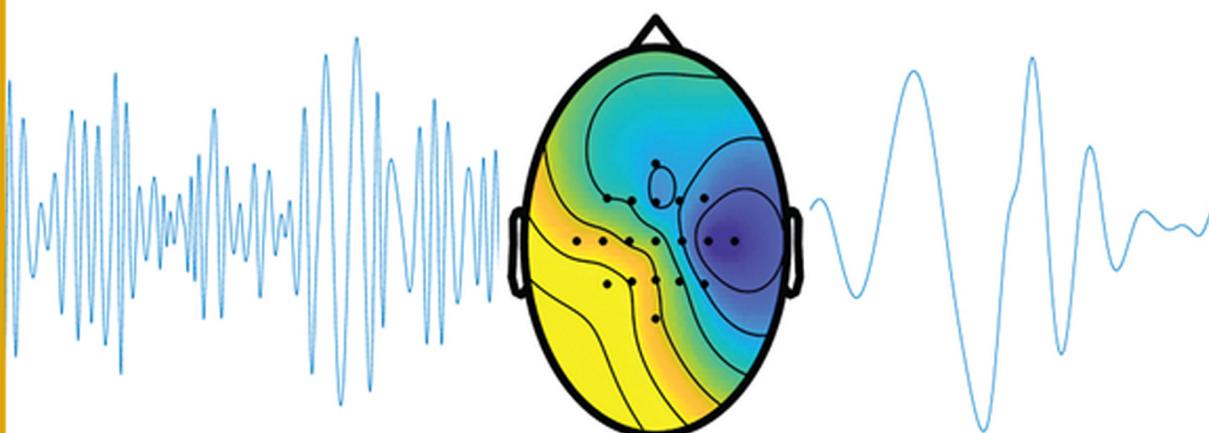


Biomedical Signal Analysis

RANGARAJ M. RANGAYYAN
SRIDHAR KRISHNAN

Third Edition



IEEE Press Series in Biomedical Engineering
Metin Akay, Series Editor



IEEE Engineering in Medicine
and Biology Society, Sponsor

IEEE PRESS

WILEY

Biomedical Signal Analysis

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

IEEE Press Editorial Board
Sarah Spurgeon, *Editor-in-Chief*

Moeness Amin
Jón Atli Benediktsson
Adam Drobot
James Duncan

Ekram Hossain
Brian Johnson
Hai Li
James Lyke
Joydeep Mitra

Desineni Subbaram Naidu
Tony Q. S. Quek
Behzad Razavi
Thomas Robertazzi
Diomidis Spinellis

Biomedical Signal Analysis

Third Edition

Rangaraj M. Rangayyan

University of Calgary
Calgary, AB
Canada

Sridhar Krishnan

Toronto Metropolitan University
Toronto, ON
Canada



IEEE Press Series in Biomedical
Engineering

IEEE PRESS
WILEY

Copyright © 2024 by The Institute of Electrical and Electronics Engineers, Inc.
All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey. Published
simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Trademarks: Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data Applied for:

Hardback: 9781119825852

Cover Design: Wiley Cover
Image: © Sri Krishnan

DEDICATION

*Mátr dévô bhava
Pitr dévô bhava
Áchárya dévô bhava*

Look upon your mother as your God
Look upon your father as your God
Look upon your teacher as your God

— from the sacred Vedic hymns of the *Taittireeya Upanishad* of India.

This book is dedicated to the fond memory of
my mother Srimati Padma Srinivasan Rangayyan
and my father Sri Srinivasan Mandayam Rangayyan,
and to all of my teachers,
in particular, Professor Ivaturi Surya Narayana Murthy.
Rangaraj

This book is dedicated to
my parents, mentors, students,
and my wife Mahitha,
and to our children Sibi and Sarvi.
Sridhar

CONTENTS

About the Authors	xvi
Foreword by Prof. Willis J. Tompkins	xviii
Foreword by Prof. Alan V. Oppenheim	xix
Preface	xxii
Acknowledgments	xxviii
Symbols and Abbreviations	xxxi
About the Companion Website	xxxix
1 Introduction to Biomedical Signals	1
1.1 The Nature of Biomedical Signals	1
1.2 Examples of Biomedical Signals	4
1.2.1 The action potential of a cardiac myocyte	5
1.2.2 The action potential of a neuron	9
1.2.3 The electroneurogram (ENG)	10
1.2.4 The electromyogram (EMG)	12
1.2.5 The electrocardiogram (ECG)	20
1.2.6 The electroencephalogram (EEG)	29
1.2.7 Event-related potentials (ERPs)	35
1.2.8 The electrogastrogram (EGG)	36
1.2.9 The phonocardiogram (PCG)	37
	vii

1.2.10	The carotid pulse	40
1.2.11	The photoplethysmogram (PPG)	41
1.2.12	Signals from catheter-tip sensors	43
1.2.13	The speech signal	44
1.2.14	The vibroarthrogram (VAG)	48
1.2.15	The vibromyogram (VMG)	52
1.2.16	Otoacoustic emission (OAE) signals	52
1.2.17	Bioacoustic signals	52
1.3	Objectives of Biomedical Signal Analysis	52
1.4	Challenges in Biomedical Signal Analysis	55
1.5	Why Use Computer-aided Monitoring and Diagnosis?	58
1.6	Remarks	60
1.7	Study Questions and Problems	60
1.8	Laboratory Exercises and Projects	62
	References	63
2	Analysis of Concurrent, Coupled, and Correlated Processes	71
2.1	Problem Statement	71
2.2	Illustration of the Problem with Case Studies	72
2.2.1	The ECG and the PCG	72
2.2.2	The PCG and the carotid pulse	73
2.2.3	The ECG and the atrial electrogram	73
2.2.4	Cardiorespiratory interaction	75
2.2.5	Heart-rate variability	75
2.2.6	The EMG and VMG	77
2.2.7	The knee-joint and muscle-vibration signals	77
2.3	Application: Segmentation of the PCG	78
2.4	Application: Diagnosis and Monitoring of Sleep Apnea	79
2.4.1	Monitoring of sleep apnea by polysomnography	80
2.4.2	Home monitoring of sleep apnea	80
2.4.3	Multivariate and multiorgan analysis	82
2.5	Remarks	85
2.6	Study Questions and Problems	85
2.7	Laboratory Exercises and Projects	86
	References	86
3	Filtering for Removal of Artifacts	91
3.1	Problem Statement	91
3.2	Random, Structured, and Physiological Noise	92
3.2.1	Random noise	92
3.2.2	Structured noise	98
3.2.3	Physiological interference	98
3.2.4	Stationary, nonstationary, and cyclostationary processes	99

3.3	Illustration of the Problem with Case Studies	101
3.3.1	Noise in event-related potentials	102
3.3.2	High-frequency noise in the ECG	102
3.3.3	Motion artifact in the ECG	102
3.3.4	Power-line interference in ECG signals	103
3.3.5	Maternal ECG interference in fetal ECG	105
3.3.6	Muscle-contraction interference in VAG signals	105
3.3.7	Potential solutions to the problem	106
3.4	Fundamental Concepts of Filtering	106
3.4.1	Linear shift-invariant filters and convolution	107
3.4.2	Transform-domain analysis of signals and systems	117
3.4.3	The pole–zero plot	123
3.4.4	The Fourier transform	125
3.4.5	The discrete Fourier transform	126
3.4.6	Convolution using the DFT	131
3.4.7	Properties of the Fourier transform	133
3.5	Synchronized Averaging	135
3.6	Time-domain Filters	139
3.6.1	Moving-average filters	139
3.6.2	Derivative-based operators to remove low-frequency artifacts	145
3.6.3	Various specifications of a filter	152
3.7	Frequency-domain Filters	153
3.7.1	Removal of high-frequency noise: Butterworth lowpass filters	154
3.7.2	Removal of low-frequency noise: Butterworth highpass filters	161
3.7.3	Removal of periodic artifacts: Notch and comb filters	162
3.8	Order-statistic Filters	169
3.9	The Wiener Filter	171
3.10	Adaptive Filters for Removal of Interference	180
3.10.1	The adaptive noise canceler	181
3.10.2	The least-mean-squares adaptive filter	184
3.10.3	The RLS adaptive filter	185
3.11	Selecting an Appropriate Filter	190
3.12	Application: Removal of Artifacts in ERP Signals	193
3.13	Application: Removal of Artifacts in the ECG	196
3.14	Application: Maternal–Fetal ECG	197
3.15	Application: Muscle-contraction Interference	199
3.16	Remarks	202
3.17	Study Questions and Problems	202
3.18	Laboratory Exercises and Projects	208
	References	209
4	Detection of Events	213
4.1	Problem Statement	213

4.2	Illustration of the Problem with Case Studies	214
4.2.1	The P, QRS, and T waves in the ECG	214
4.2.2	The first and second heart sounds	215
4.2.3	The dicrotic notch in the carotid pulse	215
4.2.4	EEG rhythms, waves, and transients	215
4.3	Detection of Events and Waves	218
4.3.1	Derivative-based methods for QRS detection	218
4.3.2	The Pan–Tompkins algorithm for QRS detection	220
4.3.3	Detection of the P wave in the ECG	224
4.3.4	Detection of the T wave in the ECG	226
4.3.5	Detection of the dicrotic notch	228
4.4	Correlation Analysis of EEG Rhythms	228
4.4.1	Detection of EEG rhythms	228
4.4.2	Template matching for EEG spike-and-wave detection	231
4.4.3	Detection of EEG rhythms related to seizure	234
4.5	Cross-spectral Techniques	235
4.5.1	Coherence analysis of EEG channels	235
4.6	The Matched Filter	237
4.6.1	Derivation of the transfer function of the matched filter	237
4.6.2	Detection of EEG spike-and-wave complexes	241
4.7	Homomorphic Filtering	242
4.7.1	Generalized linear filtering	244
4.7.2	Homomorphic deconvolution	244
4.7.3	Extraction of the vocal-tract response	245
4.8	Application: ECG Rhythm Analysis	253
4.9	Application: Identification of Heart Sounds	254
4.10	Application: Detection of the Aortic Component of S2	256
4.11	Remarks	259
4.12	Study Questions and Problems	259
4.13	Laboratory Exercises and Projects	261
	References	262
5	Analysis of Waveshape and Waveform Complexity	267
5.1	Problem Statement	267
5.2	Illustration of the Problem with Case Studies	268
5.2.1	The QRS complex in the case of bundle-branch block	268
5.2.2	The effect of myocardial ischemia on QRS waveshape	268
5.2.3	Ectopic beats	268
5.2.4	Complexity of the EMG interference pattern	268
5.2.5	PCG intensity patterns	269
5.3	Analysis of ERPs	269
5.4	Morphological Analysis of ECG Waves	269
5.4.1	Correlation coefficient	270

5.4.2	The minimum-phase correspondent and signal length	270
5.4.3	ECG waveform analysis	274
5.5	Envelope Extraction and Analysis	277
5.5.1	Amplitude demodulation	278
5.5.2	Synchronized averaging of PCG envelopes	280
5.5.3	The envelopogram	281
5.6	Analysis of Activity	283
5.6.1	The <i>RMS</i> value	283
5.6.2	Zero-crossing rate	285
5.6.3	Turns count	285
5.6.4	Form factor	286
5.7	Application: Normal and Ectopic ECG Beats	287
5.8	Application: Analysis of Exercise ECG	288
5.9	Application: Analysis of the EMG in Relation to Force	290
5.10	Application: Analysis of Respiration	292
5.11	Application: Correlates of Muscular Contraction	294
5.12	Application: Statistical Analysis of VAG Signals	295
5.12.1	Acquisition of knee-joint VAG signals	297
5.12.2	Estimation of the PDFs of VAG signals	297
5.12.3	Screening of VAG signals using statistical parameters	299
5.13	Application: Fractal Analysis of the EMG in Relation to Force	302
5.13.1	Fractals in nature	302
5.13.2	Fractal dimension	303
5.13.3	Fractal analysis of physiological signals	304
5.13.4	Fractal analysis of EMG signals	305
5.14	Remarks	306
5.15	Study Questions and Problems	307
5.16	Laboratory Exercises and Projects	309
	References	310
6	Frequency-domain Characterization of Signals and Systems	317
6.1	Problem Statement	318
6.2	Illustration of the Problem with Case Studies	318
6.2.1	The effect of myocardial elasticity on heart sound spectra	318
6.2.2	Frequency analysis of murmurs to diagnose valvular defects	319
6.3	Estimation of the PSD	321
6.3.1	Considerations in the computation of the ACF	321
6.3.2	The periodogram	323
6.3.3	The need for averaging PSDs	325
6.3.4	The use of windows: spectral resolution and leakage	326
6.3.5	Estimation of the ACF from the PSD	330
6.3.6	Synchronized averaging of PCG spectra	331
6.4	Measures Derived from PSDs	333

6.4.1	Moments of PSD functions	334
6.4.2	Spectral power ratios	337
6.5	Application: Evaluation of Prosthetic Heart Valves	337
6.6	Application: Fractal Analysis of VAG Signals	339
6.6.1	Fractals and the $1/f$ model	339
6.6.2	FD via power spectral analysis	341
6.6.3	Examples of synthesized fractal signals	341
6.6.4	Fractal analysis of segments of VAG signals	342
6.7	Application: Spectral Analysis of EEG Signals	345
6.8	Remarks	349
6.9	Study Questions and Problems	350
6.10	Laboratory Exercises and Projects	351
	References	353
7	Modeling of Biomedical Signal-generating Processes and Systems	357
7.1	Problem Statement	357
7.2	Illustration of the Problem	358
7.2.1	Motor-unit firing patterns	358
7.2.2	Cardiac rhythm	358
7.2.3	Formants and pitch in speech	359
7.2.4	Patellofemoral crepitus	360
7.3	Point Processes	360
7.4	Parametric System Modeling	365
7.5	Autoregressive or All-pole Modeling	369
7.5.1	Spectral matching and parameterization	374
7.5.2	Optimal model order	377
7.5.3	AR and cepstral coefficients	384
7.6	Pole-Zero Modeling	384
7.6.1	Sequential estimation of poles and zeros	387
7.6.2	Iterative system identification	389
7.6.3	Homomorphic prediction and modeling	393
7.7	Electromechanical Models of Signal Generation	395
7.7.1	Modeling of respiratory sounds	396
7.7.2	Modeling sound generation in coronary arteries	400
7.7.3	Modeling sound generation in knee joints	402
7.8	Electrophysiological Models of the Heart	404
7.8.1	Electrophysiological modeling at the cellular level	405
7.8.2	Electrophysiological modeling at the tissue and organ levels	410
7.8.3	Extensions to the models of the heart	412
7.8.4	Challenges and future considerations in modeling the heart	414
7.9	Application: Heart-rate Variability	416
7.10	Application: Spectral Modeling and Analysis of PCG Signals	418
7.11	Application: Coronary Artery Disease	421

7.12	Remarks	423
7.13	Study Questions and Problems	424
7.14	Laboratory Exercises and Projects	425
	References	426
8	Adaptive Analysis of Nonstationary Signals	431
8.1	Problem Statement	432
8.2	Illustration of the Problem with Case Studies	432
8.2.1	Heart sounds and murmurs	432
8.2.2	EEG rhythms and waves	433
8.2.3	Articular cartilage damage and knee-joint vibration	433
8.3	Time-variant Systems	435
8.3.1	Characterization of nonstationary signals and dynamic systems	436
8.4	Fixed Segmentation	438
8.4.1	The short-time Fourier transform	438
8.4.2	Considerations in short-time analysis	441
8.5	Adaptive Segmentation	445
8.5.1	Spectral error measure	445
8.5.2	ACF distance	450
8.5.3	The generalized likelihood ratio	450
8.5.4	Comparative analysis of the ACF, SEM, and GLR methods	452
8.6	Use of Adaptive Filters for Segmentation	452
8.6.1	Monitoring the RLS filter	453
8.6.2	The RLS lattice filter	456
8.7	The Kalman Filter	463
8.8	Wavelet Analysis	474
8.8.1	Approximation of a signal using wavelets	474
8.9	Bilinear TFDs	479
8.10	Application: Adaptive Segmentation of EEG Signals	485
8.11	Application: Adaptive Segmentation of PCG Signals	489
8.12	Application: Time-varying Analysis of HRV	490
8.13	Application: Analysis of Crying Sounds of Infants	493
8.14	Application: Wavelet Denoising of PPG Signals	493
8.15	Application: Wavelet Analysis for CPR Studies	494
8.16	Application: Detection of Ventricular Fibrillation in ECG Signals	499
8.17	Application: Detection of Epileptic Seizures in EEG Signals	503
8.18	Application: Neural Decoding for Control of Prostheses	505
8.19	Remarks	506
8.20	Study Questions and Problems	507
8.21	Laboratory Exercises and Projects	507
	References	508

9	Signal Analysis via Adaptive Decomposition	515
9.1	Problem Statement	517
9.2	Illustration of the Problem with Case Studies	517
9.2.1	Separation of the fetal ECG from a single-channel abdominal ECG	517
9.2.2	Patient-specific EEG channel selection for BCI applications	518
9.2.3	Detection of microvolt T-wave alternans in long-term ECG recordings	518
9.3	Matching Pursuit	518
9.4	Empirical Mode Decomposition	520
9.4.1	Variants of empirical mode decomposition	521
9.5	Dictionary Learning	523
9.6	Decomposition-based Adaptive TFD	525
9.7	Separation of Mixtures of Signals	531
9.7.1	Principal component analysis	533
9.7.2	Independent component analysis	539
9.7.3	Nonnegative matrix factorization	542
9.7.4	Comparison of PCA, ICA, and NMF	546
9.8	Application: Detection of Epileptic Seizures Using Dictionary Learning Methods	553
9.9	Application: Adaptive Time–Frequency Analysis of VAG Signals	560
9.10	Application: Detection of T-wave Alternans in ECG Signals	568
9.11	Application: Extraction of the Fetal ECG from Single-channel Maternal ECG	572
9.12	Application: EEG Analysis for Brain–Computer Interfaces	577
9.12.1	NMF-based channel selection	579
9.12.2	Feature extraction	579
9.13	Remarks	586
9.14	Study Questions and Problems	586
9.15	Laboratory Exercises and Projects	586
	References	587
10	Computer-aided Diagnosis and Healthcare	595
10.1	Problem Statement	596
10.2	Illustration of the Problem with Case Studies	596
10.2.1	Diagnosis of bundle-branch block	596
10.2.2	Normal or ectopic ECG beat?	597
10.2.3	Is there an alpha rhythm?	598
10.2.4	Is a murmur present?	598
10.2.5	Detection of sleep apnea using multimodal biomedical signals	598
10.3	Pattern Classification	599
10.4	Supervised Pattern Classification	600
10.4.1	Discriminant and decision functions	600
10.4.2	Fisher linear discriminant analysis	601

10.4.3	Distance functions	605
10.4.4	The nearest-neighbor rule	605
10.4.5	The support vector machine	606
10.5	Unsupervised Pattern Classification	607
10.5.1	Cluster-seeking methods	607
10.6	Probabilistic Models and Statistical Decision	611
10.6.1	Likelihood functions and statistical decision	611
10.6.2	Bayes classifier for normal patterns	613
10.7	Logistic Regression Analysis	614
10.8	Neural Networks	615
10.8.1	ANNs with radial basis functions	617
10.8.2	Deep learning	620
10.9	Measures of Diagnostic Accuracy and Cost	620
10.9.1	Receiver operating characteristics	623
10.9.2	McNemar's test of symmetry	625
10.10	Reliability of Features, Classifiers, and Decisions	627
10.10.1	Separability of features	628
10.10.2	Feature selection	630
10.10.3	The training and test steps	631
10.11	Application: Normal versus Ectopic ECG Beats	633
10.11.1	Classification with a linear discriminant function	633
10.11.2	Application of the Bayes classifier	637
10.11.3	Classification using the K -means method	637
10.12	Application: Detection of Knee-joint Cartilage Pathology	637
10.13	Application: Detection of Sleep Apnea	644
10.14	Application: Monitoring Parkinson's Disease Using Multimodal Signal Analysis	647
10.15	Strengths and Limitations of CAD	650
10.16	Remarks	656
10.17	Study Questions and Problems	657
10.18	Laboratory Exercises and Projects	658
	References	659
	Index	665

ABOUT THE AUTHORS

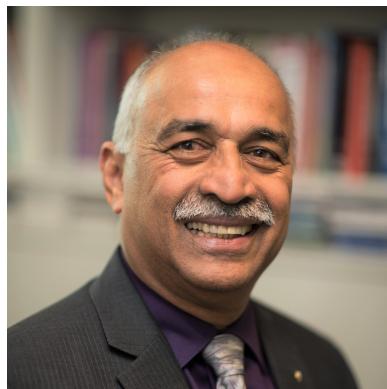


Photo credit: Skogen Photography
for the University of Calgary.

Rangaraj M. Rangayyan is Professor Emeritus of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, Canada. He received the Bachelor of Engineering degree in Electronics and Communication Engineering in 1976 from the University of Mysore at the People's Education Society College of Engineering, Mandya, Karnataka, India, and the Ph.D. in Electrical Engineering from the Indian Institute of Science, Bangalore, Karnataka, India, in 1980. He served the University of Manitoba, Winnipeg, Manitoba, Canada and the University of Calgary in research, academic, and administrative positions from 1981 to 2016. His research interests are in digital signal and image processing, biomedical signal and image analysis, and computer-aided diagnosis.

Dr. Rangayyan has published more than 170 papers in journals and 270 papers in proceedings of conferences. According to Google Scholar, Dr. Rangayyan's publications have attracted > 18,000 citations with an h-index > 66. He has supervised or cosupervised 27 Master's theses, 17 Doctoral theses, and more than 50 researchers at various levels. He has been recognized with the 1997 and 2001 Research Excellence Awards of the Department of Electrical and Computer Engineering, the 1997 Research Award of the Faculty of Engineering, appointment as "University Professor" (2003 to 2013) at the University of Calgary, and Outstanding Teaching Performance Award of the Schulich School of Engineering (2016). He is the author of two textbooks: "Biomedical Signal Analysis" (IEEE/ Wiley, 2002, 2015) and "Biomedical Image Analysis" (CRC, 2005). He has coauthored and coedited several books, including "Color Image Processing with Biomedical

Applications" (SPIE, 2011). He was recognized with the 2013 IEEE Canada Outstanding Engineer Medal, the IEEE Third Millennium Medal (2000), and elected as Fellow, IEEE (2001); Fellow, Engineering Institute of Canada (2002); Fellow, American Institute for Medical and Biological Engineering (2003); Fellow, SPIE (2003); Fellow, Society for Imaging Informatics in Medicine (2007); Fellow, Canadian Medical and Biological Engineering Society (2007); Fellow, Canadian Academy of Engineering (2009); and Fellow, Royal Society of Canada (2016).

Dr. Rangayyan's research has been featured in many newsletters, magazines, and newspapers, as well as in several radio and television interviews. He has been invited to present lectures in more than 20 countries and has held Visiting or Honorary Professorships with the University of Liverpool, Liverpool, UK; Tampere University of Technology, Tampere, Finland; Universitatea Politehnica Bucureşti, Bucharest, Romania; Universidade de São Paulo, São Paulo, Brasil; Universidade Estadual Paulista, Sorocaba, São Paulo, Brasil; Cleveland Clinic Foundation, Cleveland, OH, USA; Indian Institute of Science, Bangalore, Karnataka, India; Indian Institute of Technology, Kharagpur, West Bengal, India; Manipal Institute of Technology, Manipal, Karnataka, India; Amity University, Noida, India; Beijing University of Posts and Telecommunications, Beijing, China; Xiamen University, Xiamen, Fujian, China; Kyushu University, Fukuoka, Japan; University of Rome Tor Vergata, Rome, Italy; and École Nationale Supérieure des Télécommunications de Bretagne, Brest, France. He has been recognized as a Distinguished Lecturer by the IEEE Engineering in Medicine and Biology Society (EMBS), the University of Toronto, and the Hong Kong Institution of Engineers.

For further details, please visit his website <https://rangayyan.ca>



Sridhar Krishnan received the B.E. degree in Electronics and Communication Engineering from the College of Engineering, Guindy, Anna University, India, in 1993, and the M.Sc. and Ph.D. degrees in Electrical and Computer Engineering from the University of Calgary, Calgary, Alberta, Canada, in 1996 and 1999, respectively. He joined the Department of Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University — TMU (formerly Ryerson University), Toronto, Ontario, Canada, in July 1999, and currently, he is a Professor in the Department. He was TMU's Founding Program Director of the Undergraduate Biomedical Engineering Program, and also the Founding Co-Director of the Institute for Biomedical Engineering, Science and Technology (iBEST). Dr. Krishnan is an Affiliate Scientist with the University Health Network and the Keenan Research Centre, St. Michael's Hospital, Toronto. He held the Canada Research Chair position (2007–2017) in Biomedical Signal Analysis. He has published 405 papers in refereed journals and conference proceedings, and filed/obtained 16 patents/invention disclosures. He is currently serving as a scientific advisor to six technological start-ups in the areas of digital health, wearables, and AI.

Dr. Krishnan is a recipient of the Outstanding Canadian Biomedical Engineer Award from the Canadian Medical and Biological Engineering Society, Achievement in Innovation Award from Innovate Calgary, Sarwan Sahota Distinguished Scholar Award from TMU, Young Engineer Achievement Award from Engineers Canada, New Pioneers Award in Science and Technology from Skills for Change, and Exemplary Service Award from IEEE Toronto Section. He is a Fellow of the Canadian Academy of Engineering and a registered professional engineer in the Province of Ontario.

For further details, please visit his website
<https://www.ecb.torontomu.ca/people/Krishnan.html>

FOREWORD BY PROF. WILLIS J. TOMPKINS

I have known Raj Rangayyan for more than 30 years. Our research and teaching careers in our respective universities both focused on the acquisition and analysis of signals from the human body. In 1993, I published a textbook called “Biomedical Digital Signal Processing” that I had developed to support the courses that I taught. Subsequently, about a decade later, in 2002, Raj published the first edition of his seminal book, “Biomedical Signal Analysis.”

In my book, I had focused mostly on the analysis of a single physiological signal, the electrocardiogram. However, the human body produces a myriad of signals, not just electrical but also thermal, acoustical, pressure, vibratory, and others. In his first edition, Raj summarized time and frequency domain tools to analyze many human biological signals from basic action potentials in myocytes to the diversity of signals produced by the physiological subsystems of the human body.

When I read the first edition, it was clear to me that Raj had produced the most complete work, both in breadth and depth, on the analysis of signals that are generated by the human body. His second edition was greatly expanded, adding new topics and analyses, growing from the first edition’s 512 pages to 672 pages.

Now, the third edition, written with the help of coauthor, Sridhar Krishnan, provides revisions of some of the early chapters but also adds substantial new material in the later chapters. At the end of chapters, the book also includes comprehensive sets of study questions, problems, laboratory exercises, and projects to facilitate and enhance learning. This book is an excellent resource for teaching courses on biomedical signal analysis to senior-level and graduate-level engineering students with good background in signals and systems.

WILLIS J. TOMPKINS, PH.D., FIEEE, FAIMBE, FBMES

Professor Emeritus

Department of Biomedical Engineering

University of Wisconsin-Madison

Madison, Wisconsin, USA

July, 2023

FOREWORD BY PROF. ALAN V. OPPENHEIM

I'm delighted to have this opportunity to express my thoughts about the Third Edition of Biomedical Signal Analysis and more broadly about the field of signal processing. When asked, I was totally unfamiliar with the previous editions of the book and had only a cursory familiarity with biomedical signals and issues related to that specific application of signal processing. My entire academic career as a researcher and teacher, spanning about seven decades, has been primarily focused on the theoretical aspects of signal processing and the applications to speech, radar, communications, and image analysis and processing. Being invited to write a foreword for this edition has been an opportunity for me to delve more deeply into how signal processing has been and can be used for the analysis and processing of biomedical signals and data. More broadly, it also offers me the opportunity to comment on the audience that I see this book as being best matched to and to also express some personal thoughts and comments about the field of signal processing.

Biomedical signal analysis, either formally or informally, has a long and rich history going back centuries and even millennia. Listening to acoustic signals from the heart and lungs and introducing and analyzing echoes from acoustic and higher frequency signals penetrating the body have long been well-established noninvasive diagnostic methodologies. And over many centuries, these techniques have been significantly enhanced by the invention of various transducer and sensor technologies, such as the stethoscope, X-ray and ultrasound imaging, and the MRI. With modern technology, an increasing number of highly sophisticated transducers and sensors are being developed and introduced to generate and capture biomedical signals and data for analysis on-line (that is, in real time) or off-line for screening and diagnostic purposes. As transducer, sensor, and signal processing algorithms advance, there are an increasing number of low-cost sophisticated devices for home and personal use. Overall, the field of biomedical signal analysis has become an increasingly important application area for utilizing sophisticated signal processing methods and tools and for the development of new signal processing methodologies with applications beyond this specific class of signals.

In the preface to the previous editions and this new edition, the authors indicate, as the intended audience, engineering students in their final year of undergraduate studies; specifically, that

“Electrical Engineering students with a good background in Signals and Systems will be well prepared for the material in this book. A course on Digital Signal Processing or Digital Filters would form a useful link to the material in the present book, but a capable student without this background should be able to gain a basic understanding of the subject matter.”

From my perspective, I see great value and potential hazards for some audiences. As the authors point out, for students and practitioners with a strong background in signal processing and who are just becoming involved with this application area, this book provides an excellent high-level introduction to a wide variety of biomedical signals as well as an overview of a wide variety of signal processing methodologies with rich examples of how these might be or are being applied to this class of signals. It does not nor does it claim to present these methodologies in any depth. It assumes that the reader either has the necessary background or is capable of acquiring it. Students with the background of a previous undergraduate course in signals and systems will likely be equipped to understand the signal processing terminology in the earlier chapters of this book. Signal processing concepts such as Wiener filtering, time–frequency analysis, and wavelets are more typically discussed in more advanced courses. Many of the signal processing concepts referred to in this book can easily appear simple and familiar on the surface, but their effective use ultimately depends on a relatively sophisticated understanding of the techniques, the underlying assumptions, and their limitations.

Many of the basic tools of signal processing are developed from a mathematical formulation of the objectives of the processing. While signal processing technology is firmly grounded in mathematical analysis, its effective use in practical environments is an art. An important component of the art of signal processing is in understanding the objectives, the assumptions in the development of the tools, and the tradeoffs involved. There now exist a variety of signal processing toolboxes that are more or less “plug and play,” that is, relatively straightforward to apply to a data set. The art is in choosing which to use, how to set the parameters, and how to interpret the results. For example, filtering, as discussed in Chapter 3, is one of the fundamental sets of techniques in signal processing. The most typically used digital filter designs (Butterworth, Chebychev, elliptic IIR filters, data truncation and windowing, Parks–McClellan, and Savitzky–Golay FIR filters) are all “optimum” designs for filtering and data smoothing, but with different optimality criteria and different assumptions about the data and about the objectives of the processing. Each introduces different trade-offs between time-domain and frequency-domain characteristics. Consequently, in utilizing any filter design package, it is essential for the user to understand carefully the assumptions and trade-offs associated with the various filter designs. As I often like to comment:

“anything’s optimum if you pick the error criterion correctly. And just because it’s optimum doesn’t mean it’s good.”

Another basic set of tools in signal processing is directed at or based on characterizing the frequency content in signals, that is spectral analysis as discussed in Chapter 6 and illustrated in a number of other chapters. There are many available software packages for use in spectral analysis of biomedical signals, but here again, they are developed based on underlying assumptions and objectives. Some level of stationarity in the data is, of course, one of them, and trade-offs and assumptions about the length of the data record and the underlying spectral content is another. Here again, spectral analysis of data has a long history, with many of the standard procedures more or less optimum under different formulations of optimality. For example, there is a significant difference in how one should approach spectral analysis of a data set in attempting to identify a narrowband signal in a data set versus characterizing the spectral content in a broadband signal in the presence of noise.

For data that is nonstationary in the underlying assumptions, adaptive and time–frequency analysis methods are an important part of the signal processing toolset. Many of these are discussed or

mentioned in the latter chapters of the book (for example, Chapters 7, 8, and 9), but again, the reader is cautioned to understand carefully the basis for and the underlying assumptions of these methods before applying them from a readily available toolbox to their particular data sets. In a purely theoretical sense, we can choose to characterize a signal in either the time domain or, through the Fourier transform, the frequency domain. While intuitively we can refer to a signal as having “time-varying frequency content,” a precise description of what we mean by that is often elusive. Theoretically, you’re either in the time domain or in the frequency domain, not wandering somewhere in between. While appropriately many biomedical signals are best characterized through some notion of “time-varying frequency content,” considerable care is required in interpreting what is meant by that and which tools are appropriate in a given context.

In summary, I see this book as a wonderful resource for students and practitioners who have a relatively strong signal processing background and who are working or are beginning to work with biomedical signals. I would also emphasize the cautionary note that the effective use of signal processing techniques and toolboxes is an art, and having a solid understanding of these tools is essential for their effective use. Creatively and artfully applying the tools, and perhaps modifying them, require a good understanding of the theory and mathematics behind them.

ALAN V. OPPENHEIM, Sc.D., FIEEE

*Ford Professor of Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
July, 2023*

PREFACE

From the Second to the Third Edition

The field of biomedical signal analysis has been advancing rapidly over the past few years. More and more techniques are being developed to analyze not only the well-known signals from the previous century, but new types of biomedical signals are being explored, acquired, studied, and analyzed for various novel applications. Courses on biomedical instrumentation and signal analysis are increasingly common and popular elements in engineering curricula.

The previous editions of the book were adopted for study by several students, teachers, and researchers around the world. Keeping in mind the appreciative comments received, we have maintained the first six chapters of the book with minimal change. We have also maintained the style and spirit of the original book.

The minor modifications in the new edition include the following: additional discussions and illustrations related to the neuron, action potential, and photoplethysmography (Chapter 1); the discrete Fourier transform, frequency response, pole–zero plots, and the relationships between various representations of signals, systems, and transforms (Chapter 3); and turning points, zero-crossings, and turns count (Chapter 5). Several equations and figures have been revised and reformatted for improved comprehension. A few sections have been relocated and revised for improved connection to related sections.

The major changes in the Third Edition are present in the last four chapters, which represent thoroughly revised, expanded, updated, and reorganized versions of the last three chapters in the Second Edition. Substantial new material has been added on modeling and analysis of nonstationary and multicomponent signals as well as pattern recognition and diagnostic decision. Detailed discussions have been added on the Kalman filter, dictionary learning, electrophysiological modeling, detection of epileptic seizures, analysis of ventricular fibrillation, and diagnosis of Parkinson’s disease. Additional discussions on the strengths and limitations of computer-aided diagnosis are provided at the end of Chapter 10.

The following list describes some of the new topics, techniques, and applications presented in the Third Edition:

- Mathematical and electrophysiological modeling of the heart at the cellular, tissue, and organ levels.
- The Kalman filter and dictionary-learning methods for sophisticated analysis of nonstationary, multicomponent, and multisource signals.
- Detection of ventricular fibrillation and wavelet analysis for studies on cardiopulmonary resuscitation.
- Neural decoding for control of prostheses using the Kalman filter.
- Techniques for adaptive decomposition of multicomponent and multisource signals, including dictionary learning, nonnegative matrix decomposition, and decomposition-based adaptive time–frequency distributions.
- Detection of epileptic seizures using dictionary learning.
- Detection of T-wave alternans in ECG signals.
- Extraction of the fetal ECG from single-channel maternal ECG.
- EEG analysis for brain–computer interfaces.
- Detection of sleep apnea.
- Monitoring Parkinson’s disease using multimodal signal analysis.

References appear at the end of each chapter to facilitate chapter-by-chapter access to pdf files through digital libraries. The pdf files include hyperlinks to sections, figures, equations, references, and websites cited for efficient navigation.

In order to control and limit the number of pages in the book, the page size has been increased. This change has facilitated improved formatting and layout of the text and figures. In spite of the addition of substantial new material, the number of pages in the Third Edition is almost the same as that in the Second Edition.

Expanded and improved teaching and learning resources are available at the companion website: www.wiley.com/go/rangayyan3e, and also at <https://github.com/srikrishnan1972/Biomedical-Signal-Analysis>

The Intended Audience

As with the previous editions, the Third Edition is directed at engineering students in their final (senior) year of undergraduate studies or in their graduate studies. Electrical Engineering students with a good background in signals and systems will be well prepared for the material in the book. Students in other engineering disciplines, or in computer science, physics, mathematics, or geosciences, should also be able to appreciate the material in the book. A course on digital signal processing or digital filters would form a useful link to the material in the present book, but a capable student without this background should be able to gain a basic understanding of the subject matter. The introductory materials on systems, filters, and transforms provided in Chapter 3 should assist the reader without formal training on the same topics. Practicing engineers, computer scientists, information technologists, medical physicists, and specialists in machine learning, artificial intelligence, and data processing working in diverse areas such as telecommunications, seismic and geophysical applications, biomedical applications, and hospital information systems may find the book useful

in their quest to learn advanced techniques for signal analysis. They could draw inspiration from other applications of signal processing or analysis, and satisfy their curiosity regarding computer applications in medicine, digital healthcare, and computer-aided medical diagnosis.

Teaching and Learning Plans

The book starts with an illustrated introduction to biomedical signals in Chapter 1. Chapter 2 continues the introduction, with emphasis on the analysis of multiple channels of correlated signals.

Chapter 3 deals exclusively with filtering of signals for removal of artifacts as an important step before signal analysis. The basic properties of systems and transforms as well as signal processing techniques are reviewed and described where required. The chapter is written as a mix of theory and application so as to facilitate easy comprehension of the basics of signals, systems, and transforms. The emphasis is on the application of filters to particular problems in biomedical signal analysis. A large number of illustrations are included to provide a visual representation of the problem and the effectiveness of the various filtering methods described.

Chapter 4 presents techniques that are particularly useful in the detection of events in biomedical signals. Analysis of waveshape and waveform complexity of events and components of signals is the focus of Chapter 5. Techniques for frequency-domain characterization of biomedical signals and systems are presented in Chapter 6. A number of diverse examples are provided in all of the chapters. Attention is directed to the characteristics of the problems that are encountered when analyzing and interpreting biomedical signals, rather than to any specific diagnostic application with particular signals.

The material in the book up to and including Chapter 6 provides more than adequate material for a one-semester (13-week) course at the senior (fourth-year) engineering level. Our own teaching experience indicates that this material will require about 36 – –40 hours of lectures. It would be desirable to augment the lectures with about 12 hours of tutorials (problem-solving sessions) and 10 laboratory sessions.

Modeling biomedical signal-generating processes and systems for parametric representation and analysis is the subject of Chapter 7. Chapters 8 and 9 deal with adaptive analysis of nonstationary, multicomponent, and multisource signals. The topics in these chapters are of high mathematical complexity and are not suitable for undergraduate courses. Some sections may be selected and included in a first course on biomedical signal analysis if there is particular interest in these topics. Otherwise, the three chapters could be left for self-study by those in need of the techniques, or included in an advanced course.

Chapter 10 presents the final aspect of biomedical signal analysis, and provides an introduction to pattern classification, diagnostic decision-making, computer-aided diagnosis, and computer-aided healthcare. Although this topic is advanced in nature and could form a graduate-level course on its own, the material is introduced so as to draw the entire exercise of biomedical signal analysis to its concluding stage of diagnostic decision and healthcare. It is recommended that a few sections from this chapter be included even in a first course on biomedical signal analysis so as to give the students a flavor of the end result.

Each chapter includes a number of study questions and problems to facilitate preparation for tests and examinations. A number of laboratory exercises are also provided at the end of each chapter, which could be used to formulate hands-on exercises with real-life signals. Data files related to the problems and exercises at the end of each chapter are available at the site

<https://github.com/srikrishnan1972/Biomedical-Signal-Analysis>

It is strongly recommended that the first one or two laboratory sessions in the course include visits to a local hospital, health sciences center, or clinical laboratory to view and experience procedures related to biomedical signal acquisition and analysis in a practical (clinical) setting. Signals acquired from fellow students and instructors could form interesting and motivating material for laboratory

exercises, and may be used to supplement the data files provided. A few workshops by physiologists, neuroscientists, and cardiologists should also be included in the course so as to provide the students with a nonengineering perspective on the subject.

Practical experience with real-life signals is a key element in understanding and appreciating biomedical signal analysis. This aspect could be difficult and frustrating at times, but provides professional satisfaction and educational fun!

RANGARAJ MANDAYAM RANGAYYAN

Calgary, Alberta, Canada

SRIDHAR KRISHNAN

Toronto, Ontario, Canada

November, 2023

Excerpts from the Preface to the Second Edition

The first edition of this book has been received very well around the world. Professors at several universities across North America, Europe, Asia, and other regions of the world are using the book as a textbook. A low-cost paperback edition for selected regions of the world and a Russian edition have been published. I have received several messages and comments from many students, professors, and researchers via mail and at conferences with positive feedback about the book. I am grateful to IEEE and Wiley for publishing and promoting the book and to the many users of the book for their support and feedback.

I have myself used the book to teach my course ENEL 563 Biomedical Signal Analysis at the University of Calgary. In addition to positive responses, I have received suggestions from students and professors on revising the book to provide additional examples and including advanced topics and discussions on recent developments in the book. I also made notes identifying parts of the book that could be improved for clarity, augmented with details for improved comprehension, and expanded with additional examples for better illustrations of application. I have also identified a few new developments, novel applications, and advanced techniques for inclusion in the second edition to make the book more interesting and appealing to a wider readership.

