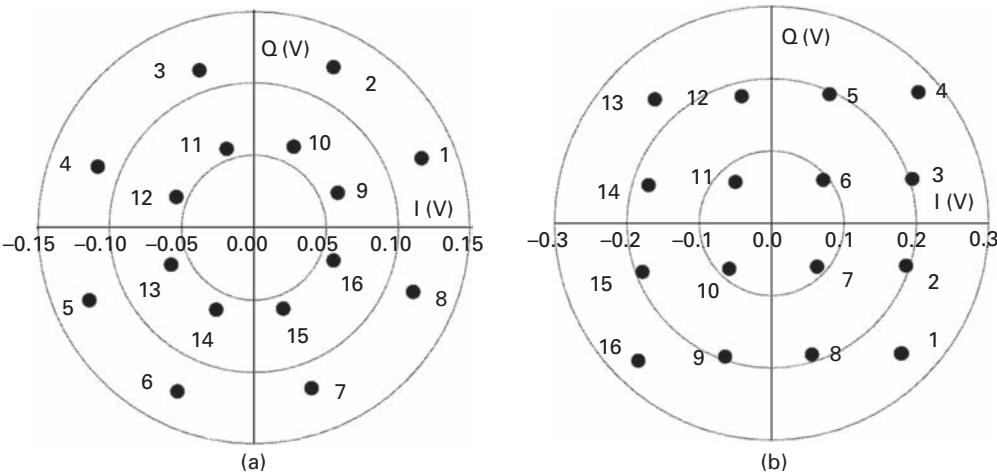


Figure 24.10 Dual-ring 16-Star QAM and Square 16-QAM demodulation results.

The $\mathbf{I}$ vector mapping in the baseband complex plane is presented for each of the 16 symbols. The resulting error vector represents the difference between the actual demodulated and the ideal symbol vectors. To remove the dependence on system gain distribution, related magnitude and phase errors are calculated for each symbol [20]. For both demodulation schemes, these errors are less than 5 percent, showing good demodulation performances without using any calibration techniques.

24.4 Proposed transceiver architectures

According to FCC definition, the transmission bandwidth of UWB signals should be greater than 500 MHz or larger than 20 percent of the central frequency. This open definition does not specify any air interface or modulation for UWB. In the early stages, time-domain impulse radio (IR) dominated UWB technology and still plays a crucial role today. However, driven by the standardization activities, conventional modulation schemes, such as Single Carrier (SC), Frequency Division Multiplexing (FDM), or Orthogonal Frequency Division Multiplexing (OFDM), have also appeared [21].

As known, the OFDM technique partitions a UWB channel in a group of non-selective narrowband channels (using a simple modulation technique such as QPSK), which makes it robust against large delay spreads, by preserving orthogonality in the frequency domain. For multi-carrier systems, two approaches can be used: (i) a high number of carriers (16, 32, 64, or 128) with a corresponding relatively low bit-rate/carrier, or (ii) a small number of carriers (2, 4, or 8) with a corresponding higher bit-rate/carrier. However, at 60 GHz, phase noise and carrier offset will degrade the multi-carrier system performances. Due to the complexity of these architectures, in the authors' opinion, only the second approach is suitable for low-cost 60 GHz UWB WLANs.

The available bandwidth, together with the efficient reuse of spectrum (due to the strong signal attenuation at 60 GHz) makes flexibility, simplicity, and cost, the most critical points of these communication systems. It was demonstrated that the transmissions using SC advanced modulation techniques and directive antennas can achieve comparable performances with OFDM, for a 60 GHz indoor channel [21]. These modulations, such as M-ary QAM and M-ary PSK, will considerably increase the spectral efficiency.

The simulation results of a 60 GHz high-speed multi-port heterodyne receiver have been published in [12]. A bit-rate up to 400 Mb/s has been achieved using a 16 QAM modulation and an IF of 900 MHz. The proposed architecture enables the design of compact and low-cost wireless millimeter wave communication receivers for future high-speed WLANs, according to the IEEE 802.15.3c standard. However, to cope with the Gb/s bit-rates, the millimeter wave circuits' bandwidth must be increased to a few GHz, and the IF must be at least 2.45 GHz.

Due to the high free-space loss at the carrier frequency, transmission on ranges up to 10 meters can be expected. The Friis path loss equation [22] shows that, for equal antenna gains, path loss increases with the square of the carrier frequency. Therefore, 60 GHz communications have an additional 22 dB of path loss when compared to an equivalent

5 GHz system. However, antenna dimensions are inversely proportional to carrier frequency. Therefore, more antennas can be placed within a fixed area and the resulting antenna array will improve the overall antenna gain. The directive pattern of a beam forming antenna array improves the channel multipath profile by limiting the spatial extent of the transmitting and receiving antenna patterns to the dominant transmission path. This aspect opens up new opportunities for wireless system design. The use of smart antennas will also improve the link budget and will reduce the transmitter power [23].

A consequence of the confinement to smaller cells is that the channel dispersion is smaller than the values encountered at lower frequencies, because the echo paths are shorter on average. However, movements of the portable stations, as well as the movement of objects in an environment, cause Doppler effects, relatively severe at 60 GHz, because they are proportional to the carrier frequency. For example, if a person moves at a walking speed of 1.5 m/s, the Doppler spread result is 1200 Hz.

In the case of a single carrier, the use of directive antenna arrays based on multi-ports, a heterodyne multi-port I/Q down-converter, and a multi-port modulator, is considered very promising for indoor low-cost UWB communications.

Figure 24.11 shows a simplified block diagram of a single-carrier millimeter wave multi-port direct conversion transceiver. A MIMO architecture using two phased arrays based on Butler matrices is proposed. This solution appears optimal because a few discrete beam directions are generally sufficient for indoor WLANs. A 20 GHz microwave oscillator and a frequency multiplier generate the 60 GHz signal. The digital signal

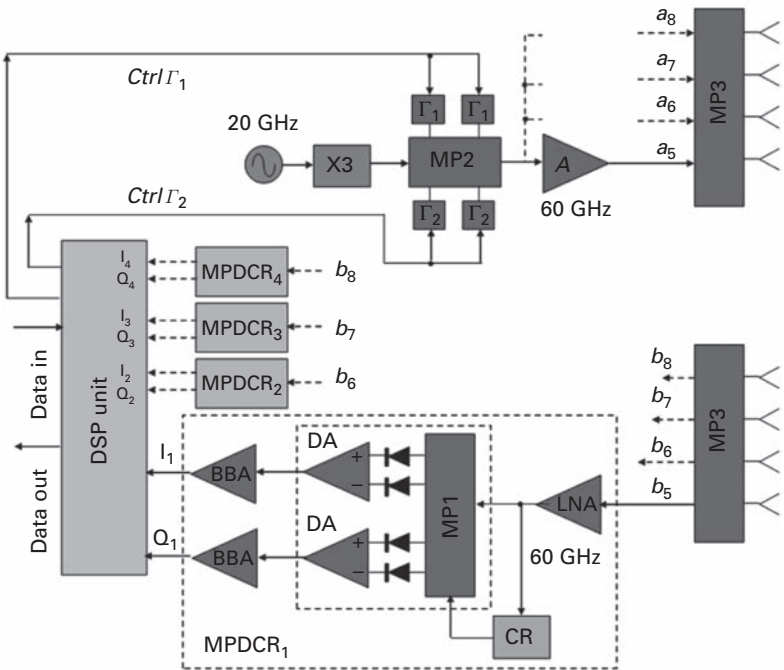


Figure 24.11 Simplified block diagram of a single carrier multi-port direct conversion transceiver.

processing (DSP) unit modulates the carrier using a multi-port direct modulator MP2. According to the 60 GHz related standard, the amplifier's maximum output power is +10 dBm.

Identical multi-port antenna arrays MP3 are used for transmitter and receiver units. Four multi-port direct conversion receivers (MP DCR_i) are connected directly to the receiver's multi-port antenna array. Each receiver is composed of an LNA, a carrier recovery circuit (CR), a multi-port down-converter (MP1, power detectors, and differential amplifiers (DA)), and two baseband amplifiers (BBA). The DSP unit will dynamically select the strongest quadrature received signal (the optimum direction of arrival) and will activate the corresponding millimeter wave amplifier to feed the transmitter's multi-port antenna array MP3.

Analog-to-digital (A/D) converters are used at the DSP input. It is important to underline that the received baseband signal data rate (frequency) is significantly lower than the frequency of the millimeter wave signal. In addition, the signal level is significantly increased due to the overall gain of the millimeter wave receivers. That's why the A/D converters are not connected directly to the multi-port antenna array output, and, consequently, the processing of the millimeter wave signal cannot be moved to the digital side. We note that the baseband analog signal levels are tens of millivolts and their frequency is related to data rate (hundreds of MHz). The baseband signal processing is the optimum and unique solution with today's technology.

The access to the in-phase (I) and quadrature (Q) signals will enable significant additional capabilities, increasing the phase measurement accuracy and offering a straightforward correspondence between the baseband phasor rotation frequency and the Doppler shift, if the same oscillator is used in the receiver part. The carrier recovery circuit is used as reference signal and will compensate the Doppler shift in a hardware approach.

Figure 24.12 shows the proposed heterodyne architecture of a multi-port transceiver. Due to the increased gain of the receiver, omni-directional antennas can also be used. An additional millimeter wave oscillator is introduced in the receiver part, due to the non-zero IF. The Doppler effects and the inherent frequency shift between millimeter wave oscillators are compensated using a phase-locked loop (PLL) circuit, operating at IF. The second down-conversion can easily use conventional means, due to the relatively low operating frequency. To cope with data-rates of 500 Mb/s, the IF of the heterodyne receiver is chosen at 900 MHz. If the data-rate is increased to 1 Gb/s, the IF can be chosen at 2.45 GHz.

As seen in the [previous section](#) (Figure 24.10), in the case of a perfect synchronism, the demodulation errors, without any multi-port calibration, are less than 5 percent. However, a digital compensation technique is implemented in the DSP unit. For example, based on the minimum "distance" between each demodulated quadrature symbol and the ideal map of symbols (corresponding to each modulation scheme), the inherent mismatch in demodulated symbol mapping is compensated. Furthermore, the data output is generated according to each specific modulation chart (e.g. see [Tables 24.1 to 24.4](#) for QPSK, 8PSK, Square 16-QAM, and 16-Star QAM) using digital-to-analog (D/A) converters.

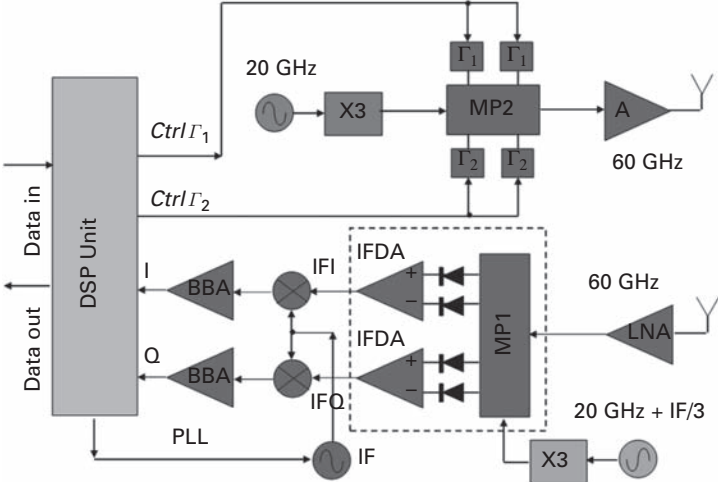


Figure 24.12 Simplified block diagram of a single carrier multi-port heterodyne transceiver.

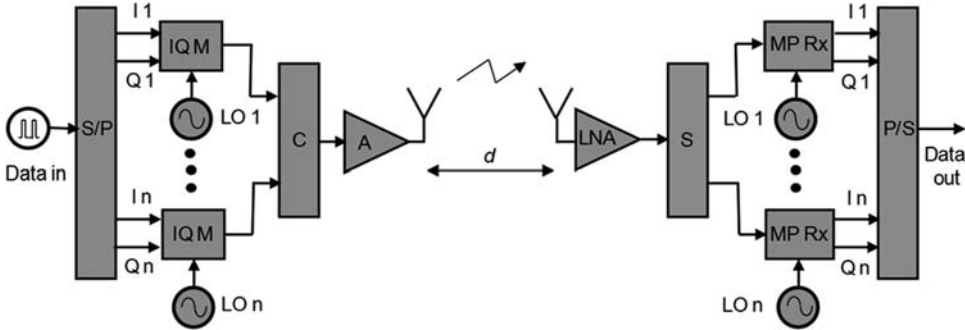


Figure 24.13 Simplified block diagram of an FDM wireless link.

The wireless link is implemented using a FDM system based on UWB millimeter wave multi-port circuits, as seen in Figure 24.13. The channel plan is related to the requested bit-rate, modulation technique, and available band. A serial to parallel converter (S/P) with $2n$ parallel outputs, n millimeter wave LO, n quadrature modulators, and n way millimeter wave combiner (C), are used to generate the FDM signal. This signal is amplified (A) and transmitted through an antenna. After being amplified by the LNA, the received signal is split (S) and coherently demodulated, using n multi-port receivers (MP Rx). Finally, a parallel-to-serial data converter (P/S) generates the output data stream. Due to the UWB characteristics, identical multi-port circuits can be used for both direct modulation and demodulation.

Direct Sequence (DS) UWB is often referred to as impulse, baseband, or zero carrier technique. It operates by sending Gaussian low-power shaped pulses, coherently received by the receiver. Considering that the system operates using pulses, the transmission spreads out, typically, many hundreds of MHz or even several GHz, over a wide

bandwidth. To enable data to be carried, DS UWB transmissions can be modulated in multiple ways. For example, Pulse Position Modulation (PPM) encodes the information by modifying the time interval and, hence, the position of the pulses; and Binary Phase Shift Keying (BPSK) reverses the phase of the pulse to modulate the data to be transmitted. Therefore, in order to use a larger bandwidth with reduced power consumption, a new method based on the transposition of Impulse Radio Ultra Wideband (IR UWB) signals at 60 GHz-band can also be taken into account [24]–[26].

As a low-cost 60 GHz IR UWB proposal, presented in Figure 24.14, the transmitter part can be implemented using an oscillator, a millimeter wave switch, and an amplifier. A pulse generator (first PG) generates sub-nanoseconds pulses (for example pulse width around 350 ps, in order to reach 3 GHz bandwidth). The 60 GHz carrier will be digitally pulse position modulated (PPM) using a millimeter wave switch. After amplification, Gaussian pulses are emitted over several GHz of bandwidth, centered into the 60 GHz band. In order to implement the receiver, either a mixer or a detector can be used. If a mixer is used, a millimeter wave LO is needed. The mixer can be implemented using the low-cost, low-power consumption multi-port down-converter. The oscillator is not required when a topology with a detector is chosen, as presented in the figure. Therefore, the receiver is composed of three main modules: a low-noise amplifier, a 60 GHz detector, and a correlator. A pulse generator (second PG) is used to control the sample and hold (S/H) circuit. The main advantage of this architecture is that no phase information is needed, and thus, no sophisticated coherent stable sources or carrier recovery circuits are involved. It is to be noted that the pulses can also be modulated using the BPSK, with minimal architectural changes in both transmitter and receiver modules. In order to transmit data information, instead of modifying the position of pulses as for PPM, the phase of sub-nanosecond pulses will be reversed at the transmitter. The receiver must be able to observe these 180° phase changes; therefore, a multi-port based phase-detector can be successfully used. It is to be noted that, as usual in block diagrams, the DSP include required A/D and D/A converters.

