

# Fixed-Point Signal Processing

# Synthesis Lectures on Signal Processing

## Editor

José Moura, *Carnegie Mellon University*

## Fixed-Point Signal Processing

Wayne T. Padgett and David V. Anderson  
2009

## Advanced Radar Detection Schemes Under Mismatched Signal Models

Francesco Bandiera, Danilo Orlando, Giuseppe Ricci  
2009

## DSP for MATLAB™ and LabVIEW™ IV: LMS Adaptive Filtering

Forester W. Isen  
2009

## DSP for MATLAB™ and LabVIEW™ III: Digital Filter Design

Forester W. Isen  
2008

## DSP for MATLAB™ and LabVIEW™ II: Discrete Frequency Transforms

Forester W. Isen  
2008

## DSP for MATLAB™ and LabVIEW™ I: Fundamentals of Discrete Signal Processing

Forester W. Isen  
2008

## The Theory of Linear Prediction

P. P. Vaidyanathan  
2007

## Nonlinear Source Separation

Luis B. Almeida  
2006

## Spectral Analysis of Signals: The Missing Data Case

Yanwei Wang, Jian Li, Petre Stoica  
2006

© Springer Nature Switzerland AG 2022  
Reprint of original edition © Morgan & Claypool 2009

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Fixed-Point Signal Processing

Wayne T. Padgett and David V. Anderson

ISBN: 978-3-031-01405-5    paperback

ISBN: 978-3-031-02533-4    ebook

DOI 10.1007/978-3-031-02533-4

A Publication in the Springer series

*SYNTHESIS LECTURES ON SIGNAL PROCESSING*

Lecture #9

Series Editor: José Moura, *Carnegie Mellon University*

Series ISSN

Synthesis Lectures on Signal Processing

Print 1932-1236    Electronic 1932-1694

# Fixed-Point Signal Processing

Wayne T. Padgett

Rose-Hulman Institute of Technology

David V. Anderson

Georgia Institute of Technology

*SYNTHESIS LECTURES ON SIGNAL PROCESSING #9*

## ABSTRACT

This book is intended to fill the gap between the “ideal precision” digital signal processing (DSP) that is widely taught, and the limited precision implementation skills that are commonly required in fixed-point processors and field programmable gate arrays (FPGAs). These skills are often neglected at the university level, particularly for undergraduates. We have attempted to create a resource both for a DSP elective course, and for the practicing engineer with a need to understand fixed-point implementation. Although we assume a background in DSP, Chapter 2 contains a review of basic theory, and Chapter 3 reviews random processes to support the noise model of quantization error. Chapter 4 details the binary arithmetic that underlies fixed-point processors, and then introduces fractional format for binary numbers. Chapter 5 covers the noise model for quantization error and the effects of coefficient quantization in filters. Because of the numerical sensitivity of IIR filters, they are used extensively as an example system in both Chapters 5 and 6. Fortunately, the principles of dealing with limited precision can be applied to a wide variety of numerically sensitive systems, not just IIR filters. Chapter 6 discusses the problems of product roundoff error, and various methods of scaling to avoid overflow. Chapter 7 discusses limit cycle effects and a few common methods for minimizing them.

There are a number of simple exercises integrated into the text to allow you to test your understanding. Answers to the exercises are included in the footnotes. A number of Matlab examples are provided in the text. They generally assume access to the Fixed-Point Toolbox. If you lack access to this software, consider either purchasing or requesting an evaluation license from The Mathworks. The code listed in the text and other helpful Matlab code is also available at <http://www.morganclaypool.com/page/padgett> and <http://www.rose-hulman.edu/~padgett/fpsp>. You will also find Matlab exercises designed to demonstrate each of the four types of error discussed in Chapters 5 and 6. Simulink examples are also provided on the web site.