New Material in the Second Edition

In view of the success of the first edition, I have not made any major change in the organization and style of the book. Notwithstanding a tighter format to reduce white space and control the total number of pages, the second edition of the book remains similar to the first edition in terms of organization and style of presentation. New material has been inserted into the same chapters as before, thereby expanding the book. The new topics have been chosen with care not only to fit with the structure and organization of the book but also to provide additional support material and advanced topics that can be assimilated and appreciated in a first course or an advanced study of the subject area.

Some of the substantial and important additions made to the book deal with the following topics:

- analysis of the variation of parameters of the electromyogram with force;
- illustrations of the electroencephalogram with application to sleep analysis and prediction of epileptic seizures;
- details on the theory of linear systems and numerical examples related to convolution;
- details on the z -transform and the Fourier transform along with additional examples of Fourier spectra and spectral analysis of biomedical signals;
- details on linear filters and their characteristics, such as the impulse response, transfer function, and pole-zero diagrams;
- description and demonstration of nonlinear order-statistic filters;

- derivation of the matched filter;
- derivations related to the complex cepstrum;
- details on random processes and their properties;
- wavelets and the wavelet transform with biomedical applications;
- fractal analysis with biomedical applications;
- time–frequency distributions and analysis of nonstationary signals with biomedical applications;
- principal component analysis, independent component analysis, and blind source separation with biomedical applications;
- monitoring of sleep apnea;
- analysis of various types of bioacoustic signals that could bear diagnostic information; and
- methods for pattern analysis and classification with illustrations of application to biomedical signals.

Discussions related to the topics listed above are spread throughout the book with several new references added to assist the reader in further studies. Many more problems and projects have been added at the ends of the chapters.

The first edition of the book (2002) has 516 pages (plus xxxv pages of front matter) with nine chapters, 538 numbered equations (with many more equations not numbered but as parts of procedures), 232 numbered figures (many with multiple subfigures), and 265 references. The second edition (2015) has 672 pages (plus xlivi pages of front matter) in a more compact layout than the first edition, with nine chapters, 814 numbered equations (and many more equations not numbered but as parts of procedures), 370 numbered figures (many with multiple subfigures), and 505 references. The discussions on some of the new topics added were kept brief in order to control the size of the book; regardless, the second edition is approximately 50% larger than the first edition in several aspects.

RANGARAJ MANDAYAM RANGAYYAN

Calgary, Alberta, Canada

February, 2015

Excerpts from the Preface to the First Edition: Background and Motivation

The establishment of the clinical electrocardiograph (ECG) by the Dutch physician Willem Einthoven in 1903 marked the beginning of a new era in medical diagnostic techniques, including the entry of electronics into healthcare. Since then, electronics, and subsequently computers, have become integral components of biomedical signal analysis systems, performing a variety of tasks from data acquisition and preprocessing for removal of artifacts to feature extraction and interpretation. Electronic instrumentation and computers have been applied to investigate a host of biological and physiological systems and phenomena, such as the electrical activity of the cardiovascular system, the brain, the neuromuscular system, and the gastric system; pressure variations in the cardiovascular system; sound and vibration signals from the cardiovascular, the musculoskeletal, and the respiratory systems; and magnetic fields of the brain, to name a few.

The primary step in investigations of physiological systems requires the development of appropriate sensors and instrumentation to transduce the phenomenon of interest into a measurable electrical signal. The next step of analysis of the signals, however, is not always an easy task for a physician or life-sciences specialist. The clinically relevant information in the signal is often masked by noise and interference, and the signal features may not be readily comprehensible by the visual or auditory systems of a human observer. Heart sounds, for example, have most of their energy at or below the threshold of auditory perception of most humans; the interference patterns of a surface electromyographic signal are too complex to permit visual analysis. Some repetitive or attention-demanding tasks, such as on-line monitoring of the ECG of a critically ill patient with cardiac rhythm problems, could be uninteresting and tiring for a human observer. Furthermore, the variability present in a given type of signal from one subject to another, and the interobserver variability inherent in subjective analysis performed by physicians or analysts make consistent understanding or evaluation of any phenomenon difficult, if not impossible. These factors created the need not only for improved instrumentation,

but also for the development of methods for objective analysis via signal processing algorithms implemented in electronic hardware or on computers.

Processing of biomedical signals, until a few years ago, was mainly directed toward filtering for removal of noise and power-line interference; spectral analysis to understand the frequency characteristics of signals; and modeling for feature representation and parameterization. Recent trends have been toward quantitative or objective analysis of physiological systems and phenomena via signal analysis. The field of biomedical signal analysis has advanced to the stage of practical application of signal processing and pattern analysis techniques for efficient and improved noninvasive diagnosis, on-line monitoring of critically ill patients, and rehabilitation and sensory aids for the handicapped. Techniques developed by engineers are gaining wider acceptance by practicing clinicians, and the role of engineering in diagnosis and treatment is gaining much-deserved respect.

The major strength in the application of computers in biomedical signal analysis lies in the potential use of signal processing and modeling techniques for quantitative or objective analysis. Analysis of signals by human observers is almost always accompanied by perceptual limitations, interpersonal variations, errors caused by fatigue, errors caused by the very low rate of incidence of a certain sign of abnormality, environmental distractions, and so on. The interpretation of a signal by an expert bears the weight of the experience and expertise of the analyst; however, such analysis is almost always subjective. Computer analysis of biomedical signals, if performed with the appropriate logic, has the potential to add objective strength to the interpretation of the expert. It thus becomes possible to improve the diagnostic confidence or accuracy of even an expert with many years of experience. This approach to improved healthcare could be labeled as *computer-aided diagnosis*.

Developing an algorithm for biomedical signal analysis, however, is not an easy task; quite often, it might not even be a straightforward process. The engineer or computer analyst is often bewildered by the variability of features in biomedical signals and systems, which is far higher than that encountered in physical systems or observations. Benign diseases often mimic the features of malignant diseases; malignancies may exhibit a characteristic pattern, which, however, is not always guaranteed to appear. Handling all of the possibilities and degrees of freedom in a biomedical system is a major challenge in most applications. Techniques proven to work well with a certain system or set of signals may not work in another seemingly similar situation.

The Problem-solving Approach

The approach I have taken in presenting material in this book is primarily that of development of algorithms for problem solving. Engineers are often said to be (with admiration, I believe) problem solvers. However, the development of a problem statement and gaining of a good understanding of the problem could require a significant amount of preparatory work. I have selected a logical series of problems, from the many case studies I have encountered in my research work, for presentation in the book. Each chapter deals with a certain type of a problem with biomedical signals. Each chapter begins with a statement of the problem, followed immediately with a few illustrations of the problem with real-life case studies and the associated signals. Signal processing, modeling, or analysis techniques are then presented, starting with relatively simple “textbook” methods, followed by more sophisticated research approaches directed at the specific problem. Each chapter concludes with one or more applications to significant and practical problems. The book is illustrated copiously with real-life biomedical signals and their derivatives.

The methods presented in the book are at a fairly high level of technical sophistication. A good background in signal and system analysis as well as probability, random variables, and stochastic processes is required in order to follow the procedures and analysis. Familiarity with systems theory and transforms such as the Laplace and Fourier, the latter in both continuous and discrete versions, will be assumed. We will not be getting into details of the transducers and instrumentation techniques essential for biomedical signal acquisition; instead, we will be studying the problems present in the signals after they have been acquired, concentrating on how to solve the problems. Concurrent or prior study of the physiological phenomena associated with the signals of specific interest, with a clinical textbook, is strongly recommended.

RANGARAJ MANDAYAM RANGAYYAN

Calgary, Alberta, Canada
September, 2001

ACKNOWLEDGMENTS

Acknowledgments: Third Edition

We thank students, teachers, and colleagues around the world who have used the book and provided suggestions for improvement. In particular, we thank Benjamin Lavoie, Faraz Oloumi, Behnaz Ghoraani, Dharmendra Gurve, Amanda Dy, Muhammad Farhat Kaleem, Karthikeyan Umapathy, April Khademi, Binh Nguyen, Michael Nigro, Mahitha Krishnan, Martin Ivanov, Abdelrahman Abdou, and Alice Rueda for their assistance in the revision of the book to its Third Edition.

We are grateful to Professor Alan V. Oppenheim, Massachusetts Institute of Technology, and Professor Willis J. Tompkins, University of Wisconsin-Madison, for their inspiring books, academic leadership, contributions to the fields of signal processing and biomedical engineering, and their comments on our book provided in the Foreword.

We thank Mary Hatcher and her colleagues at Wiley for their support and help in the publication of the Third Edition of our book.

We thank our families for their continuing support.

RANGARAJ MANDAYAM RANGAYYAN

Calgary, Alberta, Canada

SRIDHAR KRISHNAN

*Toronto, Ontario, Canada
November, 2023*

Acknowledgments: Second Edition

In addition to the various people and sources acknowledged in the first edition and in captions of figures, I thank the following for their assistance with and contributions to the second edition: Faraz Oloumi, Alexander Kalinichenko, Douglas Bell, Tingting Mu, Fábio José Ayres, Shantanu Banik, Thanh Minh Cabral, Sridhar Krishnan, Yunfeng Wu, Suxian Cai, Adrian D.C. Chan, Anca Lazăr, Karthikeyan Umapathy, Ronald Platt, April Khademi, Markad V. Kamath, Rajeev Agarwal, Maarten De Vos, Hisham Alshaer, Jayasree Chakraborty, Ashis Kumar Dhara, Ivan Cruz Aceves, Behnaz Ghoraani, Paola Casti, Mehrnaz Shokrollahi, Zahra Moussavi, and T. Douglas Bradley.

I am grateful to the University of Calgary for facilitating and supporting my academic activities. I am grateful to IEEE Press and Wiley for publishing and promoting this book.

As always, I am grateful to my family — my wife Mayura, my daughter Vidya, and my son Adarsh — for their love and support.

I hope that this book will assist those who seek to enrich their lives and those of others with the exciting field of biomedical signal analysis.

RANGARAJ MANDAYAM RANGAYYAN

Calgary, Alberta, Canada

February, 2015

Acknowledgments: First Edition

To write a book on my favorite subject of biomedical signal analysis has been a long-cherished ambition of mine. Writing this book has been a major task with many facets: challenging, yet yielding more knowledge; tiring, yet stimulating the thirst to understand and appreciate more; difficult, yet satisfying when a part was brought to a certain stage of completion.

A number of very important personalities have shaped me and my educational background. My mother, Srimati Padma Srinivasan Rangayyan, and my father, Sri Srinivasan Mandayam Rangayyan, encouraged me to keep striving to gain higher levels of education and to set and achieve higher goals all the time. I have been very fortunate to have been taught and guided by a number of dedicated teachers, the most important of them being Professor Ivaturi Surya Narayana Murthy, my Ph.D. supervisor, who introduced me to the topic of this book at the Indian Institute of Science, Bangalore, Karnataka, India. It is with great respect and admiration that I dedicate this book as a humble offering to their spirits.

My basic education was imparted by many influential teachers at Saint Joseph's Convent, Saint Joseph's Indian High School, and Saint Joseph's College in Mandya and Bangalore, Karnataka, India. My engineering education was provided by the People's Education Society College of Engineering, Mandya, affiliated with the University of Mysore. I express my gratitude to all of my teachers.

My association with clinical researchers at the University of Calgary and the University of Manitoba has been invaluable in furthering my understanding of the subject matter of this book. I express my deep gratitude to Cyril Basil Frank, Gordon Douglas Bell, Joseph Edward Leo Desautels, Leszek Hahn, and Reinhard Kloiber of the University of Calgary, and Richard Gordon and George Collins of the University of Manitoba, Winnipeg, Manitoba, Canada.

My understanding and appreciation of the subject of biomedical signal analysis has been boosted by the collaborative research and studies performed with my many graduate students, postdoctoral fellows, research associates, and colleagues. I would like to place on record my gratitude to Sridhar Krishnan, Naga Ravindra Mudigonda, Margaret Hilary Alto, Ricardo José Ferrari, Liang Shen, Roseli de Deus Lopes, Antonio César Germano Martins, Marcelo Knörich Zuffo, Begoña Acha Piñero, Carmen Serrano Gotarredona, Sylvia Delgado Olabarriaga, Christian Roux, Basel Solaiman, Olivier Menut, Denise Guliato, Mihai Ciuc, Vasile Buzuloiu, Titus Zaharia, Constantin Vertan, Sarah Rose, Salahuddin Elkadiki, Kevin Eng, Nema Mohamed El-Faramawy, Arup Das, Farshad Faghah, William Alexander Rolston, Yiping Shen, Zahra Marjan Kazem Moussavi, Joseph Provine, Hieu Ngoc Nguyen, Djamel Boulfelfel, Tamer Farouk Rabie, Katherine Olivia Ladly, Yuanting Zhang, Zhi-Qiang Liu, Raman Bhalachandra Paranjape, Joseph André Rodrigue Blais, Robert

Charles Bray, Gopinath Ramaswamaiah Kuduvalli, Sanjeev Tavathia, William Mark Morrow, Timothy Chi Hung Hon, Subhasis Chaudhuri, Paul Soble, Kirby Jaman, Atam Prakash Dhawan, and Richard Joseph Lehner. In particular, I thank Sridhar and Naga for assisting me in preparing illustrations and examples; Sridhar for permitting me to use sections of his M.Sc. and Ph.D. theses; and Sridhar, Naga, Hilary, and Ricardo for careful proofreading of the drafts of the book. Sections of the book were reviewed by Robert Clark, Martin Paul Mintchev, Sanjay Srinivasan, and Abu Bakarr Sesay, University of Calgary; and Ioan Tăbuș, Tampere Technical University, Tampere, Finland; I express my gratitude to them for their comments and advice.

The book has benefited significantly from illustrations and text provided by a number of researchers worldwide, as identified in the references and permissions cited. I thank them all for enriching the book with their gifts of knowledge and kindness. I thank Bert Unterberger for drafting some of the illustrations in the book.

The research projects that have provided me with the background and experience essential in order to write the material in this book have been supported by many agencies. I thank the Natural Sciences and Engineering Research Council of Canada, the Alberta Heritage Foundation for Medical Research, the Alberta Breast Cancer Foundation, the Arthritis Society of Canada, the Nickle Family Foundation of Calgary, Control Data Corporation, the University of Calgary, the University of Manitoba, and the Indian Institute of Science for supporting my research projects.

I thank the Killam Foundation for awarding me a Resident Fellowship to facilitate work on this book. I gratefully acknowledge support from the Alberta Provincial Biomedical Engineering Graduate Programme, funded by a grant from the Whitaker Foundation, toward student assistantship for preparation of exercises and illustrations for this book and the related course ENEL 563 Biomedical Signal Analysis at the University of Calgary. I am pleased to place on record my gratitude for the generous support from the Department of Electrical and Computer Engineering and the Faculty of Engineering at the University of Calgary in terms of supplies, services, and relief from other duties.

My association with the IEEE Engineering in Medicine and Biology Society (EMBS) in many positions has benefited me considerably in numerous ways. In particular, the period as an Associate Editor of the *IEEE Transactions on Biomedical Engineering* was very rewarding, as it provided me with a wonderful opportunity to work with many leading researchers and authors of scientific articles. I thank IEEE EMBS for lending professional support to my career on many fronts. I am grateful to the IEEE Press, in particular, Metin Akay, Series Editor, IEEE Press Series in Biomedical Engineering, for inviting me to write this book.

Writing this book has been a monumental task, often draining me of all of my energy. The infinite source of inspiration and recharging of my energy has been my family — my wife Mayura, my daughter Vidya, and my son Adarsh. While supporting me with their love and affection, they have had to bear the loss of my time and effort at home. I express my sincere gratitude to my family for their love and support, and record their contribution toward the preparation of this book.

It is my humble hope that this book will assist those who seek to enrich their lives and those of others with the wonderful powers of biomedical signal analysis. Electrical and Computer Engineering is indeed a great field in the service of humanity!

RANGARAJ MANDAYAM RANGAYYAN

Calgary, Alberta, Canada

September, 2001

SYMBOLS AND ABBREVIATIONS

Note: Boldfaced letters represent the vectorial or matrix form of the variable indicated by the corresponding italicized letters. Variables or symbols used within limited contexts are not listed: they are described within their contexts. The mathematical symbols listed may stand for other entities or variables in different applications; only the common associations are listed for ready reference.

a_k	autoregressive model or filter coefficients
au	arbitrary units
$aV\{F, L, R\}$	augmented ECG leads
Ag	silver
$AgCl$	silver chloride
A_z	area under the ROC curve
ACF	autocorrelation function
ADC	analog-to-digital converter
AHI	apnea–hypopnea index
AI	aortic insufficiency
AI	artificial intelligence
AM	amplitude modulation
ANC	adaptive noise cancellation
ANN	artificial neural network
ANS	autonomic nervous system
AO	aorta, aortic (valve or pressure)
AP	action potential
AR	interval between atrial activity and the corresponding QRS
AR	autoregressive (model or filter)
ARMA	autoregressive, moving-average (model or filter)
AS	aortic stenosis
ASD	atrial septal defect
AV	atrioventricular
A2	aortic component of the second heart sound

<i>b</i>	bit
<i>b_l</i>	moving-average model or filter coefficients
<i>bpm</i>	beats per minute
BCG	ballistocardiogram
BCI	brain-computer interfacing
BMI	brain-machine interface
BP	blood pressure
BSS	blind source separation
cps	cycles per second
C	covariance matrix
<i>Ca</i>	calcium
<i>C_i</i>	the <i>i</i> th class in a pattern classification problem
<i>Cl</i>	chlorine
<i>C_{xy}</i>	covariance between <i>x</i> and <i>y</i>
CA	constant area
CAD	computer-aided diagnosis
CBR	content-based retrieval
CCF	cross-correlation function
CD	compact disk
CL	cycle length
CNS	central nervous system
<i>CO</i>	carbon monoxide
CP	carotid pulse
CPAP	continuous positive airway pressure
CPR	cardiopulmonary resuscitation
CSA	central sleep apnea
CSD	cross-spectral density, cross-spectrum
<i>CV</i>	coefficient of variation
CWT	continuous wavelet transform
<i>diag</i>	diagonal of a matrix
D	dicrotic notch in the carotid pulse
DAC	digital-to-analog converter
DC	direct current; zero frequency
DCT	discrete cosine transform
DFT	discrete Fourier transform
DL	deep learning
DM	diastolic murmur
DSP	digital signal processing
DW	dicrotic wave in the carotid pulse
DWT	discrete wavelet transform
<i>e(n), E(ω)</i>	model or estimation error
ECG	electrocardiogram, electrocardiography
ECoG	electrocorticogram
EEG	electroencephalogram
EEMD	ensemble empirical mode decomposition
EGG	electrogastrogram
EHR	Electronic Health Record
EM	electromagnetic
EMD	empirical mode decomposition
EMG	electromyogram
ENG	electroneurogram
EOG	electrooculogram
EP	energy parameter
ERP	event-related potential
ESP	energy spread parameter

E_x	total energy of the signal x
$E[]$	statistical expectation operator
f	frequency variable, usually in Hertz
fBm	fractional Brownian motion
f_c	cutoff frequency (usually at -3 dB) of a filter in Hertz
f_s	sampling frequency in Hertz
FD	fractal dimension
FDI	first dorsal interosseus
FDM	finite-difference method
FEM	finite-element method
FF	form factor
FFT	fast Fourier transform
FI	fluctuation intensity
FICA	fast independent component analysis
FIR	finite impulse response (filter)
FL	frequency localized
FLDA	Fisher linear discriminant analysis
FM	frequency modulation
FN	false negative
FNF	false-negative fraction
FP	false positive
FP	frequency parameter
FPF	false-positive fraction
FPR	false-positive rate
FSP	frequency spread parameter
FT	Fourier transform
g	gram
GEASI	geodesic-based earliest activation sites identification
GLR	generalized likelihood ratio
GTFD	generalized time–frequency distribution
h	hour
$h(t), h(n)$	impulse response of a filter
H	entropy
H	Hurst coefficient
H	as a superscript: Hermitian (complex-conjugate) matrix transposition
Hg	mercury
$H(s), H(z)$	transfer function of a filter
$H(s)$	Laplace transform of $h(t)$
$H(z)$	z -transform of $h(n)$
$H(\omega)$	frequency response of a filter
$H(\omega)$	Fourier transform of $h(t)$
HIS	Hospital Information System
HL7	Health Level-7
HR	heart rate
HRV	heart-rate variability
HSS	hypertrophic subaortic stenosis
Hz	Hertz
i	index of a series or discrete-time signal
ICA	independent component analysis
ICT	information and communication technologies
IFT	inverse Fourier transform
IHE	Integrating the Healthcare Enterprise
IIR	infinite impulse response (filter)
IMF	instantaneous mean frequency
IMF	intrinsic mode function

IMU	inertial measurement unit
IoT	Internet of Things
IPI	interpulse interval
ISO	International Standards Organization
ISTFT	inverse short-time Fourier transform
j	index of a series or discrete-time signal
j	$\sqrt{-1}$
J	Joule
JM	Jeffries–Matusita
k -NN	k nearest neighbors
K	kurtosis
K	potassium
K	Kalman gain
KLD	Kullback–Leibler distance or divergence
KLT	Karhunen–Loëve transform
l	liter
ln	natural logarithm (base e)
L_{ij}	loss function in pattern classification
LA	left atrium
LED	light emitting diode
LHS	left-hand side
LMS	least mean squares
LOO	leave one out
LP	linear prediction (model)
LSI	linear shift-invariant
LTI	linear time-invariant
LV	left ventricle
m	mean
m	mean vector of a pattern class
mA	milliamperes
min	minute
mm	millimeter
ms	millisecond
mV	millivolt
M	number of samples
MA	moving average
MCI	muscle-contraction interference
MEMD	multivariate empirical mode decomposition
MI	mitral insufficiency
ML	machine learning
MMSE	minimum mean-squared error
MP	Matching Pursuit
MPC	minimum-phase correspondent
MPTFD	Matching Pursuit time–frequency distribution
MR	mitral regurgitation
MS	mitral stenosis
MS	mean-squared (value)
MSE	mean-squared error
MU	motor unit
MUAP	motor-unit action potential
MVC	maximal voluntary contraction
nA	nanoamperes
N	number of samples
N	filter order
Na	sodium

NCFS	neighborhood component feature selection
NMF	nonnegative matrix factorization
<i>NPV</i>	negative predictive value
NREM	non-rapid eye movement
OAE	otoacoustic emission
OLS	orthogonal least-squares
OMPTFD	optimized Matching Pursuit time–frequency distribution
OSA	obstructive sleep apnea
OSI	Open Systems Interconnection
p_k	pole of a model
$p(x)$	probability density function of the random variable x
$p(x C_i)$	likelihood function of class C_i or state-conditional PDF of x
ppm	pulses per minute
pps	pulses per second
P	atrial contraction wave in the ECG
P	percussion wave in the carotid pulse
P	model order or number of poles
$P(x)$	probability of the event x
$P(C_i x)$	posterior probability that the observation x is from class C_i
Pa	Pascal
PA	predictive area
PCA	principal component analysis
PCG	phonocardiogram
PDA	patent ductus arteriosus
PDF	probability density function
PFP	patellofemoral pulse trains or signals
PI	pulmonary insufficiency
PLP	posterior leaflet prolapse
PNS	parasympathetic nervous system
PPC	physiological patellofemoral crepitus
PPG	photoplethysmogram
PPV	positive predictive value
PQ	isoelectric segment in the ECG before ventricular contraction
PS	pulmonary stenosis
PSA	power spectral analysis
PSD	power spectral density, power spectrum
PSG	polysomnography
PVC	premature ventricular contraction
P2	pulmonary component of the second heart sound
Q	model order or number of zeros
QDA	quadratic discriminant analysis
QRS	ventricular contraction wave in the ECG
QRSTA	area under the QRS and T waves
r, r	reference input to an adaptive filter
$r_j(\mathbf{x})$	average risk or loss in pattern classification
rad	radians
rad/s	radians per second
$\mathbb{R}^{M \times N}$	$M \times N$ space of real values
RA	right atrium
RBF	radial basis function
RBFN	radial basis function network
REM	rapid eye movement
RF	radio-frequency
RHS	right-hand side
RK4	Runge–Kutta family

RLS	recursive least-squares
RLSL	recursive least-squares lattice
RMS	root mean squared (value)
ROC	receiver operating characteristics
ROC	region of convergence
ROSC	return of spontaneous circulation
RPCA	robust principal component analysis
RR	interval between two successive QRS waves in an ECG
RSPWVD	reassigned smoothed pseudo Wigner–Ville distribution
RV	right ventricle
<i>s</i>	second
<i>s</i>	Laplace-domain variable
sec	second (in figures from other sources)
<i>sgn</i>	signum (sign)
<i>S</i>	skewness
$S(\omega), S(k)$	auto- or cross-spectral density; power spectral density
SaO_2	level of oxyhemoglobin in blood
SA	sinoatrial
SCD	sudden cardiac death
SD	standard deviation
SEM	spectral error measure
SEP	somatosensory evoked potential
SL	signal length
SM	systolic murmur
SMUAP	single-motor-unit action potential
SNR	signal-to-noise ratio
SNS	sympathetic nervous system
SpO_2	level of oxyhemoglobin in blood
SPWVD	smoothed pseudo Wigner–Ville distribution
ST	isoelectric segment in the ECG during ventricular contraction
STFT	short-time Fourier transform
SVM	support vector machine
SWVD	smoothed Wigner–Ville distribution
S1	first heart sound
S2	second heart sound
S3	third heart sound
S4	fourth heart sound
S^+	sensitivity of a test
S^-	specificity of a test
<i>t</i>	time variable
T	ventricular relaxation wave in the ECG
T	tidal wave in the carotid pulse
<i>T</i>	sampling interval
<i>T</i>	as a superscript: vector or matrix transposition
T^+	positive test result
T^-	negative test result
TCR	turns count rate
TF	time–frequency
TFD	time–frequency distribution
Th	threshold
TI	tricuspid insufficiency
TN	true negative
TNF	true-negative fraction
TP	true positive
TPF	true-positive fraction

TPR	true-positive rate
Tr	trace of a matrix (sum of the diagonal entries)
TS	tricuspid stenosis
TSE	total squared error
TV	television
TWA	T wave alternans
V	Volt
V1 – V6	chest leads for ECG
VAG	vibroarthrogram
VCG	vectorcardiography
VLF	very low frequency
VMG	vibromyogram
VSD	ventricular septal defect
w	filter tap weight; weighting function
\mathbf{w}	filter weight vector
WHO	World Health Organization
WP	wavelet packet
WT	wavelet transform
WVD	Wigner–Ville distribution
$x(t), x(n)$	a signal in the time domain; usually denotes input
\mathbf{x}	vectorial representation of the signal $x(n)$
\mathbf{x}	a feature vector in pattern classification
$X(f), X(\omega)$	Fourier transform of $x(t)$
$X(k)$	discrete Fourier transform of $x(n)$
$X(s)$	Laplace transform of $x(t)$
$X(z)$	z -transform of $x(n)$
$X(\tau, \omega)$	short-time Fourier transform or time–frequency distribution of $x(t)$
$y(t), y(n)$	a signal in the time domain; usually denotes output
\mathbf{y}	vectorial representation of the signal $y(n)$
$Y(f), Y(\omega)$	Fourier transform of $y(t)$
$Y(k)$	discrete Fourier transform of $y(n)$
$Y(s)$	Laplace transform of $y(t)$
$Y(z)$	z -transform of $y(n)$
z	the z -transform variable
z^{-1}	unit delay operator in discrete-time systems
z_l	zeros of a system
\mathbf{z}	a prototype feature vector in pattern classification
ZCR	zero-crossing rate
ZT	the z -transform
1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
I, II, III	limb leads for ECG
α	an EEG wave
β	an EEG wave
β	spectral component
γ	an EEG wave
γ_i	reflection coefficient
γ_{xy}	correlation coefficient between x and y
Γ_{xy}	coherence between x and y
δ	an EEG wave
δ	Dirac delta (impulse) function
ε	total squared error
η	a random variable or noise process
θ	an angle

θ	a threshold
θ	an EEG wave
θ, Θ	cross-correlation function
λ	forgetting factor in the RLS filter
μ	the mean (average) of a random variable
μ	a rhythmic wave in the EEG
μ	step size in the LMS filter and projected gradient NMF
μV	microvolt
μm	micrometer
μs	microsecond
ρ	correlation coefficient
σ	the real part of the Laplace variable s (Neper frequency)
σ	the standard deviation of a random variable
σ^2	the variance of a random variable
τ	a time interval, delay, or shift
ϕ, Φ	autocorrelation
ω	frequency variable in radians per second
Ω	frequency variable in radians per second
*	when in-line: convolution
*	as a superscript: complex conjugation
-	average or normalized version of the variable
\wedge	complex cepstrum of the signal, if a function of time
\wedge	complex logarithm of the signal, if a function of frequency
\sim	estimate of the variable under the symbol
', ''	first and second derivatives of the preceding function
\forall	for all
\in	belongs to or is in (the set)
$ $	absolute value or magnitude of
$ \ $	norm
\angle	argument of, angle of
$\lceil \rceil$	ceiling
$\lfloor \rfloor$	floor
$<, >$ or \cdot	dot (inner) product
\odot	element-wise matrix multiplication
\oslash	element-wise matrix division

ABOUT THE COMPANION WEBSITE

This book is accompanied by a companion website

www.wiley.com/go/rangayyan3e



The website includes teaching and learning resources. Each chapter in the book includes a number of study questions and problems to facilitate preparation for tests and examinations. A number of laboratory exercises are also provided at the end of each chapter, which could be used to formulate hands-on exercises with real-life signals. Data files related to the problems and exercises at the end of each chapter are available on the companion website and the online repository

<https://github.com/srikrishnan1972/Biomedical-Signal-Analysis>

CHAPTER 1

INTRODUCTION TO BIOMEDICAL SIGNALS

1.1 The Nature of Biomedical Signals

Living organisms are made up of many component *systems* — the human body, for example, includes the nervous system, the cardiovascular system, and the musculoskeletal system, among others. Each system is made up of several subsystems that carry on many *physiological processes*. For example, the cardiac system performs the important task of rhythmic pumping of blood throughout the body to facilitate the delivery of nutrients, as well as pumping blood through the pulmonary system for oxygenation of the blood itself.

Physiological processes are complex phenomena, including nervous or hormonal stimulation and control; inputs and outputs that could be in the form of physical material, neurotransmitters, or information; and action that could be mechanical, electrical, or biochemical. Figure 1.1 shows a schematic representation of a generic physiological system to support the present discussion. Most physiological processes are accompanied by or manifest themselves as *signals* that reflect their nature and activities. Such signals could be of many types, including biochemical in the form of hormones and neurotransmitters, electrical in the form of potential or current, and physical in the form of pressure or temperature.

Diseases or defects in a biological system cause alterations in its normal physiological processes, leading to *pathological processes* that affect the performance, health, and general well-being of the system. A pathological process is typically associated with signals that are different in some aspects from the corresponding normal signals. If we possess a good understanding of a system of interest, it becomes possible to observe the corresponding signals and assess the state of the system. The task is not difficult when the signal is simple and appears at the outer surface of the body. For example, most infections cause a rise in the temperature of the body, which may be sensed easily, albeit in a

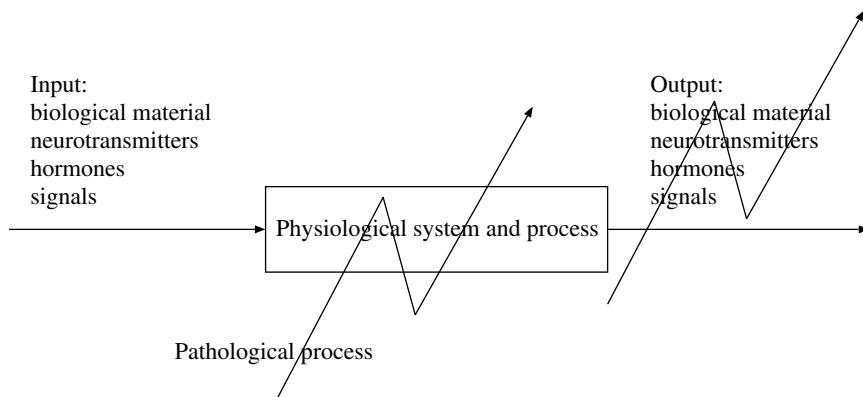


Figure 1.1 Schematic representation of a generic physiological system with various types of possible inputs and outputs. The effect of a pathological process is depicted by the zigzag line across the system and the list of possible outputs. When both the input and output are signals, the physiological system shown may be viewed as a typical signal processing system in electrical engineering.

relative and *qualitative* manner, via the palm of one's hand. Objective or *quantitative* measurement of temperature requires an instrument, such as a thermometer.

A single measurement x of temperature is a *scalar* and represents the thermal state of the body at a *particular or single instant of time t* and at a particular position. If we record the temperature continuously in some form, such as a magnetic tape, we obtain a *signal as a function of time*; such a signal may be expressed in *continuous-time* or *analog* form as $x(t)$. When the temperature is measured at *discrete* instants of time, it may be expressed in *discrete-time* form as $x(nT)$ or $x(n)$, where n is the index or measurement sample number of the array of values, and T represents the uniform interval between the time instants of measurement. A discrete-time signal that can take amplitude values only from a limited list of *quantized* levels is called a *digital signal*.

(*Note:* The sampling of an analog signal has important implications on the accuracy of its representation and analysis as a discrete-time signal [1–3]. Quantization of an analog signal introduces errors in its digital representation; the use of optimal quantizers minimizes such errors [2, 4, 5]. The use of digital filters and computers with short word lengths also introduces limitations and errors in digital signal processing (DSP) [2, 3]. The distinction between discrete-time and digital signals [3, 6] and their processing may not be important in most practical applications with currently available computers.)

In intensive-care monitoring, the tympanic (eardrum) temperature may sometimes be measured using an infrared sensor. Occasionally, when catheters are being used for other purposes, a temperature sensor may also be introduced into an artery or the heart to measure the *core* temperature of the body. It then becomes possible to obtain a continuous measurement of temperature, although only a few samples taken at intervals of a few minutes may be stored for subsequent analysis. Figure 1.2 illustrates representations of temperature measurements as a scalar, an array, and a signal plotted as a function of time. It is obvious that the graphical representation facilitates easier and faster comprehension of trends in the temperature than the numerical format. Long-term recordings of temperature can facilitate the analysis of temperature-regulation mechanisms [7, 8].

Let us now consider another basic measurement in healthcare and monitoring: blood pressure (BP). Each measurement consists of two values — the systolic pressure and the diastolic pressure. BP is measured in millimeters of mercury (*mm of Hg*) in clinical practice, although the international standard unit for pressure is the *Pascal*, with $1 \text{ Pa} = 0.0075 \text{ mm of Hg}$. A single BP measurement could thus be viewed as a *vector* $\mathbf{x} = [x_1, x_2]^T$ with two components: x_1 indicating the systolic pressure and x_2 indicating the diastolic pressure. When BP is measured at a few instants of time,

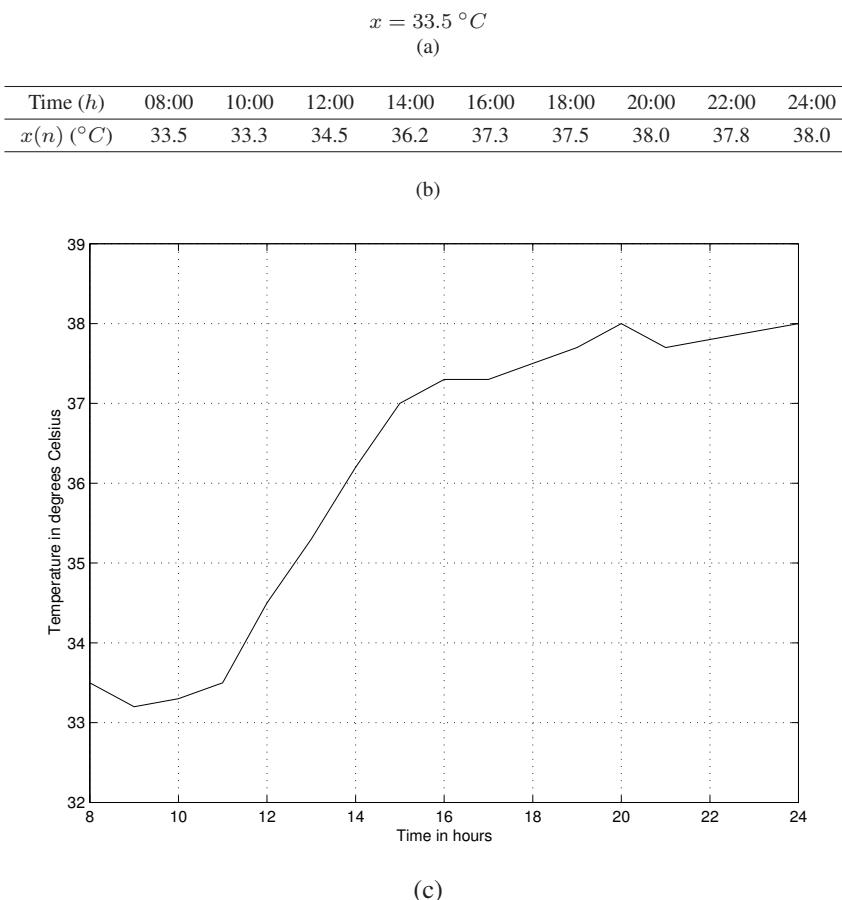


Figure 1.2 Measurements of the temperature of a patient presented as (a) a scalar with one temperature measurement x at an unspecified instant of time, (b) an array $x(n)$ made up of several measurements at different instants of time, and (c) a plot of the signal $x(n)$ or $x(t)$. The horizontal axis of the plot represents time in *hours*; the vertical axis gives temperature in *degrees Celsius*. Data courtesy of Foothills Hospital, Calgary.

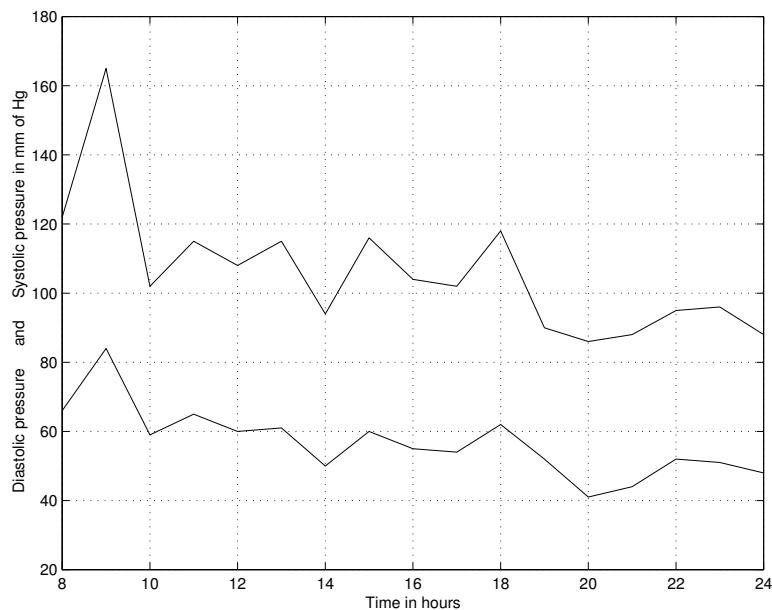
we obtain an array of vectors $\mathbf{x}(n)$. In intensive-care monitoring and surgical procedures, a pressure transducer may sometimes be inserted into a blood vessel along with other devices. It then becomes possible to obtain the arterial systolic and diastolic BP as a continuous-time recording, although the values may be transferred to a computer and stored only at sampled instants of time that are several seconds or minutes apart. (Note that, although the systolic and diastolic pressure values are quoted together for the same nominal time instant, they do not occur together and are separated by at least a part of a cardiac cycle corresponding to systole; in practice, their measurement may be separated by several cardiac cycles or a few seconds.) The signal may then be expressed as a function of time $\mathbf{x}(t)$. Figure 1.3 shows BP measurements as a single two-component vector, as an array of vectors, and as a plot as a function of time. It is clear that the plot as a function of time facilitates rapid observation of trends in the pressure.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \text{Systolic} \\ \text{Diastolic} \end{bmatrix} = \begin{bmatrix} 122 \\ 66 \end{bmatrix}$$

(a)

Time (h)	08:00	10:00	12:00	14:00	16:00	18:00	20:00	22:00	24:00
Systolic	122	102	108	94	104	118	86	95	88
Diastolic	66	59	60	50	55	62	41	52	48

(b)



(c)

Figure 1.3 Measurements of the BP of a patient presented as (a) a single pair or vector of systolic and diastolic measurements \mathbf{x} in mm of Hg at an unspecified instant of time, (b) an array $\mathbf{x}(n)$ made up of several measurements at different instants of time, and (c) a signal $\mathbf{x}(t)$ or $\mathbf{x}(n)$. Note the use of boldface \mathbf{x} to indicate that each measurement is a vector with two components. The horizontal axis of the plot represents time in hours; the vertical axis gives the systolic pressure (upper trace) and the diastolic pressure (lower trace) in mm of Hg . Data courtesy of Foothills Hospital, Calgary.