The proposed architectures enable the design of compact and low-cost wireless millimeter wave transceivers for future UWB wireless communication systems.

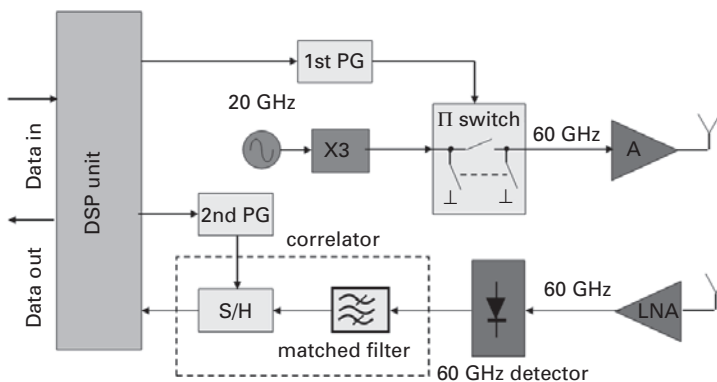


Figure 24.14 Simplified block diagram of an impulse-radio transceiver.

24.5 Advanced system simulations

Advanced Design System software of Agilent Technologies is used to validate multi-port transceiver architectures. Both single-carrier and multi-carrier approaches are discussed in this chapter. The simulation block diagrams are drawn according to the schematics presented in the [previous section](#).

In the first instance, a single-carrier architecture is analyzed. [Figure 24.15](#) shows the spectrum of the QPSK signal used in the simulations. The symbol rate per carrier (SRC) is selected at 500 MS/s, corresponding to a data rate of 1 Gb/s. The main lobe is centered, in this case, at 62 GHz.

According to well known Friis equation [22], the free space LOS attenuation is equal to 88 dB for $d = 10$ m range. In this system analysis, the antenna gains are set at 10 dBi. As in the multi-port analysis section, the LNA gain and the noise figure (NF) are 21 dB and 3.8 dB, respectively, a common value for today's 60 GHz integrated amplifiers. To obtain data output squared waves, and, consequently, perfect demodulated constellations, limiters are used in the baseband (BB) stage of the receiver. We note that this approach can be used only for BPSK or QPSK modulation schemes.

As seen in [Figure 24.16\(a\)](#), [\(b\)](#), and [\(c\)](#), for a pseudo-random bit sequence of 200 ns length, the output demodulated I signals have the same bit sequence as that transmitted. Equivalent results are obtained for Q signals.

[Figure 24.17](#) shows the BER variation versus the E_b/N_0 ratio, where E_b is the average energy of a modulated bit, and N_0 is the noise power spectral density. Obviously, this six-port receiver architecture using a single-carrier scheme has excellent BER performances, close to the theoretical one (dotted line). This result proves its high potential to be used in wireless HD (High Definition) communications.

A second analysis based on four carrier FDM architecture to validate the use of 60 GHz UWB multi-ports in multi-carrier systems is performed. A 4 Gb/s data-link is analyzed using the QPSK modulation of equidistant carriers at 60, 61, 62, and 63 GHz, respectively, and an SRC of 500 MS/s. The data-rate can be further increased, if needed, using

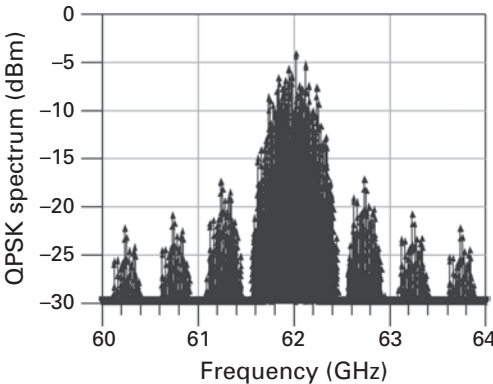


Figure 24.15 Spectrum of the single carrier QPSK signal.

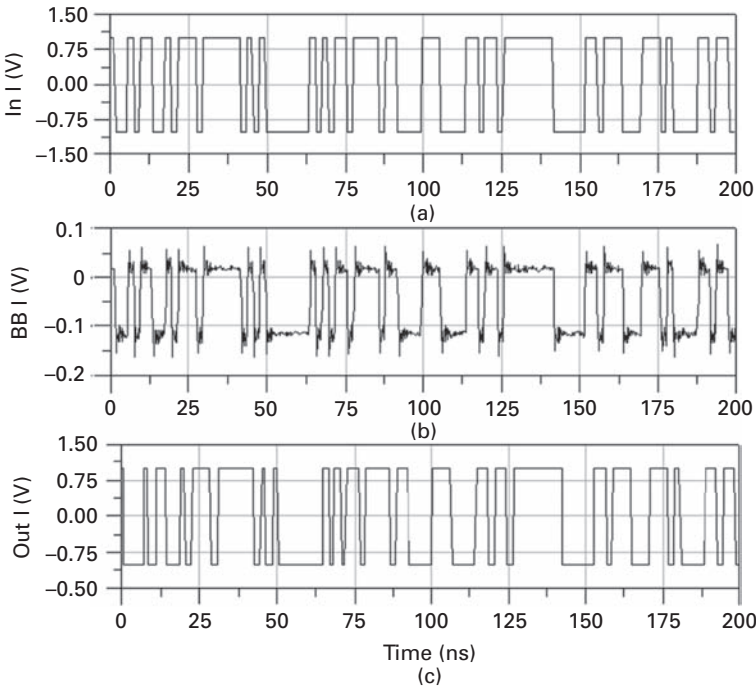


Figure 24.16 Demodulation results of 1 Gb/s QPSK pseudo-random signal: (a) transmitted, (b) received baseband signal, after differential amplifiers, (c) demodulated, at the limiter output.

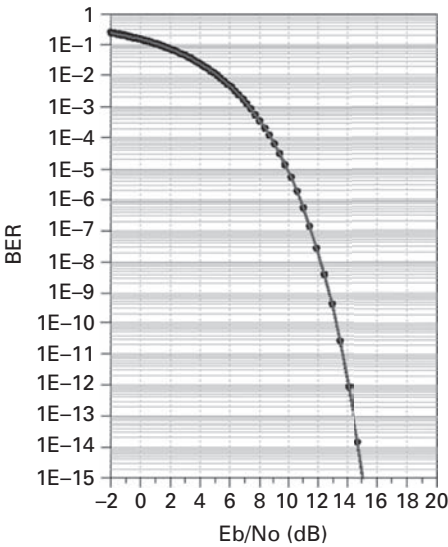


Figure 24.17 BER results for 1 Gb/s QPSK signal.

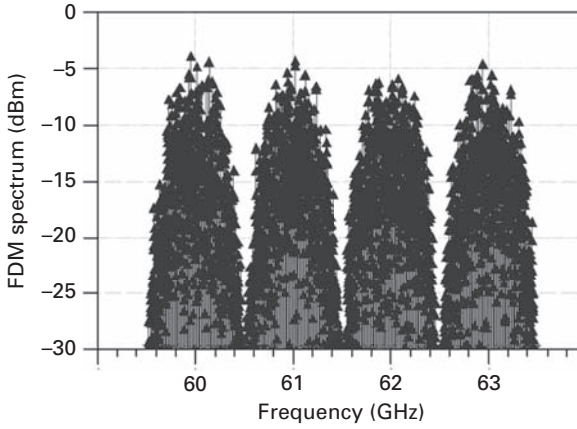


Figure 24.18 Spectrum of the four carrier FDM signal.

the whole spectrum of 7 GHz (57–64 GHz) or the OFDM technique. It is to be noted that our multi-port model is implemented starting from 60 GHz, due to the available instrument's measurement capabilities (60–90 GHz).

According to the schematic presented in Figure 24.13, an ADS envelope simulation is performed. In order to cover the whole FDM spectrum of 4 GHz, a simulation step of $1/(14 \cdot \text{SRC})$ is considered. The FDM spectrum, at the transmitting antenna input, is shown in Figure 24.18. As requested for the FDM, the sub-channel spectrums are not overlapped.

In order to evaluate the wireless link quality, a BER analysis is performed for each sub-channel. As seen in Figure 24.19, BER results are presented as a function of E_b/N_0 ratio. All sub-channels have quasi similar BER results, demonstrating once again the UWB performances of the multi-port.

Furthermore, because the transmitted symbols are uncorrelated, the global BER can be estimated using an analytical approach. The BER average is considered as the sum of the BER related to each sub-channel [27] and is obtained using the following equation:

$$P_{\text{Sys}} = \frac{1}{N \cdot \log_2 M} \cdot \sum_{i=1}^N (BER)_i. \quad (24.34)$$

In the previous equation, P_{Sys} is the global error probability (or BER) of the proposed system, N is the number of carriers which are used, M is related to the modulation levels or number of bits per symbol (for QPSK, $M = 2$), and $(BER)_i$ is the BER of each sub-carrier.

Figure 24.20 shows the BER average curve of the system at 4 Gb/s, as well as the BER corresponding to the single-carrier communication system at 1 Gb/s. It is obvious that, by using the multi-carrier modulation techniques, for a $BER = 10^{-9}$, E_b/N_0 should be incremented by about 2 dB, which is not considered a critical disadvantage. In the meanwhile, a noticeable high data rate of 4 Gb/s is attained.

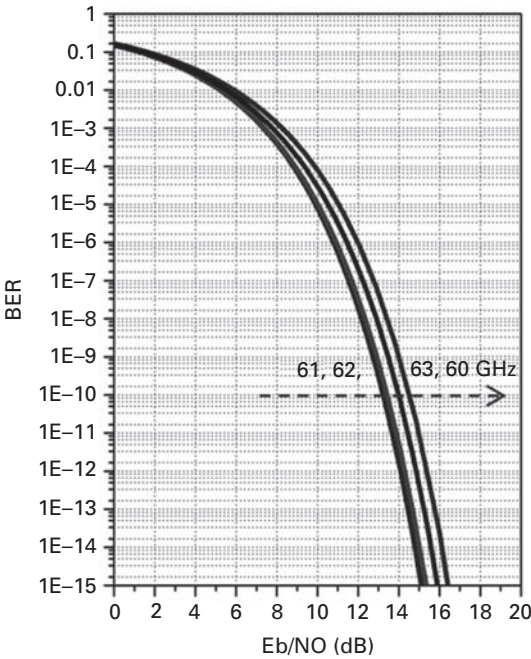


Figure 24.19 BER results for individual FDM sub-channels.

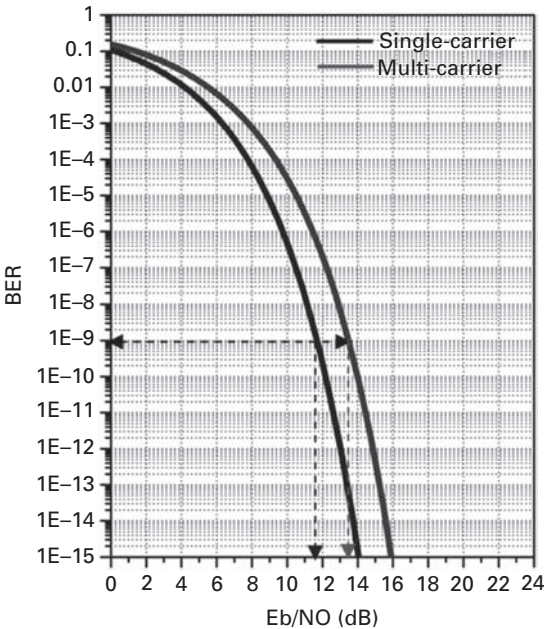


Figure 24.20 BER calculation for FDM system.

24.6 Conclusion

The chapter illustrates the interferometric concept proposed for V-band UWB multi-gigabit/second communication systems. As presented, multi-port circuits can successfully be used in quadrature down-converters, antenna arrays, and direct modulation of millimeter wave signals. An important advantage in using multi-ports is the reduced LO power requested for down-conversion. This is particularly true in millimeter wave applications where the received RF signal is considerably low, reducing both the cost of LO and the leakage between LO and the RF input. Millimeter wave antenna arrays and direct modulators can also take advantage of this interferometric approach with compact and reliable designs.

Single- and multi-carrier architectures are analyzed using advanced system simulations. Despite amplitude and phase imbalances, due to inherent design and fabrication errors of the circuits, the UWB multi-port systems provide excellent demodulation results over the 60 GHz band. For example, the proposed wireless system is able to transmit 4 Gb/s data-rate up to 10 m range, providing a BER of 10^{-9} , as requested for an uncoded HDTV wireless transmission in home or office environments.

Future works will benefit from rapid advances in MMIC technologies to design interferometric millimeter wave transceivers for advanced communication, radar and imaging systems.

References

- [1] S.B. Cohn, N.P. Weinhouse, An Automatic Microwave Phase Measurement System. *Microwave Journal*, 7:2 (1964), 49–56.
- [2] G.F. Engen, C. A. Hoer, Application of an Arbitrary 6-Port Junction to Power-Measurement Problems. *IEEE Transactions on Instrumentation and Measurement*, 21:11 (1972), 470–474.
- [3] G.F. Engen, The Six-Port Reflectometer. An Alternative Network Analyzer. *IEEE Transactions on Microwave Theory and Techniques*, 25:12 (1977), 1075–1077.
- [4] G.F. Engen, An Improved Circuit for Implementing the Six-Port Technique of Microwave Measurements. *IEEE Transactions on Microwave Theory and Techniques*, 25:12 (1977), 1080–1083.
- [5] E. Moldovan, R.G. Bosisio, Wu Ke, S. O. Tatu, *Microwave and Millimeter Wave Technologies*, ch. 19 Multi-Port Technology and Applications, 363–388, 1st edn. (Vienna: In-Tech, 2010).
- [6] J. Li, Ke Wu, R. G. Bosisio, A Collision Avoidance Radar Using Six-Port Phase/Frequency Discriminator (SPFD). *Proceedings of IEEE Microwave Theory and Techniques Symposium*, (1994), 1553–1556.
- [7] J. Li, R. G. Bosisio, Ke Wu, Computer and Measurement Simulation of a New Digital Receiver Operating Directly at Millimeter-Wave Frequencies. *IEEE Transactions on Microwave Theory and Techniques*, 43:12 (1995), 2766–2772.
- [8] J. Li, R. G. Bosisio, Ke Wu, Dual-Ton Calibration of Six-Port Junction and Its Application to the Six-Port Direct Digital Millimetric Receiver. *IEEE Transactions on Microwave Theory and Techniques*, 44:1 (1996), 93–99.