## KEYWORDS

binary arithmetic, clipping, coefficient quantization, digital signal processing (DSP), field programmable gate array (FPGA), fixed-point, IIR filter, limit cycles, Matlab, numerical sensitivity, overflow,  $Q15$ , quantization error, quantization noise, roundoff noise, saturation, scaling, second order sections (SOS), signal to noise ratio (SNR), Simulink, uniform noise, wrapping

# Contents

Notes .....	xi
<b>1 Getting Started .....</b>	<b>1</b>
1.1 Design Flow .....	1
1.2 Tools .....	2
<b>2 DSP Concepts .....</b>	<b>3</b>
2.1 Basic Systems Theory .....	3
2.1.1 Linear, Time-Invariant Systems	3
2.1.2 Difference Equations	4
2.1.3 Convolution	4
2.2 z-Transform .....	5
2.2.1 Poles and Zeros	7
2.3 Block Diagrams and Filter Implementation .....	9
2.3.1 Transpose Filters	10
2.3.2 Cascaded Second-order Sections	11
2.4 Frequency Response .....	13
2.4.1 Frequency Response from the z-transform	14
2.4.2 Frequency Response Examples	14
<b>3 Random Processes and Noise .....</b>	<b>19</b>
3.1 Random Variables .....	19
3.1.1 Expectations and Moments	21
3.1.2 Stationary and Ergodic Processes	23
3.1.3 Definitions and Properties	24
3.2 Random Processes and Fourier Analysis .....	25
3.2.1 Fourier Transform of Correlation and Covariance	25

3.2.2	Power Spectral Density	26
3.2.3	Filtering a Random Sequence	27
3.2.4	Filtering a White Random Sequence	27
3.2.5	Periodograms	28
<b>4</b>	<b>Fixed Point Numbers</b>	<b>31</b>
4.1	Binary Arithmetic	31
4.1.1	Unsigned Binary Representation	31
4.1.2	Addition	34
4.1.3	Subtraction	34
4.1.4	Multiplication	35
4.1.5	Division	35
4.1.6	Signed Binary Representation	36
4.2	Q-Format	41
4.3	Fixed-Point Arithmetic	43
4.3.1	Multiplication	44
4.3.2	Addition	44
4.3.3	Rounding	45
4.4	An FIR Filter Example	45
4.4.1	Quantization Example - Computing $y[0]$	46
4.4.2	Quantization Example - Computing $y[1]$	47
4.4.3	Quantization Example - Results	47
4.4.4	Matlab Example	48
4.5	Floating-Point	49
4.6	Block Floating-Point	50
<b>5</b>	<b>Quantization Effects: Data and Coefficients</b>	<b>53</b>
5.1	Four Types of Error	53
5.2	Data Quantization	54
5.2.1	Analog-to-Digital Conversion	54
5.2.2	Ranges	57

5.2.3	Quantization Noise Power	59
5.2.4	Signal-to-Noise Ratio	60
5.2.5	Saturation and Overflow	62
5.3	Coefficient Quantization	67
5.3.1	Significant Factors in Coefficient Quantization	68
5.3.2	2nd-Order Coupled Form Structure	69
5.3.3	Direct Form IIR Filters - Coefficient Quantization Problems	70
5.3.4	Cascaded Second-Order Section Filters	73
6	Quantization Effects - Round-Off Noise and Overflow	83
6.1	Round-Off Noise	83
6.1.1	Calculation Example	87
6.2	Overflow and Scaling	90
6.2.1	Overflow Review	90
6.2.2	Scaling	91
6.2.3	Norms	91
6.2.4	$L_{1t}$ Scaling	92
6.2.5	$L_{\infty\omega}$ Scaling	94
6.2.6	$L_2$ Norm Scaling	95
6.2.7	Scale Factor Comparisons	95
6.2.8	Cascaded Second Order Sections	96
6.3	Design Example	101
6.3.1	Scaling Calculations	101
6.3.2	Output Noise Calculations	105
6.3.3	Noise Spectrum Calculations	107
6.3.4	SNR Calculations	107
6.3.5	Comparing Different Choices	109
7	Limit Cycles	113
	Glossary	117
	Bibliography	119



# Notes

## ACKNOWLEDGMENTS

We would like to gratefully acknowledge the support of Texas Instruments and The Mathworks in the development of this material.

## PERMISSIONS

Figure 5.8 is adapted from [26]. Permission is pending.

Figures 5.14 through 5.19 are adapted from [23]. Permission is pending.

## PRELIMINARY

This is a preliminary version of this book.

Wayne T. Padgett and David V. Anderson  
September 2009