1.2 Examples of Biomedical Signals

The preceding example of body temperature as a signal is a simple example of a *biomedical signal*. Regardless of its simplicity, we can appreciate its importance and value in the assessment of the well-being of a child with a fever or that of a critically ill patient in a hospital. The origins and nature of a few other biomedical signals of various types are described in the following sections, with brief indications of their usefulness in diagnosis. Further detailed discussions on some of the signals are provided in the context of their analysis for various purposes in the chapters that follow.

This book is limited to the analysis of commonly encountered biomedical signals expressed as functions of time. Signals encountered in molecular biology and nanobioscience are not included. The topics of biomedical imaging and the analysis of biomedical images [9, 10] are not considered.

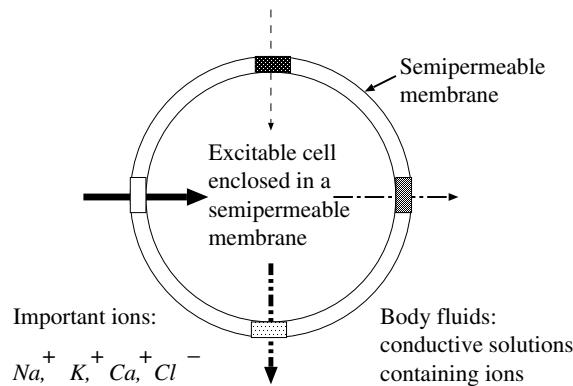
1.2.1 The action potential of a cardiac myocyte

The action potential is the basic component of all bioelectrical signals. It provides information on the nature of physiological activity at the single-cell level.

The action potential is the electrical signal that accompanies the mechanical contraction of a single muscle cell when stimulated by an electrical current (neural or external) [11–18]. The action potential is caused by the flow of sodium (Na^+), potassium (K^+), chloride (Cl^-), and other ions across the cell membrane.

Action potentials are also associated with signals and messages transmitted in the nervous system with no accompanying contraction. Hodgkin and Huxley [11, 12] conducted pioneering work on recording action potentials from a nerve fiber; see Sections 1.2.2 and 7.8.1. Recording an action potential requires the isolation of a single cell, and microelectrodes with tips of the order of a few micrometers to stimulate the cell and record the response [13].

Resting potential: A nerve or muscle cell is encased in a semipermeable membrane that permits selected substances to pass through while others are kept out. Body fluids surrounding cells are conductive solutions containing charged atoms known as ions. Figure 1.4 gives a schematic illustration of a cell and its characteristics.



Selective permeability: some ions can move in and out of the cell easily, whereas others cannot, depending upon the state of the cell and the voltage-gated ion channels.

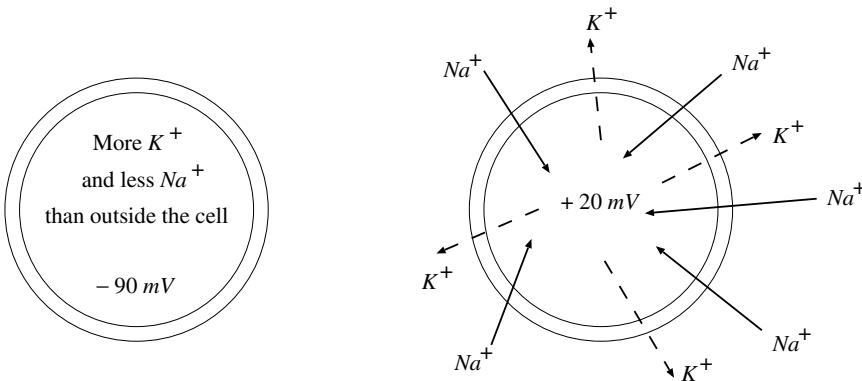
Figure 1.4 Schematic representation of a cell and its characteristics. The parts of the cell membrane with different shades and the corresponding arrows of different types and thickness represent, in a schematic manner, the variable permeability of the membrane to different ions.

In their resting state, the membranes of excitable cells readily permit the entry of K^+ and Cl^- ions, but effectively block the entry of Na^+ ions (the permeability for K^+ is 50–100 times that for Na^+). Various ions seek to establish a balance between the inside and the outside of a cell according to charge and concentration. The inability of Na^+ to penetrate a cell membrane results in the following [14]:

- Na^+ concentration inside the cell is far less than that outside.
- The outside of the cell is more positive than the inside of the cell.
- To balance the charge, additional K^+ ions enter the cell, causing higher K^+ concentration inside the cell than outside.

- Charge balance cannot be reached due to differences in membrane permeability for the various ions.
- A state of equilibrium is established with a potential difference, with the inside of the cell being negative with respect to the outside.

Figure 1.5 shows a schematic representation of a cell in its resting state. A cell in its resting state is said to be *polarized*. Most cells maintain a *resting potential* of the order of -60 to -100 mV until some disturbance or stimulus upsets the equilibrium.



(a) At rest: permeability for K^+
50 to 100 times that for Na^+ .
The cell is polarized.

(b) Depolarization: triggered by a stimulus;
fast Na^+ channels open.

Figure 1.5 Schematic representation (a) of a cell in its resting or polarized state and (b) the process of depolarization of the cell due to a stimulus.

Depolarization: When a cell is excited by ionic currents or an external stimulus, the membrane changes its characteristics and begins to allow Na^+ ions to enter the cell. This movement of Na^+ ions constitutes an ionic current, which further reduces the membrane barrier to Na^+ ions. This leads to an avalanche effect: Na^+ ions rush into the cell. K^+ ions try to leave the cell as they were in higher concentration inside the cell in the preceding resting state, but cannot move as fast as the Na^+ ions. The net result is that the inside of the cell becomes positive with respect to the outside due to an imbalance of K^+ ions. A new state of equilibrium is reached after the rush of Na^+ ions stops. This change represents the beginning of the *action potential*, with a peak value of about +20 mV for most cells. An excited cell displaying an action potential is said to be *depolarized*; the process is called *depolarization*. Figure 1.5 shows a schematic representation of a cell undergoing the process of depolarization.

Repolarization: After a certain period of being in the depolarized state the cell becomes polarized again and returns to its resting potential via a process known as *repolarization*. Repolarization occurs by processes that are analogous to those of depolarization, except that instead of Na^+ ions, the principal ions involved in repolarization are K^+ ions [16]. Membrane depolarization, while increasing the permeability for Na^+ ions, also increases the permeability of the membrane for K^+ ions via a specific class of ion channels known as voltage-dependent K^+ channels. Although this may appear to be paradoxical at first glance, the key to the mechanism for repolarization lies in the time-dependence and voltage-dependence of the membrane permeability changes for K^+ ions compared with those for Na^+ ions. The permeability changes for K^+ during depolarization occur considerably more slowly than those for Na^+ ions; hence, the initial depolarization is caused

by an inrush of Na^+ ions. However, the membrane permeability changes for Na^+ spontaneously decrease near the peak of the depolarization process, while those for K^+ ions are beginning to increase. Hence, during repolarization, the predominant membrane permeability is for K^+ ions. Because K^+ concentration is much higher inside the cell than outside, there is a net efflux of K^+ from the cell, which makes the inside more negative, thereby effecting repolarization back to the resting potential.

It should be noted that the voltage-dependent K^+ permeability change is due to a distinctly different class of ion channels than those that are responsible for setting the resting potential. A mechanism known as the $Na^+ - K^+$ pump extrudes Na^+ ions in exchange for transporting K^+ ions back into the cell. However, this transport mechanism carries very little current in comparison with ion channels, and therefore, makes a minor contribution to the repolarization process. The $Na^+ - K^+$ pump is essential to reset the $Na^+ - K^+$ balance of the cell, but the process occurs on a longer time scale than the duration of an action potential.

Nerve and muscle cells repolarize rapidly, with an action potential duration of about 1 – 5 ms. Heart muscle cells repolarize slowly, with an action potential duration of 150 – 300 ms.

The action potential is always the same for a given cell, regardless of the method of excitation or the intensity of the stimulus beyond a threshold: This is known as the *all-or-none* or all-or-nothing phenomenon. After an action potential, there is a period during which a cell cannot respond to any new stimulus, known as the *absolute refractory period* (about 1 ms in nerve cells [19]). This is followed by a *relative refractory period* (3 – 5 ms in nerve cells), when another action potential may be triggered by a much stronger stimulus than in the normal situation [19]. Figure 1.6 shows the various phases or intervals of the action potential of a cardiac (ventricular) myocyte.

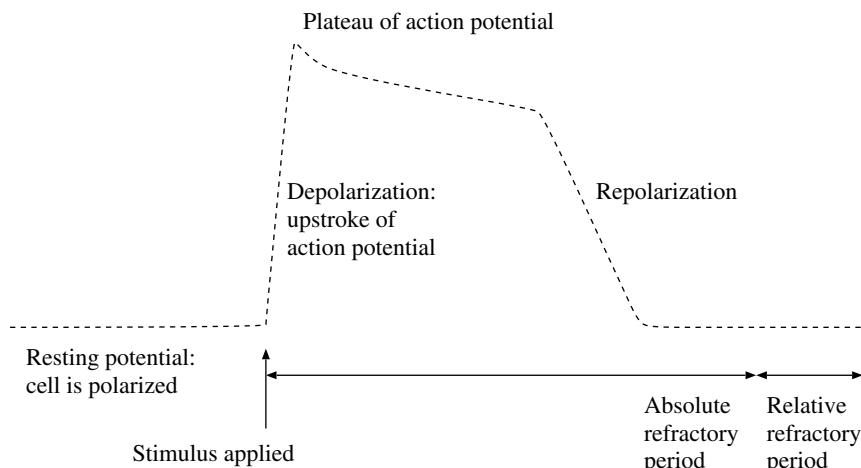


Figure 1.6 Illustration of the various phases or intervals of the action potential of a cardiac (ventricular) myocyte. The absolute refractory period includes the duration of the action potential.

Figure 1.7 shows action potentials recorded from individual rabbit ventricular and atrial myocytes (muscle cells) [16]. Figure 1.8 shows a ventricular myocyte in its relaxed and fully contracted states. The tissues were first incubated in digestive enzymes, principally collagenase, and then dispersed into single cells using gentle mechanical agitation. The recording electrodes were glass patch pipettes; a whole-cell, current-clamp recording configuration was used to obtain the action potentials. The cells were stimulated at low rates (once per 8 s); this is far less than physiological rates. Moreover, the cells were maintained at 20 °C, rather than at body temperature. Nevertheless, the major features of the action potentials shown are similar to those recorded under physiological conditions.

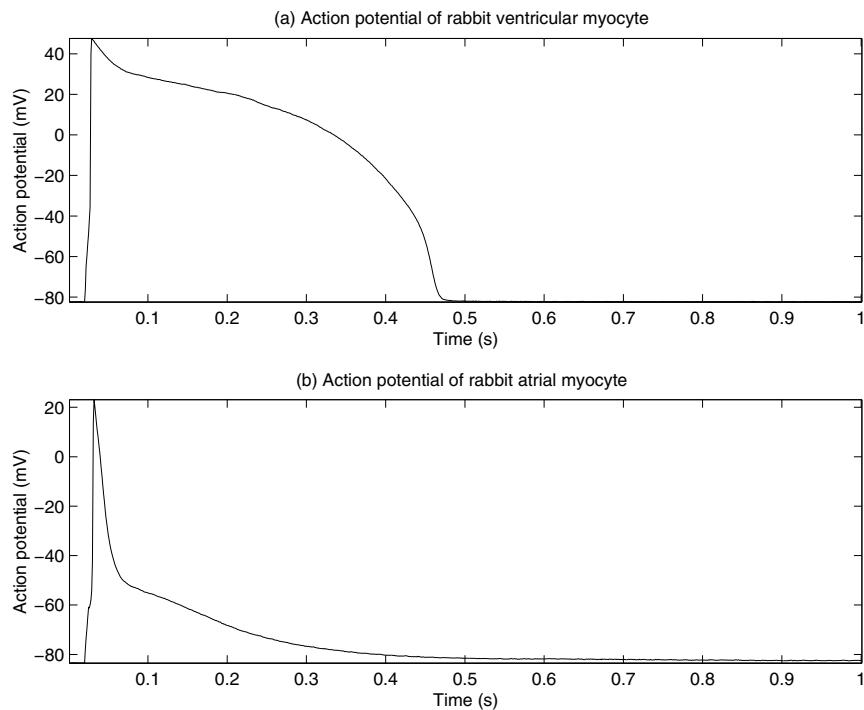


Figure 1.7 Action potentials of rabbit ventricular and atrial myocytes. Data courtesy of R. Clark, Department of Physiology and Biophysics, University of Calgary.

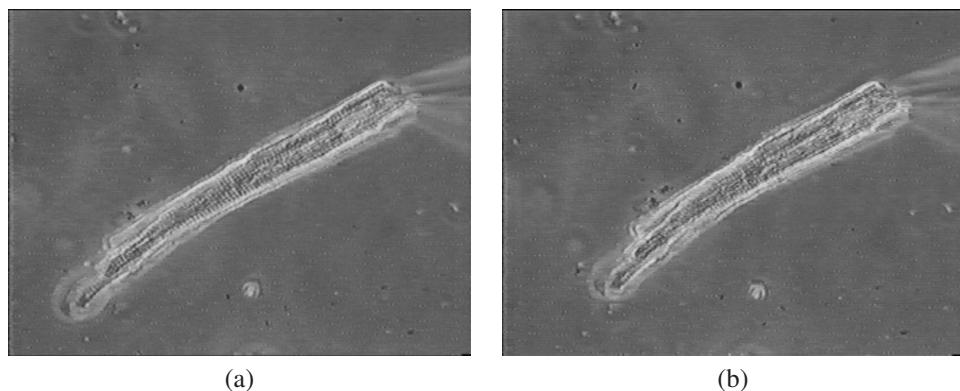


Figure 1.8 A single ventricular myocyte of a rabbit in its (a) relaxed state and (b) fully contracted state. The length of the myocyte is approximately $25 \mu\text{m}$. The tip of the glass pipette, faintly visible at the upper-right end of the myocyte, is approximately $2 \mu\text{m}$ wide. Images courtesy of R. Clark, Department of Physiology and Biophysics, University of Calgary.

The resting membrane potential of the cells (from 0 to 20 ms in the plots in Figure 1.7) is about -83 mV . A square pulse of current, 3 ms in duration and 1 nA in amplitude, was passed through the recording electrode and across the cell membrane, causing the cell to depolarize rapidly. The ventricular myocyte exhibits a depolarized potential of about $+40\text{ mV}$; it then slowly declines back to the resting potential level over an interval of about 500 ms. The initial rapid depolarization of the atrial cell is similar to that of the ventricular cell, but does not overshoot zero membrane potential as much as the ventricular action potential; repolarization occurs much more quickly than the case for the ventricular cell.

Figure 1.9 shows a schematic representation of a cell as a system. When a stimulus is applied, the cell provides a response or output. Depending on the nature of the cell, such as a nerve cell or a muscle cell, the output could be an action potential, a twitch, or contraction (change in length). When several muscle cells are stimulated, the net result is the development of force.

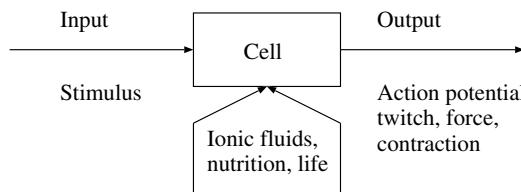


Figure 1.9 Schematic representation of a cell as a system. Upon receiving an input of a stimulus, the cell provides a response that could cause an action potential, a twitch, contraction, or force. The term “life” indicates that the cell must be alive naturally, or maintained so under laboratory conditions, in order to function. Ionic fluids and nutrition provide the support and environment that are essential for the cell.

See Section 7.8 for further discussions on this subject.

1.2.2 The action potential of a neuron

The neuron is one of the basic units or structures of the nervous system. The basic function of a neuron may be considered to be information processing and transmission. Figure 1.10 shows a schematic representation of a neuron and its constituent parts. A neuron receives inputs from other neurons or cells through its dendrites. When adequate inputs or stimuli are received, the neuron is triggered and generates an action potential. The signal generated by the neuron is communicated to other neurons or cells through its axon. A few examples of the functions of networks of neurons are: collection of signals from sensors, such as the eyes and ears, followed by coding and transmission to the brain; processing, analysis, and interpretation of signals and information received in the brain; and communication of activation or control signals from the brain and spinal cord to muscles and other systems.

In studies of the central nervous system (CNS), it is desirable to record the action potentials of isolated neurons *in situ*. Hodgkin and Huxley [11, 12] conducted pioneering studies on recording action potentials from the giant axon of the squid; they proposed mathematical and electrical circuit models for the generation of action potentials. Figure 1.11 shows the first recording of the action potential of a neuron published by Hodgkin and Huxley [11]. While the action potential demonstrates the upstroke of depolarization and the return to resting potential via repolarization, the shape and duration are different from those of the action potentials of the cardiac myocytes shown in Figure 1.7. Hodgkin and Huxley also noted variations in the characteristics of the action potentials of neurons due to temperature. See Section 7.8.1 for further related discussions.

Drake et al. [20] described the design and performance of a multisite microprobe system to record isolated and interrelated neuronal activity *in vivo*. Figure 1.12 shows a sample recording obtained from a rat’s brain (cerebral cortex) using the microprobe. It was observed that the waveforms and

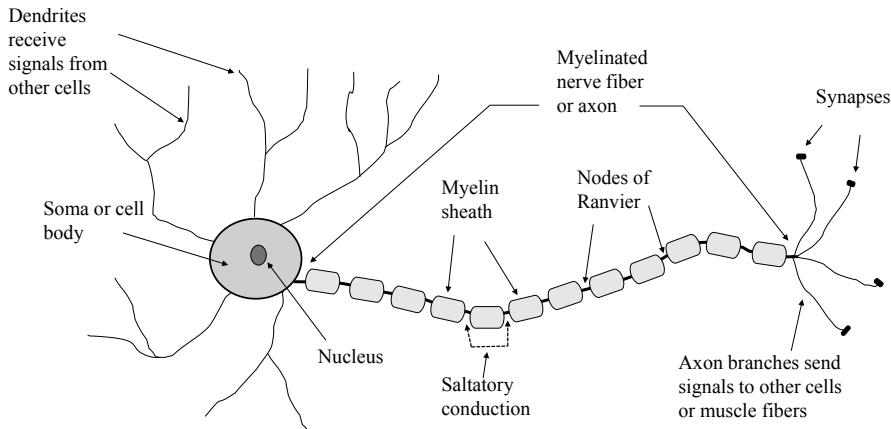


Figure 1.10 Schematic representation of a neuron.

durations of neuronal action potentials recorded with the microprobe were similar to those measured using single-needle microelectrodes [21]. Note that the durations of the neuronal action potentials are much shorter than those of cardiac myocytes. The study demonstrated the possibility of observing simultaneous spatiotemporal patterns of extracellular single-unit action potentials, which could be used to investigate the columnar organization and layering of cortical tissue. Furthermore, it was indicated that the microprobe could be useful in the exploration of the characteristics of extracellular fields around an active neuron.

Propagation of an action potential: An action potential propagates along a muscle fiber or an unmyelinated nerve fiber as explained in the following sentences [22]. Once initiated by a stimulus, the action potential propagates along the whole length of a fiber without decrease in amplitude by progressive depolarization of the membrane. Current flows from a depolarized region through the intracellular fluid to adjacent inactive regions, thereby depolarizing them. Current also flows through the extracellular fluids, through the depolarized membrane, and back into the intracellular space, completing the local circuit. The energy to maintain conduction is supplied by the fiber itself.

Myelinated nerve fibers are covered by an insulating sheath of *myelin*; see Figure 1.10. The sheath is interrupted every few millimeters by spaces known as the *nodes of Ranvier*, where the fiber is exposed to the interstitial fluid. Sites of excitation and changes of membrane permeability exist only at the nodes, and current flows by jumping from one node to the next in a process known as *saltatory conduction*.

1.2.3 The electroneurogram (ENG)

The ENG is an electrical signal observed as a stimulus and the associated action potential propagate over the length of a nerve. It may be used to measure the velocity of propagation (or conduction velocity) of a stimulus or action potential in a nerve [13, 23, 24]. ENGs may be recorded using concentric needle electrodes or silver–silver-chloride electrodes ($Ag - AgCl$) at the surface of the body.

The conduction velocity in a peripheral nerve may be measured by stimulating a motor nerve and measuring the related activity at two points that are a known distance apart along its course. In order to minimize muscle contraction and other undesired effects, the experimental limb is held in

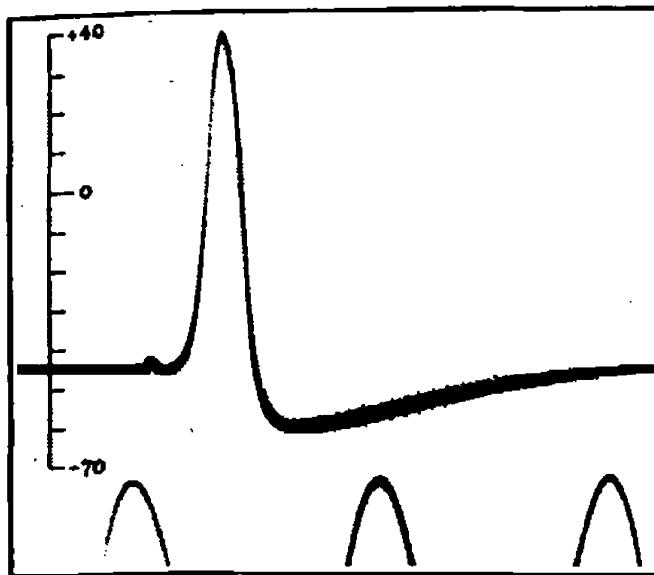


Figure 1.11 First known tracing of an action potential recorded from the axon of a squid by Hodgkin and Huxley [11]. The vertical axis is in the range -70 to $+40$ mV and the time calibration waveform shown at the bottom is a sinusoid of frequency 500 Hz . Reproduced with permission from A.L. Hodgkin and A.F. Huxley. Action potentials recorded from inside a nerve fibre, *Nature*, 144:710–711, 1939. ©Springer Nature.

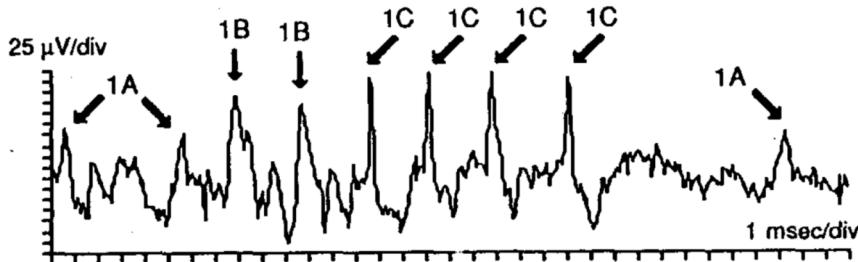


Figure 1.12 Neuronal action potentials recorded from a rat's brain using a multisite microprobe system. The action potentials marked 1A have low amplitude, just above the baseline neural noise level. The action potentials marked 1B possess broader waveforms and could possibly be composed of two superimposed action potentials. The action potentials marked 1C are clearly isolated, have the same amplitude, and are likely from the same cell. Reproduced with permission from K.L. Drake, K.D. Wise, J. Farraye, D.J. Anderson, and S.L. BeMent, Performance of planar multisite micropoles in recording extracellular single-unit intracortical activity, *IEEE Transactions on Biomedical Engineering*, 35(9):719–732, 1988. ©IEEE.

a relaxed posture and a strong but short stimulus is applied in the form of a pulse of about 100 V amplitude and 100 – 300 μs duration [13, 23, 24]. The difference in the latencies of the ENGs recorded over the associated muscle gives the conduction time. Knowing the separation distance between the stimulating and recording sites, it is possible to determine the conduction velocity in the nerve [13, 23, 24]. ENGs have amplitudes of the order of 10 μV and are susceptible to power-line interference and instrumentation noise.

Figure 1.13 illustrates the ENGs recorded in a study of nerve conduction velocity. The stimulus was applied to the ulnar nerve near the wrist. The ENGs were recorded at the wrist (marked “Wrist” in the figure), just below the elbow (BEElbow), and just above the elbow (AEElbow) using surface

electrodes, amplified with a gain of 2,000, and filtered to the bandwidth $10 - 10,000\text{ Hz}$. The three traces in the figure indicate increasing latencies with respect to the stimulus time point, which is at the left-hand margin of the plots. The responses shown in the figure are normal, indicate a BElbow–Wrist latency of 3.23 ms , and result in a nerve conduction velocity of 64.9 m/s .

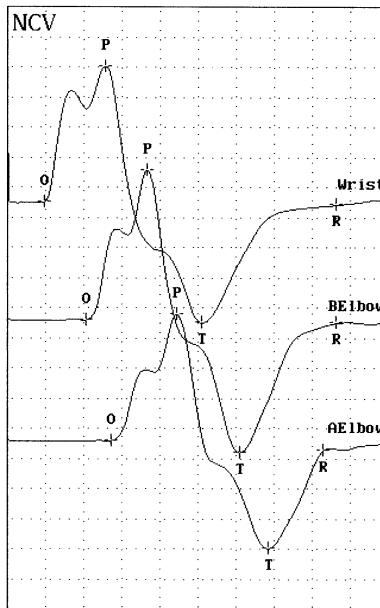


Figure 1.13 Nerve conduction velocity (NCV) measurement via electrical stimulation of the ulnar nerve. The grid boxes represent 3 ms in width and $2\text{ }\mu\text{V}$ in height. AEElbow: above the elbow. BEElbow: below the elbow. O: onset. P: peak. T: trough. R: recovery of baseline. Courtesy of M. Wilson and C. Adams, Alberta Children's Hospital, Calgary.

Typical values of propagation rate or nerve conduction velocity are [13, 22, 25]:

- $45 - 70\text{ m/s}$ in nerve fibers;
- $0.2 - 0.4\text{ m/s}$ in heart muscle;
- $0.03 - 0.05\text{ m/s}$ in the time-delay fibers between the atria and the ventricles (AV node).

Neural diseases may cause a decrease in conduction velocity.

1.2.4 The electromyogram (EMG)

Skeletal muscle fibers are considered to be twitch fibers because they produce a mechanical twitch response for a single stimulus and generate a propagated action potential. Skeletal muscles are made up of collections of *motor units*, each of which consists of an anterior horn cell (motoneuron or motor neuron), its axon, and all muscle fibers innervated by that axon [26–28]. A motor unit is the smallest muscle unit that can be activated by volitional effort. The constituent fibers of a motor unit are activated synchronously. Component fibers of a motor unit extend lengthwise in loose bundles along the muscle. In cross-section, the fibers of a given motor unit are interspersed with the fibers of other motor units [13, 22, 29]. Figure 1.14 shows a schematic representation of two motor units. Figure 1.15 (top panel) illustrates a motor unit in schematic form along with related models [29]. When stimulated by a neural signal, each motor unit contracts and causes an electrical

signal that is the summation of the action potentials of all of its constituent cells. This is known as the *single-motor-unit action potential* (SMUAP, or simply MUAP) and may be recorded using needle electrodes inserted into the muscle or region of interest.

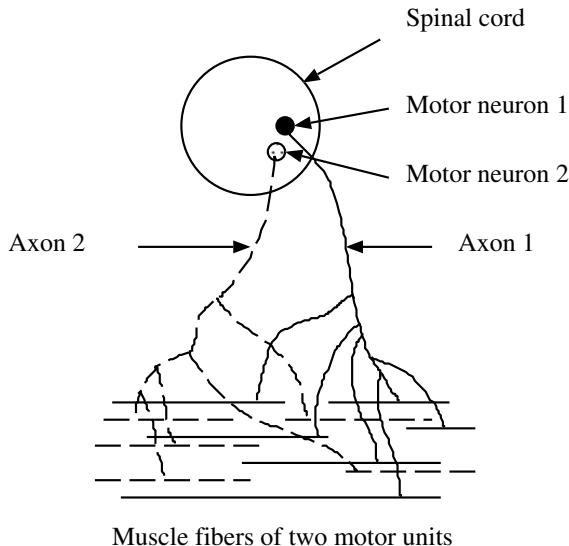


Figure 1.14 Schematic representation of two motor units, one in solid line and the other in dashed line.

Muscles for gross movement or large force have hundreds of muscle fibers per motor unit; muscles for precise or fine movement have fewer fibers per motor unit. The number of muscle fibers per motor nerve fiber (or motor unit) is known as the *innervation ratio*. The mechanical output (contraction) of a muscle is the net result of stimulation and contraction of several of its motor units.

Various experiments have been conducted to estimate the innervation ratios of several muscles in animals and human beings, *in vivo* and using autopsy samples, with some variations in the methods and definitions used. Goodgold and Eberstein [22] and Kimura [19] provide tables including the following examples. The platysma (a large sheet-like muscle spanning parts of the pectoral muscle, deltoid, clavicle, and neck) has 1,826 large nerve fibers controlling 27,100 muscle fibers in 1,096 motor units, leading to an estimate of about 25 muscle fibers per motor unit. The first dorsal interosseous (FDI) muscle (on the back of the palm of the hand and the index finger) has 199 large nerve fibers and 40,500 muscle fibers in 119 motor units, with about 340 muscle fibers per motor unit. The medial gastrocnemius (calf muscle of the leg) has 965 large nerve fibers and 1,120,000 muscle fibers in 579 motor units, with about 1,934 muscle fibers per motor unit.

Laryngeal muscles have been estimated to have only two or three muscle fibers per motor unit [22]. Gath and Stålberg [30] used a multielectrode probe for *in situ* measurement of the innervation ratios of human muscles; they estimated the number of muscle fibers per motor unit to be 72 in the brachial biceps, 70 in the deltoid, and 124 in the tibialis anterior. Brown and Harvey [27] studied the muscles of cats' eyes and noted that not more than 10 muscle fibers are supplied by a single nerve fiber in the extrinsic ocular muscles. See Goodgold and Eberstein [22], Kimura [19], Buchthal and Schmalbruch [26], and Brown et al. [28] for related discussions.

Normal SMUAPs are usually biphasic or triphasic, 3 – 15 ms in duration, 100 – 300 μ V in amplitude, and appear with frequency in the range of 6 – 30/s [13, 22]. Schematic representations of monophasic, biphasic, and triphasic waveforms are shown in Figure 1.16. The shape of a recorded SMUAP depends on the type of the needle electrode used, its position with respect to the active

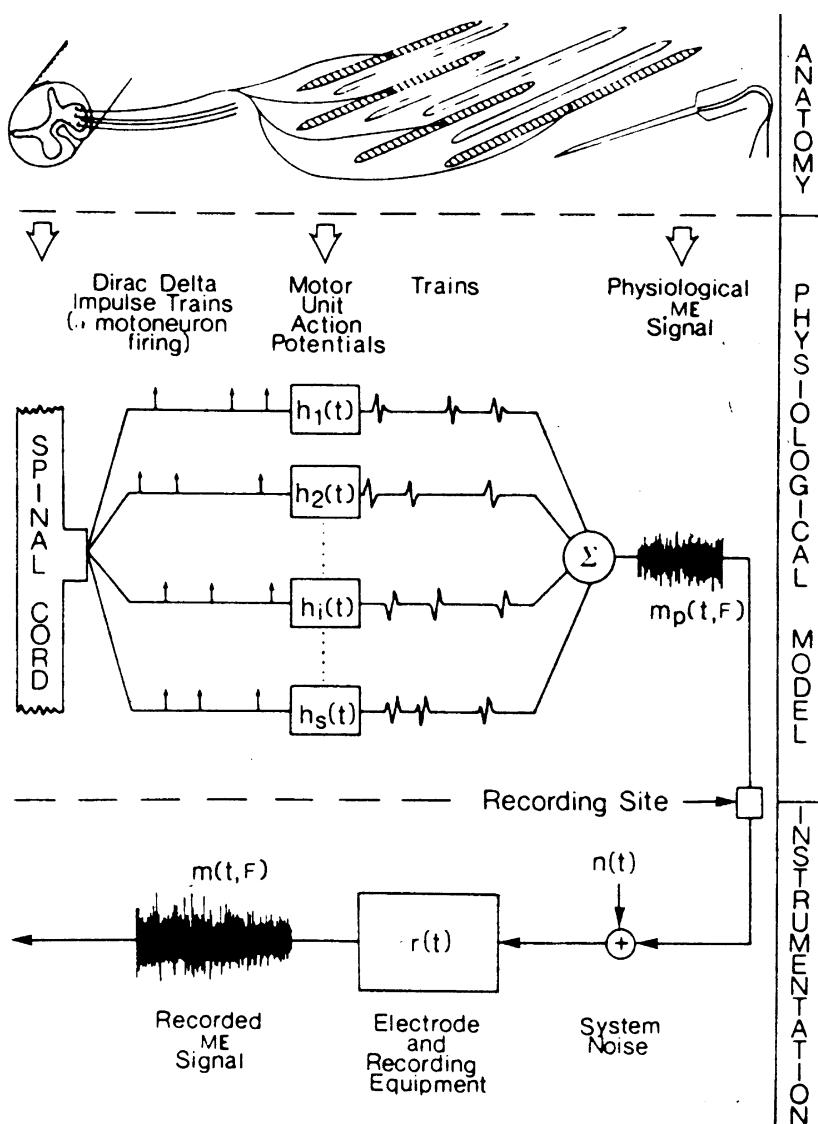


Figure 1.15 Schematic representation of a motor unit and model for the generation of EMG signals. Top panel: A motor unit includes an anterior horn cell or motor neuron (illustrated in a cross-section of the spinal cord), an axon, and several connected muscle fibers. The hatched fibers belong to one motor unit; the nonhatched fibers belong to other motor units. A needle electrode is also illustrated. Middle panel: The firing pattern of each motor neuron is represented by an impulse train. Each system $h_i(t)$ shown represents a motor unit that is activated and generates a train of SMUAPs. The net EMG is the sum of several SMUAP trains. Bottom panel: Effects of instrumentation on the EMG signal acquired. The observed EMG is a function of time t and muscular force produced F . ME: myoelectric. Reproduced with permission from C.J. de Luca, Physiology and mathematics of myoelectric signals, *IEEE Transactions on Biomedical Engineering*, 26:313–325, 1979. ©IEEE.

motor unit, and the projection of the electrical field of the activity on to the electrodes. Figure 1.17 illustrates simultaneous recordings of the activities of a few motor units from three channels of needle electrodes [31]. Although the SMUAPs are biphasic or triphasic, the same SMUAP displays variable shape from one channel to another. (*Note:* The action potentials in Figure 1.7 are monophasic; the first two SMUAPs in Channel 1 in Figure 1.17 are biphasic, and the third SMUAP in the same signal is triphasic.)

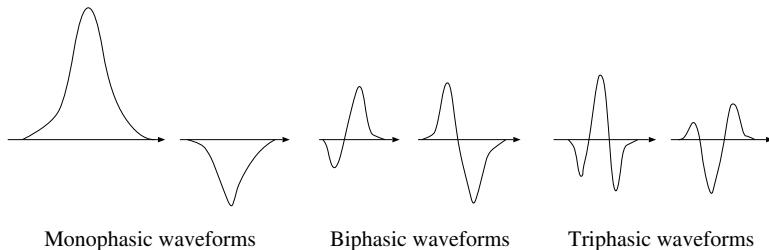


Figure 1.16 Schematic representations of monophasic, biphasic, and triphasic waveforms.

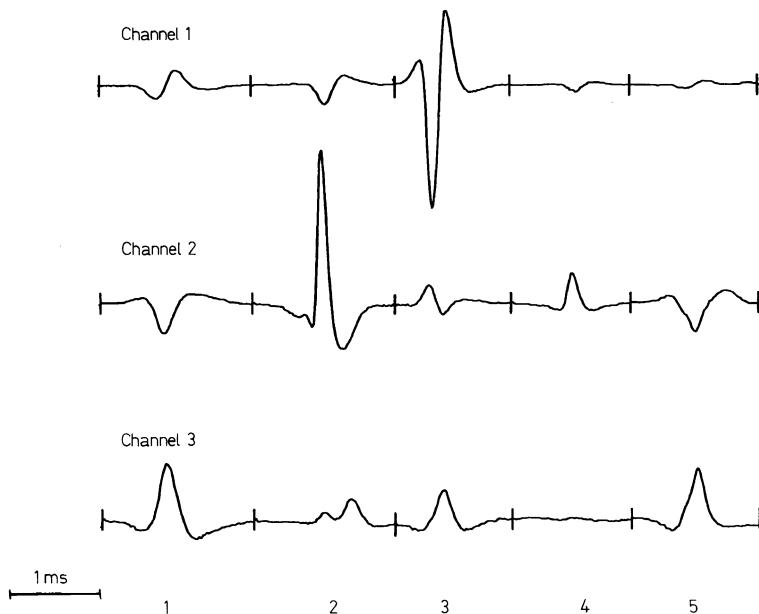


Figure 1.17 SMUAP trains recorded simultaneously from three channels of needle electrodes. Observe the different shapes of the same SMUAPs (aligned vertically across the three channels) projected on to the axes of the three channels. Three different motor units are active over the duration of the signals illustrated. Reproduced with permission from B. Mambrito and C.J. de Luca, Acquisition and decomposition of the EMG signal, in *Progress in Clinical Neurophysiology*, Volume 10: *Computer-aided Electromyography*, Editor: J.E. Desmedt, pp 52–72, 1983. ©S. Karger AG, Basel, Switzerland.

The shape of SMUAPs is affected by disease. Figure 1.18 illustrates SMUAP trains of a normal subject and those of patients with neuropathy and myopathy. Neuropathy causes slow conduction and/or desynchronized activation of the fibers within a motor unit, and a polyphasic SMUAP with an amplitude that is larger than normal. The same motor unit may be observed to fire at higher rates than normal before more motor units are recruited. Myopathy involves loss of muscle fibers

in motor units, with the related motoneuron and nerves presumably intact. *Splintering* of SMUAPs occurs due to asynchrony in activation as a result of patchy destruction of fibers (such as the case in muscular dystrophy), leading to polyphasic SMUAPs. More motor units may be observed to be recruited at low levels of effort.

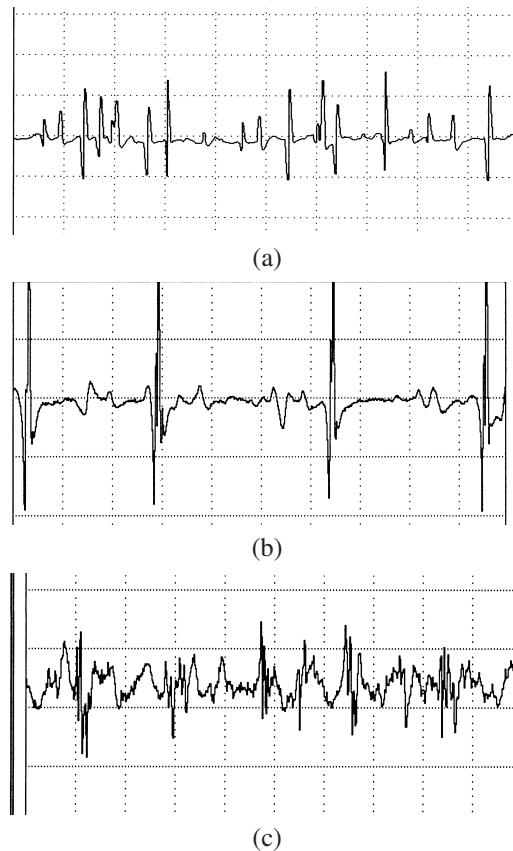


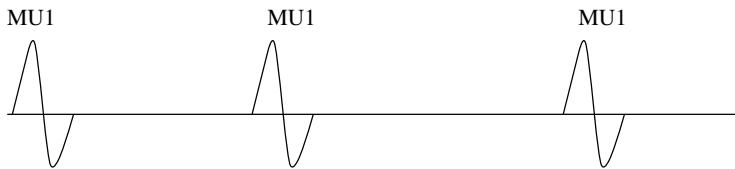
Figure 1.18 Examples of SMUAP trains. (a) From the right deltoid of a normal subject, male, 11 years; the SMUAPs are mostly biphasic, with duration in the range 3 – 5 ms. (b) From the deltoid of a six-month-old male patient with brachial plexus injury (neuropathy); the SMUAPs are polyphasic and large in amplitude (800 μ V), and the same motor unit is firing at a relatively high rate at low-to-medium levels of effort. (c) From the right biceps of a 17-year-old male patient with myopathy; the SMUAPs are polyphasic and indicate early recruitment of more motor units at a low level of effort. The signals were recorded with gauge 20 needle electrodes. The width of each grid box represents a duration of 20 ms; its height represents an amplitude of 200 μ V. Courtesy of M. Wilson and C. Adams, Alberta Children's Hospital, Calgary.