- [9] S. O. Tatu, E. Moldovan, Ke Wu, R. G. Bosisio, A New Direct Millimeter Wave Six-Port Receiver. *IEEE Transactions on Microwave Theory and Techniques*, **49**:12 (2001), 2517–2522.
- [10] S. O. Tatu, E. Moldovan, G. Brehm, Ke Wu, R. G. Bosisio, Ka-Band Direct Digital Receiver. *IEEE Transactions on Microwave Theory and Techniques*, **50**:11 (2002), 2436–2442.
- [11] S. O. Tatu, E. Moldovan, Ke Wu, R. G. Bosisio, T. Denidni, Ka-Band Analog Front-End for Software-Defined Direct Conversion Receiver. *IEEE Transactions on Microwave Theory and Techniques*, **53**:9 (2005), 2678–2776.
- [12] S. O. Tatu, E. Moldovan, V-Band Multiport Heterodyne Receiver for High-Speed Communication Systems. *EURASIP Journal on Wireless Communications and Networking*, Article ID 34358 (2007).
- [13] D. Hammou, N. Khaddaj Mallat, E. Moldovan, *et al.*, V-band Six-Port Down-conversion Techniques, *International Symposium on Signals, Systems, and Electronics, ISSSE*, conference CD, IEEE catalog Number 07EX1869C, (2007), 379–382.
- [14] R. G. Bosisio, Y. Y. Zhao, X. Y. Xu, *et al.*, New-Wave Radio, *IEEE Microwave Magazine*, **2** (2008), 89–100.
- [15] K. Leong, T. Itoh, Advanced and Intelligent RF Front-End Technology, *Wireless Communication Technology*, Conference Proceedings, (2003), 190–193.
- [16] E. Moldovan, S. O. Tatu, S. Affes, A 60 GHz Multi-Port Front-End Architecture with Integrated Phased Antenna Array, *Microwave and Optical Technology Letters*, **50**:5 (2008), 1371–1376.
- [17] W. L. Stutzman, G. A. Thiele, *Antenna Theory and Design*, 2nd edn. (New York: Artech House, 1997).
- [18] E. Moldovan, R. G. Bosisio, Ke Wu, W-Band Multiport Substrate-Integrated Waveguide Circuits. *IEEE Transactions on Microwave Theory and Techniques*, **54**:2 (2006), 625–632.
- [19] S. O. Tatu, E. Moldovan, S. Affes, *et al.*, Six-Port Interferometric Technique for Accurate W-band Phase Noise Measurements. *IEEE Transactions on Microwave Theory and Techniques*, **56**:6, (2008), 1372–1379.
- [20] S. O. Tatu, R. I. Cojocaru, E. Moldovan, Interferometric Quadrature Down-Converter for 77 GHz Automotive Radar: Modeling and Analysis, *7th European Radar Conference*, (2010), 125–128.
- [21] C. Park, T. S. Rappaport, Short-range Wireless Communications for Next-generation Networks: UWB 60 GHz millimeter-wave WPAN, and ZigBee. *IEEE Wireless Communications*, **14**: 4 (2007), 70–78.
- [22] D. M. Pozar, *Microwave and RF Design of Wireless Systems*, (New York: John Wiley and Sons, 2001).
- [23] S. O. Tatu, E. Moldovan, S. Affes, Low-cost Transceiver Architectures for 60 GHz Ultra Wideband WLANs. *International Journal of Digital Media Broadcasting*, Article ID 382695 (2009).
- [24] M. Devulder, N. Deparis, I. Telliez, 60 GHz UWB Transmitter for use in WLAN Communications. *International Symposium on Signals, Systems, and Electronics, ISSSE 2007*, Conference Proceedings, (2007), 371–374.
- [25] N. Deparis, A. Boé, C. Loyez, N. Rolland, P. A. Rolland, UWB-IR Transceivers for Millimeter Wave WLAN. *32nd IEEE Industrial Electronics Conference* (2006), 4785–4789.
- [26] N. Deparis, A. Bendjaballah, A. Boe, *et al.*, Transposition of Baseband UWB Signal at 60 GHz for High Data Rate Indoor WLAN. *IEEE Microwave and Wireless Components Letters*, **15**: 10 (2005), 609–611.
- [27] J. G. Proakis, *Digital Communications*, 3rd edn. (Boston: McGraw-Hill, 1995).

25 Cross-layer design and digital front-end for cognitive wireless link

Matthieu Gautier, Guillaume Villemaud, Cedric Lévy-Bencheton,
Dominique Noguet, and Tanguy Risset

25.1 Introduction to flexible cognitive wireless link

This chapter presents some algorithms and technology that are used in the Physical (PHY) layer entity of a flexible or cognitive radio. In the context of this chapter, cognitive radio refers to a wireless air interface that implements the cognitive cycle introduced by Mitola in his Ph.D. thesis [20]. This cycle is illustrated on [Figure 25.1](#) and is composed of four main steps: the “Sense” acquires relevant information from the radio environment, the “Analyse” and “Decide” steps represent all that implies some intelligence, “Analyse” for interpreting the observation of the “Sense” step, “Decide” for learning, planning, and decision making, and finally the “Act” step reconfigures the transceiver’s communication parameters, with the transceivers designed with software defined radio (SDR) principles in order to be as flexible as possible.

Mitola underlines the strong need for defining a novel wireless air interface (i.e., Media Access Control MAC and PHY layers) based on cognitive radio. This air interface should support the new functionalities described in [Figure 25.1](#); the MAC layer (or the cognitive manager) operating the “Analyse” and “Decide” steps and the PHY layer operating the “Sense” and “Act” steps.

In order to be as flexible as possible, the communication (i.e. “Act”) step must quickly adapt the characteristics of the transmission. The system should be compatible with several existing wireless standards. More generally, multimode systems are addressing this flexibility and are becoming common in current mobile terminals. Multimode radio based architecture provides undoubtedly the most convenient way to implement a cognitive system.

Thus, the scope of this chapter is the PHY layer of a cognitive radio and we address some issues relative to the definition of a flexible wireless link that could perform both the sensing techniques and a multimode communication. The discussed transceiver (TRX) architecture is described in [Figure 25.2](#). It is composed of three main blocks: the Physical Sensing, the Physical Communication and the Analog RF (radio frequency).

The PHY Sensing block deals with the integration of the sensing operations in order to respond to the needs of the new cognitive radio system. It uses the bandwidth provided by the Analog RF block in order to perform the sensing techniques used to detect any available spectrum.



Figure 25.1 The cognitive cycle.

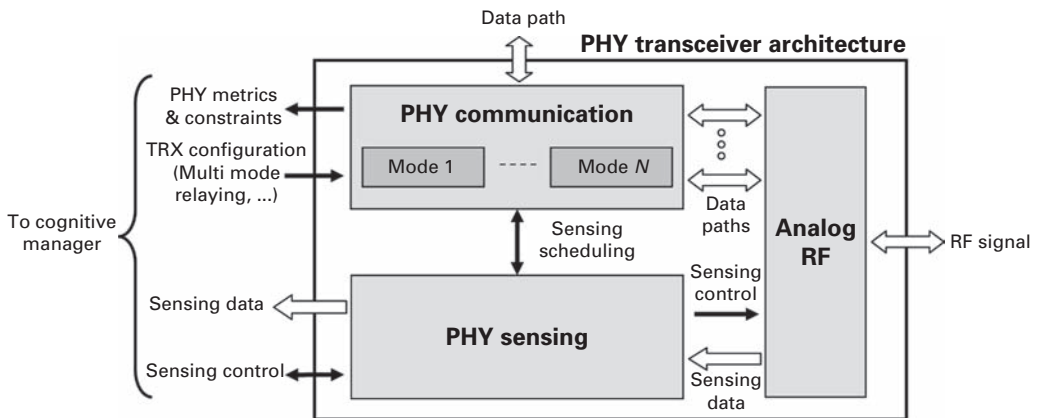


Figure 25.2 Cognitive radio physical transceiver architecture.

The PHY Communication block includes all the functions required for the baseband processing of the data to be transmitted/received over the air. It is expected to be backward compatible to some existing wireless standards and to exhibit the flexibility and the multimode characteristics.

The Analog RF includes all the functions required for the RF processing involved in the up-conversion and down-conversion of the baseband signals, up to the antenna(s). To support both sensing and communication steps, it must exhibit also frequency agility or wideband characteristics.

In Figure 25.2, the thick white-filled arrows represent either communication or sensing data paths. The thinner black arrows represent control interfaces. Two internal control interfaces between the main functional blocks can be identified. The control interface between the PHY Communication and PHY Sensing blocks is a bidirectional interface used for sensing scheduling. The scheduling parameters could be for example [7]:

- Interleaved Spectrum Sensing: the cognitive system periodically senses the active transmission bands, interrupting the data communication.
- Spectrum Reservation Sensing: the cognitive system exploits the silent frequency sensing windows reserved in the active transmission band.

The control interface between the Physical Sensing and Analog RF blocks is used to configure the Analog RF block when the PHY sensing step is active.

The external controls are connected to the Cognitive Manager. From the PHY communication block, the Cognitive Manager receives the transmission related metrics and constraints of the transceiver. These include, but are not restricted to the Channel State Information, the analog RF specifications, spectrum aggregation capabilities, etc.

The interface between the PHY communication block and the Cognitive Manager is used to configure the transmission-related parameters of the transceiver. These include, but are not restricted to, the following parameters: operation mode (standard compatible mode or cognitive radio mode), list of channels to use and associated specifications, transmit power information, etc.

The Sensing Control interface between the PHY Sensing and the Cognitive Manager is a bidirectional interface that outputs the physical sensing constraints and inputs the control for different types of sensing.

Digital implementation of such an architecture must also be defined. Indeed, the sensing block and the multimode characteristic increase considerably the complexity of the hardware part. A way to respond to this challenge is the SDR based system. The main idea is to replace the ever-increasing embedded radio chips by a single, general purpose, processor.

Based on this architecture, many issues have to be addressed in order to define all the blocks:

- Concerning the PHY communication block, an important issue is the decision of switching between different modes that cannot fully be determined RSSI (received signal strength indicator) or SNR (signal-to-noise ratio) measurements. If we consider that the same architecture will be used for different modes or standards depending on local conditions, it is crucial to be able to define some ways to accurately evaluate the reconfiguration gain. Numerous researches (like [6, 4, 23]) are dealing with an efficient use of several possible interface and hardware resources. But a key point is to take into account the global cost of communicating in a certain mode when considering a reconfiguration of the architecture. In the [second section](#) of this chapter, this problem is assessed in a particular scenario: multimode relaying.

On a higher-level view of a flexible network (not necessarily cognitive), multimode relaying is supposed to be an efficient way of reducing the global cost or consumption of this network. What could be obviously considered is that, in a cellular-like system, using nodes in good receiving conditions of the gateway (or base-station or access point) to relay the information for users in bad conditions with a low power standard is a relevant technique. But this is not so easy to manage because it appears that the decision is tightly related to the targeted hardware and the radio channel behavior if you consider the global energy consumption, taking into account not only the PHY layer, but also MAC behavior and a precise evaluation of digital and RF parts consumption

Furthermore it quickly appears that it is impossible to draw generic conclusions for this kind of flexible system without integrating a precise modeling of the architecture and the radio channel conditions. Consequently, the [second section](#) of this chapter will give some insights on how to evaluate such multimode relaying schemes from a global

consumption point-of-view. Hence, it means that ignoring MAC mechanisms or sensing activities could result in false conclusions.

- Concerning the PHY Sensing block, the definition and implementation of algorithms suitable for detecting the primary users (the incumbents) of the scanned band is a key issue. The scanned band should be theoretically all the radio spectrum. However, in practice the cognitive device will scan only a part of the spectrum.

Today, an example of one hot topic for cognitive sensing implementation is the TV White Spaces where candidate bands for cognitive communication are the TV bands. Indeed, the utilization rate of TV bands shows some bands that are not used. In the US, the FCC (Federal Communications Commission) has begun to allow unlicensed usage of TV bands, provided that these unlicensed systems do not cause harmful interference with incumbents [11]. In parallel, an important effort was also made for standardization through the IEEE 802.22 standard [24] or the ECMA 392 standard [10], which target the same frequency bands.

In essence, cognitive radios are radios able to detect the vacant spectrum, also called spectrum holes or white spaces that are considered as band opportunities to be used. Whenever priority rules exist between the cognitive system and incumbents, as in the case of the TV band example mentioned above, the opportunistic system is supposed to leave the band once the service is completed or whenever a primary user requests this band.

Three different mechanisms have been categorized to detect the presence of incumbents [26]: beaconing, geolocation, and spectrum sensing. While the first two techniques require that the incumbents collaborate with the detector (either by sending a beacon or through database registration), the latter does not assume any change at the incumbents' part but only requires the opportunistic system's ability to detect the presence of signal with a very high sensitivity. Sensing methods for opportunistic radio are surveyed in [21].

The [third section](#) of this chapter assesses the study case of TV white space spectrum sensing and presents some solutions as well as their hardware implementation.

Organized into the three sections, this chapter is devoted to two distinguished studies that provide the solutions for defining and implementing a flexible wireless link. Starting with a high-level approach, [Section 25.2](#) will present a way of evaluating the power consumption in a multimode relay considering both MAC and PHY layers. [Section 25.3](#) presents a more deeply focused technical study and also gives deep-level solutions to performing sensing in the TV bands. In the [final section](#), we will present a summary and further discussions.

25.2 A high-level optimization for SDR: multimode relaying

An interesting property of SDR systems is their ability to rapidly change the radio protocol used. This offers the possibility of choosing a communication mode according to a particular objective: quality of service, users' preferences, access cost, or energy reduction.

Still, SDR does not solve all problems. How can operators ensure permanent network connectivity without increasing costs? *Relaying* communication is a way to minimize

cost and energy consumption. A *relay* is a device transmitting data from another user device to an access point. The use of such a relay permits the user to emit at a lower energy mode. Such relays can be deployed by operators or be mobile. In the latter case, the users' terminals act as potential relays.

Therefore in this section we rely on this kind of new scheme for wireless networks in order to illustrate the increasing complexity of system evaluation and global optimization when dealing with the flexible nodes with SDR capabilities. We present an analysis of multimode terminals' energy consumption in the case of multimode relaying with SDR terminals [18]. This analysis includes a careful study of all elements (related to the physical layer) involved in the energy consumption of radio protocols: analog and digital parts. Most existing works on relaying are mainly based on mono-mode relaying. Moreover, few of these works take into account algorithmic *and* RF consumption, thus we will study not only the transmission power but also numerical and analogical energy consumptions. We focus our study on IEEE 802.11g (WiFi) [1], a WLAN standard, 3GPP UMTS [26], a voice and data long-range mobile communication mode, and IEEE 802.15.4, or Zigbee, a low-power WPAN [2].

Evaluating energy consumption is a difficult task when a number of important terminals are concerned, which must be done by simulation (analytical modeling is still too imprecise) and the simulated scenario must be sufficiently simple to provide meaningful information. Examples presented here are quite simple but emphasize that a complete and realistic evaluation of multi-mode networks requires very complex simulation frameworks.

25.2.1 Power consumption model

It is crucial to represent the energy consumption by an accurate model, especially when dealing with simulations. This model is presented and explained hereafter. It evaluates the energy consumption of a multimode SDR terminal on the basis of several key factors.

The physical layer of an SDR terminal is composed of two parts:

- The numerical part realizes the baseband operations and the signal processing. It is represented by the “PHY Communication block” in Figure 25.2.
- The radio part performs the analog signal transformations. It is represented by the “Analog RF” in Figure 25.2.

The energy consumption of an SDR terminal, denoted E_{term} , is the sum of the contribution of these two parts.

At a given moment, an SDR terminal communicates only on a subset of all available modes: the *active modes*. Hence, the energy consumption of a multi-mode SDR terminal is determined by its active modes.

Let us consider an application sending data to a destination. The SDR terminal chooses one mode, m_j , among all other active modes to transmit these data over the air. Each mode is defined by PHY parameters (data coding, modulation, and so on) and by the MAC layer, with a certain frame size. Since different modes are likely to propose different frame size, it is mandatory to evaluate the mode energy per (user data) bit; we call this

energy E_{bit} . Consequently, the numerical energy consumption, E_{num} , and the radio energy consumption, E_{rf} , are also expressed for one (user data) bit, and:

$$E_{bit} = E_{num} + E_{rf}. \quad (25.1)$$

This per-bit approach allows the selection of the least consuming mode, regardless of the amount of transmitted data.

25.2.1.1 Numerical energy model

In an SDR, the physical layer is composed of the programmable blocks which define the protocols to perform for transmission and reception. Each block is tunable, with the parameters defining the exact protocol version. The protocol execution time is expressed in clock cycles. The execution time depends on the algorithm's implementation (i.e. its number of operations) and the processor architecture (i.e. the number of operations realized per processor cycle). Hence, two different terminals may have a different consumption for the same protocol, and two different implementations of the same algorithm on the same terminal may have different energy consumption, too. In the following we evaluate the number of elementary operations for each protocol, which is an upper bound of the number of clock cycles.

In order to evaluate the energy consumption of a block for one mode, say for instance FFT (fast Fourier transform) in 802.11a, it is important to count the number of operations and then to divide this number by the number of useful transmitted bits. The resulting number of operations per bit, or *bitop*, is fixed for each block. The next step is to evaluate the *bitop* for each mode, both in transmission and in reception. The numerical energy consumption per bit, E_{num} (in joules per bit), depends on this *bitop*. It is evaluated as follows [27]:

$$E_{num} = K E_{cpu} = K A_C V_{dd}^2, \quad (25.2)$$

where K is the *bitop* for a mode (in transmission or reception), and $E_{cpu} = A_C V_{dd}^2$ (in joules) represents the target terminal consumption; E_{cpu} depends on the terminal architecture; V_{dd}^2 (in volts) is the processor input voltage, as defined by the processor specifications; A_C (in farads) represents the processor switching capacitance. The difficult task is to evaluate K . In the following we provide some *bitop* values for the protocols under study.

25.2.1.2 Radio energy model

The radio energy consumption for a mode depends on two parameters:

- the analog radio-frequency front-end, $P_{frontend}$ (in watts), based on the transceiver circuit's power consumption,
- the transmission power output, P_{out} (in watts), related to the mode and channel conditions.

The radio energy consumption, E_{rf} (in joules per bit), is evaluated as follows [25]:

$$E_{rf} = \frac{1}{R} (P_{frontend} + \theta P_{out}), \quad (25.3)$$

with R the mode data bitrate (in bit per second), and $\theta = 1$ in transmission, 0 otherwise. Since $1/R$ represents the time to send or receive one data bit, E_{rf} is expressed in joules per bit.

The values of $P_{frontend}$ depends on the front-end specifications. In general, it remains constant for a mode, with independent values for transmission and reception; P_{out} could be fixed or variable, depending on the network rules.

25.2.1.3 Multimode terminal energy consumption

The energy consumption of a multimode terminal τ_i , denoted $E_{term}(\tau_i)$ (in joules), is the sum of E_{bit} from all its active modes. For example, the consumption of τ_i in 802.11g (WiFi) and 802.15.4 (Zigbee) is expressed as follows:

$$\begin{cases} E_{term}^{TX}(\tau_i) = E_{bit}^{TX}(Wifi) + E_{bit}^{TX}(Zigbee) \\ E_{term}^{RX}(\tau_i) = E_{bit}^{RX}(Wifi) + E_{bit}^{RX}(Zigbee) \end{cases}, \quad (25.4)$$

$$\Rightarrow E_{term}(\tau_i) = E_{num}(\tau_i) + E_{rf}(\tau_i), \quad (25.5)$$

with $E_{bit}^{TX}(m_j)$ and $E_{bit}^{RX}(m_j)$ the energy consumption for one bit in a given mode m_j , respectively in transmission and reception.

25.2.1.4 Global energy consumption

We define the *global energy consumption*, E_{global} (in joules), as the sum of all terminals' energy consumption:

$$E_{global} = \sum_{i=1}^n E_{term}(\tau_i), \quad (25.6)$$

with $E_{term}(\tau_i)$ the energy consumption of terminal τ_i , and n the number of energy constrained terminals.

25.2.2 Examples of multimode relay

The energy model presented previously is used to evaluate the energy performance of multimode relays. Based on the energy consumption comparison of two scenarios, the conditions and parameters influencing the power consumption of multimode relay communication are highlighted. These scenarios allow us to understand when and how a multimode relay is useful in reducing the global energy consumption. The study of simple scenarios helps to determine key parameters influencing the energy consumption of relays. Firstly, an 802.11g-to-UMTS relay, realized from a strictly analytical point of view, shows that unlimited relaying may not be useful. Secondly, network simulations of an 802.15.4-to-802.11g relay focus on the impact of the MAC layer on the energy consumption.

25.2.2.1 Scenarios explained

The scenarios chosen for performance evaluation compare a multimode relay scenario with direct connections, depicted in [Figure 25.3](#). In these scenarios, a single access point

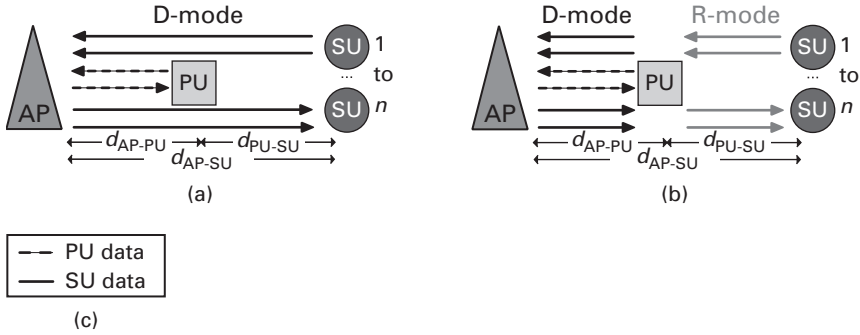


Figure 25.3 Schematic representation of the scenarios studied.

(AP) communicates with energy-constrained users on a mode denoted *D-mode*. Mobile terminals are represented by a primary user (PU) and several secondary users (SU) which communicate with each other on *R-mode*. Since all user terminals have multimode capacities, they can either communicate with AP on *D-mode*, or with each other on *R-mode*. Moreover, PU can act as a multimode relay for any SU on *R-mode*. *R-mode* and *D-mode* will be different for each scenario.

We will estimate when and how multimode relaying brings an energy reduction by comparing these two scenarios:

- S_{direct} , which refers to the direct connections scenario, one PU and n SUs are directly connected to AP (Figure 25.3(a)).
- S_{relay} , which refers to the multimode relay scenario, one PU communicates with AP and relays the n SUs on dedicated connections (Figure 25.3(b)).

In both scenarios, the PU moves between the AP and the SUs, both of which remain fixed in position.

In the two scenarios, the PU and the SUs are SDR. Their numerical processing consumes $E_{cpu} = 0.14$ nJ per operation. The energy consumption can be evaluated through measurements, or by following the front-end specifications. Here, the terminal energy is derived from the specification of circuits designed for multimode (from the ARM 968E-S specifications [3]). The radio front-end is a multimode front-end, whose consumption depends on the selected mode. Since this part evaluates the global energy consumption of multimode relays for selected modes of 802.11g [1], 802.15.4 [2] and UMTS [26], the associated values of P_{frontend} are presented in Table 25.1. UMTS and 802.11g [5] employ the same multiband front-end, whereas 802.15.4 relies on a simpler set of circuits [28].

We have chosen to adapt the transmission power output, P_{out} to the channel conditions. In both scenarios, each mode is associated a receiver sensitivity, $RXSens$ (in dB), which determines the threshold for an accurate signal reception and decoding. For a mode with a channel path-loss PL (in dB), the value of P_{out} (in dB) follows:

$$P_{\text{out(dB)}} = RXSens - PL \quad (RXSens < 0, PL < 0). \quad (25.7)$$

Table 25.1 P_{frontend} for the considered SDR (802.11g and UMTS use the same front-end).

	802.11g and UMTS	802.15.4
Transmission (mW)	338	1
Reception (mW)	198.8	1

The energy consumption evaluation is performed for two examples of these scenarios: an 802.11g-to-UMTS relay and an 802.15.4-to-802.11 g relay. Each evaluation presents a set of parameters influencing the energy consumption of multimode relays.

25.2.2.2 PHY layer only evaluation: UMTS-WiFi scenario

We can now try to evaluate first the energy consumption of a multimode relay, with terminals capable of communicating in 802.11g at 54 Mbps and in UMTS at 384 kbps. The global energy consumption evaluation focuses on the PHY layer, with no consideration of upper layers, no interference, and no retransmissions. The terminal power consumption has been presented in Table 25.1. The *bitop* is evaluated following generic algorithms, and is presented in Table 25.2. The numerical and radio consumptions have been evaluated for one bit, following (25.2), (25.3), and the *bitop* from Table 25.2. The UMTS signal suffers a COST Hata path-loss in a metropolitan environment [8]. The 802.11g signal suffers an ITU-R path-loss considering an indoor office environment with three walls [16]. *RXsens* is set to -71dBm in 802.11g at 54Mbps and to -106dBm in UMTS at 384kbps. Moreover, P_{out} is not limited. Table 25.3 presents this energy per bit for selected modes at a distance of 50 m. For additional details, please refer to [18].

Figure 25.4 presents the energy consumption of S_{relay} , normalized by S_{direct} , when each terminal sends one bit to AP. In Figure 25.4(a), the SU is fixed at $d_{\text{AP-SU}} = 1\text{ km}$, and the PU moves toward the SU. In Figure 25.4(b), the distance between the PU and the SU is fixed at $d_{\text{PU-SU}} = 50\text{ m}$, and both terminals move together from AP to reach the cell border ($d_{\text{AP-SU}} = 1\text{ km}$). Both figures also represent a realization on independent Rayleigh fading channels with standard deviation $\sigma = 0.5$.

As shown in Figure 25.4(a), for a mobile PU, multimode relay has no real impact on the global energy consumption. When the PU and the SU are close to each other, and far from AP, the energy gains are around 5 percent better than direct connections. This behavior could be explained by several factors.

The first noticeable factor comes from the difference in the numerical complexity of both modes (as per Table 25.2). Whereas 802.11g relies on OFDM modulation (Orthogonal Frequency Division Multiplexing), UMTS is based on WCDMA (Wideband Code Division Multiple Access). The OFDM multiple subcarriers rely on Fourier Transform to reach a high bitrate. The WCDMA spreads the coded data by a certain factor (e.g. 4 times in uplink for UMTS at 384 kbps). Due to this spreading, the energy per bit is much higher in UMTS, as represented in Table 25.3. This is especially true for the reception part.

Table 25.2 Number of operations per bit (*bitop*) for the selected modes.

	802.11g at 54 Mbps	UMTS at 384 kbps
Transmission (<i>bitop</i>)	142	745
Reception (<i>bitop</i>)	655	5308

Table 25.3 Energy consumption E_{num} and E_{rf} (in nJ per bit) for the selected modes, at 50 m.

	802.11g at 54 Mbps		UMTS at 384 kbps	
	E_{num}	E_{rf}	E_{num}	E_{rf}
Transmission at 50 m (nJ per bit)	0.02	0.34	0.104	0.88
Reception (nJ per bit)	0.09	0.004	0.744	0.518

The high radio cost of UMTS comes from the front-end active time. Indeed, for the same front-end, UMTS at 384 kbps has higher energy consumption per bit than 802.11g at 54 Mbps (as shown in Table 25.1). It follows that the communication on fast modes is more likely to reduce the terminal energy consumption.

These primary results highlight a trade-off problem between the data rate and its energy consumption. Modes with a higher data rate do not necessarily mean a higher consumption. However, such a trade-off operates on two levels. Firstly, the numerical complexity must be carefully studied: expensive operations lead to energy requirements too high to be useful. Secondly, the radio front-end activity time should be as short as possible. Nevertheless, a too short active time may decrease the reliability when mobile. In addition, the baseband throughput should be sufficiently high to realize operations in the allocated time.

Moreover, depending on the channel conditions, a terminal may connect via a multi-mode relay in an opportunistic manner. For example, in Figure 25.4(b) at $d_{PU-SU} = 600$ m: even though a multimode relay has a negative effect on the global energy consumption, relaying with the presented channel realization leads to energy savings (here, S_{relay} consumption is 7 percent lower than S_{direct} for a particular Rayleigh fading). This behavior remains variable, as it requires a good channel from all links at the same time.

25.2.2.3 PHY+MAC evaluation: WiFi+ZigBee scenario

The previous subsection stated some important points in considering a multimode relay with SDR terminals. This subsection will now explain the influence of the MAC layer, and the necessity to employ network simulation for a more realistic energy evaluation. What is illustrated is a scenario using an 802.15.4-to-802.11g relay, with the CSMA/CA implemented in 802.11g.