Gradation of muscular contraction: Muscular contraction levels are controlled in two ways:

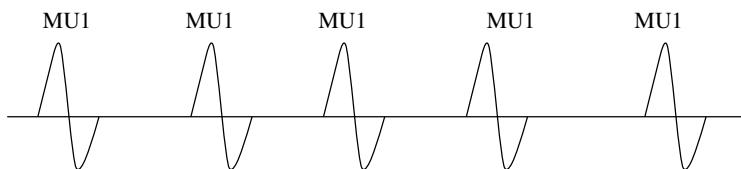
- *Spatial recruitment*, by activating new motor units with increasing effort; and
- *Temporal recruitment*, by increasing the frequency of discharge (firing rate) of each motor unit with increasing effort.

Figure 1.19 gives a schematic illustration of spatiotemporal recruitment of motor units and the related EMG signals. Motor units are activated at different times and at different frequencies causing asynchronous contraction. The twitches of individual motor units sum and fuse to form tetanic contraction and increased force. Weak volitional effort causes motor units to fire at about 5 – 15 pps

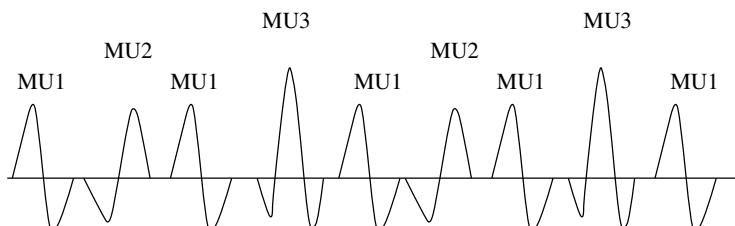
(pulses per second). As greater tension is developed, an *interference pattern* EMG is obtained, with the constituent and active motor units firing in the range of 25 – 50 pps. Grouping of MUAPs has been observed as fatigue develops, leading to decreased high-frequency content and increased amplitude in the EMG [29].



(a) At the beginning with low effort, only motor unit MU1 is firing at a low rate.



(b) At a slightly higher level of effort, with temporal recruitment, the firing rate of MU1 is increased. No other motor unit has been recruited yet.



(c) At an even higher level of effort, with spatial recruitment, new motor units MU2 and MU3 have been brought into action. MU1 continues to fire at the same rate as in (b).

Figure 1.19 Schematic representation of spatiotemporal recruitment of motor units and the resulting EMG signals. To keep the illustration simple, it is assumed that the MUAPs do not overlap.

Spatiotemporal summation of the MUAPs of all of the active motor units gives rise to the EMG of the muscle; see Figure 1.19. EMG signals recorded using surface electrodes are complex signals including interference patterns of several MUAP trains and are difficult to analyze. An EMG signal indicates the level of activity of a muscle, and may be used to diagnose neuromuscular diseases such as neuropathy and myopathy.

Figure 1.20 illustrates an EMG signal recorded from the crural diaphragm of a dog using fine-wire electrodes sewn in-line with the muscle fibers and placed 10 mm apart [32]. The signal represents one period of breathing (inhalation being the active part as far as the muscle and EMG are concerned). It is seen that the overall level of activity in the signal increases during the initial phase of inhalation. Figure 1.21 shows the early parts of the same signal on an expanded time scale. SMUAPs are seen at the beginning stages of contraction, followed by increasingly complex interference patterns of several MUAPs.

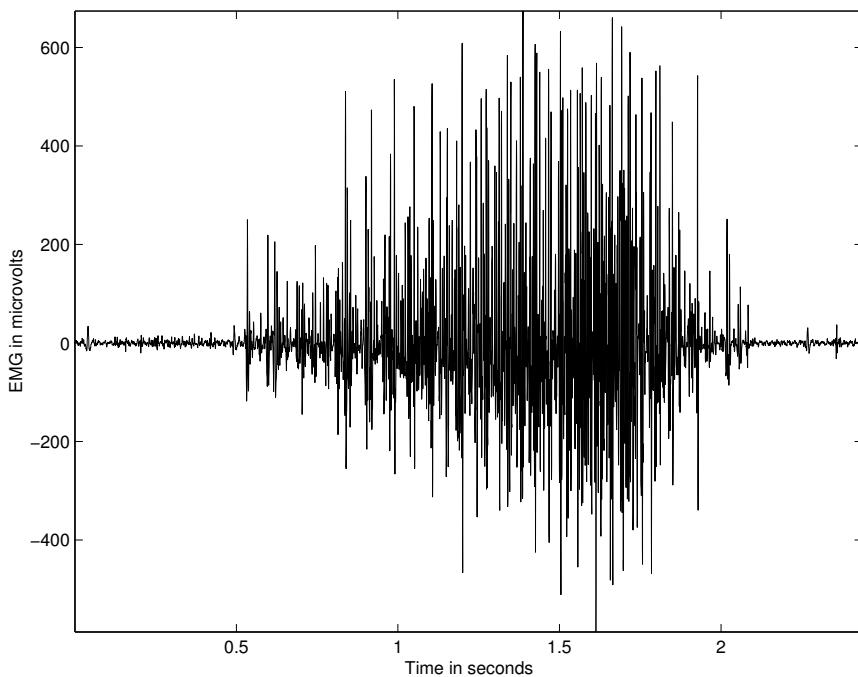


Figure 1.20 EMG signal recorded from the crural diaphragm muscle of a dog using implanted fine-wire electrodes; see also Figure 1.21. Data courtesy of R.S. Platt and P.A. Easton, Department of Clinical Neurosciences, University of Calgary.

Figure 1.22 shows an EMG signal recorded using surface electrodes placed on the forearm of a subject. In this experiment, designed to study the variation of the EMG signal with respect to the force exerted by a muscle, the subject performed a series of contractions using a gripping device equipped with a force transducer. (The output of the transducer was not calibrated in units of force.) The subject was instructed to squeeze the gripping device to the maximum extent possible; the corresponding output was noted as the level of maximal voluntary contraction (MVC). The subject was then asked to relax the muscle, and squeeze the device again in five steps, between approximately 20% and 100% MVC, each contraction lasting for about 2 – 3 s, with resting periods of about 2 – 3 s between each contraction. The EMG and force signals were lowpass filtered with the cutoff frequency at 1 kHz, highpass filtered with the cutoff frequency at 10 Hz, and recorded at the sampling rate of 2 kHz per channel. In the plot shown in Figure 1.22, the force signal has been normalized such that the minimum is zero and the maximum is 100; thus, the force exerted is expressed in %MVC. It is evident that the dynamic range (peak-to-peak swing) and power of the EMG signal increase as the level of force exerted increases. Figure 1.23 shows an expanded view of the period 10 – 12 s of the force and EMG signals, where the increasing trend of the EMG signal with increasing force is clearly seen.

Signal processing techniques for the analysis of EMG signals are described in Sections 5.2.4, 5.6, 5.9, 5.10, 5.11, 5.13.4, 7.2.1, and 7.3. See Nikolic and Krarup [33] for the description of methods for decomposition of EMG signals into their constituent MUAPs and firing patterns for quantitative analysis. EMG signals are useful in the control of prosthetic devices [34, 35].

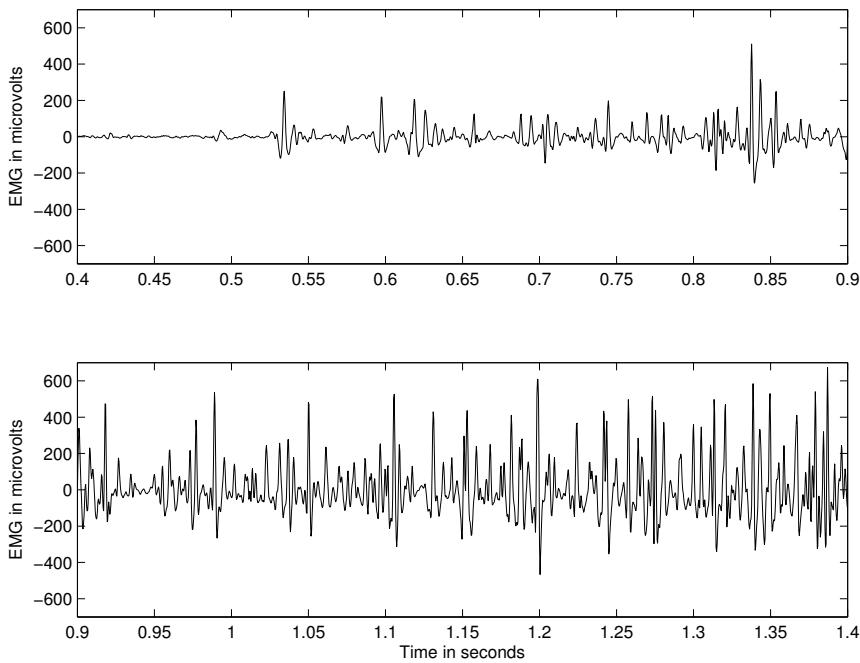


Figure 1.21 The initial part of the EMG signal in Figure 1.20 shown on an expanded time scale. Observe the SMUAPs at the initial stages of contraction, followed by increasingly complex interference patterns of several MUAPs. Data courtesy of R.S. Platt and P.A. Easton, Department of Clinical Neurosciences, University of Calgary.

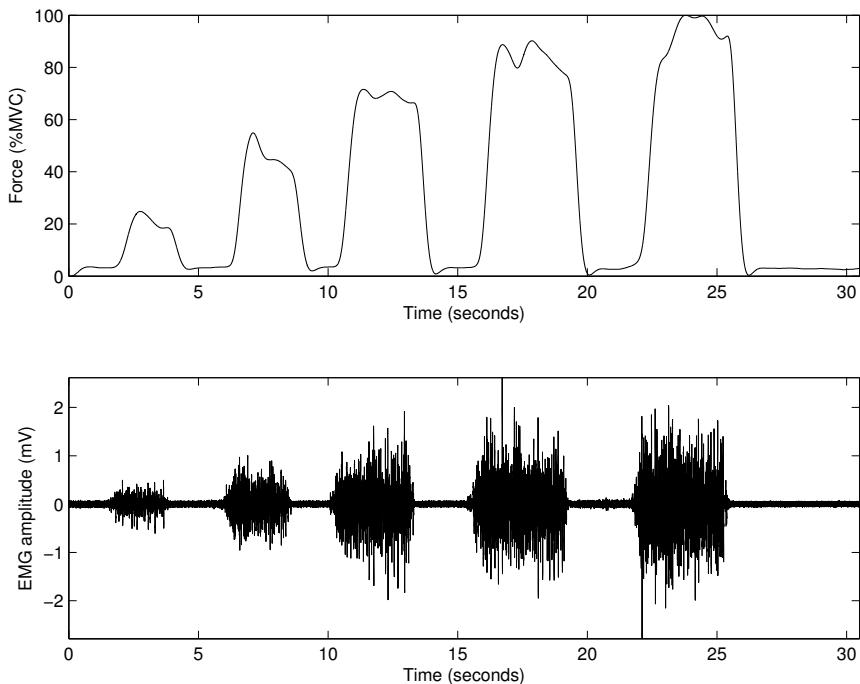


Figure 1.22 Force signal (upper plot) and the EMG signal (lower plot) recorded from the forearm muscle of a subject using surface electrodes; see also Figure 1.23. Data courtesy of Shantanu Banik.

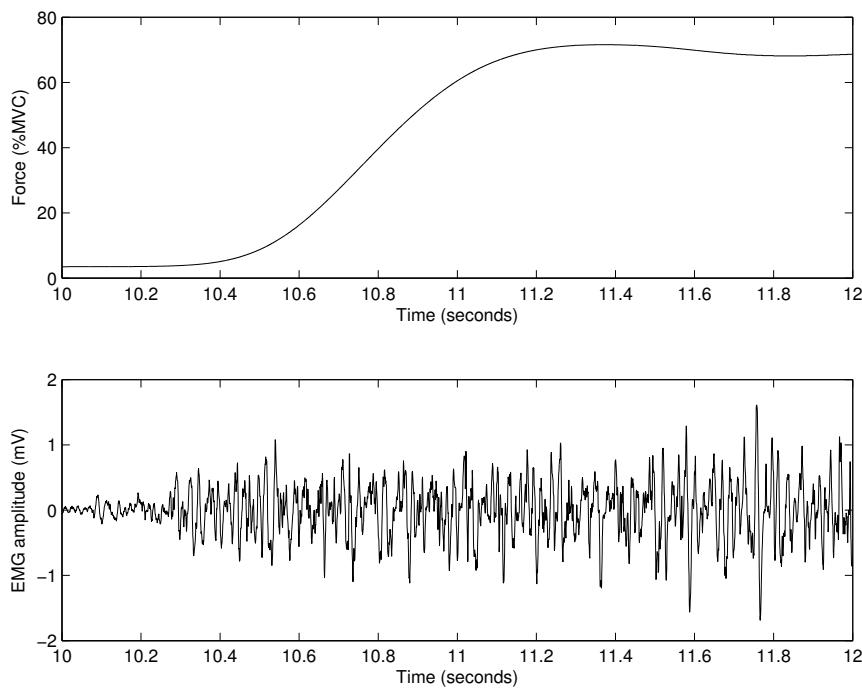


Figure 1.23 Expanded view of the part of the EMG (lower plot) and force signals (upper plot) in Figure 1.22 over the period 10 – 12 s. Observe the increasing levels of the range and power of the EMG signal at the initial stages of contraction. Data courtesy of Shantanu Banik.

1.2.5 The electrocardiogram (ECG)

The ECG is the electrical manifestation of the contractile activity of the heart and can be recorded fairly easily with surface electrodes on the limbs or chest. The ECG is the most commonly known, recognized, and used biomedical signal. Original investigations on recording of the human ECG were conducted by Waller [36] and Einthoven [37] in the late 1800s and the early 1900s.

The rhythm of the heart in terms of beats per minute (*bpm*) may be easily estimated by counting the readily identifiable waves in the ECG signal. More important is the fact that the ECG weshape is altered by cardiovascular diseases and abnormalities, such as myocardial ischemia, myocardial infarction, ventricular hypertrophy, and conduction problems.

The heart: The heart is a four-chambered pump with two atria for collection of blood and two ventricles to pump out blood. Figure 1.24 shows a schematic representation of the four chambers and the major vessels connecting to the heart. The resting or filling phase of a cardiac chamber is called *diastole*; the contracting or pumping phase is called *systole*.

The right atrium (or auricle) collects deoxygenated blood from the superior and inferior vena cavae. During atrial contraction, blood is passed from the right atrium to the right ventricle through the tricuspid valve. During ventricular systole, the deoxygenated blood in the right ventricle is pumped out through the pulmonary valve to the lungs for oxygenation.

The left atrium receives oxygenated blood from the lungs, which is passed on during atrial contraction to the left ventricle via the mitral valve. The left ventricle is the largest and most important cardiac chamber. The left ventricle contracts the strongest among the cardiac chambers, as it has to pump out the oxygenated blood through the aortic valve and the aorta against the pressure of the rest of the vascular system of the body. Due to the higher level of importance of contraction of the ventricles, the terms systole and diastole are applied to the ventricles by default.

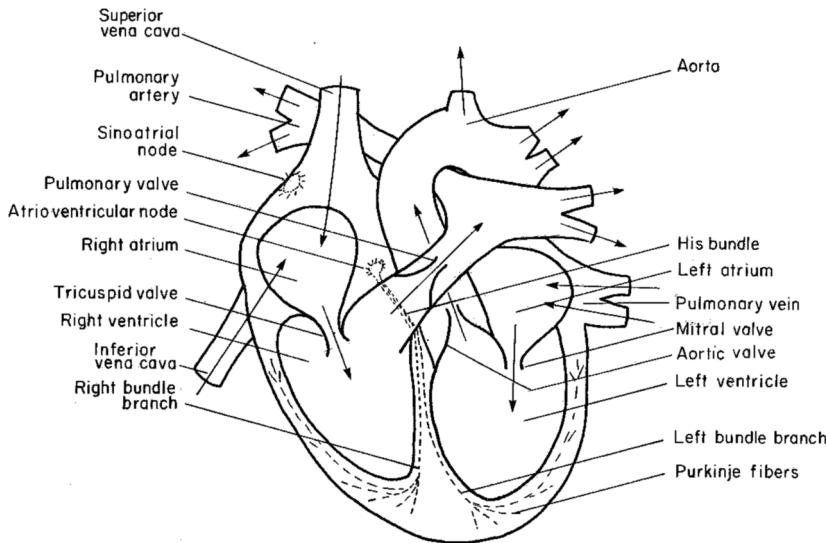


Figure 1.24 Schematic representation of the chambers, valves, vessels, and conduction system of the heart.

The heart rate (HR) or cardiac rhythm is controlled by specialized pacemaker cells that form the sinoatrial (SA) node located at the junction of the superior vena cava and the right atrium [25]. The firing rate of the SA node is controlled by the autonomic nervous system (ANS) leading to the delivery of the neurotransmitters acetylcholine (for vagal stimulation, causing a reduction in heart rate) or epinephrine (for sympathetic stimulation, causing an increase in the heart rate). The normal (resting) heart rate is about 70 bpm. The heart rate is lower during sleep, but abnormally low heart rates below 60 bpm during activity could indicate a disorder called *bradycardia*. The instantaneous heart rate could reach values as high as 200 bpm during vigorous exercise or athletic activity; a high resting heart rate could be due to illness, disease, or cardiac abnormalities, and is termed *tachycardia*.

The electrical system of the heart: Coordinated electrical events and a specialized conduction system intrinsic and unique to the heart play major roles in the rhythmic contractile activity of the heart. The SA node is the basic, natural, cardiac pacemaker that triggers its own train of action potentials. The action potential of the SA node propagates through the rest of the heart, causing a particular pattern of excitation and contraction (see Figure 1.25). A schematic ECG signal is shown in Figure 1.26 with labels showing the names and durations of the various component waves. Also shown in the figure are representations of the action potentials of the atrial and ventricular myocytes. The sequence of events and waves in a cardiac cycle is as follows [25]:

1. The SA node fires.
2. Electrical activity is propagated through the atrial musculature at comparatively low rates, causing slow-moving depolarization (contraction) of the atria. This results in the P wave in the ECG (see Figures 1.26 and 1.27). Due to the slow contraction of the atria and their relatively small size, the P wave is a slow, low-amplitude wave, with an amplitude of about 0.1 – 0.2 mV and a duration of about 60 – 80 ms.
3. The excitation wave faces a propagation delay at the atrioventricular (AV) node, which results in a normally isoelectric segment of about 60 – 80 ms after the P wave in the ECG, known as

the PQ segment. The pause assists in the completion of the transfer of blood from the atria to the ventricles.

4. The AV node fires.
5. The His bundle, the bundle branches, and the Purkinje system of specialized conduction fibers propagate the stimulus to the ventricles at a high rate.
6. The wave of stimulus spreads rapidly from the apex (at the bottom) of the heart upwards, causing rapid depolarization (contraction) of the ventricles. This results in the QRS wave of the ECG — a sharp biphasic or triphasic wave of about 1 mV amplitude and 80 ms duration (see Figure 1.27).
7. The plateau portion of the action potential causes a normally isoelectric segment of about $100 - 120\text{ ms}$ after the QRS, known as the ST segment; see Figure 1.26. This is because ventricular muscle cells possess a relatively long action potential duration of $300 - 350\text{ ms}$ (see Figure 1.7).
8. Repolarization (relaxation) of the ventricles causes the slow T wave, with an amplitude of $0.1 - 0.3\text{ mV}$ and duration of $120 - 160\text{ ms}$ (see Figure 1.27).

A summary of the various events described above, represented as parts of a cardiac cycle in relation to an ECG signal, is given in Figure 1.28.

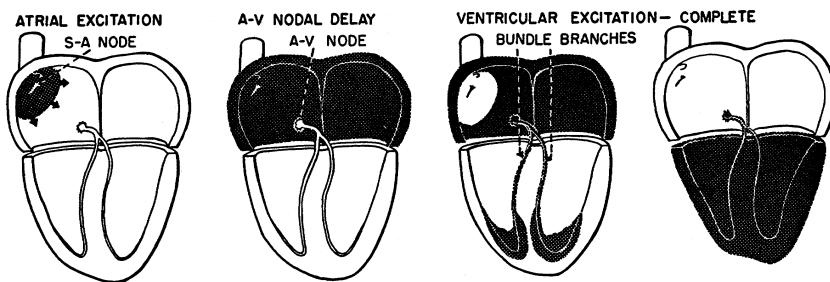


Figure 1.25 Propagation of the excitation pulse through the heart. Reproduced with permission from R.F. Rushmer, *Cardiovascular Dynamics*, 4th edition, ©W.B. Saunders, Philadelphia, PA, 1976.

It should be noted that, whereas Figure 1.26 shows only one representative action potential each of an atrial myocyte and a ventricular myocyte, multitudes of such action potentials are produced in rapid succession. The action potentials of various cardiac myocytes project on to the axis of the ECG being recorded with varying scale factors (including negative values or reversal) and phase (or time delay). Thus, the shapes of various action potentials as projected on to the recording axis could be widely different (see Figure 1.17). For the same reasons, the waveform of an externally recorded ECG varies with the positions of the leads used, which essentially determine the recording axis.

Any disturbance in the regular rhythmic activity of the heart is called *arrhythmia*. Cardiac arrhythmia may be caused by irregular firing patterns from the SA node, or by abnormal and additional pacing activity from other parts of the heart. Many parts of the heart possess inherent rhythmicity and pacemaker properties, for example, the SA node, the AV node, the Purkinje fibers, atrial tissue, and ventricular tissue. If the SA node is inactive or depressed, any one of the above tissues may take over the role of the pacemaker or introduce *ectopic* beats. Different types of abnormal rhythm (arrhythmia) result from variations in the site and frequency of impulse formation. Premature ventricular contractions (PVCs) caused by ectopic foci on the ventricles upset the regular rhythm and may lead to ventricular dissociation and fibrillation — a state of disorganized contraction of the

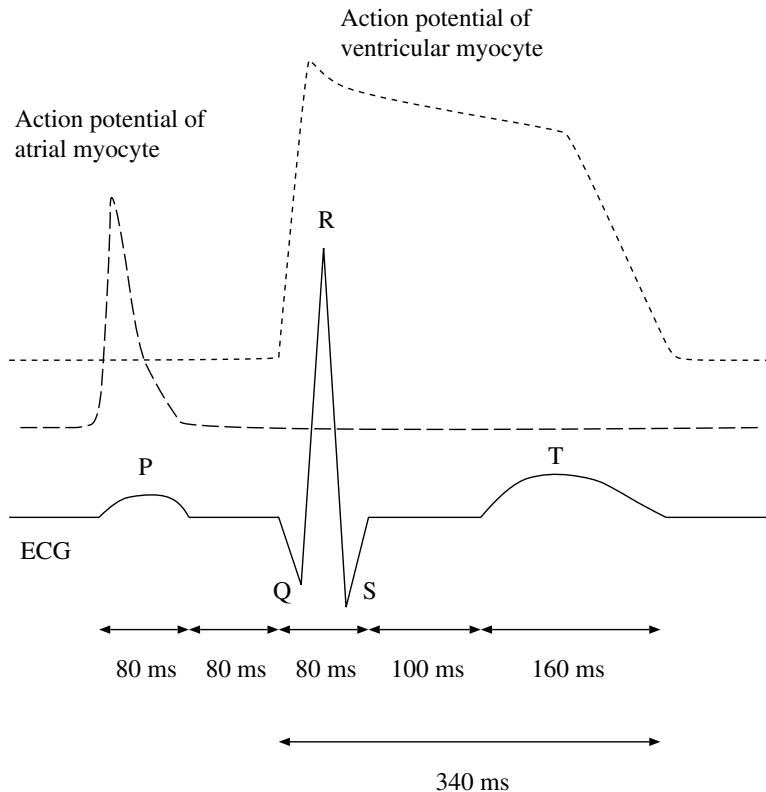


Figure 1.26 Schematic representations of an ECG signal and the action potentials of atrial and ventricular myocytes. See also Figure 1.7.

ventricles independent of the atria — resulting in no effective pumping of blood and possibly death. The waveshapes of PVCs are usually substantially different from those of the normal beats of the same subject due to the different conduction paths of the ectopic impulses and the associated abnormal contraction events. Figure 1.29 shows an ECG signal with a few normal beats and two PVCs. (See Figure 10.10 for an illustration of ventricular bigeminy, where every second pulse from the SA node is replaced by a PVC with a full compensatory pause.)

The QRS waveshape is affected by conduction disorders; for example, bundle-branch block causes a widened and possibly jagged QRS. Figure 1.30 shows the ECG signal of a patient with right bundle-branch block. Observe the wider-than-normal QRS complex, which also displays a waveshape that is significantly different from normal QRS waves. Ventricular hypertrophy (enlargement) could also cause a wider-than-normal QRS.

The ST segment, which is normally isoelectric (flat and in-line with the PQ segment) may be elevated or depressed due to myocardial ischemia (reduced blood supply to a part of the heart muscles caused by a block in the coronary arteries) or due to myocardial infarction (total lack of blood supply leading to dead myocardial tissue or scar incapable of contraction). Many other diseases cause specific changes in the ECG waveshape: the ECG is an important signal that is useful in heart-rate (rhythm) monitoring and the diagnosis of cardiovascular diseases.

ECG signal acquisition: In clinical practice, the standard 12-lead ECG is obtained using four limb leads and chest leads in six positions [25, 38]. The right leg is used to place the reference (ground) electrode. The left arm, right arm, and left leg are used to obtain leads I, II, and III. The illustration in Figure 1.31 shows the limb leads used to acquire the commonly used lead II ECG.

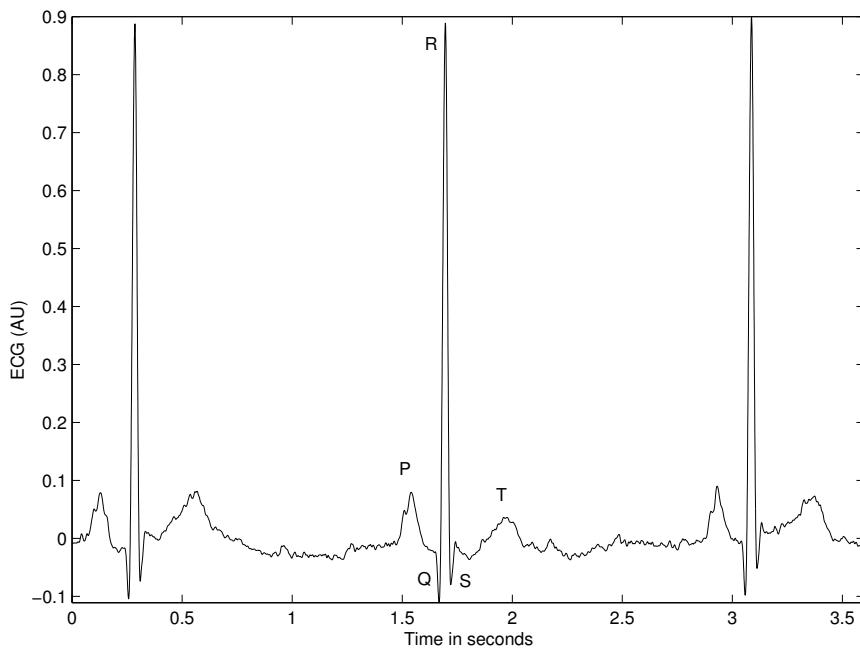


Figure 1.27 A typical ECG signal (male subject of age 24 years). (Note: Signal values may not be calibrated, that is, specified in actual units, in some situations. Sometimes, the calibration details may be lost in the acquisition and processing steps. As is the case in this plot, signal values in plots in this book are in arbitrary units (AU) or normalized units unless specified.)

A combined reference known as *Wilson's central terminal* is formed by combining the left arm, right arm, and left leg leads; it is used as the reference for chest leads. The *augmented* limb leads known as aVR, aVL, and aVF (aV for augmented vector, R for the right arm, L for the left arm, and F for the left leg or foot) are obtained by using the exploring electrode on the limb indicated by the name of the lead, with the reference being Wilson's central terminal without the exploring limb lead.

Figure 1.32 shows the directions of the axes formed by the six limb leads. The hypothetical equilateral triangle formed by leads I, II, and III is known as *Einthoven's triangle*. The center of the triangle represents Wilson's central terminal. Schematically, the heart is assumed to be placed at the center of the triangle. The six leads measure projections of the three-dimensional (3D) cardiac electrical vector on to the axes illustrated in Figure 1.32. The six axes sample the $0^\circ - 180^\circ$ range in steps of approximately 30° . The projections facilitate viewing and analysis of the electrical activity of the heart from different perspectives in the frontal plane.

The six chest leads (written as V1–V6) are obtained from six standardized positions on the chest [25] with Wilson's central terminal as the reference. The positions for placement of the precordial (chest) leads are indicated in Figure 1.33. The V1 and V2 leads are placed at the fourth intercostal space just to the right and left of the sternum, respectively. V4 is recorded at the fifth intercostal space at the left midclavicular line. The V3 lead is placed half-way between the V2 and V4 leads. The V5 and V6 leads are located at the same level as the V4 lead, but at the anterior axillary line and the midaxillary line, respectively. The six chest leads permit viewing the cardiac electrical vector from different orientations in a cross-sectional plane: V5 and V6 are most sensitive to left-ventricular activity; V3 and V4 depict septal activity best; V1 and V2 reflect well activity in the right-half of the heart.

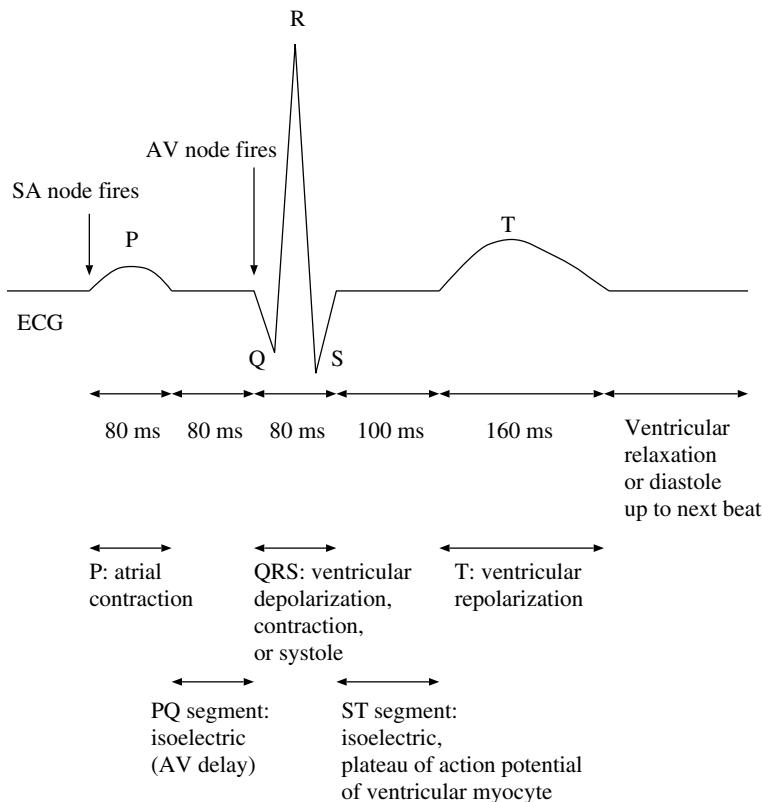


Figure 1.28 Summary of the parts and waves of a cardiac cycle as seen in an ECG signal.

In spite of being redundant, the 12-lead system serves as the basis of the standard clinical ECG. Clinical ECG interpretation is mainly empirical. A compact and efficient system has been proposed for *vectorcardiography* or VCG [25, 39], where loops inscribed by the 3D cardiac electrical vector in three mutually orthogonal planes, namely, the frontal, horizontal, and sagittal planes, are plotted and analyzed. Regardless, the 12-lead scalar ECG is the most commonly used procedure in clinical practice.

Because an external ECG is a projection of the internal 3D cardiac electrical vector, external ECG recordings are not unique. Some of the lead interrelationships are [25, 38]:

- $\text{II} = \text{I} + \text{III}$,
- $\text{aVL} = (\text{I} - \text{III}) / 2$,

which are depicted in Figures 1.34 and 1.35.

Some of the important features of the standard clinical ECG are:

- A rectangular calibration pulse of 1 mV amplitude and 200 ms duration is applied to produce a pulse of 1 cm height on the plot.
- The typical recording speed used is 25 mm/s, resulting in a graphical scale of 0.04 s/mm or 40 ms/mm. Then, the width of the calibration pulse is 5 mm.
- The ECG signal peak value is normally about 1 mV.

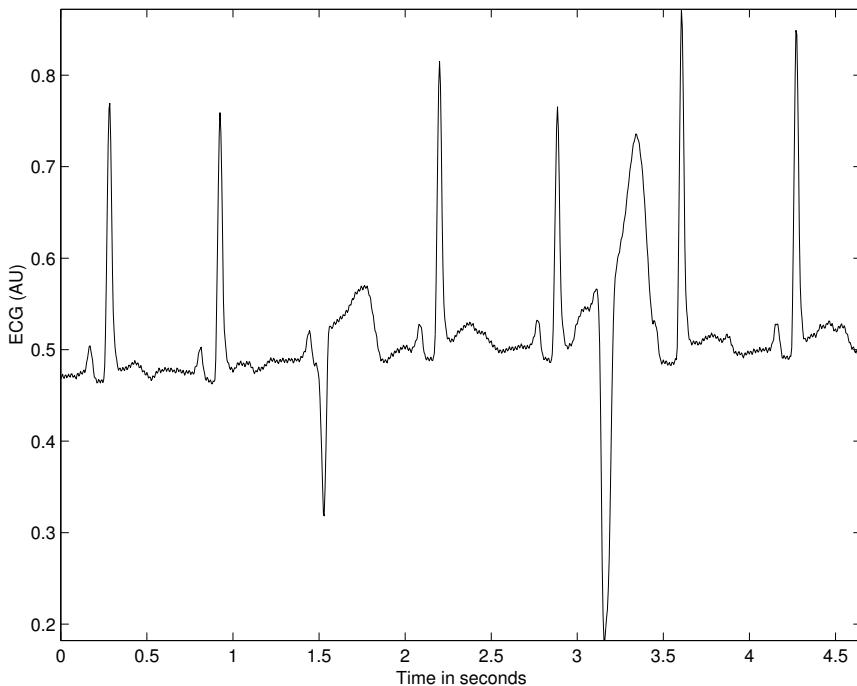


Figure 1.29 ECG signal with PVCs. The third and sixth beats are PVCs. The first PVC has blocked the normal beat that would have appeared at about the same time instant, but the second PVC has not blocked any normal beat triggered by the SA node. Data courtesy of G. Groves and J. Tyberg, Department of Physiology and Biophysics, University of Calgary.

- The amplifier gain used is 1,000.
- Clinical ECG is usually filtered to a bandwidth of about $0.05 - 100\text{ Hz}$, with a recommended sampling rate of 500 Hz for diagnostic ECG. Distortions in the shape of the calibration pulse may indicate improper filter settings or a poor signal acquisition system.
- ECG for heart-rate monitoring could use a reduced bandwidth $0.5 - 50\text{ Hz}$ and a lower sampling rate 100 Hz .
- High-resolution ECG requires a greater bandwidth of $0.05 - 500\text{ Hz}$.

Some of the features mentioned above are related to recording of the ECG on paper and may not be relevant to digital or computer recording of the ECG.

For details of other systems of ECG leads and electrodes, see Rushmer [25], Friedman [40], Goldberger [39], Malmivuo and Plonsey [41], Webster [13], and Draper et al. [42].

Figure 1.36 shows the 12-lead ECG of a normal male adult. The system used to obtain the illustration records three channels at a time: leads I, II, III; aVR, aVL, aVF; V1, V2, V3; and V4, V5, V6 are recorded in the three available channels simultaneously. Other systems may record one channel at a time. Observe the variable shapes of the ECG waves from one lead to another. A well-trained cardiologist is able to deduce the 3D orientation of the cardiac electrical vector by analyzing the waveshapes in the six limb leads. Cardiac defects, if any, may be localized by analyzing the waveshapes in the six chest leads.

Figure 1.37 shows the 12-lead ECG of a patient with right bundle-branch block with secondary repolarization changes. The increased QRS width and distortions in the QRS shape indicate the effects of asynchronous activation of the ventricles due to the bundle-branch block.

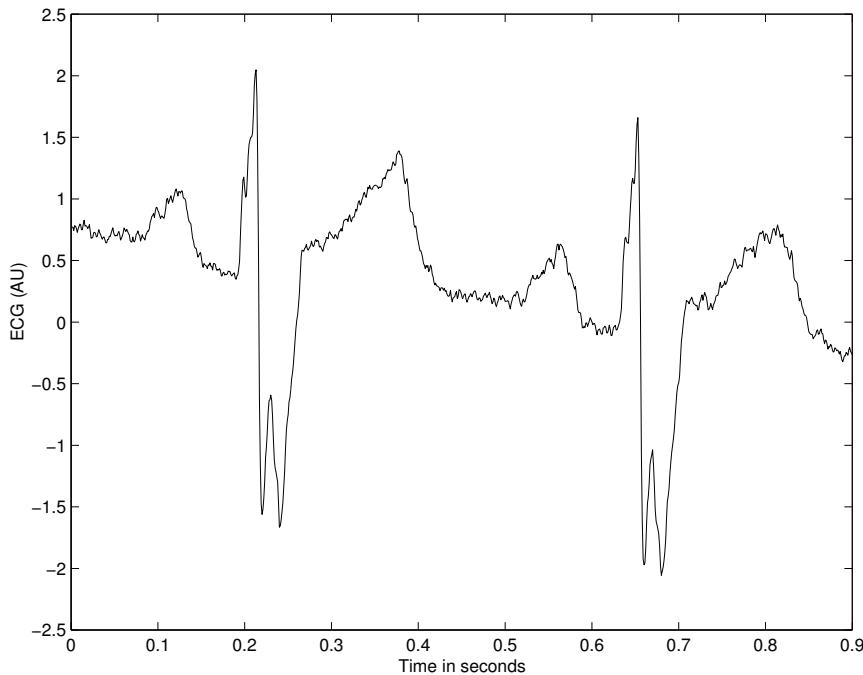


Figure 1.30 ECG signal of a patient with right bundle-branch block and hypertrophy (male patient of age 3 months). The QRS complex is wider than normal, and displays an abnormal, jagged waveform due to desynchronized contraction of the ventricles. (The signal also has a baseline drift, which has not been corrected.)

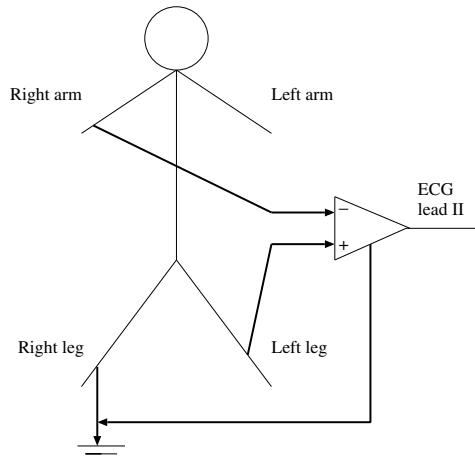


Figure 1.31 Limb leads used to acquire the commonly used lead II ECG. *Note:* The labeling of the left or right side refers to the corresponding side of the patient or subject, as in medical convention, and not the side of the reader.

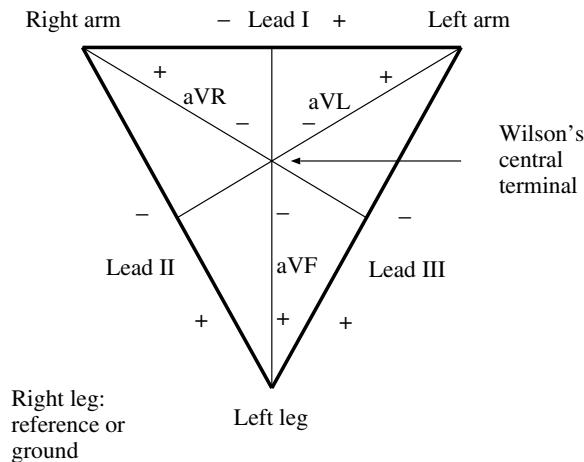


Figure 1.32 Einthoven's triangle and the axes of the six ECG leads formed by using four limb leads.

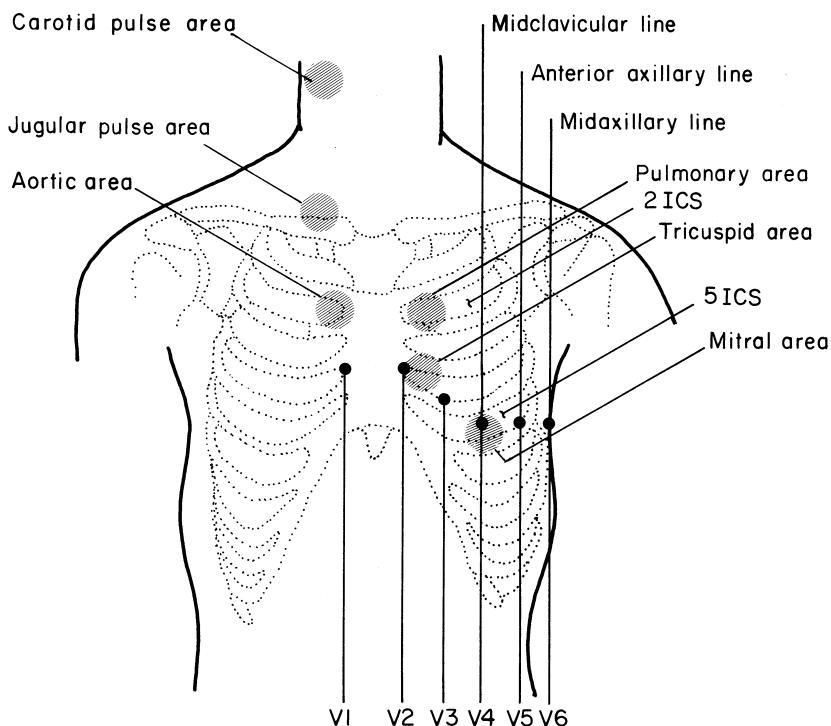


Figure 1.33 Positions for placement of the precordial (chest) leads V1–V6 for ECG, auscultation areas for heart sounds, and pulse transducer positions for the carotid and jugular pulse signals. ICS: intercostal space.