The terminal front-end power consumption has been presented in Table 25.1. The *bitop* has been evaluated from generic algorithms and is presented in Table 25.4 for 802.11g at 6 Mbps and 802.15.4 at 20 kbps, respectively. The numerical and radio

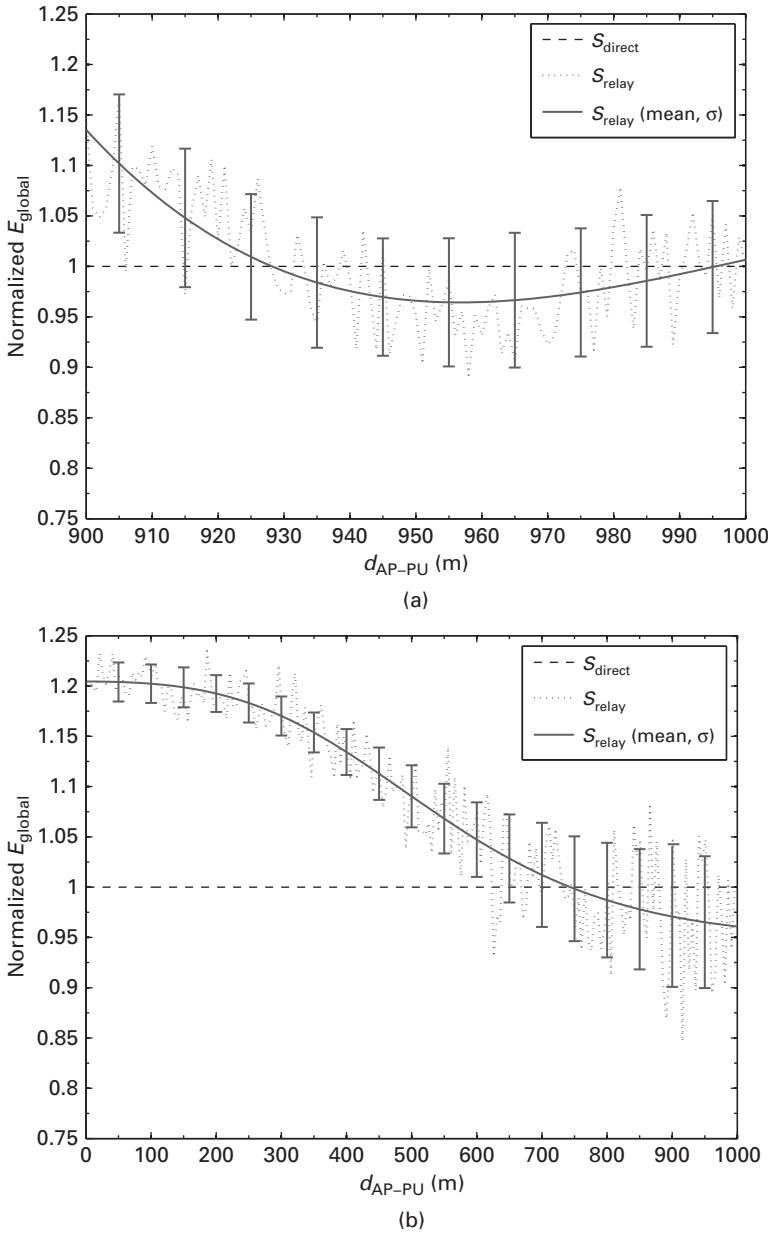


Figure 25.4 Comparison of energy per bit in various scenarios, normalized to scenario S_{direct} , in 802.11g-to-UMTS (one PU and one SU).

consumptions are evaluated for one bit, following (25.2), (25.3) and the *bitop* from Table 25.2. The 802.15.4 signal suffers a Friis path-loss with a path-loss exponent of 3.1 [12]. The 802.11g signal suffers an ITU-R path-loss considering an indoor office environment with three walls [16]. The receiver sensitivity is set to -92dBm in

Table 25.4 Number of operations per bit (*bitop*) for the selected modes.

	802.11g at 6 Mbps	802.15.4 at 20 kbps
Transmission	308	94
Reception	3570	2694

Table 25.5 Energy consumption E_{num} and E_{rf} (in nJ per bit) for the selected modes, at 50 m.

	802.11g at 6 Mbps		802.15.4 at 20 kbps	
	E_{num}	E_{rf}	E_{num}	E_{rf}
Transmission @ 50m	0.043	0.13	0.013	0.46
Reception	0.5	0.033	0.37	0.05

802.15.4 at 20 kbps, -87dBm in 802.11g at 6 Mbps. Moreover, the radio output power P_{out} ranges from -20 to 0dBm in 802.15.4, and from -20 to $+20\text{dBm}$ in 802.11g.

Table 25.5 presents the energy per bit for selected modes at a distance of 50 m.

Network simulations provide a way to evaluate the energy consumption with an MAC layer. In multimode simulations, it is crucial to store the original data size, so that the energy per bit can be recovered in a multimode relay. Here, an application runs on all terminals and sends 4000 data bits to AP. The 802.11g uses a CSMA/CA MAC layer, which relies on control packet to transmit data. The MAC frame size adds 84 bytes (672 bits) of cumulative header in 802.11g, and 40 bytes (200 bits) in 802.15.4. Moreover, the control packets in 802.11g have the following size: 288 bits for RTS, 224 bits for CTS, 208 bits for ACK.

Figure 25.5 presents the global energy consumption of an 802.15.4-to-802.11g relay, in uplink (from the SU to AP). These results have been obtained using the WSNNet simulator [29]. The PU moves from $d_{AP-PU} = 5\text{ m}$ to $d_{AP-SU} = 30\text{ m}$, and SUs are placed at regular intervals on a circle centered at 15 m, with a radius of 15 m.

For more details, Figure 25.5(a) presents the impact of different parameters on the global energy consumption:

- Passive overhearing: occurs when a terminal receives data not intended for itself.
- Control packet: the 802.11g MAC layer establishes a communication between the PU and AP by sending small control packets (called RTS, CTS, and ACK).

By default, the terminals listen on all modes. This behaviour leads to a “jump” at distance $d_{AP-PU} = 15\text{ m} = 1/2d_{AP-SU}$. Before this limit, the multimode relay brings energy gains of up to 7 percent compared to direct communications. After this limit, the distance between the PU and AP is longer than the distance between the PU and the SU: the SU receives all PU’s communications, even the ones sent to AP. This case of passive overhearing on the 802.11g interface leads to a higher global energy consumption. Due to the SU decoding all (unneeded) 802.11g data, the energy consumption multimode relay consumption is

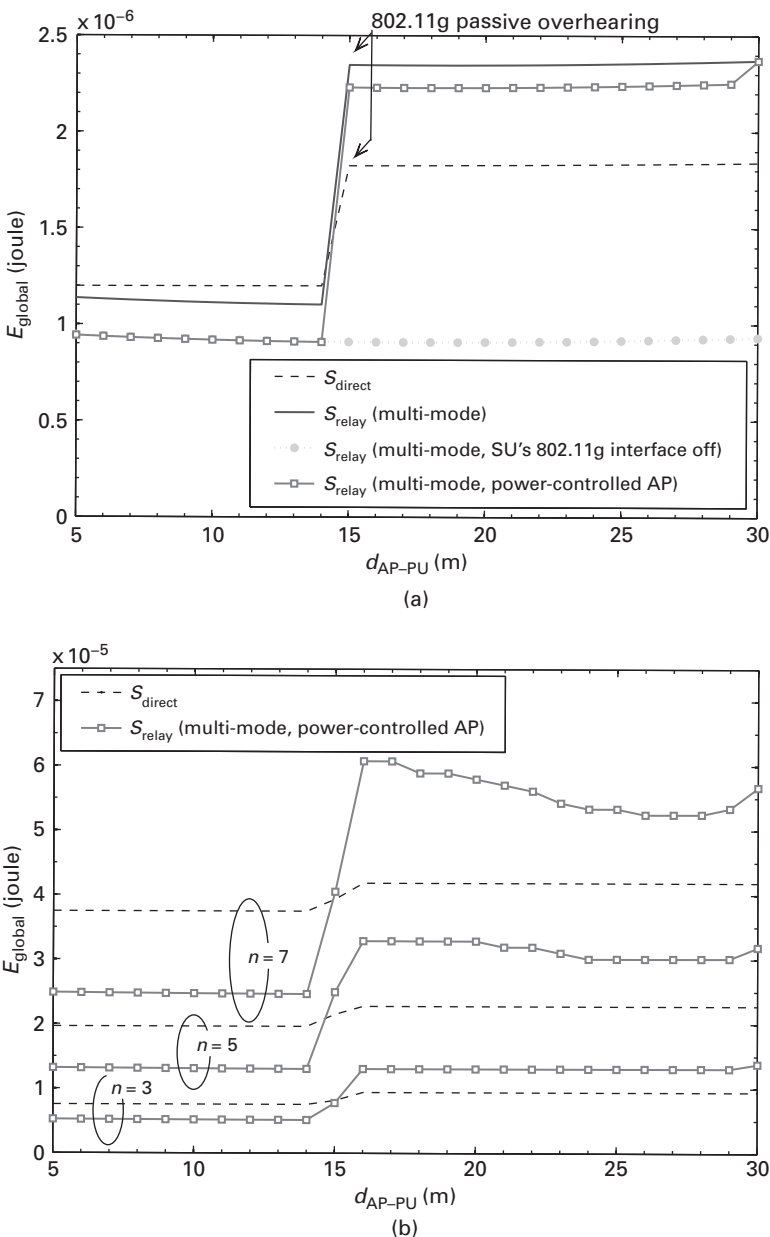


Figure 25.5 WSNet Simulation results: comparison of energy per bit for direct or 802.15.2-to-802.11g relayed communication for n SUs, in uplink showing the importance of MAC layer.

28 percent more than that of direct connections. If $d_{\text{AP-PU}} > 1/2 d_{\text{AP-SU}}$, direct connections should be prioritized.

The deactivation of SU's 802.11g interface significantly reduces the global energy consumption when the SU stops receiving the communications of the PU with AP.

However, this measure implies some constraints, since the SU loses its multimode property. One solution may be to rely on wake-up radios, where a specific signal received on the 802.15.4 interface activates the 802.11g interface. Another solution may be to perform periodical sensing on the inactive interface. These solutions allow a multimode relay to achieve an important reduction of the global energy consumption.

However, some cases require all interfaces to remain active. Figure 25.5(a) shows the impact of control packets. In this figure, the radio transmission power output is limited in such a way that control packets transmitted by the AP only reach the PU. In a large network, the AP has knowledge of all the PUs in the network. This solution shows an energy consumption 17 percent lower than a classical relay, below the half distance. This shows the importance of control packets in multimode relays, and the necessity of network simulations for a careful energy evaluation; the MAC layer has a high impact on the energy consumption.

Figure 25.5(b) shows that the energy consumption of control packets is not negligible when the number of SUs rises. With the PU close to the AP, the limitation of P_{out} for AP's control packets leads to energy gains around 30 percent for $n = 3$, and up to 35 percent for $n = 7$. Since the number of connections between the PU and the AP increases with n , the cost of passive overhearing increases because the SUs receive signals from all their neighbors and from the PU after the jump. By referring to the value of E_{bit} in reception for both modes, the energy consumption of 802.11g at 6 Mbps is higher. Hence, the energy consumed by passive overhearing (on any mode) is less important for the multimode relay; the gains increase with the number of SUs. As the number of users rises, a multimode relay becomes an interesting solution to reduce the global energy consumption. Beyond the half distance, relaying in multimode has no real interest for the same reasons as before.

25.2.3 Multimode energy consumption summary

In this section, we have shown that multimode relaying can reduce the global energy consumption under certain conditions. These conditions depend on several parameters: the terminal's position, channel conditions, and MAC control packets. The careful study of all these parameters shows that relaying at all times cannot be considered as a good solution to reduce the network consumption. Still, it cannot be ruled out, given some restrictions on the terminals, that a multimode relay brings interesting energy gains. For instance, the limitation of control packets' transmission power brings lower energy consumption than classical relay, while retaining the multimode ability of all terminals. However, this solution could lead to hidden terminal problems and a dedicated MAC layer adapted to relays might become mandatory.

The performance evaluation presented in this section has focused on the importance of PHY and MAC layers in the energy evaluation. But this study still relies on generic complexity for the digital part and supposes an unbounded digital resource. The next section thus offers a deeper insight into the problem of practical implementation of particular algorithms specially dedicated to cognitive radio. The chosen example features a performance evaluation of TV white space sensing at a lower level. It presents

algorithms and hardware implementations dedicated to this detection of the TV band's incumbents.

25.3 Implementation of the PHY sensing block: example of TV white space sensing

In this section, we assess the TV white space sensing case, which represents the very first large-scale potential deployment case of cognitive systems. Techniques to detect incumbents that operate in the UHF (Ultra High Frequency) TV band from 470 MHz to 860 MHz are thus taken as a relevant example of a process to implement a practical cognitive radio system. Incumbents can be rationalized into primary users, namely the TV transmitters and secondary licensed users, namely the PMSE (programme making and special events) equipment. In the following, the sensing algorithms for wireless microphones (PMSE) and Digital Video Broadcasting (DVB-T) signals [9] will be further detailed. A hardware implementation of the proposed DVB-T detector will also be presented.

25.3.1 Sensing algorithms for TV white space incumbents

As previously stated, a key feature for cognitive systems is the PHY sensing block which will allow the system to have a complete knowledge of the surrounding spectrum. Before detailing the question of implementing this block on a specific hardware, this section presents two possible algorithms dedicated to the detection of wireless microphone and DVB-T signals.

25.3.1.1 Wireless microphone sensing

Wireless microphones are low-power radio transmitting devices and their detection is mandatory according to the FCC rules on TV white space operation [11]. Let us first summarize the main characteristics of these systems.

They operate in the UHF band and most of them use an analog Frequency Modulation (FM). The signal has a spectral bandwidth B_s of 100 kHz, but practically most of the signal energy is concentrated in a bandwidth of 40 kHz. The transmit power is a few tens of mW. The coverage area is therefore relatively low, typically about 500 meters for the most powerful ones. Thus, their detection could be difficult to achieve. The signal from the microphone $s(t)$ can be modeled as follows:

$$s(t) = A \cos \left(2\pi f_0 t + \frac{f_d}{s_m} \int x(\tau) d\tau \right), \quad (25.8)$$

where f_0 is the carrier frequency, f_d the frequency deviation of the FM modulation, and $x(t)$ the modulating signal having an amplitude s_m .

The analyzed UHF TV channel B_c is wider (6 to 8 MHz depending on countries) compared to the frequency band occupied by the microphone signal (100 kHz). Besides, specific features of the FM signal are difficult to model (unlike the case of the OFDM

autocorrelation, see the next subsection). Therefore, the proposed algorithm consists of a frequency domain narrowband energy detector.

The received RF signal is first transposed down to baseband before processing the algorithm. Thus, the process begins with a signal sampling at B_c . The samples are stored in a buffer and a N_{FFT} -points FFT algorithm is processed over the buffer. The tone spacing is B_c/N_{FFT} , and has to be less than the signal bandwidth:

$$\frac{B_c}{N_{FFT}} < B_s \Leftrightarrow \frac{B_c}{B_s} < N_{FFT}. \quad (25.9)$$

Figure 25.6(a) shows an example of a narrowband signal ($B_s = 40$ kHz) in the considered UHF channel. In this example, $B_c = 8$ MHz (European value), $B_c/B_s = 200$ and $N_{FFT} = 16384$.

The average and maximum power are computed (P_{av} and P_{max} respectively). A sliding window, as large as the searched signal bandwidth, sweeps the spectrum, and the average power over this window is computed. The maximum value P_b is stored. If the signal is present in the channel, the window position which produces P_b has to contain also the maximum power. Thus a threshold Th is computed as follows:

$$Th = \frac{P_{max} + P_{av}}{2} \quad (25.10)$$

The maximum value P_b is compared to the threshold. If it exceeds this threshold, the signal is considered to be present. When there is only noise (Figure 25.6(b)), $P_{av} = P_b$ and thus the threshold is greater than the max because obviously $Th = \frac{P_{max} + P_b}{2} > P_b$ is always true.

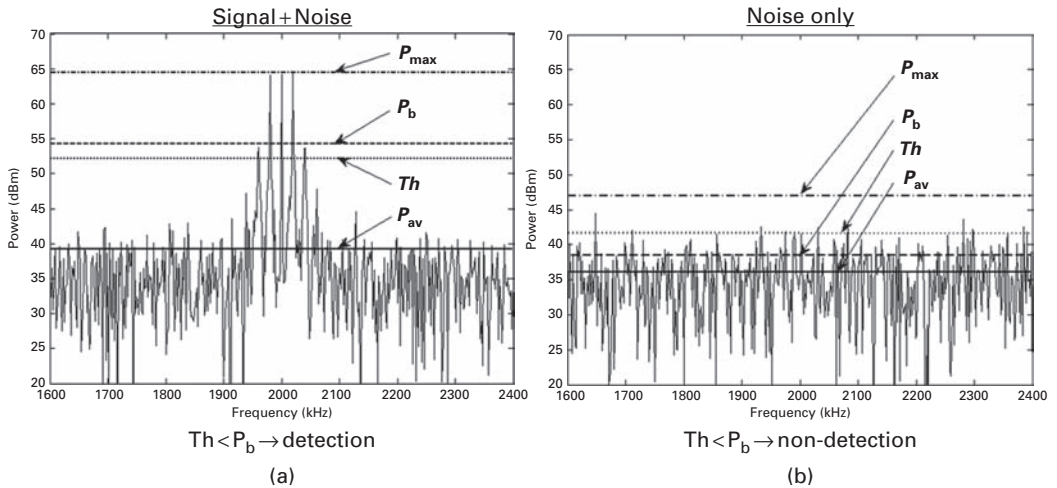


Figure 25.6 Values computed by the detection algorithm when the signal is present (a) and when there is only noise (b).

25.3.1.2 Cyclostationarity-based OFDM sensing

We now give more details on a second sensing algorithm dedicated to more evolved waveforms. This algorithm will also serve as the example in the next section on hardware implementation.

For the DVB-T signal sensing, the proposed technique is based on the a-priori knowledge that the DVB-T physical layer is defined by an OFDM modulation. The algorithm, described in [17], aims at detecting the cyclostationarity of the DVB-T signal through the analysis of the Fourier decomposition of its second-order momentum. It exploits the structure of the OFDM symbols which contain the same pattern at its beginning and end; the so-called cyclic prefix. By computing the autocorrelation of the incoming signal with a lag corresponding to the symbol duration, the cyclic prefix is emphasized while the rest of the correlation tends to zero. This is due to the fact that the data portion of the OFDM symbols is uncorrelated over consecutive symbols. Thus, the mathematical expectation of the correlation signal is time periodic, also referred to as the cyclostationary nature of the OFDM signal.

Let us now consider the autocorrelation of this signal:

$$R_s(u, m) = E\{s[u + m]s^*[u]\}. \quad (25.11)$$

Under the condition that all subcarriers are used, the autocorrelation of an OFDM signal is written as [17]:

$$R_s(u, m) = R_s(u, 0) + R_s(u, N)\delta(m - N) + R_s(u, -N)\delta(m + N), \quad (25.12)$$

with N being the number of subcarriers.

The first term of (25.12) is the power of the received signal. Energy detectors, which analyze this term only, provide poor performance at low SNR. To increase the performance of the detector at low SNR, we focus on the last two terms of (25.12) to build a cost function. The terms $R_s(u, N)$ and $R_s(u, -N)$ correspond to the correlation induced by the cycle prefix. It can be shown [15] that $R_s(u, N)$ is a periodic function of u which characterizes the signal s . $R_s(u, N)$ has a period $\alpha_0^{-1} = N + D$ with D being the length of the cyclic prefix.

As this function depends on u in a periodic way, the signal is not a stationary but a cyclostationary signal. Its autocorrelation function can be written as a Fourier series:

$$R_s(u, N) = R_s^0(N) + \sum_{k=-\frac{N+D}{2}, k \neq 0}^{\frac{N+D}{2}-1} R_s^{k\alpha_0}(N) e^{2i\pi k\alpha_0 u}. \quad (25.13)$$

In (25.13), $R_s^{k\alpha_0}(N)$ is the cycle correlation coefficient at cycle frequency $k\alpha_0$ and at time lag N . This term can be estimated as follows:

$$R_s^{k\alpha_0}(N) = \frac{1}{U} \sum_{u=0}^{U-1} s(u + N)s^*(u) e^{-2i\pi k\alpha_0 u}, \quad (25.14)$$

where U is the observation time.

The basic idea behind the cyclostationarity detector is to analyze this Fourier decomposition and assess the presence of the signal by setting a cost function related to one or more of these cyclic frequencies. This cost function is compared to some reference value. This technique was introduced in a more general context in the late 1980s by Gardner [14, 15]. Recent papers have applied this approach to the opportunistic radio context [13, 19, 17]. They mainly differ in the way the harmonics are considered. In our study, the proposed cost function exploits both the fundamental and several harmonics as expressed in (25.15):

$$J_s(K_s) = \frac{1}{2K_s + 1} \sum_{k=-K_s}^{K_s} |R_s^{ka_0}(N)|^2, \quad (25.15)$$

where K_s is the number of harmonics that are considered.

It can be observed that the cost function is only built upon $R_s(u, N)$, while $R_s(u, -N)$ is omitted. Indeed, it is quite easy to prove that $|R_s^{ka_0}(N)|^2 = |R_s^{ka_0}(-N)|^2$.

A hardware integration of this algorithm is presented in the following section in the particular case of DVB-T signals.

25.3.2 Hardware implementation of DVB-T detectors

In this section, the problem of hardware implementation of the cyclostationarity detector is described for the particular case of detecting DVB-T signals.

25.3.2.1 Adaptation of the cyclostationarity detector

The DVB-T standard defines four FFT sizes: $N=1024$, 2048, 4096, or 8192 and $B_s=8$ MHz. The cyclic prefix over FFT size ratio D/N can also vary: 1/32, 1/16, 1/8, 1/4. Considering all the configurations leads to very highly complex hardware architecture. However, in practice, deployment considers a smaller set of parameters depending on the country. For instance, in France, the set of parameters used is $N=8192$ and $D/N=1/32$.

A key characteristic which will be exploited in the architecture design stems from the broadcast nature of the DVB-T signal. This means that detector sensitivity can be increased significantly by a very long integration time. This is a relevant feature since sensitivity requirements for primary user detection are very challenging (typically $\text{SNR}=-10\text{dB}$ [24]). It also changes the way that the reference signal is used to define the decision threshold. When undertaking this calibration phase, the secondary system needs to consider a reference noise value which is independent of the signal presence.

When considering long (ideally infinite) integration time, the autocorrelation function $R_s(u, N)$ defined in Subsection 25.3.1.2 tends to a rectangular signal as depicted in Figure 25.7, the cyclic ratio being $D/(N + D)$. In this case, the Fourier coefficient is written as:

$$R_s^{ka_0}(N) = \frac{A}{2\pi \cdot k} \left[\sin \frac{2\pi \cdot k \cdot D}{N + D} + j \left(1 - \cos \frac{2\pi \cdot k \cdot D}{N + D} \right) \right]. \quad (25.16)$$

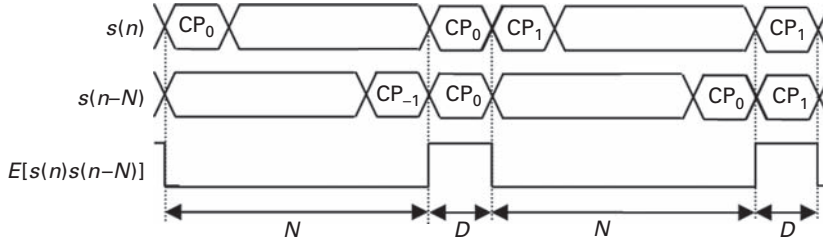


Figure 25.7 Ideal autocorrelation signal of an OFDM symbol burst.

Each coefficient power is given by:

$$\begin{cases} |R_s^0(N)|^2 = \left(\frac{A \cdot D}{N + D}\right)^2 & k = 0 \\ |R_s^{ka_0}(N)|^2 = 2 \left(\frac{A}{2\pi k}\right)^2 \left(1 - \cos \frac{2\pi \cdot k \cdot D}{N + D}\right) & k \neq 0 \end{cases} \quad (25.17)$$

Decision threshold computation

First, the reference noise level has to be computed from the observation in order to be compared with the signal cost function $J_s(K_s)$. It is obvious from Equation (25.17) that $R_s^{ka_0}(N) = 0$ when $k = l(N/D + 1)$, $l = \{1, 2, \dots, +\infty\}$. Figure 25.8 plots the Fourier coefficients of a rectangular signal when $N/D = 32, 16, 8$, and 4 , respectively. It can therefore be concluded that Fourier harmonics $l(N/D + 1)$ are not impacted by the presence of the signal and can thus be used for calibration purposes to define the reference noise level. Similarly to (25.16), a cost function J_n could be defined in order to compute the noise level:

$$J_n(K_n) = \frac{1}{2K_n} \sum_{\substack{l = -K_n \\ l \neq 0}}^{K_n} \left| R_s^{l(\frac{N}{D}+1)a_0}(N) \right|^2, \quad (25.18)$$

with K_n being the number of harmonics that are considered.

As a comparison, calibration based on input power computation is not relevant as this estimator is strongly impacted by the presence of the signal.

For example, considering the French set of parameters ($D/N = 1/32$) and considering the first four signal harmonics $\{-3; +3\}$ and one noise harmonic (i.e. $K_s = 3$ and $K_n = 1$), the decision threshold V can be expressed as follows:

$$V = \frac{J_s(K_s)}{J_n(K_n)} = \frac{2}{7} \frac{\sum_{k=-3}^3 |R_s^{ka_0}|^2}{|R_s^{-33a_0}|^2 + |R_s^{33a_0}|^2}. \quad (25.19)$$

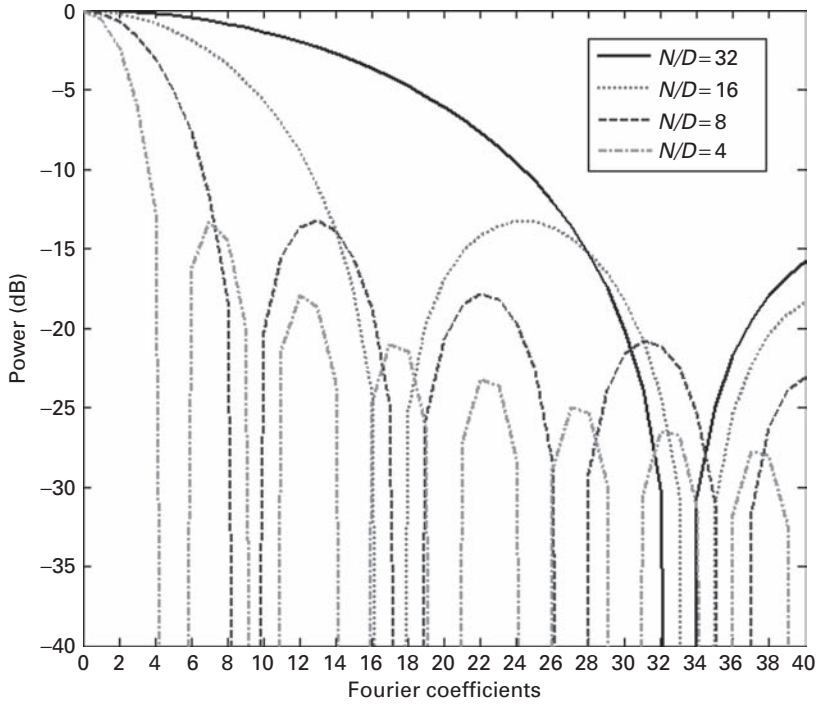


Figure 25.8 Fourier coefficient values for $N/D = 32, 16, 8$, and 4 .

Filter's characteristics

This technique holds only for infinite integration time to guarantee the rectangular shape of the autocorrelation estimator (Figure 25.7). Whenever a finite integration is performed, the convergence of the integrator needs to be considered. The integrator is a first-order IIR (Infinite Impulse Response) filter, and its indicial response is given by:

$$s_{ind}(k) = 1 - \left(\frac{n-1}{n} \right)^k \quad (k \geq 0), \quad (25.20)$$

where n can be tuned to adjust the raising time of the filter.

The raising time k_r (in number of symbols) to reach 90 percent is then given by:

$$k_r = \frac{\ln(10\%)}{\ln(n/n-1)}. \quad (25.21)$$

For large n values, the expression in Equation (25.21) tends to $2.3n$. Estimator performance is improved by increasing the integration ability of the filter. However, this is at the cost of long integration time. Thus, this approach is to be considered for “always on” kind of systems, such as DVB-T broadcast signals to guarantee reliable detection under low SNR conditions.

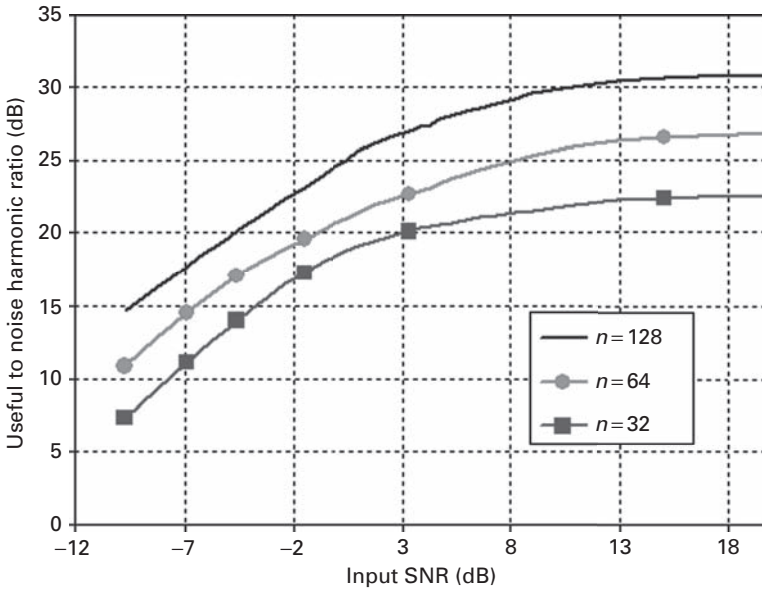


Figure 25.9 Detection threshold according to the input SNR.

Figure 25.9 shows the decision threshold V as a function of the input SNR (under additive white Gaussian noise conditions) for several values of n . The aim of the curve is to show how an increase in the integration time impacts the performance of the detector for a given threshold value. For instance, when an SNR of -7 dB is targeted and for a threshold set to 15, no detection is possible when considering $n = 32$. However, when n is set to 128, a reliable behavior is achieved. Setting n to 64 results in non-reliable decisions. From this graph, a trade-off between SNR detection condition and integration time can be made.