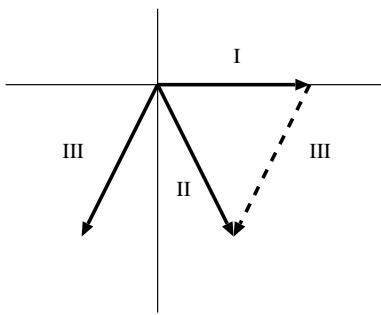


Figure 1.34 Vectorial relations between ECG leads I, II, and III. See also Figure 1.32.

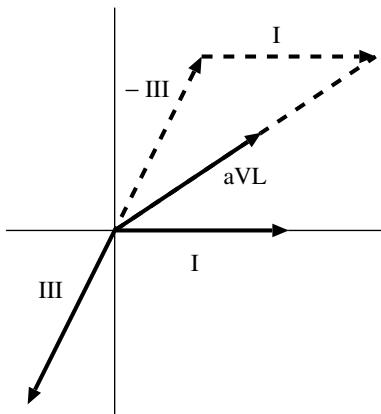


Figure 1.35 Vectorial relations between ECG leads I, III, and aVL. See also Figure 1.32.

See Section 7.8 for discussions on electrophysiological modeling of the heart. Signal processing techniques to filter ECG signals are presented in Sections 3.3, 3.6, 3.7, 3.9, and 3.13. Detection of ECG waveforms is discussed in Sections 4.2.1, 4.3.2, 4.3.3, and 4.8. Analysis of ECG waveform shape and classification of beats are dealt with in Sections 5.2.1, 5.2.2, 5.2.3, 5.4, 5.7, 5.8, 10.2.1, and 10.11. Analysis of heart-rate variability (HRV) is described in Sections 7.2.2, 7.9, and 8.12. See Sections 3.14, 9.7.2, and 9.11 for discussions on fetal and maternal ECG. For reviews of computer applications in ECG analysis, see Jenkins [43, 44] and Cox et al. [45].

1.2.6 The electroencephalogram (EEG)

The EEG (popularly known as *brain waves*) represents the electrical activity of the brain [46–48]. A few important aspects of the organization of the brain are as follows. The main parts of the brain are the cerebrum, the cerebellum, the brain stem (including the midbrain, pons medulla, and the reticular formation), and the thalamus (between the midbrain and the hemispheres). Figure 1.38 gives a schematic representation of the various parts and functional areas of the human brain. The cerebrum is divided into two hemispheres, separated by a longitudinal fissure across which there is a large connective band of fibers known as the corpus callosum. The outer surface of the cerebral hemispheres, known as the cerebral cortex, is composed of neurons (gray matter) in convoluted patterns, and separated into regions by fissures (sulci). Beneath the cortex lie nerve fibers that lead to other parts of the brain and the body (white matter).

Cortical potentials are generated due to excitatory and inhibitory postsynaptic potentials developed by cell bodies and dendrites of pyramidal neurons. Physiological control processes, thought

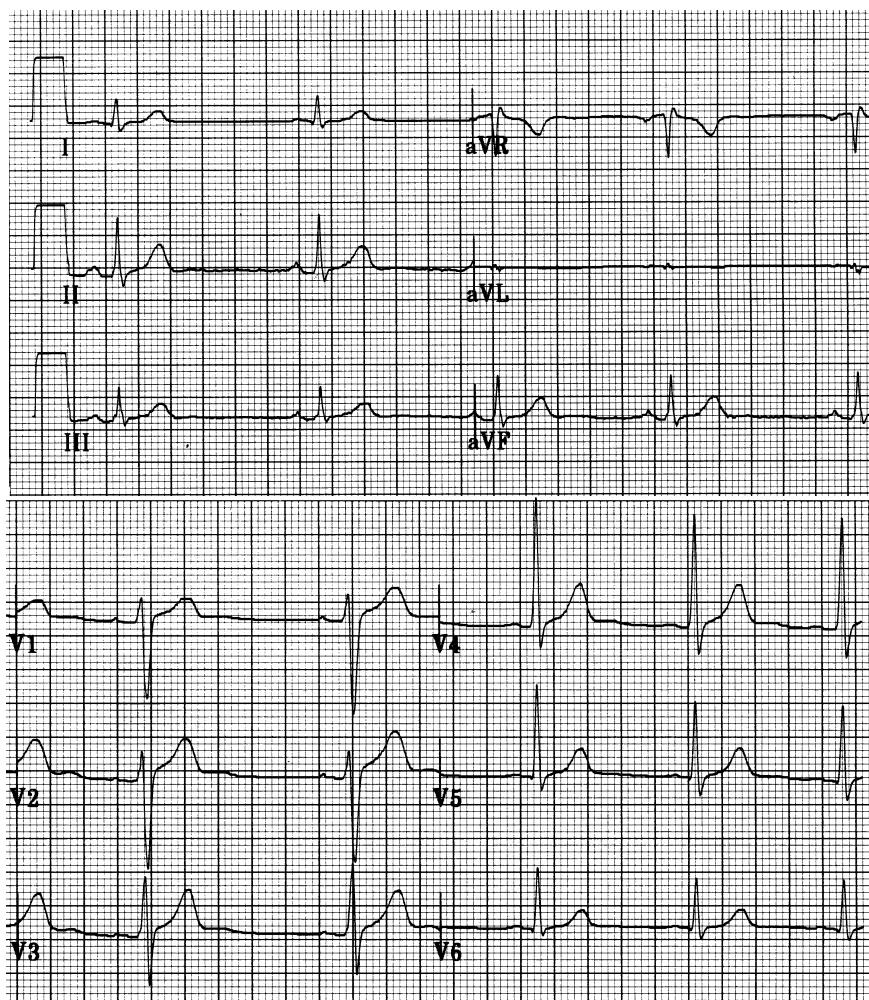


Figure 1.36 Standard 12-lead ECG of a normal male adult. Signal courtesy of E. Gedamu and L.B. Mitchell, Foothills Hospital, Calgary.

processes, and external stimuli generate signals in the corresponding parts of the brain that may be recorded using electrodes placed on the surface of the scalp. The scalp EEG is an average of the multifarious activities of many small zones of the cortical surface beneath the electrodes used.

In clinical practice, several channels of the EEG are recorded simultaneously from various locations on the scalp for comparative analysis of activities in different regions of the brain. The International Federation of Societies for Electroencephalography and Clinical Neurophysiology recommended the 10 – 20 system of electrode placement for clinical EEG recording [46], which is schematically illustrated in Figure 1.39. The name 10 – 20 indicates the fact that the electrodes along the midline are placed at 10%, 20%, 20%, 20%, 20%, and 10% of the total nasion–ionion distance; the other series of electrodes are also placed at similar fractional distances of the corresponding reference distances [46]. The interelectrode distances are equal along any anteroposterior or transverse line, and electrode positioning is symmetrical. EEG signals may be used to study the nervous system, to monitor sleep stages [49, 50], for biofeedback and control, for brain–computer interfacing (BCI) [51], and for detection or diagnosis of epilepsy (seizure) [52–54].

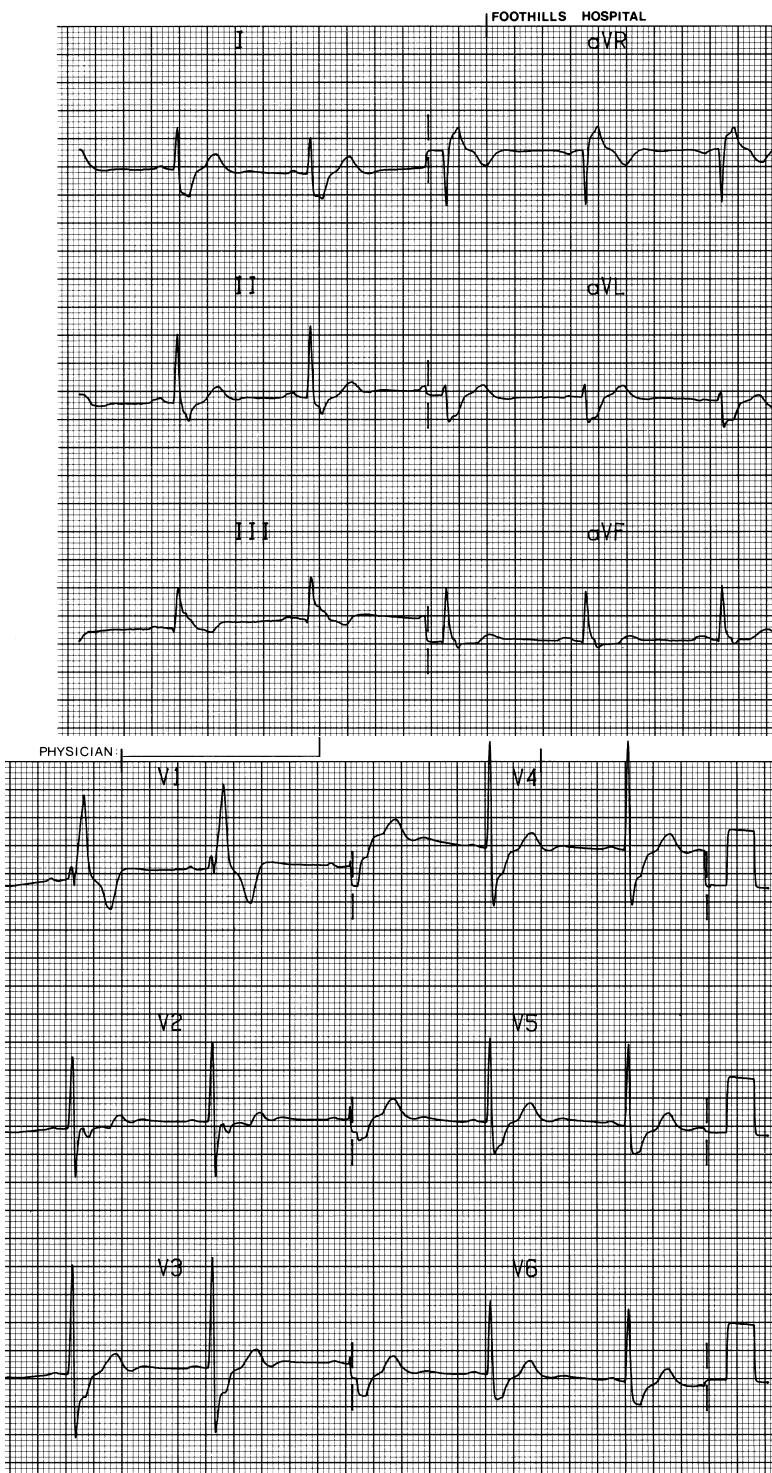


Figure 1.37 Standard 12-lead ECG of a patient with right bundle-branch block. Signal courtesy of L.B. Mitchell, Foothills Hospital, Calgary.

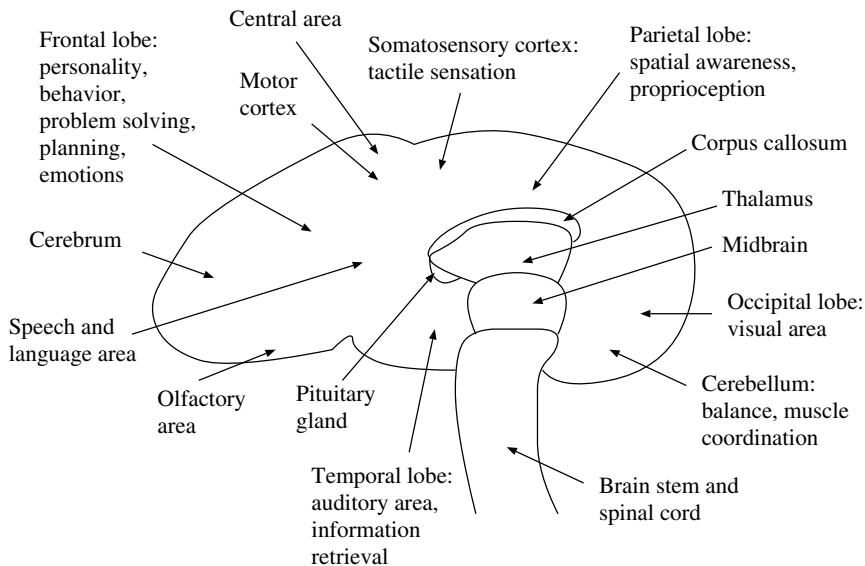


Figure 1.38 Schematic diagram showing the various parts and functional areas of the human brain.

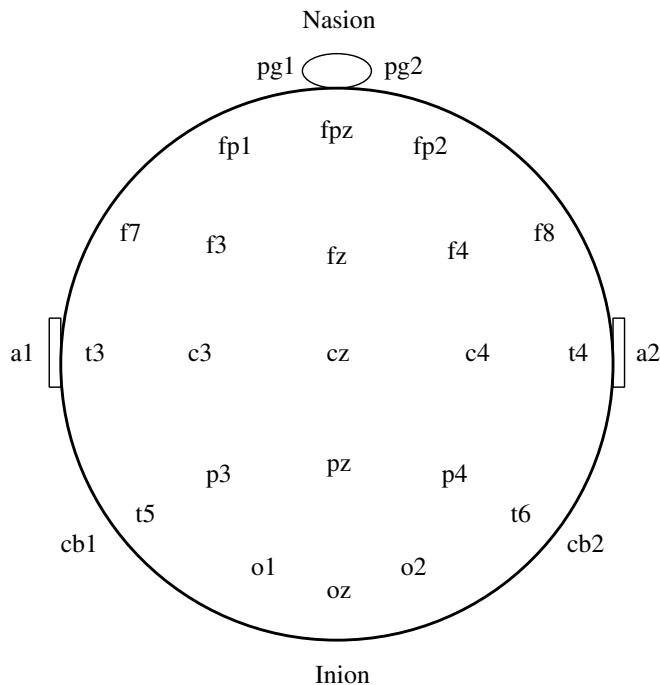


Figure 1.39 The 10 – 20 system of electrode placement for EEG recording [46]. Notes regarding channel labels: pg, nasopharyngeal; a, auricular (earlobes); fp, prefrontal; f, frontal; p, parietal; c, central; o, occipital; t, temporal; cb, cerebellar; z, midline; odd numbers on the left, even numbers on the right of the subject.

Typical EEG instrumentation settings used are lowpass filtering at 75 Hz , and recording at $100\text{ }\mu\text{V/cm}$ and 30 mm/s for $10 - 20$ minutes over $8 - 16$ simultaneous channels. Monitoring of sleep EEG and detection of transients related to epileptic seizures may require multichannel EEG acquisition over several hours. Special EEG techniques include the use of needle electrodes and nasopharyngeal electrodes, recording the electrocorticogram (ECoG) from an exposed part of the cortex, and the use of intracerebral electrodes. Evocative techniques for recording the EEG include initial recording at rest (eyes open and eyes closed), hyperventilation (after breathing at 20 respirations per minute for $2 - 4$ minutes), photic stimulation (with $1 - 50$ flashes of light per second), auditory stimulation with loud clicks, sleep (different stages), and pharmaceuticals or drugs.

EEG signals exhibit several patterns of rhythmic or periodic activity. (*Note:* The term *rhythm* stands for different phenomena or events in the ECG and the EEG.) The commonly used terms for EEG frequency (f) bands are:

- Delta (δ): $0.5 \leq f < 4\text{ Hz}$;
- Theta (θ): $4 \leq f < 8\text{ Hz}$;
- Alpha (α): $8 \leq f \leq 13\text{ Hz}$; and
- Beta (β): $f > 13\text{ Hz}$.

Figure 1.40 illustrates traces of EEG signals with the rhythms listed above.

EEG rhythms are associated with various physiological and mental processes [47,48]. The alpha rhythm is the principal resting rhythm of the brain; it is common in wakeful, resting adults, especially in the occipital area, with bilateral synchrony. Auditory and mental arithmetic tasks with the eyes closed lead to strong alpha waves, which are suppressed when the eyes are opened (that is, by a visual stimulus); see Figure 1.40 (e) [46].

In addition to the commonly studied rhythms mentioned above, the gamma rhythm is defined as activity in the range $30 - 80\text{ Hz}$. The gamma rhythm is considered to be related to responses induced by various types of sensory input or stimuli, active sensory processes involving attention, and short-term memory processes [55]; see Mantini et al. [56] and Rennie et al. [57] for related discussions.

The alpha wave is replaced by slower rhythms at various stages of sleep. Theta waves appear at the beginning stages of sleep; delta waves appear at deep-sleep stages. High-frequency beta waves appear as background activity in tense and anxious subjects. The depression or absence of the normal (expected) rhythm in a certain state of a subject could indicate abnormality. The presence of delta or theta (slow) waves in a wakeful adult may be considered to be abnormal. Focal brain injury and tumors lead to abnormal slow waves in the corresponding regions. Unilateral depression (left-right asymmetry) of a rhythm could indicate disturbances in cortical pathways. Spikes and sharp waves could indicate the presence of epileptogenic regions in the corresponding parts of the brain.

Figure 1.41 shows an example of eight channels of the EEG recorded simultaneously from the scalp of a subject. All channels display high levels of alpha activity. Figure 1.42 shows 10 channels of the EEG of a subject with spike-and-wave complexes. Observe the distinctly different waveshape and sharpness of the spikes in Figure 1.42 as compared to the smooth waves in Figure 1.41.

EEG signals also include spikes, transients, and other waves and patterns associated with various disorders of the nervous system (see Figure 4.1 and Section 4.2.4). Figure 1.43 shows a 21-channel record of a patient with a seizure starting at about the 50-s mark [58]. The signal is characterized by a recruiting theta rhythm at about 5 Hz in the channels labeled as T2, F8, T4, and T6. Artifacts are evident in the signal due to muscle activity (in T3, C3, and C4) and blinking of the eye (in Fp1 and Fp2). Increased amounts of relatively high-frequency activity are seen in several channels after the 50-s mark related to the seizure.

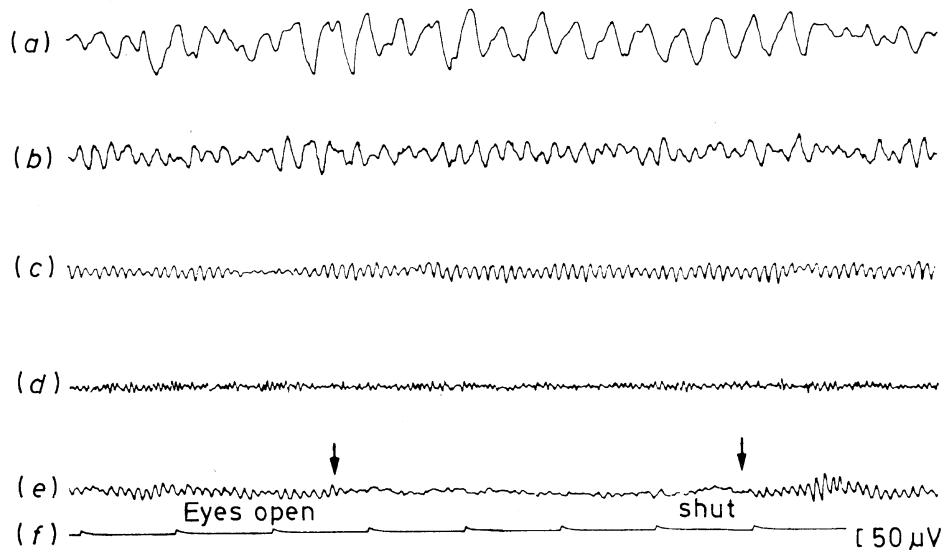


Figure 1.40 From top to bottom: (a) delta rhythm; (b) theta rhythm; (c) alpha rhythm; (d) beta rhythm; (e) blocking of the alpha rhythm by eye opening; (f) 1 s time markers and $50 \mu\text{V}$ marker. Reproduced with permission from R. Cooper, J.W. Osselton, and J.C. Shaw, *EEG Technology*, 3rd Edition, 1980. ©Butterworth Heinemann Publishers, a division of Reed Educational & Professional Publishing Ltd., Oxford, UK.

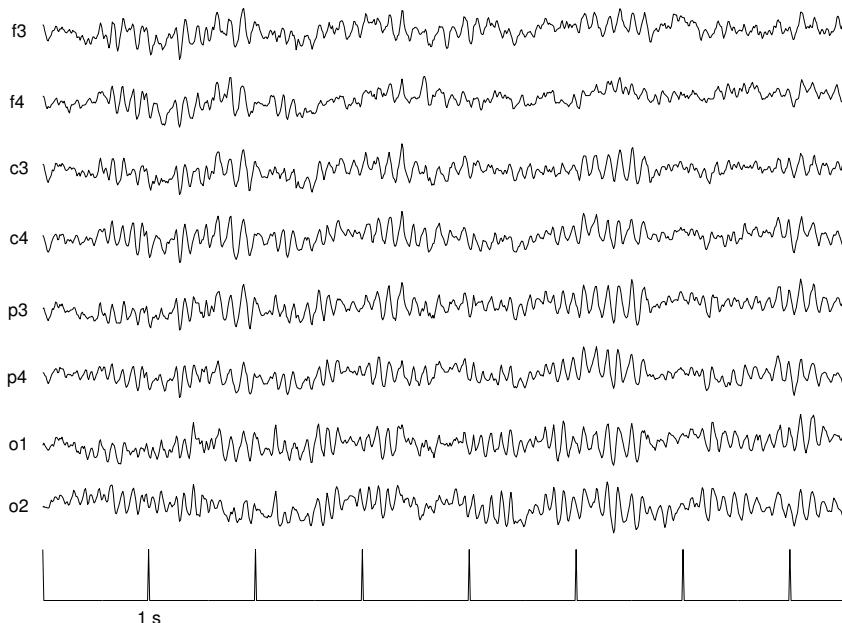


Figure 1.41 Eight channels of the EEG of a subject displaying alpha rhythm. See Figure 1.39 for details regarding channel labels. Data courtesy of Y. Mizuno-Matsumoto, Osaka University Medical School, Osaka, Japan.

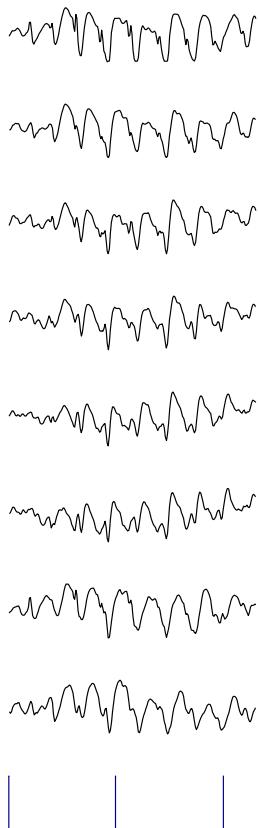


Figure 1.42 Ten channels of the EEG of a subject displaying spike-and-wave complexes. The channels shown are, from top to bottom: c3, c4, p3, p4, o1, o2, t3, t4, and time (1 s per mark). See Figure 1.39 for details regarding channel labels. Data courtesy of Y. Mizuno-Matsumoto, Osaka University Medical School, Osaka, Japan.

Detection of events and rhythms in EEG signals is discussed in Sections 4.4, 4.5, and 4.6. Spectral analysis of EEG signals is dealt with in Sections 6.3.4, 6.7, and 7.5.2. Adaptive segmentation of EEG signals is described in Sections 8.2.2, 8.5, and 8.10. Detection of seizures is discussed in Sections 8.17 and 9.8. See Section 9.12 for a discussion on EEG channel selection for BCI.

1.2.7 Event-related potentials (ERPs)

The term *event-related potential* is more general than and preferred to the term *evoked potential*, and includes the ENG or the EEG in response to light, sound, electrical, or other external stimuli. Short-latency ERPs are predominantly dependent on the physical characteristics of the stimulus, whereas longer-latency ERPs are predominantly influenced by the conditions of presentation of the stimuli.

Somatosensory evoked potentials (SEPs) are useful for noninvasive evaluation of the nervous system from a peripheral receptor to the cerebral cortex. Median nerve short-latency SEPs are obtained by placing stimulating electrodes about 2 – 3 cm apart over the median nerve at the wrist with electrical stimulation at 5 – 10 pps, each stimulus pulse being of duration less than 0.5 ms with an amplitude of about 100 V (producing a visible thumb twitch). The SEPs are recorded from the surface of the scalp. The latency, duration, and amplitude of the response are measured.

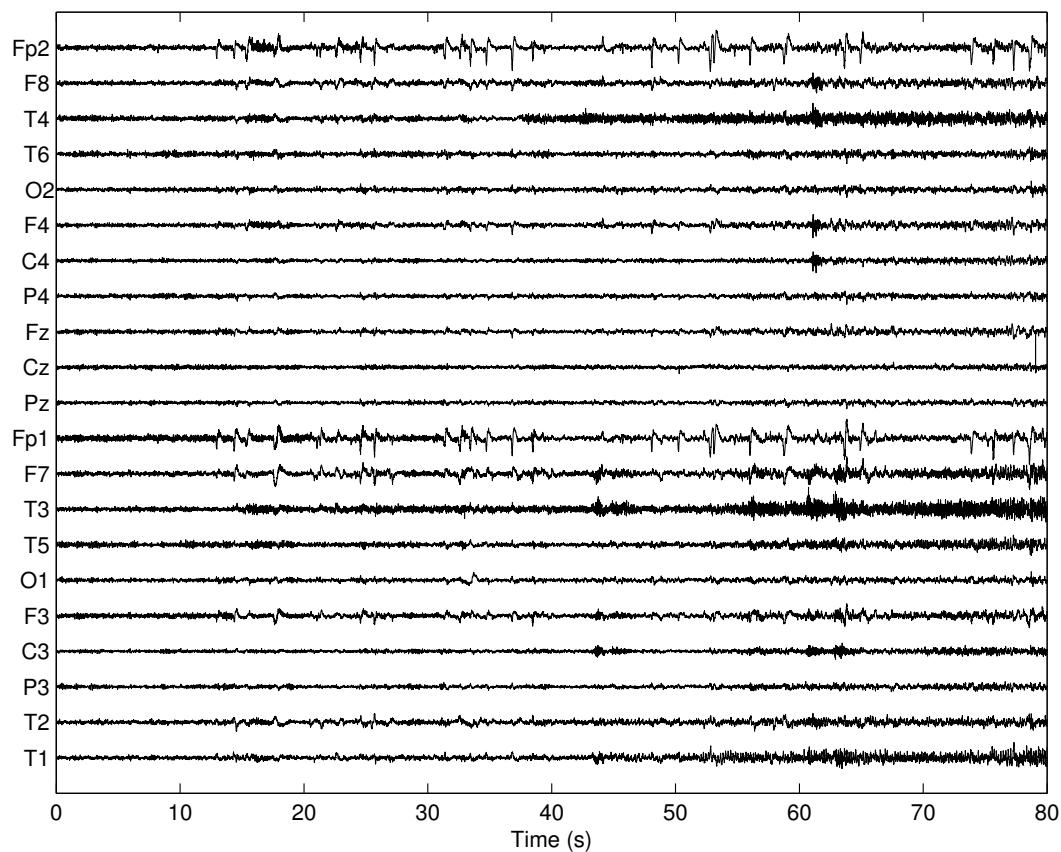


Figure 1.43 21 channels of the EEG of a subject displaying seizure activity. Data courtesy of M. De Vos, Katholieke Universiteit Leuven, Leuven, Belgium. T1 and T2 are sphenoidal electrodes.

ERPs and SEPs are weak signals typically buried in ongoing activity of the associated systems. Examples of ERPs are provided in Figures 3.2 and 3.43. Improvement of signal-to-noise ratio (*SNR*) is usually achieved by synchronized averaging and filtering. See Sections 3.5, 3.12, 8.18, and 9.12 for examples of ERPs and related signal processing methods.

1.2.8 The electrogastrogram (EGG)

The electrical activity of the stomach consists of rhythmic waves of depolarization and repolarization of its constituent smooth muscle cells [59–61]. The activity originates in the midcorpus of the stomach, with intervals of about 20 s in humans. The waves of activity are always present and are not directly associated with contractions; they are related to the spatial and temporal organization of gastric contractions.

External (cutaneous) electrodes can record the signal known as the electrogastrogram (EGG). Chen et al. [62] used the following procedures to record cutaneous EGG signals. With the subject in the supine position and remaining motionless, the stomach was localized using a 5 – MHz ultrasound transducer array, and the orientation of the distal stomach was marked on the abdominal surface. Three active electrodes were placed on the abdomen along the antral axis of the stomach with an interelectrode spacing of 3.5 cm. A common reference electrode was placed 6 cm away in the upper-right quadrant. Three bipolar signals were obtained from the three active electrodes in

relation to the common reference electrode. The signals were amplified and filtered to the bandwidth of $0.02 - 0.3 \text{ Hz}$ with 6 dB/octave transition bands, and sampled at 2 Hz .

The surface EGG is believed to reflect the overall electrical activity of the stomach, including the electrical control activity and the electrical response activity. Chen et al. [62] indicated that gastric dysrhythmia or arrhythmia may be detected via analysis of the EGG. Other researchers suggest that the diagnostic potential of the signal has not yet been established [59, 60]. Accurate and reliable measurement of the electrical activity of the stomach requires implantation of electrodes within the stomach [63], which limits its practical applicability.

1.2.9 The phonocardiogram (PCG)

The heart sound signal is perhaps the most traditional biomedical signal, as indicated by the fact that the stethoscope is the primary instrument carried and used by physicians. The PCG is a vibration or sound signal related to the contractile activity of the cardiohemic system (the heart and blood together) [25, 64–69]; it represents a recording of the heart sound signal. Recording of the PCG signal requires a transducer to convert the vibration or sound signal into an electronic signal: Microphones, pressure transducers, or accelerometers may be placed on the chest surface for this purpose. The normal heart sounds provide an indication of the general state of the heart in terms of rhythm and contractility. Cardiovascular diseases and defects cause changes to the heart sounds or additional sounds and murmurs that could be useful in their diagnosis.

The genesis of heart sounds: The externally recorded heart sounds are not caused by valve leaflet movements *per se*, as believed earlier, but by vibrations of the whole cardiovascular system triggered by pressure gradients [25]. The cardiohemic system may be compared to a fluid-filled balloon, which, when stimulated at any location, vibrates as a whole. Externally, however, heart sound components are best heard at certain locations on the chest individually, and this localization has led to the concept of *secondary sources* on the chest related to the well-known auscultatory areas: the mitral, aortic, pulmonary, and tricuspid areas [25]. The standard auscultatory areas are indicated in Figure 1.33. The mitral area is near the apex of the heart. The aortic area is to the right of the sternum, in the second intercostal space. The tricuspid area is in the fourth intercostal space on either side of the sternum. The pulmonary area lies at the left parasternal line in the second or third intercostal space [25].

A normal cardiac cycle contains two major sounds — the first heart sound (S1) and the second heart sound (S2). Figure 1.44 shows a normal PCG signal, along with the ECG and carotid pulse tracings. S1 occurs at the onset of ventricular contraction, and corresponds in timing to the QRS complex in the ECG signal.

The initial vibrations in S1 occur when the first myocardial contractions in the ventricles move blood toward the atria, sealing the AV (mitral and tricuspid) valves (see Figure 1.45). The second component of S1 begins with abrupt tension of the closed AV valves, decelerating the blood. At this stage, known as isovolumic contraction, all four of the cardiac valves are closed. Next, the semilunar (aortic and pulmonary) valves open, and the blood is ejected out of the ventricles. The third component of S1 may be caused by oscillation of blood between the root of the aorta and the ventricular walls. This is followed by the fourth component of S1, which may be due to vibrations caused by turbulence in the ejected blood flowing rapidly through the ascending aorta and the pulmonary artery.

Following the systolic pause in the PCG of a normal cardiac cycle, the second sound S2 is caused by the closure of the semilunar valves. While the primary vibrations occur in the related arteries due to deceleration of blood, the ventricles and atria also vibrate, due to transmission of vibrations through the blood, the valves, and the valve rings. S2 has two components, one due to closure of the aortic valve (A2) and the other due to closure of the pulmonary valve (P2). The aortic valve normally closes before the pulmonary valve, and hence A2 precedes P2 by a few milliseconds. The A2–P2 gap is widened in normal subjects during inspiration. The pulmonary impedance is lower during

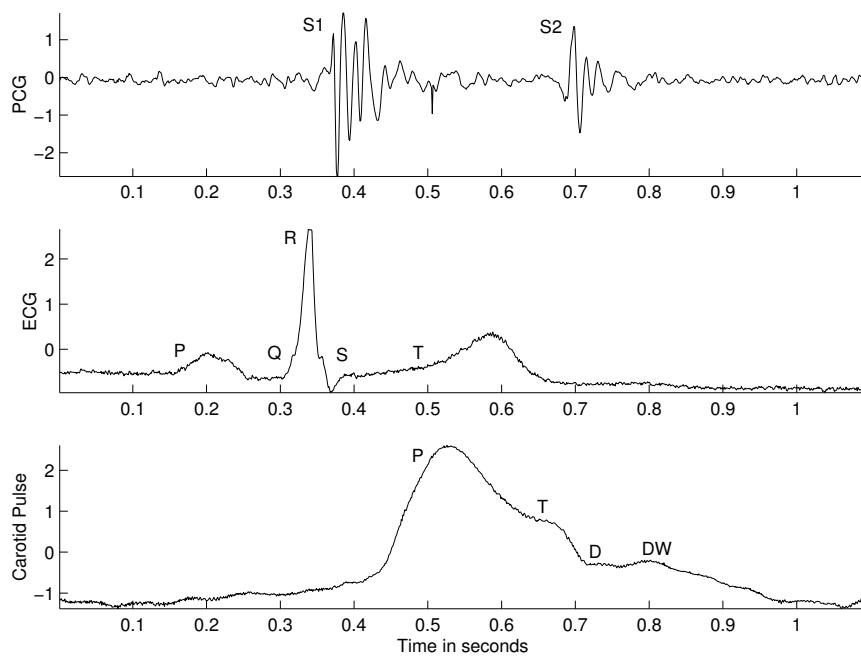


Figure 1.44 Three-channel simultaneous record of the PCG, ECG, and carotid pulse signals of a normal male adult.

inspiration as compared to the condition during expiration. The compliance of the pulmonary vessels is increased during inspiration because they are dilated due to negative thoracic pressure. These conditions cause P₂ to be delayed during inspiration as compared to the situation during expiration. (Note: The PCG signal in Figure 1.44 does not show the A₂ and P₂ components separately.) A wide gap is sustained during expiration in the case of pulmonary hypertension. Bundle-branch block could cause the A₂-P₂ gap to widen, or may even reverse the order of occurrence of A₂ and P₂.

In some cases, a third heart sound (S₃) may be heard, corresponding to sudden termination of the ventricular rapid-filling phase. Because the ventricles are filled with blood and their walls are relaxed during this part of diastole, the vibrations of S₃ contain much lower frequencies than S₁ and S₂. In late diastole, a fourth heart sound (S₄) may be heard sometimes, caused by atrial contractions displacing blood into the distended ventricles. In addition to these sounds, valvular clicks and snaps are occasionally heard.

Heart murmurs: The intervals between S₁ and S₂, and S₂ and S₁ of the next cycle (corresponding to ventricular systole and diastole, respectively), are normally silent. Murmurs, which are caused by certain cardiovascular defects and diseases, may occur in these intervals. Murmurs are high-frequency noise-like sounds that arise when the velocity of blood becomes high as it flows through an irregularity, such as a constriction, an orifice, or a baffle. Typical conditions in the cardiovascular system that cause turbulence in blood flow are valvular stenosis and insufficiency. A valve is said to be stenosed when, due to the deposition of calcium or other reasons, the valve leaflets are stiffened and do not open completely, thereby causing an obstruction or baffle in the path of the blood being ejected. A valve is said to be insufficient when it cannot close effectively and causes reverse leakage or regurgitation of blood through a narrow gap.

Systolic murmurs are caused by conditions such as ventricular septal defect (a hole in the septum or wall between the left ventricle and the right ventricle), aortic stenosis, pulmonary stenosis, mitral insufficiency, and tricuspid insufficiency. Semilunar valvular stenosis (aortic stenosis and

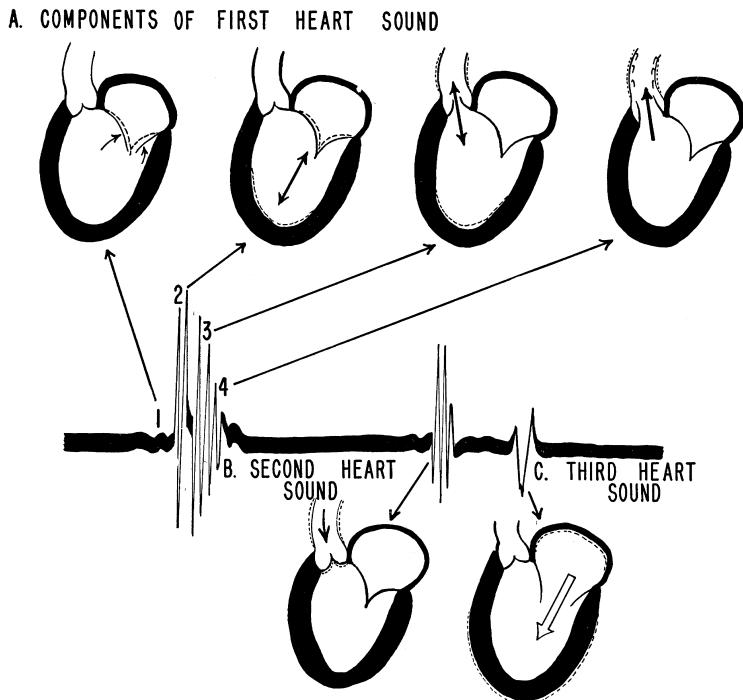


Figure 1.45 Schematic representation of the genesis of heart sounds. Only the left portion of the heart is illustrated as it is the major source of the heart sounds. The corresponding events in the right portion also contribute to the sounds. The atria do not directly contribute much to the heart sounds. Reproduced with permission from R.F. Rushmer, *Cardiovascular Dynamics*, 4th edition, ©W.B. Saunders, Philadelphia, PA, 1976.

pulmonary stenosis) causes an obstruction in the path of blood being ejected during systole. AV valvular insufficiency (mitral insufficiency and tricuspid insufficiency) causes regurgitation of blood to the atria during ventricular contraction.

Diastolic murmurs are caused by conditions such as aortic insufficiency, pulmonary insufficiency, mitral stenosis, and tricuspid stenosis. Other conditions causing murmurs are atrial septal defect and patent ductus arteriosus (an abnormal connection or shunt between the aorta and the pulmonary artery), as well as certain physiological or functional conditions that result in increased cardiac output or blood velocity.

Various features of heart sounds and murmurs, such as intensity, frequency content, and timing, are affected by many physical and physiological factors, including the recording site on the thorax, intervening thoracic structures, ventricular contractility, positions of the cardiac valves at the onset of systole, the degree of the defect present, the heart rate, and blood velocity. For example, S1 is loud and delayed in mitral stenosis; right bundle-branch block causes wide splitting of S2; left bundle-branch block results in reversed splitting of S2; acute myocardial infarction causes a pathologic S3; and severe mitral regurgitation leads to an increased S4 [64–68]. Although murmurs are noise-like events, their features aid in distinguishing between different causes. For example, aortic stenosis causes a diamond-shaped midsystolic murmur, whereas mitral stenosis causes a decrescendo-crescendo type of diastolic–presystolic murmur. Figure 1.46 illustrates the PCG, ECG, and carotid pulse signals of a patient with aortic stenosis; the PCG displays the typical diamond-shaped murmur in systole.

Recording of PCG signals: PCG signals are normally recorded using piezoelectric contact sensors that are sensitive to displacement or acceleration at the skin surface. The PCG signals illustrated in this section were obtained using a Hewlett Packard HP21050A transducer, which has a nominal bandwidth of $0.05 - 1,000\text{ Hz}$. The carotid pulse signals shown in this section were recorded using the HP21281A pulse transducer, which has a nominal bandwidth of $0 - 100\text{ Hz}$. PCG recording is normally performed in a quiet room, with the patient in the supine position with the head resting on a pillow. The PCG transducer is placed firmly at the desired position on the chest using a suction ring and/or a rubber strap.

The use of the ECG and carotid pulse signals in the analysis of PCG signals is described in Sections 2.2.1, 2.2.2, and 2.3. Segmentation of the PCG based on events detected in the ECG and carotid pulse signals is discussed in Section 4.9. A particular type of synchronized averaging to detect A2 in S2 is the topic of Section 4.10. Spectral analysis of the PCG and its applications are presented in Sections 6.2.1, 6.3.6, 6.5, and 7.11. Parametric modeling and detection of S1 and S2 are described in Sections 7.5.2 and 7.10. Modeling of sound generation in stenosed coronary arteries is discussed in Section 7.7.2. Adaptive segmentation of PCG signals with no other reference signal is considered in Section 8.11.

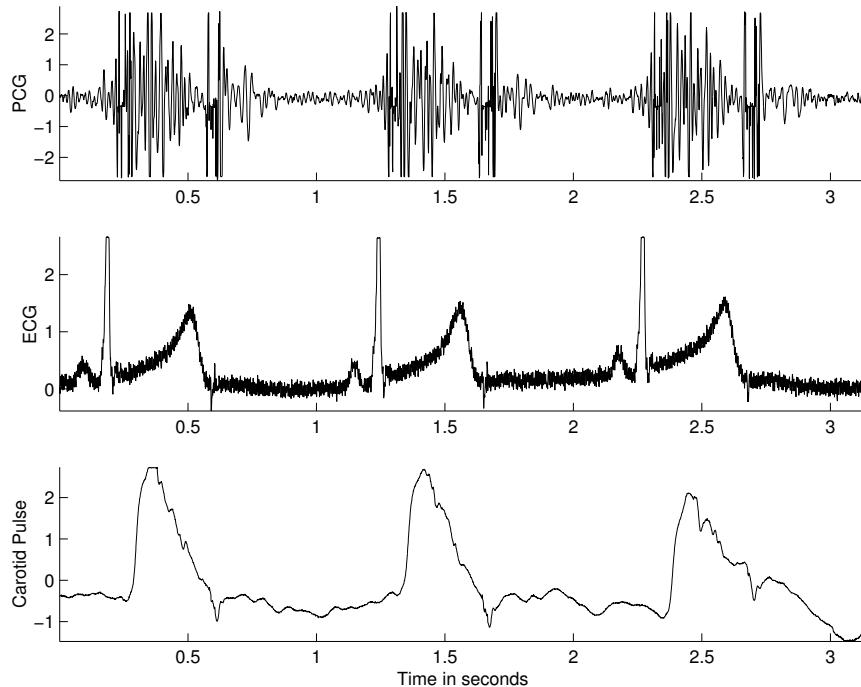


Figure 1.46 Three-channel simultaneous record of the PCG, ECG, and carotid pulse signals of a patient (female, 11 years) with aortic stenosis. Note the presence of the typical diamond-shaped systolic murmur and the split nature of S2 in the PCG.

1.2.10 The carotid pulse

The carotid pulse is a pressure signal recorded over the carotid artery as it passes near the surface of the body at the neck. It provides a pulse signal indicating the variations in arterial BP and volume with each heart beat. Because of the proximity of the recording site to the heart, the carotid pulse signal closely resembles the morphology of the pressure signal at the root of the aorta; however, it

cannot be used to measure absolute pressure [65]. The carotid pulse is a useful adjunct to the PCG and can assist in the identification of S2 and its components.

The carotid pulse rises abruptly with the ejection of blood from the left ventricle to the aorta, reaching a peak called the percussion wave (marked as P in Figure 1.44). This is followed by a plateau or a secondary wave known as the tidal wave (marked as T in Figure 1.44), caused by a reflected pulse returning from the upper body. Next, closure of the aortic valve causes a notch known as the dicrotic notch (marked as D in Figure 1.44). The dicrotic notch may be followed by the dicrotic wave (marked as DW in Figure 1.44) due to a reflected pulse from the lower body [65]. The carotid pulse trace is affected by valvular defects such as mitral insufficiency and aortic stenosis [65]. In the case of aortic stenosis, obstruction of blood flow across the aortic valve leads to prolongation of ejection, which, in turn, causes an abnormally slow initial upstroke of the carotid pulse and longer ejection time [65]. In the case of mitral insufficiency, the dicrotic wave is accentuated [65]. Regardless of some benefits provided by the carotid pulse in the diagnosis of cardiovascular defects and diseases, it is not commonly used in current clinical practice.

The carotid pulse signals shown in this section were recorded using the HP21281A pulse transducer, which has a nominal bandwidth of 0 – 100 Hz. The carotid pulse signal is usually recorded with the PCG and ECG signals. Placement of the carotid pulse transducer requires careful selection of a location on the neck as close to the carotid artery as possible, where the pulse is felt the strongest, usually by a trained technician (see Figure 1.33).

Details on intervals that may be measured from the carotid pulse and their use in segmenting the PCG are presented in Sections 2.2.2 and 2.3. Signal processing techniques for the detection of the dicrotic notch are described in Section 4.3.5. The use of the dicrotic notch for segmentation of PCG signals is explored in Sections 4.9 and 4.10. The use of the carotid pulse to average PCG spectra in systole and diastole is described in Section 6.3.6.

1.2.11 The photoplethysmogram (PPG)

The term plethysmography is used to indicate techniques used to measure volume and related changes in a part of the body of interest [70, 71]. Typically, changes over time in the volume of air or blood in a certain part of the body are of interest; examples include measuring changes in the volume of air in the lungs to study respiration and study of flow of blood in the arteries in the legs to diagnose atherosclerosis (hardening of the arteries that impedes flow of blood). Impedance plethysmography involves the measurement of the electrical conductivity or resistance of a part of the body; variations in the impedance may then be related to changes in volume of the body part or its fluid content [70, 71].

An optically produced plethysmogram called a photoplethysmogram (PPG) [72] can be used to identify changes in blood volume in a microvascular bed of tissue. A PPG signal can be recorded noninvasively by attaching a pulse oximeter device at the skin surface of an earlobe or a fingertip. Figure 1.47 shows plots of PPG signals obtained from multiple bilateral sites from a patient with unilateral peripheral artery occlusive disease of the lower extremity. The PPG signals illustrate the damping, relative delay between the legs, and reduction in amplitude of the affected side's toe pulse. The bilateral similarity of the ear and finger PPG signals is indicative of the absence of severe proximal artery dysfunction [72].

The PPG signal consists of pulsatile and superimposed (low-frequency or DC) components. The cardiac-synchronized fluctuations in blood volume caused by heart beats contribute to the AC component. Sympathetic nervous system (SNS) activity, respiration, and thermoregulation influence the DC component. The PPG can be used to monitor respiration as well as hypovolemia and other circulatory disorders because of their influence on the blood flow under the skin.

In terms of the physical principles, the interaction of light with biological tissues causes reflection, scattering, and absorption, resulting in certain PPG signal characteristics. Melanin significantly absorbs light with shorter wavelengths; red and near-infrared light easily flow through water,

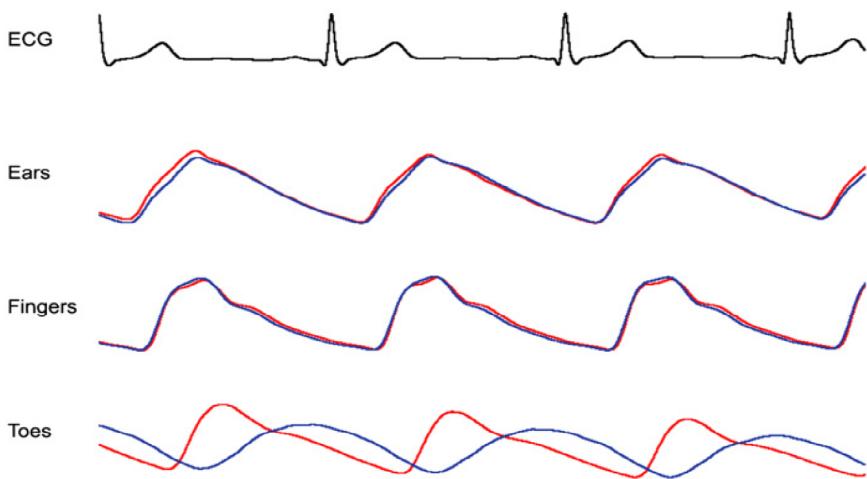


Figure 1.47 PPG signals obtained from multiple body sites (left and right earlobes, index fingers, and toes) of a patient; the corresponding ECG is also shown. Reproduced with permission from J. Allen, Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3), p.R1, 2007. ©IOP.

but ultraviolet and longer infrared light are absorbed. Therefore, PPG sensors typically use infrared wavelengths. A photodetector located opposite the light emitting diode (LED) source detects light that is transmitted through the material in the transmission mode, whereas in the reflection mode, the photodetector detects light that is backscattered or reflected off tissue, bone, and/or blood vessels. Although the transmission mode can get a usable signal, suitable measurement sites might be limited: the fingertip, nasal septum, face, tongue, and earlobe are examples of body parts where transmitted light can be easily detected and where the sensors may be easily placed for optimal performance. Effective PPG acquisition at the nasal septum, cheek, or tongue may require a topical anesthetic. The preferred PPG monitoring locations are the earlobe and fingertip, but there is little blood flow at these locations. Furthermore, the earlobe and fingertip are sensitive to the environment, such as low temperature. A fingertip sensor's interference with regular activates is its biggest drawback.

The reflection mode obviates sensor location issues, allowing for the utilization of numerous measurement sites. However, motion artifacts and pressure anomalies affect the PPG in the reflection mode. Any movement, including physical activity, can produce motion artifacts that corrupt the PPG signal and reduce the precision with which physiological parameters may be measured. Compressional deformation of the underlying arteries can occur as a result of the pressure or contact force between a PPG sensor and the measurement site. As a result, the pressure applied to the skin may have an impact on the AC component of the reflected PPG signal.

Clinical applications of PPG signals [72] include measurement of blood oxygen saturation [73], estimation of the heart rate [74, 75], measurement of BP [76], and diagnosis of arterial disease [77]. PPG signals may also be used to measure respiratory activity without the need for an additional sensor. Figure 1.48 shows a PPG signal, the respiratory signal obtained using a strain gauge in a chest belt, and the respiratory signal derived from the PPG signal [78]. Madhav et al. [78] showed that the PPG-derived respiratory signal had low error values, high correlation, and similar morphological characteristics as compared to the respiratory signal obtained using a strain gauge on the chest. The PPG signal is useful in monitoring sleep apnea (see Section 2.4). Wavelet analysis of PPG is Section 8.14.

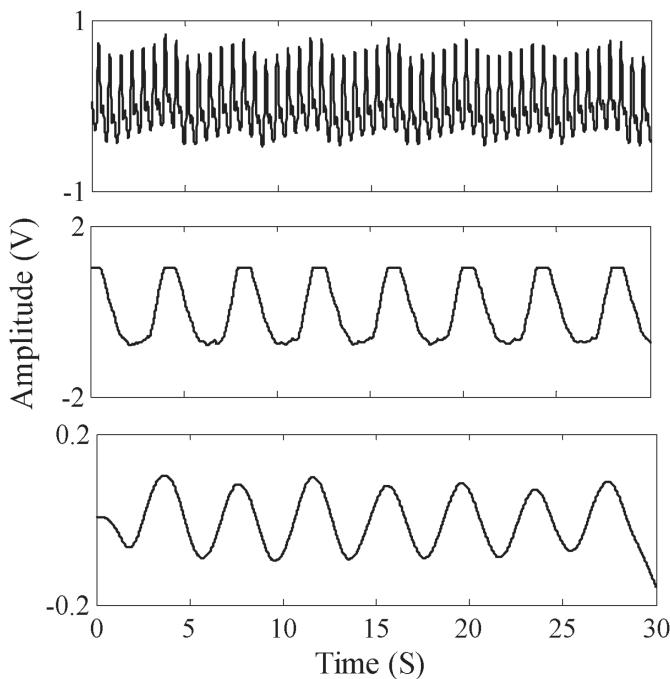


Figure 1.48 From top to bottom: PPG signal; respiratory signal obtained using a strain gauge in a chest belt; respiratory signal derived from the PPG signal. Reproduced with permission from K.V. Madhav, M.R. Ram, E.H. Krishna, N.R. Komalla, and K.A. Reddy, Robust extraction of respiratory activity from PPG signals using modified MSPCA. *IEEE Transactions on Instrumentation and Measurement*, 62(5), pp 1094-1106, 2013 ©IEEE.

1.2.12 Signals from catheter-tip sensors

For specific and detailed monitoring of cardiac function, sensors placed on catheter tips may be inserted into the cardiac chambers. It then becomes possible to acquire several signals such as left ventricular pressure, right atrial pressure, aortic pressure, and intracardiac sounds [67, 68]. While these signals provide valuable and accurate information, the procedures are invasive and are associated with certain risks.

Figures 1.49 and 1.50 illustrate multichannel aortic, left ventricular, and right ventricular pressure recordings from a dog using catheter-tip sensors. The ECG signal is also shown. Observe in Figure 1.49 that the right ventricular and left ventricular pressures increase exactly at the instant of each QRS complex. The aortic pressure peaks slightly after the corresponding increase in the left ventricular pressure. The notch (incisura) in the aortic pressure signal is due to closure of the aortic valve. (The same notch propagates through the vascular system and appears as the dicrotic notch in the carotid pulse signal.) The left ventricular pressure range ($10 - 110 \text{ mm of Hg}$) is much larger than the right ventricular pressure range ($5 - 25 \text{ mm of Hg}$). The aortic pressure range is limited to the vascular BP range of $80 - 120 \text{ mm of Hg}$.

The signals in Figure 1.50 display the effects of PVCs. Observe the depressed ST segment in the ECG signal in the figure, likely due to myocardial ischemia. (It should be noted that the PQ and ST segments of the ECG signal in Figure 1.49 are isoelectric, even though the displayed values indicate a nonzero level. On the other hand, in the ECG in Figure 1.50, the ST segment stays below the corresponding isoelectric PQ segment.) The ECG complexes appearing just after the 2 s and 3 s markers are PVCs arising from different ectopic foci, as evidenced by their markedly different waveforms. Although the PVCs cause a less-than-normal increase in the left ventricular pressure,

they do not cause a rise in the aortic pressure, as not much blood is effectively pumped out of the left ventricle during the ectopic beats.

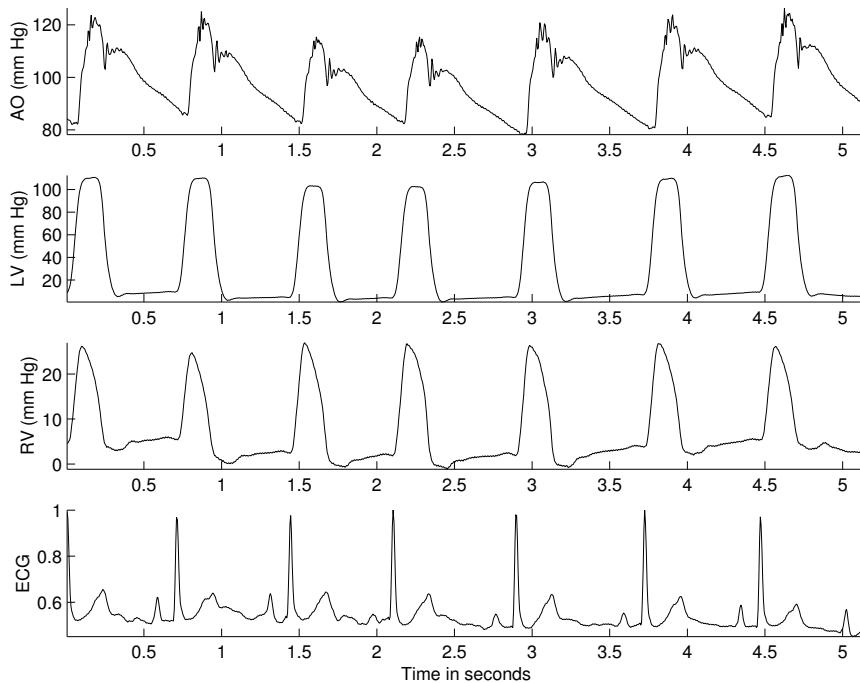


Figure 1.49 Normal ECG and intracardiac pressure signals from a dog. The pressure signals shown are AO, near the aortic valve; LV, in the left ventricle; and RV, in the right ventricle. Data courtesy of R. Sas and J. Tyberg, Department of Physiology and Biophysics, University of Calgary.

1.2.13 The speech signal

Human beings are social creatures by nature and have an innate need to communicate. Humans are endowed with a sophisticated vocal system. The speech signal is an important signal, although it is more commonly considered as a communication signal than a biomedical signal. However, the speech signal can serve as a diagnostic signal when speech and vocal-tract disorders need to be investigated [79, 80].

Speech sounds are produced by transmitting puffs of air from the lungs through the oral tract as well as the nasal tract for certain sounds [81]. Figure 1.51 shows a schematic diagram of the anatomy of the vocal tract; Figure 1.52 gives a schematic representation of the parts of the speech production system. The vocal tract starts at the vocal cords or glottis in the throat and ends at the lips and the nostrils. The shape of the vocal tract is varied to produce different types of sound units or *phonemes* which, when concatenated, form speech. In essence, the vocal tract acts as a filter that modulates the spectral characteristics of the input puffs of air. It is evident that the system is dynamic; therefore, the filter and the speech signal produced have time-varying characteristics, that is, they are nonstationary (see Section 3.2.4).

Speech sounds may be classified as voiced, unvoiced, and plosive sounds [81]. Voiced sounds involve the participation of the glottis: air is forced through the vocal cords held at a certain tension. The vocal cords vibrate and the result is a series of quasiperiodic pulses of air which is passed through the vocal tract; see Figure 1.53. The input to the vocal tract may be treated as a train of impulses or pulses that is almost periodic. The vocal tract acts as a filter: Upon convolution

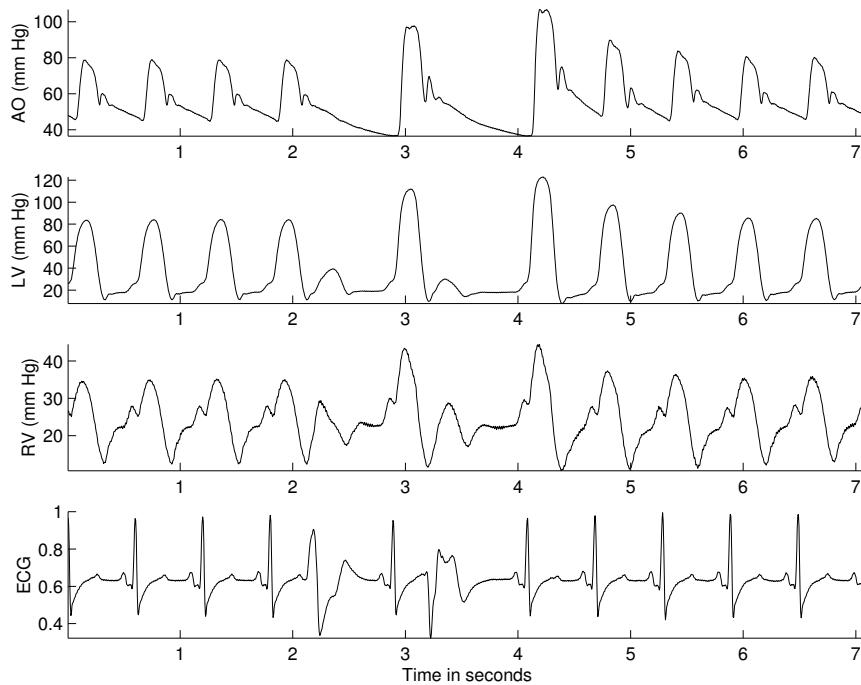


Figure 1.50 ECG and intracardiac pressure signals from a dog with PVCs. The pressure signals shown are AO, near the aortic valve; LV, in the left ventricle; and RV, in the right ventricle. Data courtesy of R. Sas and J. Tyberg, Department of Physiology and Biophysics, University of Calgary.

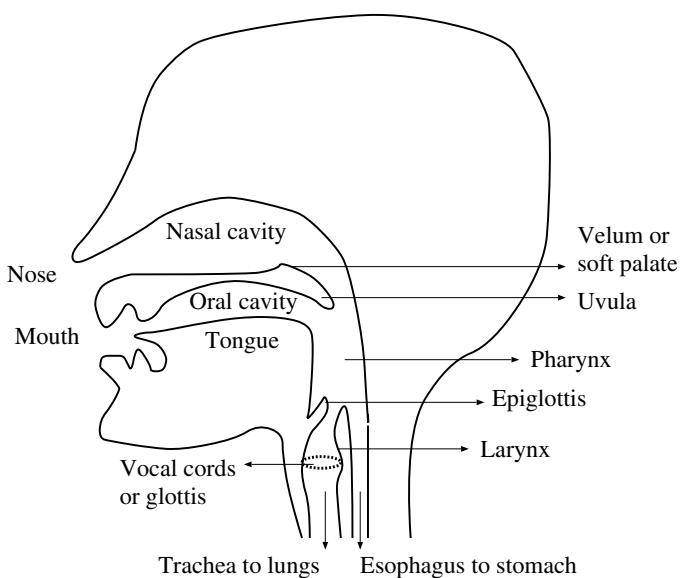


Figure 1.51 Schematic diagram of the anatomy of the vocal tract.

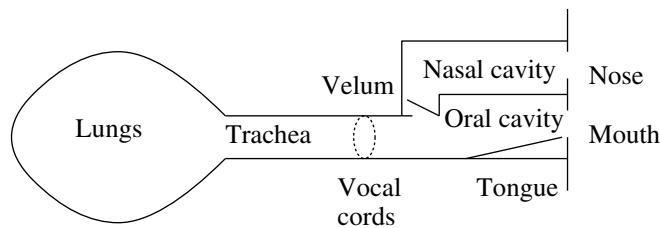


Figure 1.52 Schematic representation of the speech production system.

with the impulse response of the vocal tract, which is held steady in a certain configuration for the duration of the voiced sound desired, a quasiperiodic signal is produced with a characteristic waveshape that is repeated. All vowels are voiced sounds. Figure 1.54 shows the speech signal of the word “safety” spoken by a male subject. Figure 1.55 shows, in the upper trace, a portion of the signal corresponding to the phoneme /E/ (the letter “a” in the word). The quasiperiodic nature of the signal is evident. Features of interest in voiced signals are the pitch (average interval between the repetitions of the vocal-tract impulse response or basic wavelet) and the resonance or formant frequencies of the vocal-tract system.

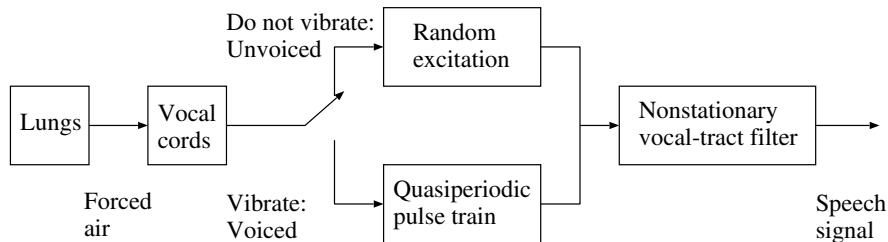


Figure 1.53 Schematic representation of the production of voiced and unvoiced speech.

An unvoiced sound (or fricative) is produced by forcing a steady stream of air through a narrow opening or constriction formed at a specific position along the vocal tract. The vocal cords do not vibrate for such sounds; see Figure 1.53. The result is a turbulent signal that appears like random noise. The input to the vocal tract is a broadband random signal, which is filtered by the vocal tract to yield the desired sound. Fricatives are unvoiced sounds, as they do not involve any activity (vibration) of the vocal cords. The phonemes /S/, /SH/, /Z/, and /F/ are examples of fricatives. The lower trace in Figure 1.55 shows a portion of the speech signal corresponding to the phoneme /S/ in the word “safety.” The signal has no identifiable structure and appears to be random (see also Figures 3.1, 3.3, and 3.4, as well as Section 3.2.4). The transfer function of the vocal tract, as evidenced by the Fourier spectrum of the signal itself, would be of interest in analyzing a fricative.

Plosives, also known as stops, involve complete closure of the vocal tract, followed by an abrupt release of built-up pressure. The phonemes /P/, /T/, /K/, and /D/ are examples of plosives. The sudden burst of activity at about 1.1 s in Figure 1.54 illustrates the plosive nature of /T/. Plosives are difficult to characterize as they are transients; their properties are affected by the preceding phoneme as well. For more details on the speech signal, see Rabiner and Schafer [81].

Parkinson’s disease, which causes tremor, rigidity, and loss of muscle control, is also known to affect speech. The changes in speech caused by the disease include reduced loudness, increased vocal tremor, and breath-related noise. Vocal impairment caused by the disease are labeled as dysphonia, which refers to the inability to produce normal vocal sounds, and dysarthria, which relates to difficulty in pronouncing words [80]. Tsanas et al. [80] and Rueda et al. [82] describe several techniques for the analysis of speech signals for the classification of Parkinson’s disease. Maciel

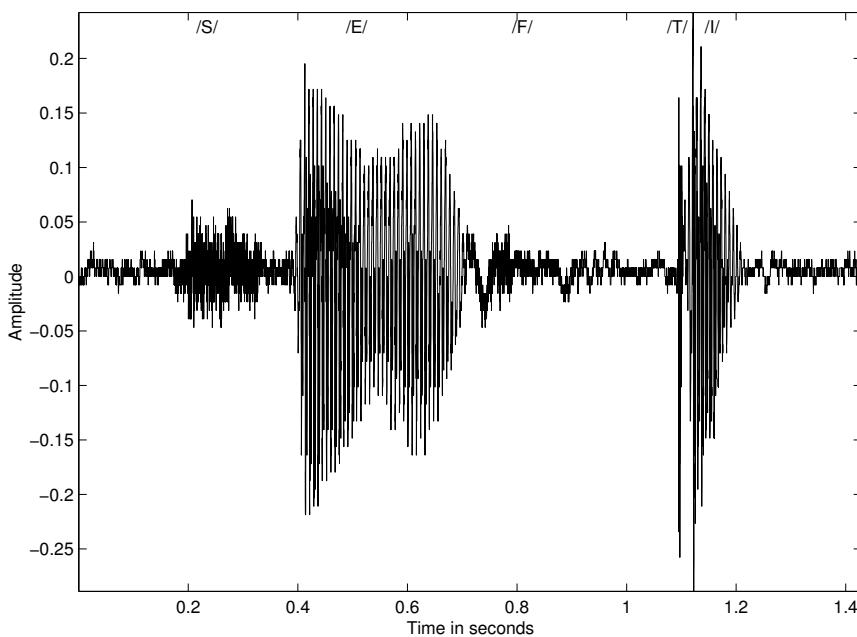


Figure 1.54 Speech signal of the word “safety” uttered by a male speaker. Approximate time intervals of the various phonemes in the word are /S/: 0.2 – 0.35 s; /E/: 0.4 – 0.7 s; /F/: 0.75 – 0.95 s; /T/: transient at 1.1 s; /I/: 1.1 – 1.2 s. Background noise is also seen in the signal before the beginning and after the termination of the speech segment, as well as during the stop interval before the plosive /T/.

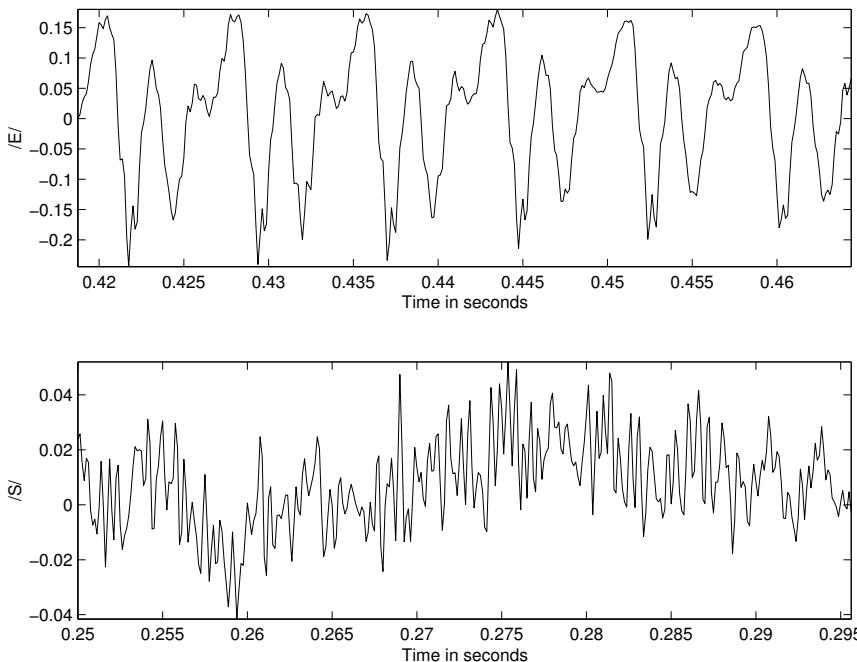


Figure 1.55 Segments of the speech signal in Figure 1.54 on an expanded scale to illustrate the quasiperiodic nature of the voiced sound /E/ in the upper trace and the almost random nature of the fricative /S/ in the lower trace.

et al. [83] describe techniques used to derive diagnostically useful parameters from speech signals. In the case of voiced speech segments, some of such measures represent temporal or spectral variations over time, perturbation of pitch, perturbation of frequency, and jitter. While most methods for the analysis of speech abnormalities use signals of sustained vowels, Umapathy et al. [84] proposed methods to decompose continuous speech signals and extract features for classification of speech pathology. See also Umapathy and Krishnan [85], Ghoraani et al. [86], and Arias-Londoño et al. [87] for additional related material.

Signal processing techniques for extraction of the vocal-tract response from voiced speech signals are described in Section 4.7.3. Frequency-domain characteristics of speech signals are illustrated in Sections 7.6.3 and 8.4.1. See Section 10.14 for discussions on multimodal signal analysis procedures for the diagnosis of Parkinson’s disease.

1.2.14 The vibroarthrogram (VAG)

Several processes related to aging and arthritis cause deterioration of various joints in the body, accompanied by pain and sounds. The following paragraphs provide descriptions of the knee joint, a few related disorders, and the nature of the related sounds.

The knee joint: As illustrated in Figure 1.56, the knee joint is formed between the femur, the patella, and the tibia. The knee joint is the largest articulation in the human body that can effectively move from 0° extension to 135° flexion, together with 20° to 30° rotation of the flexed leg on the femoral condyles. The joint has four important features: a joint cavity, articular cartilage, a synovial membrane, and a fibrous capsule [88, 89]. The knee joint is known as a synovial joint, as it contains a lubricating substance called the synovial fluid. The patella (knee cap), a sesamoid bone, protects the joint, and is precisely aligned to slide in the groove (trochlea) of the femur during leg movement. The knee joint is made up of three compartments: the patellofemoral, the lateral tibiofemoral, and the medial tibiofemoral compartments. The patellofemoral compartment is classified as a synovial gliding joint, and the tibiofemoral as a synovial hinge joint [90]. The anterior and posterior cruciate ligaments as well as the lateral and medial ligaments bind the femur and tibia together, give support to the knee joint, and limit movement of the joint. The various muscles around the joint help in the movement of the joint and contribute to its stability.

The knee derives its physiological movement and its typical rolling–gliding mechanism of flexion and extension from its six degrees of freedom: three in translation and three in rotation. The translations of the knee take place on the anterior–posterior, medial–lateral, and proximal–distal axes. The rotational motion consists of flexion–extension, internal–external rotation, and abduction–adduction.

Although the tibial plateaus are the main load-bearing structures in the knee, the cartilage, menisci, and ligaments also bear loads. The patella aids knee extension by lengthening the lever arm of the quadriceps muscle throughout the entire range of motion, and allows a better distribution of compressive stresses on the femur [91].

Articular cartilage: Two types of cartilage are present in the knee joint: the *articular cartilage*, which covers the ends of bones, and the wedge-shaped fibrocartilaginous structure called the *menisci*, located between the femur and the tibia [92]. The shock-absorbing menisci are composed of the medial meniscus and the lateral meniscus, which are two crescent-shaped plates of fibrocartilage that lie on the articular surface of the tibia.

The articular surfaces of the knee joint are the large curved condyles of the femur, the flattened condyles (medial and lateral plateaus) of the tibia, and the facets of the patella. There are three types of articulation: an intermediate articulation between the patella and the femur, as well as lateral and medial articulation between the femur and the tibia. The articular surfaces are covered by cartilage, as in all of the major joints of the body. Cartilage is vital to joint function because it protects the underlying bone during movement. Loss of cartilage function leads to pain, decreased mobility, and, in some instances, deformity and instability.

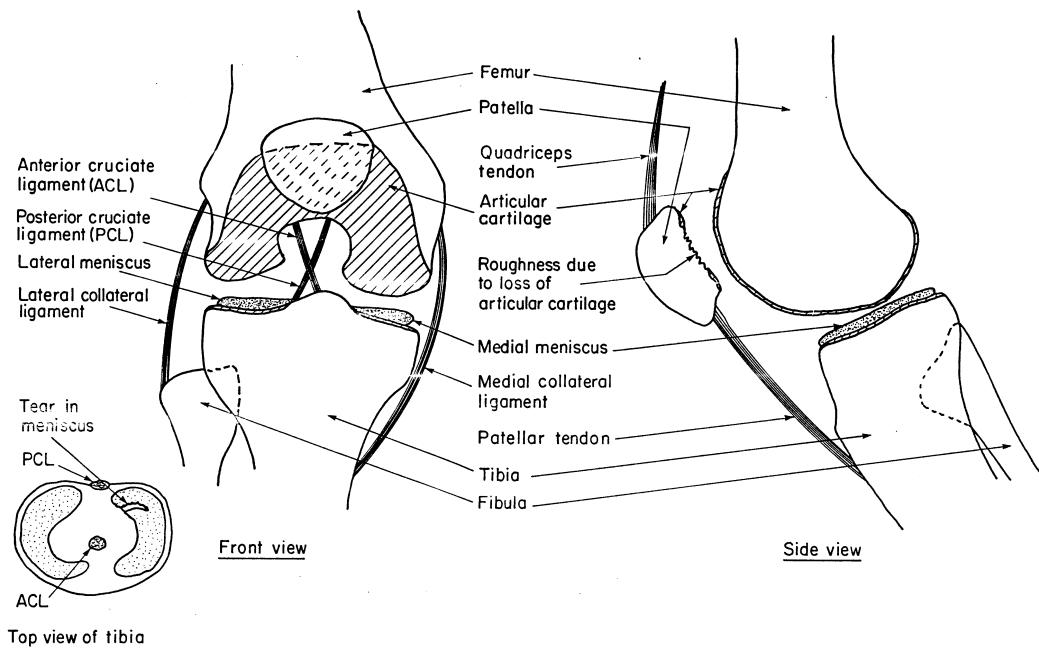


Figure 1.56 Front and side views of the knee joint (the two views are not mutually orthogonal). The inset shows the top view of the tibia with the menisci.

Knee-joint disorders: The knee is a commonly injured joint in the body. Arthritic degeneration of injured knees is a well-known phenomenon and is known to result from a variety of traumatic causes. Damage to the stabilizing ligaments of the knee or to the shock-absorbing fibrocartilaginous pads (the menisci) are two of the most common causes of deterioration of knee-joint surfaces. Impact trauma to the articular cartilage surfaces could lead to surface deterioration and secondary osteoarthritis.

Nontraumatic conditions of the knee joint include the common idiopathic condition known as chondromalacia patella (soft cartilage of the patella), in which articular cartilage softens, fibrillates, and sheds off the undersurface of the patella. Similarly, the meniscal fibrocartilage of the knee can soften, which could lead to degenerative tears and secondary changes in the regional hyaline surfaces; see Figure 8.2 for related illustrations.

Knee-joint sounds: Considerable noise is often associated with degeneration of knee-joint surfaces. The VAG is a vibration signal recorded from a joint during movement (articulation) of the joint. Normal joint surfaces are smooth and produce little or no sound, whereas joints affected by osteoarthritis and other degenerative diseases may have suffered cartilage loss and produce palpable and audible grinding vibration and sound.

Figure 1.57 shows the experimental setup used by Rangayyan et al. [93] to record VAG signals. In their studies, the subject sat on a rigid table in a relaxed position with the leg to be tested being freely suspended in air. An accelerometer (model 3115a, Dytran, Chatsworth, CA) was placed at the midpatella (knee cap) position on the surface of the knee joint. The VAG signal was recorded as the subject swung the leg over an approximate angle range of 135° (approximately full flexion) to 0° (full extension) and back to 135° in 4 s. The first half of each VAG signal corresponds, approximately, to extension, and the second half corresponds to flexion of the leg. The VAG signals were prefiltered to the bandwidth of 10 Hz to 1 kHz and digitized at the sampling rate of 2.5 kHz.

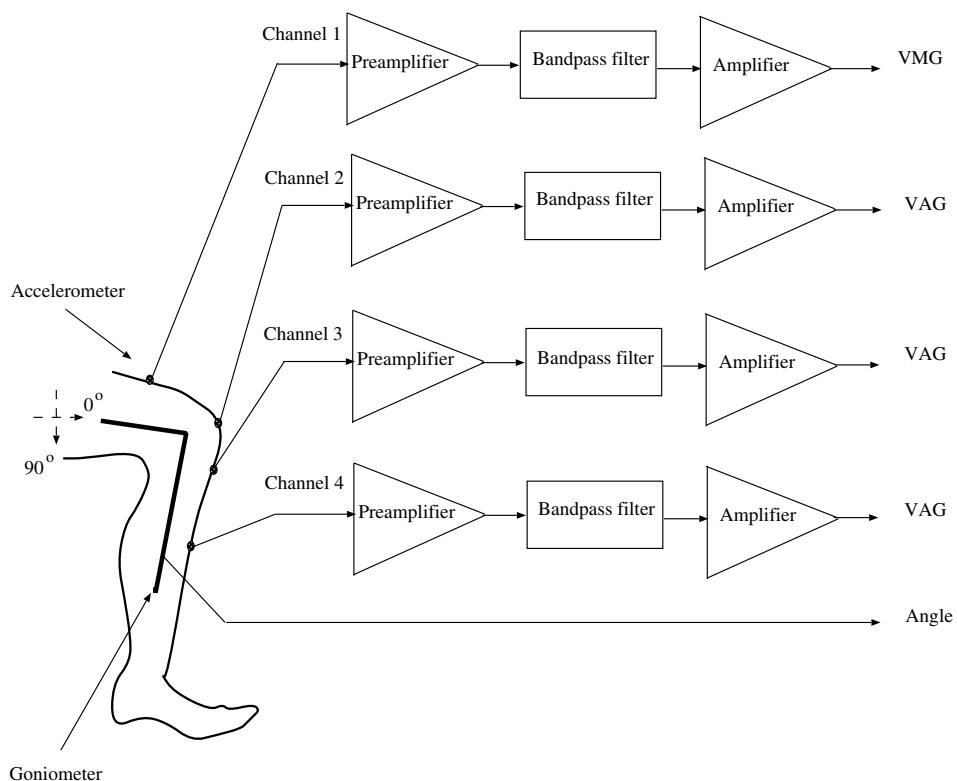


Figure 1.57 Experimental setup to measure VAG and the related muscle-contraction interference signals at various positions along the leg [93, 94]. Channels 1 to 4 correspond, in order, to the distal rectus femoris (thigh), midpatella (knee cap), tibial tuberosity, and midtibial shaft positions.

In related studies by Zhang et al. [94], additional accelerometers were placed at the distal rectus femoris (thigh), tibial tuberosity, and midtibial shaft positions to study the possibility of muscle-contraction interference in VAG signals; see Figure 1.57 as well as Sections 1.2.15, 2.2.7, and 3.3.6. The muscle-contraction or vibration signal from the rectus femoris was referred to as the vibromyogram (VMG).

The left-hand column in Figure 1.58 shows VMG signals recorded at the distal rectus femoris (thigh), midpatella, tibial tuberosity, and midtibial shaft positions of a subject during isometric contraction of the rectus femoris muscle (with no leg or knee movement). The right-hand column of the figure shows VAG signals recorded at the same positions using the same accelerometers, but during isotonic contraction (swinging movement of the leg). The top signal (a) in the right-hand column indicates the VMG signal generated at the rectus femoris during acquisition of the VAG signals; parts (b)–(d) of the right-hand column show the VAG signals.