25.3.2.2 Hardware architecture

In an operational system, the PHY sensing block has to be implemented on a real hardware platform. Then, all theoretical functions have to be expressed as logical blocks with inputs and outputs, and an explicit dimensioning of all links between blocks and data format must be carefully performed. As an illustration the cyclostationarity detector hardware architecture is shown in Figure 25.10 for the parameters $K_s = 3$ and $K_n = 1$.

First, the autocorrelation is computed on the I/Q complex samples. The IIR integrator then averages over a number of symbols tuned by setting the integration time parameter to achieve the required sensitivity. The supervisor, a Finite State Machine (FSM), then triggers the writing into a buffer that stores 8192 filter output samples (equivalent to the length of an OFDM symbol). Then, using a faster clock, the Fourier harmonics are computed sequentially. The sine generator computes sequentially the required sine function of the Fourier taps of interest. The Multiply Accumulate (MAC) function enables the Fourier coefficient to be obtained for these taps. The sequence is as follows.

Table 25.6 Complexity evaluation of the DVB-T detector.

Complexity			
Slices	RAM blocks of 18kbits	Mult	Latency
1600	122	23	Depends on n

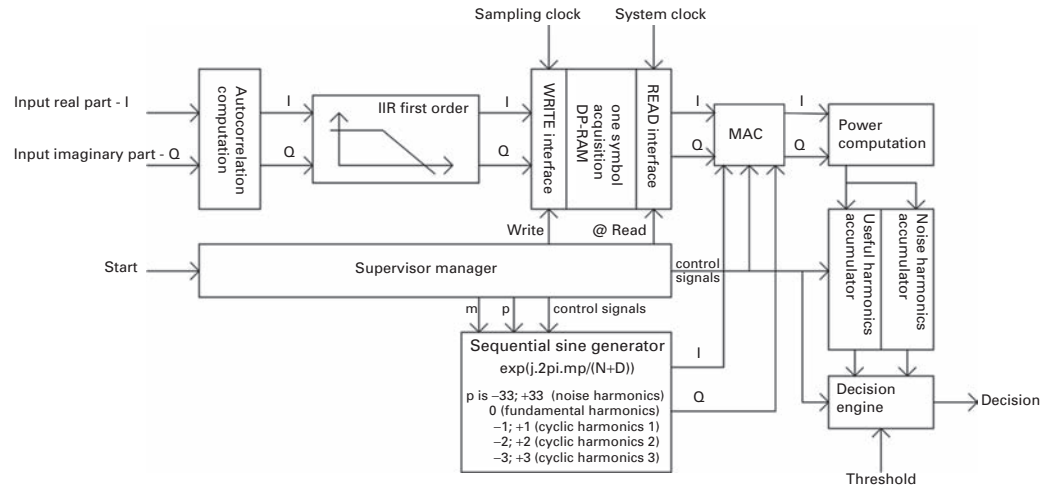


Figure 25.10 Cyclostationarity detector for DVB-T signals.

First the reference harmonics $\{-33; +33\}$ are generated to compute the noise reference power. Then the harmonics of interest for the DVB-T signal $\{0; -1; +1; -2; +2; -3; +3\}$ are calculated. The power of each harmonic is summed up to obtain the cyclostationarity estimator value. Finally the decision engine gives the final result by comparing the estimated value to the threshold value according to Equation (25.23), which provides a hard decision output of the detector.

Finally, the complexity of such a detector hardware implementation is determined on a Xilinx Virtex 5 target technology using the ISE XST synthesis tool. Results are provided in Table 25.6.

Essentially, to summarize this study, here some solutions have been proposed in order to detect the primary users of the TV band. For the wireless microphones, an energy detector in the frequency domain has been introduced in order to take into account the narrowband characteristics of the microphone signal. The cyclostationarity detectors allow the sensing of the DVB-T signals by using the cyclic prefix of the OFDM modulation in order to achieve a low SNR detection. The performance of this method is given in [17]. Finally, a hardware implementation of the cyclostationarity detector has been presented with the proposed architecture being adapted to the broadcast nature of the DVB-T signal.

Therefore, based on this operational implementation, a deeper view on the expected power consumption of this digital PHY sensing block could be offered and then could be integrated in a higher-level simulation tool. But obviously it comes that for a precise energy evaluation of a complete SDR system a global translation of all digital blocks regarding the targeted hardware is required, leading to a very complex optimization problem. Also, this implies that any decision functions for a multimode SDR system has to rely on a specific knowledge of underlying hardware.

25.4 Conclusion

Increasing the digital part of flexible terminals is an attractive way of dealing with the problems of tunability and reconfiguration. But the digital implementation of cognitive or multimode terminals comes with a lot of new problems.

Of course, if considered as a way of reusing white spaces in the spectrum, cognitive radio systems only have to cope with sensing problems in addition to a classical radio interface. But if we go further to consider a cognitive process to decide on relaying in a network in different available bands or standards, it appears that reconfiguration metrics taking into account all parts of the global consumption is still an open problem.

The wireless flexible link introduced in this chapter copes with this requirement and presents the novelty of performing both sensing and multimode capabilities. In order to go further into the definition and the implementation of the different modes, two distinguished studies are presented.

First, a generalist view of evaluating the use of the flexibility of SDR systems for multimode relaying is detailed, describing a way of considering a global power consumption in SDR terminals, with RF and digital parts, but also the importance of network level simulations. The power consumption evaluation aims to give decision metrics for switching between different modes of the systems.

The second exposes the operation of the sensing function in a cognitive system, illustrated in the TV White space. Two techniques are proposed to detect incumbents: a frequency domain energy detector improves the detection of the wireless microphones and a cyclostationarity detector is introduced for DVB-T signals. For the cyclostationarity detector, a possible implementation on FPGA platform is described and leads to a low complexity algorithm.

To go further in the definition and the implementation of the cognitive radio link introduced in the [first section](#), the power consumption model of the PHY Sensing block has to be included in the evaluation proposed in the [second section](#).

Finally, this study contains sufficient information to clearly emphasize that for any new system with a large number of degrees of freedom a specific and in-depth study of the all parts of the system is necessary. Consequently, the global optimization of such multimode and multi-link systems is still an open problem.

References

- [1] IEEE Computer Society, “802.11g part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications,” <http://standards.ieee.org/getieee802/download/802.11g-2003.pdf>, June 2003.
- [2] IEEE Computer Society, “IEEE Std 802.15.4–2006 Part 15.4: Wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (WPANs),” <http://standards.ieee.org/getieee802/download/802.15.4-2006.pdf>, September 2006.
- [3] ARM Processor, “ARM 968 E-S processor,” www.arm.com/products/CPU/ARM968E-S.html, consulted on July 2010.
- [4] L. Berlemann, R. Pabst, and B. Walke, “Multimode communication protocols enabling reconfigurable radios,” *EURASIP Journal on Wireless Communications and Networks*, vol. **2005**, no. 3, pp. 390–400, 2005.
- [5] I. Burciu, G. Villemaud, J. Verdier, and M. Gautier, “A 802.11g and UMTS simultaneous reception front-end architecture using a double IQ structure,” in Proceedings of the IEEE 69th Vehicular Technology Conference (VTC '09), Barcelona, Spain, April 2009.
- [6] D. Charlet, V. Issarny, and R. Chibout, “Energy-efficient middleware-layer multi-radio networking: An assessment in the area of service discovery,” *Computer Networks*, vol. **52**, no. 1, pp. 4–24, 2008.
- [7] C. Cordeiro, M. Ghosh, D. Cavalcanti, and K. Chappali, “Spectrum sensing for dynamic spectrum access of TV bands,” in Proceedings of the 2nd International Conference on Cognitive Radio Oriented Wireless Networks and Communications, Orlando, Florida, USA, Aug. 2007, pp. 225–233.
- [8] D. J. Cichon and T. Kürner, “Cost 231 Final Report – Chapter 4: propagation prediction models,” www.lx.it.pt/cost231/final_report.htm, 1998.
- [9] ETSI, “Digital video broadcasting (DVB) framing structure, channel coding and modulation for digital terrestrial television,” in ETSI Report, November 2004.
- [10] “Final Draft Standard ECMA-392 - MAC and PHY for operation in TV white space”, ECMA, 2009. www.ecma-international.org/publications/files/drafts/tc48-2009-061.doc
- [11] Federal Communications Commission, “FCC 04–113 – unlicensed operation in the TV broadcast bands,” in ET Docket No. 04–186, May 2004.
- [12] K. Laasonen, “Radio propagation modeling,” University of Helsinki, Technical Report, 2003.
- [13] M. Ghozzi, M. Dohler, F. Marx, and J. Palicot, “Cognitive radio: methods for detection of free Bands,” *Elsevier Science Journal, Special Issue on Cognitive Radio*, vol. **7**, Sept. 2006.
- [14] W. A. Gardner and G. Zivanovic, “Degrees of cyclostationary and their application to signal detection and estimation,” *Signal Processing*, vol. **22**, No. 3, March 1991.
- [15] W. Gardner and M. Spooner, “Signal interception: performance advantages of cyclic-feature detectors”, *IEEE Transactions on Communications*, vol. **40**, no. 1, pp. 149–159, Jan. 1992.
- [16] ITU Radiocommunication Assembly, “Recommendation ITU-R P.1238–1 : Propagation data and prediction models for the planning of indoor radiocommunication systems and radio local area networks in the frequency range 900 MHz to 100 GHz,” October 1999.
- [17] P. Jallon, “An algorithm for detection of DVB-T signals based on their second order statistics,” *EURASIP Journal on Wireless Communications and Networks*, 2008.
- [18] C. Lévy-Bencheton, D. Ammar, G. Villemaud, T. Risset, and C. Reboul, “Multi-mode relaying for energy consumption reduction,” in *Annals of Telecommunications*, in Revision.

- [19] J. Lunden, V. Koivunen, A. Huttunen, and H. Vincent Poor, "Spectrum sensing in cognitive radios based on multiple cyclic frequencies", <http://arxiv.org/abs/0707.0909>, July 2007
- [20] J. Mitola, "Cognitive radio: an integrated agent architecture for software defined radio", Ph.D. dissertation, Royal Institute of Technology, Sweden, May 2000.
- [21] D. Nogu  t *et al.* "Sensing techniques for Cognitive Radio – state of the art and trends", available at http://grouper.ieee.org/groups/scc41/6/documents/white_papers/P1900.6_WhitePaper_Sensing_final.pdf, April 2009
- [22] "OFCOM Digital Dividend consultation", available at <http://stakeholders.ofcom.org.uk/spectrum/project-pages/ddr/>, 2009.
- [23] G. K. Rauwerda, P. M. Heysters, and G. J. M. Smit, "Mapping wireless communication algorithms onto a reconfigurable architecture," *Journal of Supercomputing*, vol. **30**, no. 3, pp. 263–282, 2004.
- [24] S. Shellhammer, "Spectrum sensing in IEEE802.22," First Workshop on Cognitive Information Processing (CIP 2008).
- [25] E. Shih, S.-H. Cho, N. Ickes, *et al.*, "Physical layer driven protocol and algorithm design for energy-efficient wireless sensor networks," in Proceedings of the 7th Annual International Conference on Mobile Computing and Networking (MobiCom '01), Rome, Italy, July 2001, pp. 272–287.
- [26] 3rd Generation Partnership Project, "UMTS Physical layer procedures (TDD) (Release 8)," www.3gpp.org/ftp/Specs/html-info/25224.htm, March 2008.
- [27] A. Wang and A. Chandrakasan, "Energy-efficient DSPs for wireless sensor networks," *IEEE Signal Processing Magazine*, vol. **19**, no. 4, pp. 68–78, July 2002.
- [28] C.-C. Wang, J.-M. Huang, C.-Y. Chang, K.-T. Cheng, and C.-P. Li, "A 6.57 mw zigbee transceiver for 868/915 mhz band," in Proceedings of IEEE International Symposium on Circuits and Systems, 2006. ISCAS 2006. pp. 5195–5198.
- [29] G. Chelius, A. Fraboulet, and E. Ben Hamida, "WSNet – An event-driven simulator for large scale wireless sensor networks," <http://wsnet.gforge.inria.fr/>, 2008.

Index

- 1-bit DAC: 413
- 1-bit sigma-delta ADC: 631
- 2D LUT: 167
- 3GPP: 40
- 8VSP: 10, 98

- access point: 739
- active constellation extension (ACE): 314
- active resistor-capacitor: 672
- adaptive array antenna: 549
- adaptive DPD: 154, 298
- adaptive interference cancellation (AIC): 460
- additive white Gaussian noise (AWGN): 415, 495, 597
- adjacent channel: 641
- adjacent channel leakage ratio (ACLR): 61, 123, 144
- adjacent channel power ratio (ACPR): 144
- aliasing: 676
- amplitude and power modulation (AM and PM): 146
- amplitude clipping: 316
- amplitude modulation (AM): 146
- amplitude response: 637
- amplitude-to-amplitude (AM/AM): 281
- amplitude-to-phase (AM/PM): 281
- analog-to-digital converter (ADC): 51, 111, 450
- angles of departure (AoD): 40
- antenna discrimination: 121
- anti-aliasing filter: 338, 682
- anti-causal: 519
- assembly: 655
- ATSC: 10
- auto-correlation function: 489
- autocorrelation windowing: 517
- autocorrelation matrix: 296
- automatic gain control (AGC): 161

- back-off: 244
- bandpass sigma-delta modulator: 417
- bandpass signal: 197
- baseband (BB): 476
- base transceiver station (BTS): 120

- basic predistortion cell (BPC): 223
- basis waveforms: 168
- beamforming: 16
- bin variance: 163
- biquadratic filter: 671
- bit error rate (BER): 15, 32, 477
- bitop: 738
- blind imbalance estimation: 490
- blind signal separation (BSS): 490
- BPSK modulation : 30
- broadcast control channel (BCCH): 78
- broadcasting: 4
- bytecode: 658

- canonic signed digit: 632
- carrier diversity: 564
- carrier frequency offset (CFO): 487, 666
- carrier leakage: 161
- carrier recovery circuit: 724
- cascaded DPD: 175
- causality: 519
- CDMA: 6
- cellular network: 4, 14
- channel estimation: 541
- channel fading: 37
- channel selection: 621, 700
- channel selection filter: 644
- channel state information (CSI): 45
- CIC: 62, 352, 606, 630
- circulator: 83
- clipping: 455
- clipping-and-filtering: 321
- clipping function: 157
- clock jitter: 408
- coarse spectrum sensing: 462
- co-channel interference: 593
- code multiplexing: 578, 589
- coded orthogonal frequency division multiplex (COFDM): 12
- coding sequences: 34

- cognitive manager: 735
- cognitive radio (CR): 22, 77
- companding technique: 316
- complementary auto-correlation: 489
- complementary cumulative density function (CCDF): 152
- continuous-time: 676
- continuous wave (CW): 624
- control packet: 744
- CORDIC: 229, 390, 642
- correlation matrix: 179
- crest factor: 84, 270, 309, 310
- crest fact reduction (CFR): 135, 271
- cross-correlation: 34
- cubic nonlinearity: 194
- cycle correlation: 749
- cyclic prefix (CP): 65, 537