Rangayyan et al. [93] observed substantial differences between normal and abnormal VAG signals, with the latter including grinding and clicking sounds; see also Frank et al. [95] and Wu et al. [96]. VAG signals are difficult to analyze because they have no well-defined or recognizable waveforms and have complex nonstationary characteristics. Detection of knee-joint problems via the analysis of VAG signals could help avoid unnecessary exploratory surgery and also aid better selection of patients who would benefit from surgery [93, 95, 97–102]. Further details on the VAG signal are provided in Sections 2.2.7, 3.3.6, 5.12.1, and 8.2.3. Modeling of a specific type of VAG signal known as patellofemoral crepitus is presented in Sections 7.2.4, 7.3, and 7.7.3. Adaptive filtering of the VAG signal to remove muscle-contraction interference is described in Sections 3.10.2,

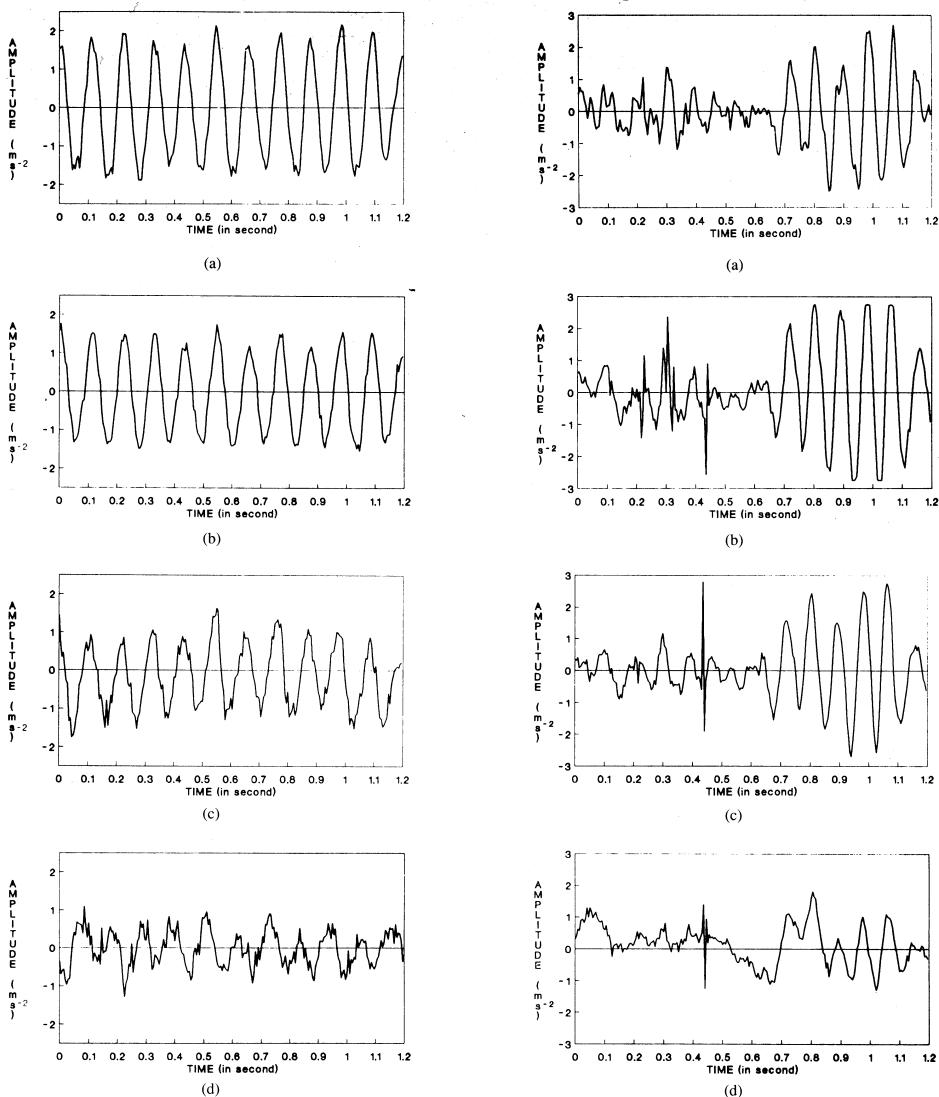


Figure 1.58 Left-hand column: VMG signals recorded simultaneously at (top-to-bottom) (a) the distal rectus femoris, (b) midpatella, (c) tibial tuberosity, and (d) midtibial shaft positions during isometric contraction (no leg or knee movement). Right-hand column: Vibration signals recorded simultaneously at the same positions as above during isotonic contraction (swinging movement of the leg). Observe the muscle-contraction interference appearing in the extension parts (second halves) of each of the VAG signals [plots (b)–(d)] in the right-hand column [94]. The recording setup is shown in Figure 1.57. Reproduced with permission from Y.T. Zhang, R.M. Rangayyan, C.B. Frank, and G.D. Bell, Adaptive cancellation of muscle-contraction interference from knee joint vibration signals, *IEEE Transactions on Biomedical Engineering*, 41(2):181–191, 1994. ©IEEE.

3.10.3, and 3.15. Adaptive segmentation of VAG signals into quasistationary segments is illustrated in Sections 8.6.1 and 8.6.2. The role of VAG signal analysis in the detection of articular cartilage pathology is discussed in Sections 5.12, 6.6, 9.9, and 10.12.

1.2.15 The vibromyogram (VMG)

The VMG is the direct mechanical manifestation of contraction of a skeletal muscle and is a vibration signal that accompanies the EMG. The signal has also been named the sound myogram, acoustic myogram, or phono-myogram. Muscle sounds or vibrations are related to variations in the dimensions (contraction) of the constituent muscle fibers (see Figure 1.8) and may be recorded using contact microphones or accelerometers placed on the muscle surface [103, 104]; see Section 1.2.14. The frequency and intensity of the VMG have been shown to vary in direct proportion to the contraction level [103, 104]. The VMG, along with the EMG, may be useful in studies related to neuromuscular control, muscle contraction, athletic training, biofeedback, and rehabilitation. VMG signal analysis, however, is not as well established or common as EMG analysis.

Simultaneous analysis of the VMG and EMG signals is discussed in Section 2.2.6. Adaptive cancellation of the VMG from knee-joint vibration signals is the topic of Sections 3.10.2, 3.10.3, and 3.15. Analysis of muscle contraction using the VMG is described in Section 5.11.

1.2.16 Otoacoustic emission (OAE) signals

The OAE signal represents the acoustic energy emitted by the cochlea either spontaneously or in response to an acoustic stimulus. The existence of this signal indicates that the cochlea not only receives sound but also produces acoustic energy [105]. The OAE signal could provide objective information on the micromechanical activity of the preneural or sensory components of the cochlea that are distal to the nerve-fiber endings. Analysis of the OAE signal could lead to improved noninvasive investigative techniques to study the auditory system. The signal may also assist in screening of hearing function and in the diagnosis of hearing impairment.

1.2.17 Bioacoustic signals

Several systems and parts of the human body produce sounds and vibrations in various bands of frequencies under both normal physiological and pathological conditions. In the preceding sections, we have studied the PCG, VMG, VAG, and OAE signals. A few other bioacoustic signals that have been studied by several researchers are breathing, tracheal, lung, and chest sounds [106–114]; snoring sounds [115, 116]; swallowing sounds [117]; gastrointestinal or bowel sounds [118]; sounds of the shoulder, temporomandibular, and hip joints [119]; cough sounds [120]; and crying sounds of infants [121]. Kompis et al. [108] describe methods for acoustic imaging of the chest using multiple transducers. Kaniunas et al. [122] proposed techniques for integrated analysis of sound signals related to the cardiac system, respiration, and snoring. See Section 8.13 for a discussion on analysis of crying sounds of infants.

1.3 Objectives of Biomedical Signal Analysis

The representation of biomedical signals in electronic form facilitates computer processing and analysis of the data. Figure 1.59 illustrates the typical steps and processes involved in computer-aided diagnosis (CAD) and therapy based on biomedical signal analysis.

Some of the major objectives of biomedical instrumentation and signal analysis [13, 14, 123–126] are the following:

- *Information gathering* — measurement and analysis of phenomena to interpret a system.

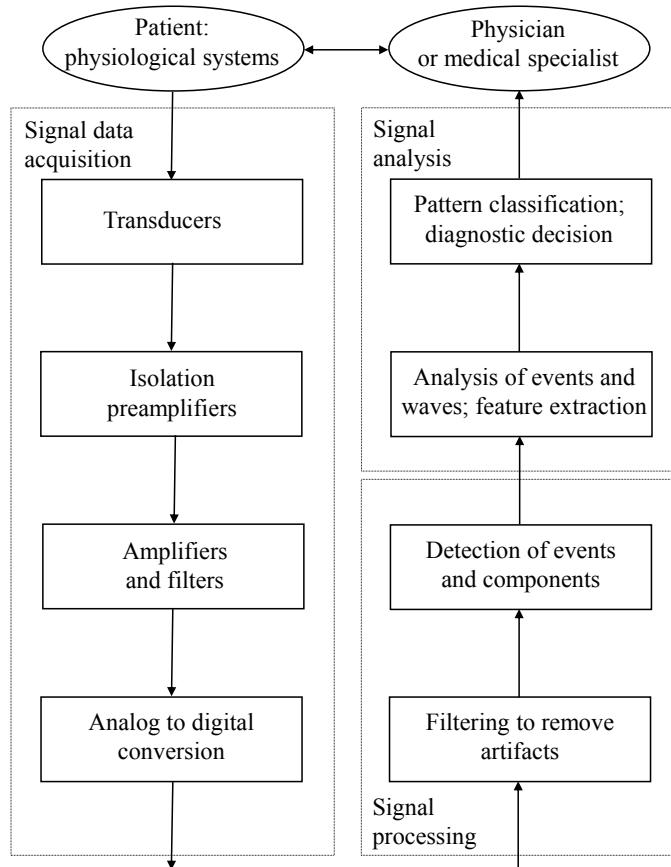


Figure 1.59 Computer-aided diagnosis and therapy based on biomedical signal analysis.

- *Diagnosis* — detection of malfunction, pathology, or abnormality.
- *Monitoring* — obtaining continuous or periodic information about a system.
- *Therapy and control* — modification of the behavior of a system based on the outcome of the activities listed above to ensure a specific result.
- *Evaluation* — objective analysis to determine the ability to meet functional requirements, obtain proof of performance, perform quality control, or measure the effect of treatment.

Signal acquisition procedures may be categorized as being invasive or noninvasive, and active or passive.

Invasive versus noninvasive procedures: Invasive procedures involve the placement of transducers or other devices inside the body, such as needle electrodes to record MUAPs, or insertion of catheter-tip sensors into the heart to record intracardiac signals. Noninvasive procedures are desirable in order to minimize risk to the subject. Recording of the ECG using limb or chest electrodes, the EMG with surface electrodes, or the PCG with microphones or accelerometers placed on the chest are noninvasive procedures.

Note that making measurements or imaging with X rays or ultrasound may be classified as invasive procedures, as they involve penetration of the body with externally administered radiation, even though the radiation is invisible and there is no visible puncturing or invasion of the body.

Active versus passive procedures: Active data acquisition procedures require external stimuli to be applied to the subject, or require the subject to perform a certain activity to stimulate the system of interest in order to elicit the desired response or signal. For example, recording an EMG signal requires contraction of the muscle of interest, such as clenching a fist; recording the VAG signal from the knee requires flexing of the leg over a certain joint angle range; and recording visual ERP signals requires the delivery of flashes of light to the subject. While these stimuli may appear to be innocuous, they do carry risks in certain situations for some subjects: Flexing the knee beyond a certain angle may cause pain for some subjects, and strobe lights may trigger epileptic seizures in some subjects. The investigator should be aware of such risks, factor them in a *risk–benefit analysis*, and be prepared to manage adverse reactions.

Passive procedures do not require the subject to perform any activity. Recording of the ECG using limb or chest electrodes, the EEG during sleep using scalp-surface electrodes, or the PCG with microphones or accelerometers placed on the chest are passive procedures, but require contact between the subject and the transducers or instruments. Note that although the procedure is passive, the system of interest is active under its own natural control in these procedures. Acquiring an image of a subject with reflected natural light (with no flash from the camera) or with the natural infrared (thermal) emission could be categorized as a passive and noncontact procedure.

Most organizations require ethics approval by specialized committees for experimental procedures involving human or animal subjects, with the aim of minimizing the risk and discomfort to the subject and maximizing the benefits to both the subjects and the investigator.

The human–instrument system: Some of the components of a *human–instrument system* [13, 14, 123–126] are the following:

- *The subject or patient:* It is important always to bear in mind that the main purpose of biomedical instrumentation and signal analysis is to provide a certain benefit to the subject or patient. All systems and procedures should be designed so as not to cause undue inconvenience to the subject and not to cause any harm or danger. In applying invasive or risky procedures, it is extremely important to perform a risk–benefit analysis and determine if the anticipated benefits of the procedure are worth placing the subject at the risks involved.
- *Stimulus or procedure of activity:* Application of stimuli to the subject in active procedures requires instruments such as strobe light generators, sound generators, and electrical pulse generators. Procedures in which the subject is required to perform some specified actions or activities should include standardized protocols of the desired activities to ensure repeatability and consistency of the experiment and the results. Such procedures require ethics approval by specialized committees.
- *Transducers:* electrodes, sensors.
- *Signal conditioning equipment:* preamplifiers, amplifiers, filters.
- *Display equipment:* oscilloscopes, strip-chart or paper recorders, computer monitors, printers.
- *Recording, data acquisition, data transmission, and data processing equipment:* analog instrumentation tape recorders, analog-to-digital converters (ADCs), digital-to-analog converters (DACs), telemetry systems, wireless data transmission and reception devices, digital tapes, compact disks (CDs), data storage devices, computers.
- *Control devices:* power supply stabilizers and isolation equipment, patient intervention systems.

The science of measurement of physiological variables and parameters is known as *biometrics*. Some of the aspects to be considered in the design, specification, or use of biomedical instruments [13, 14, 123–126] are the following:

- *Isolation of the subject or patient* — Safety is of paramount importance so that the subject is not placed at the risk of electrocution or any other harm.
- *Dynamic range of operation* — The dynamic range is defined as the minimum to the maximum values of the signal or parameter being measured.
- *Sensitivity* — This can be represented by the smallest signal variation measurable and determines the resolution of the system.
- *Linearity* — Linear transfer characteristics are desired over at least a portion of the range of operation. Any nonlinearity present may need to be corrected or taken into account at later stages of signal processing.
- *Hysteresis* — A lag in measurement due to the direction of variation of the entity being measured indicates hysteresis, which may add a bias to the measurement.
- *Frequency response* — This represents the variation of sensitivity with frequency. Most systems encountered in practice exhibit a lowpass behavior, that is, the sensitivity of the system decreases as the frequency of the input signal increases. Signal restoration techniques may be required to compensate for reduced high-frequency sensitivity.
- *Stability* — An unstable system could preclude repeatability and consistency of measurements.
- *SNR* — Artifacts due to power-line interference, grounding problems, thermal noise, and other sources could compromise the quality of the signal being acquired. A good understanding of the signal-degrading phenomena affecting the system is necessary in order to design appropriate filtering and correction procedures.
- *Accuracy* — The accuracy of a measurement could be affected by errors due to component tolerance; movement or mechanical errors; drift due to changes in temperature, humidity, or pressure; reading errors due to parallax; and zeroing or calibration errors.

The scientific and technological aspects of biomedical signal acquisition have advanced substantially over the past few decades. Systems are now available to acquire a few dozen to a few hundred channels of signals; wearable and portable systems are also available to facilitate wireless signal data acquisition from untethered and mobile subjects engaged in various activities [127–131]. Junnila et al. [132] developed a ballistocardiographic (BCG) chair, designed to look like a normal office chair, with a lightweight and flexible electromechanical film sensor to record the vibration signal related to cardiac activity along with the ECG and an impedance cardiogram. Guerrero et al. [133] used multiple pressure sensors integrated into a bed to obtain BCG signals. The envelope of the BCG signal was used to derive information related to respiration and detect sleep-related breathing disorders. Peltokangas et al. [134] describe a system with eight embroidered textile electrodes attached to a bed sheet to measure multiple channels of the ECG and facilitate monitoring of patients during sleep. Mikhelson et al. [135] reported on a system with a millimeter-wave sensor, two cameras, and a pan/tilt base not only to detect and track a subject but also to measure the subject's chest displacement and the heart rate in a noncontact manner. See Krishnan [131] for related discussions.

1.4 Challenges in Biomedical Signal Acquisition and Analysis

In spite of the long history of biomedical instrumentation and its extensive use in healthcare and research, many practical challenges and difficulties are encountered in biomedical signal acquisition, processing, and analysis [13, 14, 123–126]. The characteristics of the problems, and hence

their potential solutions, are unique to each type of signal. Particular attention should be paid to the following issues.

Accessibility of the variables to measurement: Most of the systems and organs of interest, such as the cardiovascular system and the brain, are located well within the body in protective enclosures (for good reasons!). While the ECG may be recorded using limb electrodes, the signal so acquired is only a projection of the true 3D cardiac electrical vector on to the axis of the electrodes. Such a signal may be sufficient for rhythm monitoring, but could be inadequate for more specific analysis of the cardiac system such as atrial electrical activity. Accessing the atrial electrical activity at the source requires the insertion of an electrode close to the atrial surface or within the atria.

Similarly, measurement of BP using a pressure cuff over an arm gives an estimate of the brachial arterial pressure. Detailed study of pressure variations within the cardiac chambers or arteries over a cardiac cycle would require the insertion of catheters with pressure sensors into the heart. Such invasive procedures provide access to the desired signals at their sources and often provide clear and useful signals, but carry high risks.

The surface EMG includes the interference pattern of the activities of several motor units even at low levels of muscular contraction. Acquisition of MUAPs requires access to the specific muscle layer or unit of interest by insertion of fine-wire or needle electrodes. The procedure carries risks of infection and damage to muscle fibers, and causes pain to the subject during muscular activity.

An investigator should assess the system and variables of interest carefully, and determine the minimal level of intervention that is absolutely essential to the data acquisition procedure. A trade-off between the integrity and quality of the information acquired versus the pain and risks to the subject may be needed.

Variability of the signal source: It is evident from the preceding sections that the various systems that comprise the human body are dynamic with several variables or degrees of freedom. Biomedical signals represent the dynamic activity of physiological systems and the states of their constituent variables. The nature of the processes or the variables could be deterministic or random (stochastic); a special case is that of periodicity or quasiperiodicity.

A normal ECG exhibits a regular rhythm with a readily identifiable waveshape (the QRS complex) in each period; under such conditions, the signal may be referred to as a deterministic and periodic signal. However, the cardiovascular system of a heart patient may not stay in a given state over substantial periods, and the waveshape and rhythm may vary over time.

The surface EMG is the summation of the MUAPs of the motor units that are active at the given instant of time. Depending on the level of contraction desired (at the volition of the subject), the number of active motor units varies, increasing with increasing effort. Furthermore, the firing intervals and the firing rate of each motor unit also vary in response to the level of contraction desired, and exhibit stochastic properties. While the individual MUAPs possess readily identifiable and simple monophasic, biphasic, or triphasic waveshapes, the interference pattern of several motor units firing at different rates will appear as an almost random signal with no visually recognizable waves or waveshapes.

The dynamic nature of biological systems causes most signals to exhibit stochastic and non-stationary behavior. This means that signal statistics such as mean, variance, and spectral density change with time. For this reason, signals from a dynamic system should be analyzed over extended periods of time including various possible states of the system, and the results should be placed in the context of the corresponding states.

Interrelationships and interactions among physiological systems: The various systems that compose the human body are not mutually independent; rather, they are interrelated and interact in various ways. Some of the interactive phenomena are compensation, feedback, cause-and-effect, collateral effects, ipsilateral and/or contralateral effects, loading, and take-over of function of a disabled system or part by another system or part. For example, the second heart sound exhibits a split during active inspiration in normal subjects due to reduced intrathoracic pressure and decreased venous return to the left side of the heart [65] (but not during expiration); this is due to normal

physiological processes. However, the second heart sound is split in both inspiration and expiration due to delayed right ventricular contraction in right bundle-branch block, pulmonary valvular stenosis or insufficiency, and other conditions [65]. Ignoring this interrelationship could lead to misinterpretation of the signal. See Chapter 2 for more detailed discussions on this topic.

Effect of the instrumentation or procedure on the system: The placement of transducers on and connecting a system to instruments could affect the performance or alter the behavior of the system, and could cause spurious variations in the parameters being investigated. The experimental procedure or activity required to elicit the signal may lead to certain effects that could alter signal characteristics. This aspect may not always be obvious unless careful attention is paid. For example, the placement of a relatively heavy accelerometer may affect the vibration characteristics of a muscle and compromise the integrity of the vibration or sound signal being measured. Fatigue may set in after a few repetitions of an experimental procedure, and subsequent measurements may not be indicative of the true behavior of the system; the system may need some rest between procedures or their repetitions.

Physiological artifacts and interference: One of the prerequisites for the acquisition of a good ECG signal is for the subject to remain relaxed and still with no movement. Coughing, tensing of muscles, and movement of the limbs cause the corresponding EMG to appear as an undesired artifact. In the absence of any movement by the subject, the only muscular activity in the body would be that of the heart. When chest leads are used, even normal breathing could cause the associated EMG of the chest muscles to interfere with the desired ECG. It should also be noted that breathing causes beat-to-beat variations in the RR interval, which should not be mistaken to be sinus arrhythmia (see Section 2.2.4 for more details and a related illustration). An effective solution would be to record the signal with the subject holding his or her breath for a few seconds. This simple solution is not applicable in long-term monitoring of critically ill patients or in recording the ECG of infants; signal processing procedures would then be required to remove the concomitant artifacts.

A unique situation is that of acquiring the ECG of a fetus through surface electrodes placed over the expectant mother's abdomen: the maternal ECG appears as an interference in this situation. No volitional or external control is possible or desirable to prevent the artifact in this situation, which calls for adaptive cancellation techniques using multiple channels of various signals [136]; see Chapters 3 and 9.

Another example of physiological interference or cross-talk is that of muscle-contraction interference in the recording of the knee-joint VAG signal [94]. The rectus femoris muscle is active (contracting) during the swinging movement of the leg that is required to elicit the knee-joint vibration signal. The VMG of the muscle is propagated to the knee and appears as an interference. Swinging the leg mechanically using a mechanical actuator is a possible solution; however, this represents an unnatural situation and may cause other sound or vibration artifacts from the machine. Adaptive filtering using multichannel vibration signals from various points is a feasible solution [94]; see Chapter 3.

Energy limitations: Most biomedical signals are generated at microvolt or millivolt levels at their sources. Recording such signals requires sensitive transducers and instrumentation with low noise levels. The connectors and cables need to be shielded, in order to obviate pickup of ambient electromagnetic (EM) signals. Some applications may require transducers with integrated amplifiers and signal conditioners so that the signal leaving the subject at the transducer is much stronger than ambient sources of potential interference. A Faraday cage or shield, constructed with a wire mesh or foil of a conducting material to enclose the entire data acquisition setup including the subject and all devices, may be used to prevent external EM waves from contaminating the signals being recorded. Whereas such a cage is feasible in a laboratory, it may be impractical in a clinical or hospital setting.

When external stimuli are required to elicit a certain response from a system, the level of the stimulus is constrained due to safety factors and physiological limitations. Electrical stimuli to

record the ENG need to be limited in voltage level so as to not cause local burns or interfere with the electrical control signals of the cardiac or nervous systems. Auditory and visual stimuli are constrained by the lower thresholds of detectability, and upper thresholds related to frequency response, saturation, or pain.

Patient safety: Protection of the subject or patient from electrical shock or radiation hazards is an unquestionable requirement of paramount importance. The relative levels of any other risks involved should be assessed when a choice is available between various procedures and should be analyzed against their relative benefits. Patient safety concerns may preclude the use of a procedure that may yield better signals or results than others, or require modifications to a procedure that may lead to inferior signals. Review of experimental procedures by ethics committees should consider all of these issues before providing ethics approval. Further signal processing steps would then become essential in order to improve signal quality or otherwise compensate for the initial loss.

In consideration of the various difficulties encountered in the acquisition of biomedical signals, several institutions and agencies have supported efforts to develop public databases of annotated signals. The PhysioNet [137–139] is one such resource that is useful in education and research activities related to biomedical signals.

1.5 Why Use Computer-aided Monitoring and Diagnosis?

Physicians, cardiologists, neuroscientists, and healthcare technologists are highly trained and skilled practitioners. Why then would we want to use computers or electronic instrumentation for monitoring and analysis of biomedical signals? The following points provide some arguments in favor of the application of computers to process and analyze biomedical signals. Further discussions on the strengths and limitations of CAD are provided in Section 10.15.

- Humans are highly skilled and fast in the analysis of visual patterns and waveforms, but are slow in arithmetic operations with large numbers of values. The ECG of a single cardiac cycle (heart beat) could have at least 200 numerical values; the corresponding PCG at least 2,000. If signals need to be processed to remove noise or extract a parameter, it would not be practical for a person to perform such computation. Computers can perform millions of arithmetic operations per second. It should be noted, however, that recognition of waveforms and images using mathematical procedures typically requires huge numbers of operations and procedures that could lead to slow responses in such tasks from low-level computers.
- Humans could be affected by fatigue, boredom, and environmental factors, and are susceptible to committing errors. Long-term monitoring of signals, for example, the heart rate and ECG of a critically ill patient, by a human observer watching an oscilloscope or computer display is neither economical nor feasible. A human observer could be distracted by other events in the surrounding areas and may miss short episodes or transients in the signal. Computers, being inanimate but mathematically accurate and consistent machines, can be designed to perform computationally specific and repetitive tasks.
- Analysis by humans is usually subjective and qualitative. When comparative analysis is required between the signals of a subject and another or a standard pattern, a human observer would typically provide a qualitative response. For example, if the QRS width of the ECG is of interest, a human observer may remark that the QRS of the subject is wider than the reference or normal. More specific or objective comparison to the accuracy of the order of a few milliseconds would require the use of electronic instrumentation or a computer. Derivations of quantitative or numerical features from signals with large numbers of samples would certainly demand the use of computers.

- Analysis by humans is subject to interobserver as well as intraobserver variations. Given that most analyses performed by humans are based on qualitative judgment, they are liable to vary with time for a given observer, or from one observer to another. The former could also be due to lack of diligence or due to inconsistent application of knowledge, and the latter due to variations in training and level of understanding. Computers can apply a given procedure repeatedly and whenever recalled in a consistent manner. It is further possible to encode the knowledge (to be more specific, the logic) of many experts into a single computational procedure, and thereby enable a computer with the collective intelligence of several human experts in the area of interest.
- Most biomedical signals vary slowly over time (that is, they are lowpass signals), with their bandwidth limited to a few tens to a few thousand Hz . Typical sampling rates for digital processing of most biomedical signals range from 100 Hz to a few kHz ; only a few of the signals mentioned in this chapter, such as speech and bioacoustic signals, require higher sampling rates of the order of 5 – 20 kHz . Sampling rates as above facilitate *on-line, real-time* analysis of biomedical signals with even low-end computers. The term “on-line” indicates that the patient or subject is connected to the system or computer that is acquiring and analyzing signals. The term “real-time analysis” may be used to indicate the processing of each sample of a signal before the next sample arrives, or the processing of an epoch or episode such as an ECG beat before the next one is received in its entirety in a buffer. Monitoring the heart rate of critically ill patients demands on-line and real-time ECG analysis. However, some applications do not require on-line, real-time analysis: For example, processing a VAG signal to diagnose cartilage degeneration, or analyzing a long-term ECG record obtained over several hours using an ambulatory system, usually does not demand immediate attention and results. In such cases, computers could be used for *off-line* analysis of prerecorded signals with sophisticated signal processing and time-consuming modeling techniques. The speed required for real-time processing and the computational complexities of modeling techniques in the case of off-line applications would both rule out the possibility of performance of the tasks by humans.

One of the important points to note in the preceding discussion is that *quantitative analysis* becomes possible by the application of computers to biomedical signals. The logic of medical or clinical diagnosis via signal analysis could then be *objectively* encoded and *consistently* applied in routine or repetitive tasks. However, it should be emphasized that the end goal of biomedical signal analysis should be seen as *computer-aided* diagnosis and not automated diagnosis. A physician or medical specialist typically uses a significant amount of information in addition to signals and measurements, including the general physical appearance and mental state of the patient, family history, and socioeconomic factors affecting the patient, many of which are not amenable to quantification and logical rule-based processes. Biomedical signals are, at best, indirect indicators of the state of the patient; most cases lack a direct or unique signal–pathology relationship [45]. The results of signal analysis need to be integrated with other clinical signs, symptoms, and information by a physician. Above all, the *intuition* of the specialist plays an important role in arriving at the final diagnosis. For these reasons, and keeping in mind the realms of practice of various licensed and regulated professions, liability, and legal factors, the final diagnostic decision is best left to the physician or medical specialist. It is expected that quantitative and objective analysis facilitated by the application of computers to biomedical signal analysis will lead to an accurate diagnostic decision by the physician.

The techniques used for and applications of CAD in medicine have been given various names, labels, and attributes, some of which are pattern recognition, pattern classification, machine learning (ML), deep learning (DL), and artificial intelligence (AI). Classical pattern recognition and CAD methods use symbolic representation and expert systems based on established rules and logic; in such a system, the process and reasoning behind a decision or diagnosis can be explained; see Chapter 10. DL systems are based on massive artificial neural networks (ANNs) with multiple lay-

ers (hence the label “deep”) of computational neurons, weights, and synaptic connections; such a system requires tremendous amounts of data for training and learning (referred to as “big data”) as well as enormous computational resources. A typical DL or AI system cannot provide the reasoning behind a certain result and, therefore, is referred to as a “black box” whose internal details are unknown. Whereas the labels CAD and ML accurately depict the nature of the methods of our interest in this discussion, the term “AI” gives a connotation that goes beyond the true nature of the subject matter and the procedures or systems being used. CAD incorporates, encodes, and encapsulates the knowledge, intelligence, and expertise of several professionals from multiple disciplines spanning engineering, science, and medicine. CAD arises from natural human collaborative endeavors: the label “artificial” in AI is demeaning and undermines the contributions of the various professionals involved in the exercise. It is important to use appropriate terminology to recognize, admire, and respect the contributing professionals and their subject areas. Ultimately, in life-and-death matters such as medical diagnosis and healthcare, it is best to have an appropriately qualified human expert in charge.

Further discussion on practical issues related to CAD, as well as the strengths and limitations of CAD systems, is presented in Section 10.15.

1.6 Remarks

We have taken a general look at the nature of biomedical signals in this chapter and have seen a few signals illustrated for the purpose of gaining familiarity with their typical appearance and features. Specific details of the characteristics of the signals and their processing or analysis are provided in the subsequent chapters.

We have also stated the objectives of biomedical instrumentation and signal analysis. Some of the challenges and difficulties that arise in biomedical signal investigation were discussed in order to draw attention to the relevant practical issues. The suitability and desirability of the application of computers for biomedical signal analysis were discussed, with emphasis on objective and quantitative analysis toward the end goal of CAD for improved diagnosis and therapy. The remaining chapters provide descriptions of specific signal processing techniques and applications.

On the importance of quantitative analysis: “When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of *science*.”

— Lord Kelvin (William Thomson, 1824–1907) [140]

On assumptions made in quantitative analysis: “Now, things do not, in general, run around with their measures stamped on them like the capacity of a freight-car: it requires a certain amount of investigation to discover what their measures are ... What most experimenters do take for granted before they begin their experiments is infinitely more important and interesting than any results to which their experiments lead.”

— Norbert Wiener (1894–1964) [141]

1.7 Study Questions and Problems

Note: Some of the questions may require background preparation with other sources on the ECG (for example, Rushmer [25]), the EMG (for example, Goodgold and Eberstein [22]), and biomedical instrumentation (for example, Webster [13]).

1. Give two reasons to justify the use of electronic instruments and computers in medicine.

2. State any two objectives of using biomedical instrumentation and signal analysis.
3. Distinguish between open-loop and closed-loop monitoring of a patient.
4. List three common types or sources of artifact in a biomedical instrument.
5. A nerve cell has an action potential of duration 10 ms including the refractory period. What is the maximum rate (in pulses per second) at which this cell can transmit electrical activity?
6. Consider a myocardial cell with an action potential of duration 300 ms including its refractory period. What is the maximum rate at which this cell can be activated (fired) into contraction?
7. Distinguish between spatial and temporal recruitment of motor units to obtain increasing levels of muscular activity.
8. Consider three motor units with action potentials (SMUAPs) that are of different biphasic and triphasic shapes. Consider the initial stages of contraction of the related muscle. Draw three plots of the net EMG of the three motor units for increasing levels of contraction with the spatial and temporal recruitment phenomena invoked individually and in combination. Assume low levels of contraction, and that the SMUAPs do not overlap.
9. Draw a typical ECG waveform over one cardiac cycle indicating the important component waves, their typical durations, and the typical intervals between them. Label each wave or interval with the corresponding cardiac event or activity.
10. Draw the waveform corresponding to two cycles of a typical ECG signal and indicate the following waves and periods: (a) the P, QRS, and T waves; (b) the RR interval; (c) atrial contraction; (d) atrial relaxation; (e) ventricular contraction; and (f) ventricular relaxation.
11. Explain why the P and T waves are low-frequency signals, whereas the QRS complex is a high-frequency signal. Include diagrams of action potentials and an ECG waveform in your reasoning.
12. Explain the reasons for widening of the QRS complex in the case of certain cardiac diseases.
13. Give two examples that call for the use of electronic instruments and/or computers in ECG analysis.
14. A heart patient has a regular SA node pulse (firing) pattern and an irregular ectopic focus. Over a period of 10 s, the SA node was observed to fire regularly at $t = 0, 1, 2, 3, 4, 5, 6, 7, 8$, and 9 s. The ectopic focus was observed to fire at $t = 1.3, 2.8, 6.08$, and 7.25 s.
Draw two impulse sequences corresponding to the firing patterns of the SA node and the ectopic focus. Draw a schematic waveform of the resulting ECG of the patient. Explain the source of each beat (SA node or ectopic focus) and give reasons.
15. A patient has ventricular bigeminy, where every other pulse from the SA node is replaced by a premature ventricular ectopic beat with a full compensatory pause. (See Figure 10.10 for an illustration of bigeminy.) The firing rate of the SA node is regular at 80 beats a minute, and each ectopic beat precedes the blocked SA node pulse by 100 ms. (a) Draw a schematic trace of the ECG for 10 beats, marking the time scale in detail. (b) Draw a histogram of the RR intervals for the ECG trace. (c) What is the average RR interval computed over the 10 beats?
16. Draw a typical PCG (heart sound signal) waveform over one cardiac cycle indicating the important component waves, their typical durations, and the typical intervals between them. Label each wave or interval with the corresponding cardiac event or activity.
17. Give two examples that require the application of electronic instruments and/or computers in EEG analysis.
18. Distinguish between ECG rhythms and EEG rhythms. Sketch one example of each.
19. What are the causes and characteristics of ventricular ectopic beats (PVCs)? Draw a sample ECG signal including five normal beats and two PVCs. Label parts of the signal indicating atrial contraction, ventricular contraction, and the premature nature of PVCs.
20. Describe two cardiac abnormalities that cause changes in the shape of the QRS complex. Draw the ECG waveforms for the two cases and compare their features with those of a normal ECG waveform.
21. What is the ENG? Describe an experimental procedure to acquire an ENG. Describe the typical result obtained from ENGs and a potential application.

22. Explain the cardiac events that cause the second heart sound (S2). Identify the valves and their actions associated with S2. Explain how the dicrotic notch in the carotid pulse is related to S2.
23. Draw Einthoven's triangle and indicate the axes representing the six leads of the ECG obtained using the four limb leads. Mark the positions of the four limb leads in your diagram.
24. Draw a schematic representation of an ECG signal over one cardiac cycle. Label intervals related to atrial contraction and ventricular contraction. Overlay the ECG signal with schematic representations of the action potentials of atrial and ventricular myocytes. Explain the relationships between the various parts of the action potentials and the ECG signal.
25. Draw a schematic representation of Einthoven's triangle showing the directions (polarities) of leads I, II, and III of the ECG signal. Rearrange the vectors and derive the relationship between the three leads I, II, and III using vectorial arithmetic.
26. Consider the contraction of a muscle at a low level of effort. Suppose that two motor units are active. Assume that one of the motor units has a biphasic action potential and is firing at the rate of 10 pps. Assume that the other motor unit has a triphasic action potential and is firing at the rate of 12 pps. For a duration of 0.5 s, draw a schematic sketch of the action potential trains for the two motor units individually and separately. Draw a schematic sketch of the overall or combined EMG signal of the muscle. (Assume that the action potentials do not overlap.) Mark the time axis clearly in your drawings.
27. Draw a schematic diagram of a motor unit including labels (names) for each part. Explain the generation of the EMG signal, including the following: (a) single motor unit action potential, (b) innervation ratio, (c) temporal recruitment, and (d) spatial recruitment. Include sketches of at least two sample EMG signals in your discussion.
28. Describe the events in the cardiac system that cause (a) the components of the first heart sound (S1), and (b) the components of the second heart sound (S2). Describe an abnormal condition that could cause (c) a systolic murmur (SM), and (d) a diastolic murmur (DM). Draw a schematic representation of the PCG signal and the corresponding ECG signal over two cardiac cycles and label the PCG signal with the parts related to S1, S2, SM, and DM.
29. (a) Draw a schematic graphical representation of a normal ECG signal over one cardiac cycle. Identify all the waves and their typical durations. Identify the isoelectric parts of the signal. (b) Draw an ECG signal with a wide and jagged QRS complex. (c) Draw an ECG signal with an elevated ST segment. (d) Draw an ECG signal with a depressed ST segment.
30. Draw a schematic sketch of a normal ECG signal as well as the corresponding PCG and carotid pulse signals over one cardiac cycle. Identify the following waves in the signals: P, QRS, T, S1, S2, and the dicrotic notch. Indicate the temporal relationships between the waves mentioned. Label your figure with the relationships between the waves listed above and the following events in the cardiac cycle: atrial contraction, ventricular contraction, ventricular relaxation, and closure of the aortic valve. Label the intervals where systolic and diastolic murmurs could appear.

1.8 Laboratory Exercises and Projects

Note: In exercises involving visits to a medical facility or biomedical data acquisition, attention should be paid to the following items of advice.

- Obtain the necessary ethical and other administrative permissions.
- Respect the privacy, sensitivities, priorities, and confidentiality of patients, medical professionals, and experimental subjects.
- Request a medical practitioner or technologist to explain the related procedures and processes.
- Observe the precautions related to the safety and well-being of the patients, subjects, and observers.
- Get assistance from a qualified technician in the operation of any equipment used.
- Discuss the project with a medical expert and a technologist specialized in the relevant area and obtain information on the differences between normal and abnormal patterns in the experiments conducted and the signals acquired.

- Collect a few sample signals for use in signal processing experiments.
 - Ensure that no patient identification or confidential information is taken out of the laboratory or facility.
- Attention to the points made above can help in gaining purposeful and worthwhile experience with biomedical signal acquisition and related matters.

1. Visit an ECG, EMG, or EEG laboratory in a local hospital or health sciences center. View a demonstration of the acquisition of a few biomedical signals. Request a specialist in a related field to explain how he or she would interpret the signals. Volunteer to be the experimental subject and experience first-hand a biomedical signal acquisition procedure!
2. Set up an ECG acquisition system and study the effects of the following conditions or actions on the quality and nature of the signal: loose electrodes; lack of electrode gel; the subject holding his/her breath or breathing freely during the recording procedure; and the subject coughing, talking, or squirming during signal recording. Record noise-free and noisy ECG signals under various conditions for use in exercises on filtering.
3. Using a stethoscope, listen to your own heart sounds and those of your friends. Examine the variability of the sounds with the site of auscultation. Study the effects of heavy breathing and speaking by the subject as you are listening to the heart sound signal.
4. Record speech signals of vowels (/A/, /I/, /U/, /E/, /O/), diphthongs (/EI/, /OU/), fricatives (/S/, /F/), and plosives (/T/, /P/) as well as words with all four types of sounds (for example, safety, explosive, hearty, heightened, and house). You may be able to perform this experiment with the microphone on your computer. Study the waveform and characteristics of each signal. Use the signals in exercises on filtering, segmentation, and spectral analysis.
5. Using surface electrodes placed on the forearm, record EMG signals at various levels of force exerted by clenching the fist using a device with a force transducer. Provide for rest between repetitions of the experiment to avoid the development of fatigue in the muscle. Plot and study variations in the signal's characteristics with the level of force.

References

- [1] Lathi BP. *Linear Systems and Signals*. Oxford University Press, New York, NY, 2nd edition, 2005.
- [2] Oppenheim AV, Willsky AS, and Nawab SH. *Signals and Systems*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1997.
- [3] Oppenheim AV and Schafer RW. *Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1975.
- [4] Max J. Quantizing for minimum distortion. *IEEE Transactions on Information Theory*, 6:7–12, 1960.
- [5] Lloyd SP. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.
- [6] Oppenheim AV and Schafer RW. *Discrete-time Signal Processing*. Pearson, Englewood Cliffs, NJ, third edition, 2010.
- [7] Cooper KE, Cranston WI, and Snell ES. Temperature regulation during fever in man. *Clinical Science*, 27(3):345–356, 1964.
- [8] Cooper KE. Body temperature and its regulation. In *Encyclopedia of Human Biology*, volume 2, pages 73–83. Academic, New York, NY, 1997.
- [9] Rangayyan RM. *Biomedical Image Analysis*. CRC Press, Boca Raton, FL, 2005.
- [10] Rangayyan RM, Acha B, and Serrano C. *Color Image Processing with Biomedical Applications*. SPIE, Bellingham, WA, 2011.
- [11] Hodgkin AL and Huxley AF. Action potentials recorded from inside a nerve fibre. *Nature*, 144:710–711, 1939.
- [12] Hodgkin AL and Huxley AF. Resting and action potentials in single nerve fibres. *Journal of Physiology*, 104:176–195, 1945.