- DAC dynamic range: 617
- DC calibration: 605
- DC offset: 258, 356, 684
- DC offset cancellation: 356
- decimation: 365, 680
- dedicated physical channels: (DPCH): 60
- delayed-locked loop (DLL): 60
- delta-sigma ADC: 395
- delta-sigma converter: 127
- delta-sigma modulation: 395
- device under test (DUT): 298
- differential nonlinearity (DNL): 386, 453
- differential pulse-code modulation: 413
- digital automatic gain correction (DAGC): 645
- digital down-converter (DDC): 333
- digital front-end (DFE): 52
- digital IF architecture: 610
- digital image: 401
- digital interpolation filters: 55
- digital mixer: 339, 606
- digital-to-analog converter (DAC): 51, 111
- digital predistorter (DPD): 51, 143
- digital up-converter (DUC): 72, 333, 370, 402
- direct conversion: 113
- direct current (DC): 282
- direct digital synthesizer (DDS): 338, 360
- direct down-conversion: 605
- direct memory access: 661
- direct sequence: 725
- direct sequences spread spectrum (DS-SS): 59
- direct up-conversion transmitter: 54
- dirty RF: 475, 585
- discrete Fourier transform (DFT): 20, 531
- discrete-time: 676
- discrete-time filters: 671
- dithering: 458
- dominant eigenvector: 185
- Doppler frequency: 547
- Doppler shift: 98, 547
- double IQ structure: 595

- down-conversion: 333
- down-link: 543
- down sampling: 268, 680
- duplexers: 128
- DVB-H: 106
- DVB-return channel terrestrial (DVB-RCT): 105
- DVB-SH: 107
- DVB-T: 11, 106
- dynamic biasing: 281
- dynamic deviation reduction (DDR): 251
- dynamic range: 382

- effective number of bits (ENOB): 124
- eigenvalue: 178
- eigenvector: 178
- end to end reconfigurability: 655
- enhanced data rates for GSM evolution (EDGE): 58
- envelope elimination and restoration (EE&R): 86, 281
- envelope tracking: 192
- equivalent isotropic radiated power (EIRP): 100
- equivalent number of bits (ENOB): 383, 397
- error detection: 541
- error vector magnitude (EVM): 131, 144, 280, 397, 607
- ETSI: 57

- fast Fourier transform (FFT) : 38
- feedback loop (FBL): 508
- finite impulse response (FIR): 170, 677
- finite state machine: 753
- first-order quadrature modulator: 421
- first-order sigma-delta modulator: 421
- forward error correction (FEC): 99
- fourth generation (4G): 28
- frequency correction channel (FCCH): 78
- frequency division duplex (FDD): 58
- frequency modulation: 721
- frequency multiplexing: 578
- frequency offset: 585, 587
- full-duplex (FDD): 381
- functional description language: 656
- Gaussian process: 41
- general packet radio service (GPRS): 57
- GGE: 620
- giga-operations per second (GOPS): 651
- global energy consumption: 739
- group delay: 637
- GSM: 6
- guard interval (GI): 537

- halfband: (HB): 72
- half duplex frequency division duplex (H-FDD): 65
- Hammerstein model: 205, 288
- hard clipping: 455
- hardware abstraction layer: 653
- harmonic distortion: 402
- Hermitian: 222
- Hermitian transpose: 263

-
- heterodyne architecture: 575
 - heterodyne receiver: 386
 - heterogeneous platforms: 651
 - hierarchical transmission: 100
 - high definition multimedia interface: 707
 - high definition television (HDTV): 98
 - higher-order statistics: 490
 - homodyne: 122
 - homodyne architecture: 575
 - HSDPA: 7
 - identity matrix: 41
 - image frequency: 111
 - image interference: 507
 - image rejection: 111, 606
 - image rejection ratio (IRR): 436, 480, 642
 - impulse radio: 722
 - in-band blocking: 622
 - indirect learning architecture (ILA): 514
 - infinite impulse response (IIR): 173, 677
 - inner product: 37
 - in-phase: 215
 - integral nonlinearity (INL): 385, 453
 - inter channel interference (ICI): 21
 - inter-modulation distortion (IMD): 123, 143, 282
 - inter-symbol interference (ISI): 38, 53
 - interference: 3
 - interference temperature: 78
 - interleaved spectrum sensing: 734
 - interleaving: 318
 - intermediate frequency (IF): 194, 258, 334, 476
 - International Mobile Telecommunications-2000 (IMT-2000): 57
 - interpolation: 365, 465
 - interpreter: 654
 - inverse discrete Fourier transform (IDFT): 65, 534
 - I&Q: 246, 333
 - I/Q demodulation: 340
 - I/Q demodulator: 718
 - I/Q demodulator mismatch: 508
 - I/Q imbalance: 435, 479, 502
 - IQ interface: 603
 - I/Q interleaving: 631
 - I/Q mismatch: 389, 436, 506, 585, 588
 - I/Q mismatch compensation (IQMC): 642
 - I/Q mixing: 476
 - I/Q modulation: 341
 - I&Q modulator: 54, 502
 - IQM impairment: 508
 - I/Q predistorter: 510
 - IQ receiver: 581
 - ISDB-T: 98
 - Java: 653
 - jitter: 434
 - joint predistorters: 503
 - joint tactical radio systems: 657
 - just-in-time complier: 658
 - kernel: 656
 - kernel programs: 656
 - Khan process network: 659
 - knowledge-based DPD: 159
 - Lagrange polynomial: 627
 - latency: 64
 - learning module: 159
 - least mean square (LMS): 149, 236, 263, 295, 461
 - least significant bit (LSB): 386, 453
 - least square (LS): 250, 263
 - Leslie--Singh modulator: 429
 - limiter: 455
 - linear response: 245
 - linear time-invariant (LTI): 205
 - LO jitter: 624
 - LO leakage: 397, 502, 506, 684
 - LO phase noise: 397
 - local oscillator (LO): 54, 604, 620
 - low density parity check code (LDPC): 101
 - low-IF: 123, 388, 494, 688
 - low-noise amplifier (LNA): 51, 576, 721
 - low-pass filter: 674
 - low-pass sigma-delta modulator: 415
 - LTE: 8, 28, 58
 - LUT: 214, 456
 - LUT-based DPD: 161, 214, 247
 - MAC layer: 733
 - mapping predistortion: 215
 - matrix inversion: 569
 - matrix product: 185
 - maximum ratio combining (MRC): 18, 549
 - maximum selection interpolation: 465
 - media forward link only (MediaFLO): 108
 - memory polynomial: 172, 250
 - memory polynomial model: 290
 - memoryless polynomial: 172
 - memoryless polynomial model: 199
 - memoryless nonlinear model: 287
 - memoryless nonlinearity (MLN): 175
 - microwave integrated circuit solutions (MMIC): 89
 - minimum mean square error (MMSE): 551, 584
 - mirror frequency: 442
 - mirror frequency interference: 479, 502
 - mirror transfer function: 440
 - mismatch filters: 479
 - mixing and combining: 348
 - mobile switching center (MSC): 4
 - mobile WiMAX: 56
 - model inverse structure: 255
 - model inversion: 459
 - model reference adaptive system (MRAS): 155, 256
 - modem: 603

- mode reference controller: 256
- modulator imbalance: 161
- modules: 655
- mono-antenna receiver: 576
- most significant bits (MSB): 403
- multi-antenna: 576, 585
- multiband architecture: 585
- multiband sigma-delta modulator: 433
- multi-bit quantizer: 427
- multi-carrier modulation: 532
- multicast: 102
- multi-dimensional LUT: 166
- multi-dimensional Fourier transform: 285
- multi-gigabits: 707
- multi-layer perceptrons (MLP): 200
- multimedia broadcast and multicast services (MBMS): 14, 104, 109
- multipath: 537
- multiple access interference (MAI): 35
- multiple-input multiple-output (MIMO): 4, 15, 29, 39, 58, 593
- multi-port: 707
- multi-port heterodyne receiver: 722
- multi-port interferometer: 713
- multi-port phase discriminator: 709
- multipliers: 264
- multiplier-accumulator (MAC): 372
- multiplexing: 268, 577
- multi-stage filtering: 342, 343
- multi-stage noise shaping (MASH): 429

- narrowband blocking: 613
- neural networks: 200
- near zero intermediate frequency (NZIF): 605
- noise figure: 613
- noise shaping: 417
- noise transfer function (NTF): 414
- non-line-of-sight: 593
- non-simultaneous receivers: 573
- normalized LMS: 464
- normalized mean square error (NMSE): 303
- normalized square error (NSE): 204
- nonlinear auto-regressive moving average (NARMA): 209
- nonlinear moving average (NMA): 203
- numerical control oscillator (NCO): 333, 338, 547
- Nyquist theorem: 29
- Nyquist theory: 675

- offset error: 452
- omnidirectional antenna pattern: 718
- one segment: 100
- orthogonal frequency division multiplex (OFDM): 4, 19, 37, 64, 531
- orthogonal frequency division multiple access (OFDMA): 28, 58, 540
- orthogonal polynomial: 301, 513
- orthogonal spreading codes: 591
- orthogonal variable spreading factor (OVSF): 35, 45, 59
- over-sampling: 337, 389
- over-sampling rate (OSR): 416

- PAR meter: 309
- parallel Hammerstein: 512
- parallel-Wiener model: 208
- partial transmit sequencing (PTS) : 273, 317
- passband bandwidth: 611
- passive overhearing: 744
- peak envelope power (PEP): 319
- peak-to-average ratio (PAR): 309
- peak-to-average power ratio (PAPR): 21, 84, 244
- peak windowing: 272
- per antenna rate control (PARC): 44
- phase-locked loop (PLL): 724
- phase noise: 585, 586
- phase-plane indexing: 457
- phase-quadrature: 215
- PHY layer: 733
- PHY sensing: 735
- physical layer: 32
- physical layer description language: 661
- pipeline ADC: 396
- planar antenna array: 715
- polar loop architecture: 608
- polar predistortion: 215
- polynomial filters: 516
- polynomial resampler: 627
- polyphase filter: 344
- post-processing: 483
- power added efficiency (PAE): 616
- power amplifier (PA): 51, 143, 281
- power class: 61
- power spectral density: 37, 297
- power splitters: 114
- predictive predistortion method: 220, 224
- predistortion function : 153
- primary users (PU): 77, 740
- processing component: 651
- processing gain: 127
- programmable digital radio: 653
- pseudo-random (PR): 60
- PSK modulation : 31
- pulse position modulation: 726

- QAM: 31
- QPSK modulation: 30
- QR decomposition: 517
- quadratic programming: 293
- quadrature down-conversion: 611, 708
- quadrature image rejection: 612
- quadrature modulator: 419
- quadrature sigma-delta conversion: 419
- quality of services (QoS): 56
- quantization noise: 309

- radio frequency (RF): 3, 143, 334, 475
- radio network controller (RNC): 61
- radio network system (RNS): 61
- radio virtual machine: 650
- RAKE: 60
- rate change filter: 631, 641
- rate control multi-paths (RC-MPD): 45
- real-time operating system (RTOS): 90
- receiver selectivity: 121
- receiver sensitivity: 740
- recursive least square (RLS): 186, 263, 517
- reconfigurable Khan process network: 659
- resistor: 673
- resource block (RB): 67, 121
- resource elements (RE): 67
- RF error: 531, 556
- RF predistortion: 194
- RF switch: 577
- root mean square (RMS): 144, 270
- RRC: 62

- Saleh model: 198
- sampling frequency: 55
- sample matrix inversion (SMI): 551
- scattering matrix: 709
- Schroeder algorithm: 310
- scrambler: 653
- second users (SU): 77, 740
- selecting mapping (SLM): 273, 317
- self-tuning regular (STR): 155
- serial port interface (SPI): 617
- serial to parallel converter: 725
- sigma-delta modulation: 413
- signal-to-interference-plus-noise ratio (SINR): 45
- signal-to-interference ratio (SIR): 497
- signal-to-noise ratio (SNR): 32, 309, 380, 381
- signal-to-noise-and-distortion ratio (SNDR): 383, 468
- signal transfer function (STF): 414
- simultaneous receivers: 573
- single carrier frequency division multiple access (SC-FDMA): 64, 545
- single carrier modulation: 532
- single-frequency network (SFN): 12
- single-input multiple-output (SIMO): 585, 593
- single-input single-output (SISO): 197
- singular value decomposition: 41, 206, 517
- smart memory engine: 665
- soft clipping: 455
- software defined radio (SDR): 50, 573, 650
- software radio: 51
- software radio code: 650
- solid-state PA (SSPA): 288
- source code: 658
- space time block codes (STBC): 42
- space time transmit diversity (STTD): 45
- S-parameter: 719
- spatial division multiple access (SDMA): 58
- spectrum handoff: 78
- spectrum reservation sensing: 734
- spurious free dynamic range (SFDR): 385
- square-root raised cosine filter: 523
- static control program: 660
- static filter bank: 174
- sub-sampling: 336, 390
- substrate integrated waveguide: 718
- superheterodyne: 122
- super-heterodyne receivers: 611, 692
- surface acoustic wave: 622
- switched-capacitor: 418
- switched-capacitor filters: 673
- symbol rate per carrier: 727
- symbol error rate (SER): 477
- synchronization: 518, 541
- synchronization channel (SCH): 78
- synthesizer: 694
- synthesizer pulling: 617
- system generator: 227, 362
- system-on-chip (SoC): 603

- terrestrial digital multimedia broadcasting (T-DMB): 108
- thermal conductivity: 89
- third generation (3G): 28
- threshold: 748
- time delayed neural networks (TDNN): 200
- time division duplex (TDD): 58
- time division multiple access (TDMA): 58
- time multiplexing: 577
- tone injection: 275, 314
- tone reservation: 274, 315
- total transmit power: 45
- tracking filter: 694
- training symbols: 561
- transceiver: 603
- transmit power: 132
- transmitter linearity: 121
- travelling wave tube PA (TWTA): 288
- truncated Volterra series: 515

- ultra wideband (UWB): 707
- under-sampling: 258
- underlay: 78
- unicasting: 3, 102
- universal terrestrial radio access (UTRA): 58
- up-conversion: 333
- up-link (UL): 543
- user equipment (UE): 61, 120

- V-band: 718
- vector threshold decomposition: 253
- very low IF (VLIF): 620, 642
- virtual machine: 665

-
- voltage controlled oscillators (VCO): 80, 547, 604
 - Volterra kernels: 202
 - Volterra series: 169, 202, 249
 - Volterra series model: 510

 - waveform description language: 652
 - WCDMA: 7, 28, 287
 - white space: 747
 - wideband-IF: 123
 - widow position shift: 531
 - window shift: 562

 - Wiener-Hammerstein model: 289
 - Wiener model: 176, 207, 288, 520
 - wireless microphones: 747
 - WLAN: 541

 - Z-domain transfer function: 630
 - Z-transform: 676
 - zero crossings: 153
 - zero-forcing: 17
 - zero-order hold: 402
 - zero-IF: 122, 388, 494, 620
 - zero-padded: 490