- [13] Webster JG, editor. *Medical Instrumentation: Application and Design*. Wiley, New York, NY, 3rd edition, 1998.
- [14] Cromwell L, Weibell FJ, and Pfeiffer EA. *Biomedical Instrumentation and Measurements*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1980.
- [15] Plonsey R. Action potential sources and their volume conductor fields. *Proceedings of the IEEE*, 65(5):601–611, 1977.
- [16] Clark R. *Action potentials (Personal Communication)*. University of Calgary, Calgary, Alberta, Canada, 1999.
- [17] Hille B. Membrane excitability: Action potential propagation in axons. In Patton H, Fuchs A, Hille B, Scher A, and Steiner R, editors, *Textbook of Physiology*, pages 49–79. WB Saunders, Philadelphia, PA, 21st edition, 1989.
- [18] Koester J. Action conductances underlying the action potential. In Kandel E and Schwartz J, editors, *Principles of Neural Science*, pages 53–62. Elsevier–North Holland, New York, NY, 1981.
- [19] Kimura J. *Electrodiagnosis in Diseases of Nerve and Muscle: Principles and Practice*. Oxford University Press, New York, NY, 4th edition, 2013.
- [20] Drake KL, Wise KD, Farraye J, Anderson DJ, and BeMent SL. Performance of planar multisite microprobes in recording extracellular single-unit intracortical activity. *IEEE Transactions on Biomedical Engineering*, 35(9):719–732, 1988.
- [21] Frank K and Fuortes MGF. Potentials recorded from the spinal cord with microelectrodes. *Journal of Physiology*, 130:625–654, 1955.
- [22] Goodgold J and Eberstein A. *Electrodiagnosis of Neuromuscular Diseases*. Williams and Wilkins, Baltimore, MD, 3rd edition, 1983.
- [23] Hodes R, Larrabee MG, and German W. The human electromyogram in response to nerve stimulation and the conduction velocity of motor axons: Studies on normal and on injured peripheral nerves. *Archives of Neurology and Psychiatry*, 60(4):340–365, 1948.
- [24] Clark, Jr. JW. The origin of biopotentials. In Webster JG, editor, *Medical Instrumentation: Application and Design*, pages 121–182. Wiley, New York, NY, 3rd edition, 1998.
- [25] Rushmer RF. *Cardiovascular Dynamics*. WB Saunders, Philadelphia, PA, 4th edition, 1976.
- [26] Buchthal F and Schmalbruch H. Motor unit of mammalian muscle. *Physiological Reviews*, 60(1):90–142, 1980.
- [27] Brown GL and Harvey AM. Neuro-muscular transmission in the extrinsic muscles of the eye. *Journal of Physiology*, 99:379–399, 1941.
- [28] Brown WF, Strong AM, and Snow R. Methods for estimating numbers of motor units in biceps-brachialis muscles and losses of motor units with aging. *Muscle & Nerve*, 11(5):423–432, 1988.
- [29] de Luca CJ. Physiology and mathematics of myoelectric signals. *IEEE Transactions on Biomedical Engineering*, 26:313–325, 1979.
- [30] Gath I and Stålberg E. In situ measurement of the innervation ratio of motor units in human muscles. *Experimental Brain Research*, 43:377–382, 1981.
- [31] Mambrizo B and de Luca CJ. Acquisition and decomposition of the EMG signal. In Desmedt JE, editor, *Progress in Clinical Neurophysiology, Volume 10: Computer-aided Electromyography*, pages 52–72. S. Karger AG, Basel, Switzerland, 1983.
- [32] Platt RS, Hajduk EA, Hulliger M, and Easton PA. A modified Bessel filter for amplitude demodulation of respiratory electromyograms. *Journal of Applied Physiology*, 84(1):378–388, 1998.
- [33] Nikolic M and Krarup C. EMGTools, an adaptive and versatile tool for detailed EMG analysis. *IEEE Transactions on Biomedical Engineering*, 58(10):2707–2718, 2011.
- [34] Amsüss S, Goebel PM, Jiang N, Graumann B, Paredes L, and Farina D. Self-correcting pattern recognition system of surface EMG signals for upper limb prosthesis control. *IEEE Transactions on Biomedical Engineering*, 61(4):1167–1176, 2014.

- [35] Chan FHY, Yang YS, Lam FK, Zhang YT, and Parker PA. Fuzzy EMG classification for prosthesis control. *IEEE Transactions on Rehabilitation Engineering*, 8(3):1167–1176, 2000.
- [36] Waller AD. A demonstration on man of electromotive changes accompanying the heart's beat. *Journal of Physiology*, 8(5):229–234, 1887.
- [37] Einthoven W. The different forms of the human electrocardiogram and their signification. *Lancet*, 1(4622):853–861, 1912.
- [38] Tompkins WJ. *Biomedical Digital Signal Processing*. Prentice-Hall, Upper Saddle River, NJ, 1995.
- [39] Goldberger E. *Unipolar Lead Electrocardiography and Vectorcardiography*. Lea & Febiger, Philadelphia, PA, 3rd edition, 1954.
- [40] Friedman HH. *Diagnostic Electrocardiography and Vectorcardiography*. McGraw-Hill, New York, NY, 2nd edition, 1977.
- [41] Malmivuo J and Plonsey R. *Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields*. Oxford University Press, New York, NY, 1995.
- [42] Draper HW, Peffer CJ, Stallmann FW, Littmann D, and Pipberger HV. The corrected orthogonal electrocardiogram and vectorcardiogram in 510 normal men (Frank lead system). *Circulation*, 30:853–864, 1964.
- [43] Jenkins JM. Computerized electrocardiography. *CRC Critical Reviews in Bioengineering*, 6:307–350, November 1981.
- [44] Jenkins JM. Automated electrocardiography and arrhythmia monitoring. *Progress in Cardiovascular Disease*, 25(5):367–408, 1983.
- [45] Cox Jr. JR, Nolle FM, and Arthur RM. Digital analysis of the electroencephalogram, the blood pressure wave, and the electrocardiogram. *Proceedings of the IEEE*, 60(10):1137–1164, 1972.
- [46] Cooper R, Osselton JW, and Shaw JC. *EEG Technology*. Butterworths, London, UK, 3rd edition, 1980.
- [47] Kooi KA, Tucker RP, and Marshall RE. *Fundamentals of Electroencephalography*. Harper & Row, Hagerstown, MD, 2nd edition, 1978.
- [48] Hughes JR. *EEG in Clinical Practice*. Butterworth, Woburn, MA, 1982.
- [49] Agarwal R and Gotman J. Computer-assisted sleep staging. *IEEE Transactions on Biomedical Engineering*, 48(12):1412–1423, 2001.
- [50] Johnson L, Lubin A, Naitoh P, Nute C, and Austin M. Spectral analysis of the EEG of dominant and non-dominant alpha subjects during waking and sleeping. *Electroencephalography and Clinical Neurophysiology*, 26(4):361–370, 1969.
- [51] Zimmermann-Schlatter A, Schuster C, Puhan MA, Siekierka E, and Steurer J. Efficacy of motor imagery in post-stroke rehabilitation: A systematic review. *Journal of Neuroengineering and Rehabilitation*, 5(1):1–10, 2008.
- [52] Yadav R, Swamy MNS, and Agarwal R. Model-based seizure detection for intracranial EEG recordings. *IEEE Transactions on Biomedical Engineering*, 59(5):1419–1428, 2012.
- [53] Grewal S and Gotman J. An automatic warning system for epileptic seizures recorded on intracerebral EEGs. *Clinical Neurophysiology*, 116:2460–2472, 2005.
- [54] Chisci L, Mavino A, Perferi G, Sciandrone M, Anile C, Colicchio G, and Fuggetta F. Real-time epileptic seizure prediction using AR models and support vector machines. *IEEE Transactions on Biomedical Engineering*, 57(5):1124–1132, 2010.
- [55] Whittington MA, Cunningham MO, LeBeau FEN, Racca C, and Traub RD. Multiple origins of the cortical gamma rhythm. *Developmental Neurobiology*, 71(1):92–106, 2011.
- [56] Mantini D, Perrucci MG, Del Gratta C, Romani GL, and Corbetta M. Electrophysiological signatures of resting state networks in the human brain. *Proceedings of the National Academy of Sciences*, 104(32):13170–13175, 2007.
- [57] Rennie CJ, Wright JJ, and Robinson PA. Mechanisms of cortical electrical activity and emergence of gamma rhythm. *Journal of Theoretical Biology*, 205(1):17–35, 2000.

- [58] De Vos M, Vergult A, De Lathauwer L, De Clercq W, Van Huffel S, Dupont P, Palmini A, and Van Paesschen W. Canonical decomposition of ictal scalp EEG reliably detects the seizure onset zone. *NeuroImage*, 37(3):844–854, 2007.
- [59] Verhagen MAMT, van Schelven LJ, Samsom M, and Smout AJPM. Pitfalls in the analysis of electro-gastrographic recordings. *Gastroenterology*, 117:453–460, 1999.
- [60] Mintchev MP and Bowes KL. Capabilities and limitations of electrogastrograms. In Chen JDZ and McCallum RW, editors, *Electrogastrography: Principles and Applications*, pages 155–169. Raven, New York, NY, 1994.
- [61] Mintchev MP and Bowes KL. Extracting quantitative information from digital electrogastrograms. *Medical and Biological Engineering and Computing*, 34:244–248, 1996.
- [62] Chen JDZ, Stewart Jr. WR, and McCallum RW. Spectral analysis of episodic rhythmic variations in the cutaneous electrogastrogram. *IEEE Transactions on Biomedical Engineering*, 40(2):128–135, 1993.
- [63] Mintchev MP, Stickel A, and Bowes KL. Dynamics of the level of randomness in gastric electrical activity. *Digestive Diseases and Sciences*, 43(5):953–956, 1998.
- [64] Rangayyan RM and Lehner RJ. Phonocardiogram signal processing: A review. *CRC Critical Reviews in Biomedical Engineering*, 15(3):211–236, 1988.
- [65] Tavel ME. *Clinical Phonocardiography and External Pulse Recording*. Year Book Medical, Chicago, IL, 3rd edition, 1978.
- [66] Luisada AA and Portaluppi F. *The Heart Sounds — New Facts and Their Clinical Implications*. Praeger, New York, NY, 1982.
- [67] Shaver JA, Salerni R, and Reddy PS. Normal and abnormal heart sounds in cardiac diagnosis, Part I: Systolic sounds. *Current Problems in Cardiology*, 10(3):1–68, 1985.
- [68] Reddy PS, Salerni R, and Shaver JA. Normal and abnormal heart sounds in cardiac diagnosis, Part II: Diastolic sounds. *Current Problems in Cardiology*, 10(4):1–55, 1985.
- [69] Abbas AK and Bassam R. *Phonocardiography Signal Processing*. Morgan & Claypool and Springer, 2009.
- [70] Brown CC, Giddbon DB, and Dean ED. Techniques of plethysmography. *Psychophysiology*, 1(3):253–266, 1965.
- [71] Cri  e CP, Sorichter S, Smith HJ, Kardos P, Merget R, Heise D, Berdel D, K  hler D, Magnussen H, Marek W, Mitfessel H, Rasche K, Rolke M, Worth H, and J  rres RA. Body plethysmography—its principles and clinical use. *Respiratory Medicine*, 105(7):959–971, 2011.
- [72] Allen J. Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3):R1, 2007.
- [73] Kyriacou PA. Pulse oximetry in the oesophagus. *Physiological Measurement*, 27(1):R1, 2005.
- [74] Galli A, Frigo G, Narduzzi C, and Giorgi G. Robust estimation and tracking of heart rate by PPG signal analysis. In *2017 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6. IEEE, 2017.
- [75] Patterson JAC, McIlwraith DC, and Yang GZ. A flexible, low noise reflective PPG sensor platform for ear-worn heart rate monitoring. In *2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks*, pages 286–291. IEEE, 2009.
- [76] Fortino G and Giamp   V. PPG-based methods for non invasive and continuous blood pressure measurement: An overview and development issues in body sensor networks. In *2010 IEEE International Workshop on Medical Measurements and Applications*, pages 10–13. IEEE, 2010.
- [77] Alnaeb ME, Aloabaid N, Seifalian AM, Mikhailidis DP, and Hamilton G. Optical techniques in the assessment of peripheral arterial disease. *Current Vascular Pharmacology*, 5(1):53–59, 2007.
- [78] Madhav KV, Ram MR, Krishna EH, Komalla NR, and Reddy KA. Robust extraction of respiratory activity from PPG signals using modified MSPCA. *IEEE Transactions on Instrumentation and Measurement*, 62(5):1094–1106, 2013.

- [79] Childers DG and Bae KS. Detection of laryngeal function using speech and electroglottographic data. *IEEE Transactions on Biomedical Engineering*, 39(1):19–25, 1992.
- [80] Tsanas A, Little MA, McSharry PE, Spielman J, and Ramig LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson’s disease. *IEEE Transactions on Biomedical Engineering*, 59(5):1264–1271, 2012.
- [81] Rabiner LR and Schafer RW. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, 1978.
- [82] Rueda A, Vásquez-Correa JC, Orozco-Arroyave JR, Nöth E, and Krishnan S. Empirical mode decomposition articulation feature extraction on Parkinson’s Diadochokinesia. *Computer Speech and Language*, 24(101322), 2021.
- [83] Maciel CD, Pereira JC, and Stewart D. Identifying healthy and pathologically affected voice signals [lecture notes]. *IEEE Signal Processing Magazine*, 27:120–123, January 2010.
- [84] Umapathy K, Krishnan S, Parsa V, and Jamieson DG. Discrimination of pathological voices using a time-frequency approach. *IEEE Transactions on Biomedical Engineering*, 52(3):421–430, March 2005.
- [85] Umapathy K and Krishnan S. Feature analysis of pathological speech signals using local discriminant bases technique. *Medical and Biological Engineering and Computing*, 43:457–464, 2005.
- [86] Ghoraani B, Umapathy K, Sugavaneswaran L, and Krishnan S. Pathological speech signal analysis using time-frequency approaches. *Critical Reviews in Biomedical Engineering*, 40(1):63–95, 2012.
- [87] Arias-Londoño JD, Godino-Llorente JI, Sáenz-Lechón N, Osma-Ruiz V, and Castellanos-Domínguez G. Automatic detection of pathological voices using complexity measures, noise parameters, and Mel-Cepstral coefficients. *IEEE Transactions on Biomedical Engineering*, 58(2):370–379, 2011.
- [88] Ellison AE. *Athletic Training and Sports Medicine*. American Academy of Orthopaedic Surgeons, Chicago, IL, 1984.
- [89] Moore KL. *Clinically Oriented Anatomy*. Williams/Wilkins, Baltimore, MD, 1984.
- [90] Tortora GJ. Articulations. In Wilson CM and Helfgott N, editors, *Principles of Human Anatomy*, pages 167–203. Harper and Row, New York, NY, 1986.
- [91] Frankel VH and Nordin M, editors. *Basic Biomechanics of the Skeletal System*. Lea and Febiger, Philadelphia, PA, 1980.
- [92] Nicholas JA and Hershman EB, editors. *The Lower Extremity and Spine in Sports Medicine*. CV Mosby, Missouri, KS, 1986.
- [93] Rangayyan RM, Krishnan S, Bell GD, Frank CB, and Ladly KO. Parametric representation and screening of knee joint vibroarthrographic signals. *IEEE Transactions on Biomedical Engineering*, 44(11):1068–1074, 1997.
- [94] Zhang YT, Rangayyan RM, Frank CB, and Bell GD. Adaptive cancellation of muscle contraction interference from knee joint vibration signals. *IEEE Transactions on Biomedical Engineering*, 41(2):181–191, 1994.
- [95] Frank CB, Rangayyan RM, and Bell GD. Analysis of knee sound signals for non-invasive diagnosis of cartilage pathology. *IEEE Engineering in Medicine and Biology Magazine*, pages 65–68, March 1990.
- [96] Wu YF, Krishnan S, and Rangayyan RM. Computer-aided diagnosis of knee-joint disorders via vibroarthrographic signal analysis: A review. *Critical Reviews in Biomedical Engineering*, 38(2):201–224, 2010.
- [97] Tavathia S, Rangayyan RM, Frank CB, Bell GD, Ladly KO, and Zhang YT. Analysis of knee vibration signals using linear prediction. *IEEE Transactions on Biomedical Engineering*, 39(9):959–970, 1992.
- [98] Moussavi ZMK, Rangayyan RM, Bell GD, Frank CB, Ladly KO, and Zhang YT. Screening of vibroarthrographic signals via adaptive segmentation and linear prediction modeling. *IEEE Transactions on Biomedical Engineering*, 43(1):15–23, 1996.
- [99] Krishnan S, Rangayyan RM, Bell GD, Frank CB, and Ladly KO. Adaptive filtering, modelling, and classification of knee joint vibroarthrographic signals for non-invasive diagnosis of articular cartilage pathology. *Medical and Biological Engineering and Computing*, 35(6):677–684, 1997.

- [100] Kernohan WG, Beverland DE, McCoy GF, Hamilton A, Watson P, and Mollan RAB. Vibration arthrometry. *Acta Orthopædica Scandinavica*, 61(1):70–79, 1990.
- [101] Chu ML, Gradišar IA, and Mostardi R. A noninvasive electroacoustical evaluation technique of cartilage damage in pathological knee joints. *Medical and Biological Engineering and Computing*, 16:437–442, 1978.
- [102] Wu Y. *Knee Joint Vibroarthrographic Signal Processing and Analysis*. Springer, 2015.
- [103] Zhang YT, Frank CB, Rangayyan RM, and Bell GD. A comparative study of vibromyography and electromyography obtained simultaneously from active human quadriceps. *IEEE Transactions on Biomedical Engineering*, 39(10):1045–1052, 1992.
- [104] Zhang YT, Frank CB, Rangayyan RM, and Bell GD. Relationships of the vibromyogram to the surface electromyogram of the human rectus femoris muscle during voluntary isometric contraction. *Journal of Rehabilitation Research and Development*, 33(4):395–403, 1996.
- [105] Probst R, Lonsbury-Martin B, and Martin GK. A review of otoacoustic emissions. *Journal of the Acoustical Society of America*, 89(5):2027–2067, 1991.
- [106] Pasterkamp H, Kraman SS, and Wodicka GR. Respiratory sounds: Advances beyond the stethoscope. *American Journal of Respiratory and Critical Care Medicine*, 156:974–987, 1997.
- [107] Fenton TR, Pasterkamp H, Tal A, and Chernick V. Automated spectral characterization of wheezing in asthmatic children. *IEEE Transactions on Biomedical Engineering*, 32(1):50–55, 1985.
- [108] Kompis M, Pasterkamp H, and Wodicka GR. Acoustic imaging of the human chest. *Chest*, 120(4):1309–1321, 2001.
- [109] Gnitecki J and Moussavi Z. The fractality of lung sounds: A comparison of three waveform fractal dimension algorithms. *Chaos, Solitons & Fractals*, 26(4):1065–1072, 2005.
- [110] Yadollahi A and Moussavi ZMK. A robust method for estimating respiratory flow using tracheal sounds entropy. *IEEE Transactions on Biomedical Engineering*, 53(4):662–668, 2006.
- [111] Alshaer H, Fernie GR, Maki E, and Bradley TD. Validation of an automated algorithm for detecting apneas and hypopneas by acoustic analysis of breath sounds. *Sleep Medicine*, 14(6):562–571, 2013.
- [112] Alshaer H, Fernie GR, and Bradley TD. Monitoring of breathing phases using a bioacoustic method in healthy awake subjects. *Journal of Clinical Monitoring and Computing*, 25(5):285–294, 2011.
- [113] Hadjileontiadis LJ. *Lung Sounds: An Advanced Signal Processing Perspective*. Morgan & Claypool and Springer, 2009.
- [114] Moussavi Z. *Fundamentals of Respiratory Sounds and Analysis*. Morgan & Claypool and Springer, San Rafael, CA, 2006.
- [115] Issa FG, Morrison D, Hadjuk E, Iyer A, Feroah T, and Remmers JE. Digital monitoring of sleep-disordered breathing using snoring sound and arterial oxygen saturation. *American Review of Respiratory Disease*, 148:1023–1029, 1993.
- [116] Dalmasso F and Prota R. Snoring: Analysis, measurement, clinical implications and applications. *European Respiratory Journal*, 9:146–159, 1996.
- [117] Lazareck LL and Moussavi ZMK. Classification of normal and dysphagic swallows by acoustical means. *IEEE Transactions on Biomedical Engineering*, 51(12):2103–2112, 2004.
- [118] Tomomasa T, Morikawa A, Sandler RH, Mansy HA, Koneko H, Masahiko T, Hyman PE, and Itoh Z. Gastrointestinal sounds and migrating motor complex in fasted humans. *The American Journal of Gastroenterology*, 94(2):374–381, 1999.
- [119] Gay T, Bertolami CN, and Solonche DJ. *Method and Apparatus for the Acoustic Detection and Analysis of Joint Disorders*. US Patent 4,836,218, 1989.
- [120] Pavesi L, Subburaj S, and Porter-Shaw K. Application and validation of a computerized cough acquisition system for objective monitoring of acute cough: A meta-analysis. *Chest*, 120(4):1121–1128, 2001.
- [121] Várallyay Jr. G. The melody of crying. *International Journal of Pediatric Otorhinolaryngology*, 71(11):1699–1708, 2007.

- [122] Kaniusas E, Pfützner H, and Saletu B. Acoustical signal properties for cardiac/respiratory activity and apneas. *IEEE Transactions on Biomedical Engineering*, 52(11):1812–1822, 2005.
- [123] Aston R. *Principles of Biomedical Instrumentation and Measurement*. Merrill, Columbus, OH, 1990.
- [124] Bronzino JD. *Biomedical Engineering and Instrumentation*. PWS Engineering, Boston, MA, 1986.
- [125] Bronzino JD, editor. *The Biomedical Engineering Handbook*. CRC and IEEE, Boca Raton, FL, 1995.
- [126] Cohen A. *Biomedical Signal Processing*. CRC Press, Boca Raton, FL, 1986.
- [127] *Electrical Geodesics, Inc.* www.electricalgeodesics.com, accessed on 2023-06-26.
- [128] *g.Nautilus wireless biosignal acquisition*. www.gtec.at, accessed on 2023-06-26.
- [129] *BTS bioengineering*. www.btsbioengineering.com, accessed on 2023-06-26.
- [130] *PLUX wireless biosignals S.A.* www.pluxbiosignals.com, accessed on 2023-06-26.
- [131] Krishnan S. *Biomedical Signal Analysis for Connected Healthcare*. Academic Press, New York, NY, 2021.
- [132] Junnila S, Akhbardeh A, and Värri A. An electromechanical film sensor based wireless ballistocardiographic chair: Implementation and performance. *Journal of Signal Processing Systems*, 57:305–320, 2009.
- [133] Guerrero G, Kortelainen JM, Palacios E, Bianchi AM, Tachino G, Tenhunen M, Méndez MO, and van Gils M. Detection of sleep-disordered breathing with pressure bed sensor. In *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1342–1345, Osaka, Japan, July 2013.
- [134] Peltokangas M, Verho J, and Vehkaoja A. Night-time EKG and HRV monitoring with bed sheet integrated textile electrodes. *IEEE Transactions on Information Technology in Biomedicine*, 16(5):935–942, 2012.
- [135] Mikhelson IV, Lee P, Bakhtiari S, Elmer II TW, Katsaggelos AK, and Sahakian AV. Noncontact millimeter-wave real-time detection and tracking of heart rate on an ambulatory subject. *IEEE Transactions on Information Technology in Biomedicine*, 16(5):927–934, 2012.
- [136] Widrow B, Glover Jr. JR, McCool JM, Kaunitz J, Williams CS, Hearn RH, Zeidler JR, Dong Jr. E, and Goodlin RC. Adaptive noise cancelling: Principles and applications. *Proceedings of the IEEE*, 63(12):1692–1716, 1975.
- [137] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, and Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101:e215–e220, 2000.
- [138] Penzel T, Moody GB, Mark RG, Goldberger AL, and Peter JH. The Apnea-ECG database. In *Proceedings of IEEE Computers in Cardiology*, pages 255–258, <https://www.physionet.org/content/apnea-ecg/1.0.0/>, 2000.
- [139] Moody GB, Mark RG, and Goldberger AL. PhysioNet: A web-based resource for the study of physiologic signals. *IEEE Engineering in Medicine and Biology*, pages 70–75, May/June 2001.
- [140] Bartlett J. *Familiar Quotations*. Little, Brown and Co., Boston, MA, 15th edition, 1980.
- [141] Wiener N. A new theory of measurement: A study in the logic of mathematics. In *Proceedings of the London Mathematical Society, Series 2, Volume 19*, pages 181–205, 1919.

CHAPTER 2

ANALYSIS OF CONCURRENT, COUPLED, AND CORRELATED PROCESSES

The human body is a complex integration of a number of biological systems with several ongoing physiological, functional, and possibly pathological processes. Most biological processes within a body are not independent of one another; rather, they are mutually correlated and bound together by physical or physiological control and communication phenomena. Analyzing any single process without due attention to others that are concurrent, coupled, or correlated with the process may provide only partial information and pose difficulties in the comprehension of the process. Then, the problem is, how do we recognize the existence of concurrent, coupled, and correlated phenomena? How do we obtain the corresponding signals and identify the correlated features? Unfortunately, there is no simple or universal solution or rule to apply to this problem.

Ideally, an investigator should explore the system or process of interest from all possible perspectives and use multidisciplinary approaches to identify several potential sources of related information. The multichannel signals so obtained may be electrical, mechanical, biochemical, or physical, among the many possibilities, and may exhibit interrelationships confounded by peculiarities of transduction, time delays, multipath transmission or reflection, waveform distortions, and filtering effects that may need to be accounted for in their simultaneous analysis. Events or waves in signals of interest may be nonspecific and difficult to identify and analyze. How could we exploit the concurrency, coupling, and correlation present between processes or related signals to gain better understanding of the system or systems of interest?

2.1 Problem Statement

Determine the correspondences, correlation, and interrelationships present between concurrent signals related to a common underlying physiological system or process, and identify their potential applications.

The statement above represents, of necessity at this stage of the discussion, a vague and generic problem. The case studies and applications presented in the following sections provide a few illustrative examples dealing with specific systems and problems. Signal processing techniques for the various tasks identified in the case studies are developed in the chapters that follow. Note that the examples cover a diverse range of systems, processes, and signals. A specific problem of interest will very likely not be directly related to any of the case studies presented here. It is expected that a study of the examples provided will expand the scope of an investigator's analytical skills and lead to improved solutions for the specific case of interest.

2.2 Illustration of the Problem with Case Studies

2.2.1 The ECG and the PCG

A clinical ECG record typically includes 12 channels of sequentially or simultaneously recorded signals, and can be used on its own to diagnose many cardiac diseases. This is mainly due to the simple and readily identifiable waveforms in the ECG, and the innumerable studies that have firmly established clinical ECG as a standard procedure. The PCG, on the other hand, is a more complex signal. PCG waveforms cannot be visually analyzed except for the identification of gross features such as the presence of murmurs, time delays as in a split S2, and envelopes of murmurs. An advantage with the PCG is that it may be listened to; auscultation of heart sounds is more commonly performed than visual analysis of the PCG signal. However, objective analysis of the PCG requires the identification of components, such as S1 and S2, and subsequent analysis tailored to the nature of the components.

Given a run of a PCG signal over several cardiac cycles, visual identification of S1 and S2 is possible if there are no murmurs between the sounds, and if the heart rate is low such that the S2 – S1 (of the next beat) interval is longer than the S1 – S2 interval (as expected in normal situations). At high heart rates and with the presence of murmurs or premature beats, identification of S1 and S2 could be difficult.

Problem: Identify the beginning of S1 in a PCG signal and extract the heart sound signal over one cardiac cycle.

Solution: The ECG and PCG are concurrent phenomena, with the noticeable difference that the former is electrical while the latter is mechanical (sound or vibration). It is customary to record the ECG with the PCG; see Figures 1.44 and 1.46 for examples.

The QRS wave in the ECG is directly related to ventricular contraction, as the summation of the action potentials of ventricular muscle cells (see Section 1.2.5). As the ventricles contract, the tension in the chordae tendineae and the pressure of retrograde flow of blood toward the atria seal the AV valves shut, thereby causing the initial vibrations of S1 [1] (see Section 1.2.9). Thus, S1 begins immediately after the QRS complex. Given the nonspecific nature of vibration signals and the various possibilities in the transmission of the heart sounds to the recording site on the chest, detection of S1 on its own is a difficult problem.

As shown in Sections 3.5, 4.3.1, and 4.3.2, detection of the QRS is fairly easy, given that the QRS is the sharpest wave in the ECG over a cardiac cycle; in fact, the P and T waves may be almost negligible in many ECG records. Thus, the QRS complex in the ECG is a reliable indicator of the beginning of S1 and may be used to segment a PCG record into individual cardiac cycles: from the beginning of one QRS (and thereby S1) to the beginning of the next QRS and S1. This method may be applied visually or via signal processing techniques: The former requires no further explanation but is expanded upon in Section 2.3; the latter is dealt with in Section 4.9.

2.2.2 The PCG and the carotid pulse

Identification of the diastolic segment of the PCG may be required in some applications in cardiovascular diagnosis [2]. Ventricular systole ends with the closure of the aortic and pulmonary valves, indicated by the aortic (A2) and pulmonary (P2) components of S2 (see Section 1.2.9). The end of contraction is also indicated by the T wave in the ECG, and S2 appears slightly after the end of the T wave (see Figure 1.44). S2 may be taken to mark the end of systole and the beginning of ventricular relaxation or diastole. (Note: Shaver et al. [3] and Reddy et al. [4] have included S2 in the part of their article on systolic sounds.) However, as in the case of S1, S2 is also a nonspecific vibrational wave that cannot be readily identified (even visually), especially when murmurs are present.

Given the temporal relationship between the T wave and S2, it may appear that the former may be used to identify the latter. This, however, may not always be possible in practice, as the T wave is often a low-amplitude and smooth wave and is sometimes not recorded at all (see the normal beats in Figure 1.29). ST segment elevation (as in Figure 1.29) or depression (as in Figure 1.50) may make even visual identification of the T wave difficult. Thus, the T wave is not a reliable indicator to use for identification of S2 or diastole.

Problem: Identify the beginning of S2 in a PCG signal.

Solution: Given the inadequacy of the T wave as an indicator of diastole, we need to explore other possible sources of information. Closure of the aortic valve is accompanied by deceleration and reversal of blood flow in the aorta. This causes a sudden drop in the pressure within the aorta, which is already on a downward slope due to the end of systolic compression. The sudden change in pressure causes an *incisura* or notch in the aortic pressure wave (see Figures 1.49 and 1.50). The aortic pressure signal may be obtained using catheter-tip sensors [3, 4], but the procedure would be invasive. Fortunately, the notch is transmitted through the arterial system and may be observed in the carotid pulse (see Section 1.2.10) recorded at the neck.

The dicrotic notch (D) in the carotid pulse signal bears a delay with respect to the corresponding notch in the aortic pressure signal, but has the advantage of being accessible in a noninvasive manner. (Similar events occur in the pulmonary artery, but provide no externally observable effects.) See Figures 1.44 and 1.46 for examples of three-channel PCG – ECG – carotid pulse recordings that illustrate the S2 – T – D relationships. The dicrotic notch may be used as a reliable indicator of the end of systole or beginning of diastole that may be obtained in a noninvasive manner. The average S2 – D delay has been found to be 42.6 ms with a standard deviation (*SD*) of 5 ms [5] (see also Tavel [6]), which should be subtracted from the dicrotic notch position to obtain the beginning of S2.

See Section 2.3 for related discussions. Signal processing techniques for the detection of the dicrotic notch and segmentation of the PCG are described in Sections 4.3.5, 4.9, and 4.10.

2.2.3 The ECG and the atrial electrogram

Most studies on the ECG and the PCG pay more attention to ventricular activity than to atrial activity — and, even then, more to left ventricular activity than to the right. Rhythm analysis is commonly performed using QRS complexes to obtain interbeat intervals known as RR intervals. Such analysis neglects atrial activity.

Recollect that the AV node introduces a delay between atrial contraction initiated by the SA node impulse and the consequent ventricular contraction. This delay plays a major role in the coordinated contraction of the atria and the ventricles. Certain pathological conditions may disrupt this coordination and may even cause AV dissociation [1]. It then becomes necessary to study atrial activity independent of ventricular activity and establish their association, or lack thereof. Thus, the interval between the P wave and the QRS (termed the PR interval) would be a valuable adjunct to the RR interval in rhythm analysis. Unfortunately, the atria, being relatively small chambers with weak contractile activity, cause a small and smooth P wave in the external ECG. Quite often, the P

wave may not be recorded or seen in the external ECG; see, for example, leads I and V3 – V6 in Figure 1.36.

Problem: Obtain an indicator of atrial contraction to measure the PR interval.

Solution: One of the reasons for the lack of specificity of the P wave is the effect of transmission from the atria to the external recording sites. An obvious solution would be to insert electrodes into one of the atria via a catheter and record the signal at the source. This would, of course, constitute an invasive procedure. Jenkins et al. [7, 8] and Jenkins [9, 10] proposed a unique and interesting procedure to obtain a strong and clear signal of atrial activity: they developed a pill electrode that could be swallowed and lowered through the esophagus to a position close to the left atrium (the bipolar electrode pill being held suspended by wires about 35 cm from the lips). The procedure may or may not be termed invasive, although an object is inserted into the body (and removed after the procedure), as the action required is that of normal swallowing of a tablet-like object. The gain required to obtain a good atrial signal was 2 – 5 times that used in ECG amplifiers. With a 5 – 100 Hz bandpass filter, Jenkins et al. obtained an *SNR* of 10.

Figure 2.1 shows recordings from a normal subject of the atrial electrogram from the pill electrode and an external ECG lead. Atrial contraction is clearly indicated by a sharp spike in the atrial electrogram. Measurement of the PR interval (or the AR interval, as called by Jenkins et al.) now becomes an easy task, with identification of the spike in the atrial electrogram (the “A” wave, as labeled by Jenkins et al.) being simpler than identification of the QRS in the ECG.

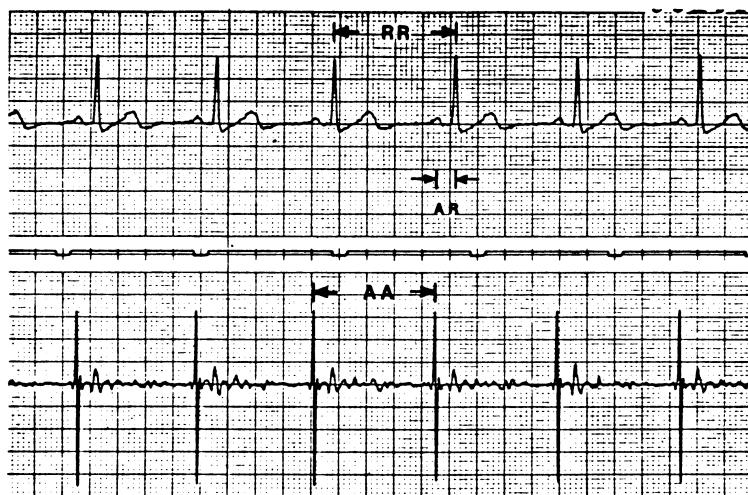


Figure 2.1 Pill-electrode recording of the atrial electrogram (lower tracing) and the external ECG (upper tracing) of a normal subject. The pulse train between the two signals indicates intervals of 1 s. Reproduced with permission from J.M. Jenkins, D. Wu, and R. Arzbaecher, Computer diagnosis of abnormal cardiac rhythms employing a new P-wave detector for interval measurement, *Computers and Biomedical Research*, 11:17–33, 1978. ©Academic Press.

Figure 2.2 shows the atrial electrogram and external ECG of a subject with ectopic beats. The PVCs have no immediately preceding atrial activity. The first PVC has blocked the conduction of the atrial activity occurring immediately after, resulting in a compensatory pause before the following normal beat. The second PVC has not blocked the subsequent atrial wave, but has caused a longer-than-normal AV delay and an aberrant conduction path, which explains the different waveshape of the consequent beat. The third PVC has not affected the timing of the following SA-node-initiated pulse, but has caused a change in waveshape in the resulting QRS-T by altering the conduction path [7–10].

Jenkins et al. developed a four-digit code for each beat, as illustrated in Figure 2.2. The first digit was coded as 0: abnormal waveshape or 1: normal waveshape, as determined by a correlation coefficient computed between the beat being processed and a normal template (see Sections 3.5, 4.4.2, 5.4.1, and 5.8). The remaining three digits encoded the nature of the RR, AR, and AA intervals, respectively, as 0: short, 1: normal, or 2: long. The absence of a preceding A wave related to the beat being analyzed was indicated by the code \times in the fourth digit (in which case the AR interval is longer than the RR interval). Figure 2.2 shows the code for each beat. Based on the code for each beat, Jenkins et al. were able to develop a computerized method to detect a wide variety of arrhythmia.

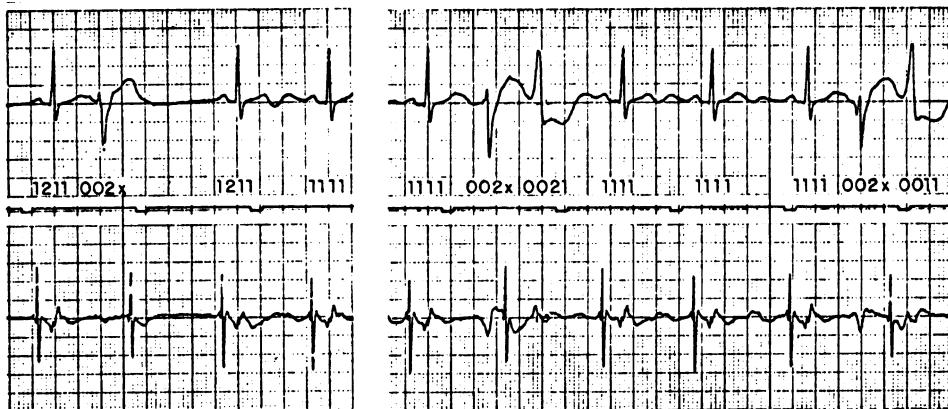


Figure 2.2 Atrial electrogram (lower tracing) and the external ECG (upper tracing) of a subject with ectopic beats. The pulse train between the two signals indicates intervals of 1 s. Reproduced with permission from J.M. Jenkins, D. Wu, and R. Arzbaecher, Computer diagnosis of abnormal cardiac rhythms employing a new P-wave detector for interval measurement, *Computers and Biomedical Research*, 11:17–33, 1978. ©Academic Press.

2.2.4 Cardiorespiratory interaction

The heart rate is affected by normal breathing due to the coupling and interaction existing between the cardiac and respiratory systems [11–16]. Figure 2.3 shows two recordings of the ECG of a normal subject taken a few minutes apart, the first with the subject breathing normally, and the second with the subject holding his breath. Baroreceptors in the aorta detect changes in the aortic transmural pressure associated with variations in the intrapleural pressure with respiration. A decrease in the intrapleural pressure during inspiration causes the vagus nerve activity to be impeded, which causes an increase in the heart rate during inspiration. Therefore, as a subject carries on breathing normally, there could be substantial variations in the heart rate and RR intervals.

Breathing also affects the transmission of the heart sounds from the cardiac chambers to the chest surface. Durand et al. [17] recorded intracardiac and chest-surface PCG signals and derived the dynamic transfer function of the heart – thorax acoustic system in dogs. Analysis of the synchronization and coupling within the cardiorespiratory system could require sophisticated analysis of several signals acquired simultaneously from the cardiac and respiratory systems [18].

2.2.5 Heart-rate variability

Even under resting and apparently steady conditions, the intervals between heart beats and the heart rate are not constant. Variability of the RR interval and heart rate is a normal and healthy

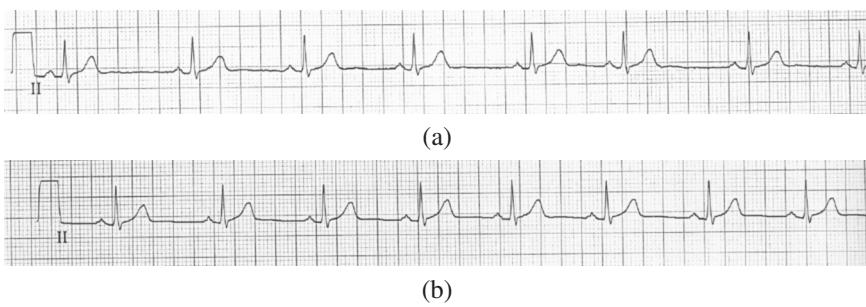


Figure 2.3 ECG signal of a subject (a) with the subject breathing normally, and (b) with the subject holding breath. Signal courtesy of E. Gedamu and L.B. Mitchell, Foothills Hospital, Calgary.

physiological phenomenon. Section 2.2.4 presented a discussion on the effects of normal breathing on the heart rate due to the coupling and interaction between the cardiac and respiratory systems.

As noted in Section 1.2.5, cardiac rhythm is controlled by the SA node and the ANS. The ANS is divided into the SNS and the parasympathetic nervous system (PNS). The SNS and PNS do not work in opposition to each other, but interact in a complex and dynamic manner, with the interactions modulated by secondary messengers [19]. The PNS can inhibit sympathetic nerve activity, and sympathetic activation can inhibit parasympathetic activation [19]. Vagal tone or tonic parasympathetic activation is stronger than the sympathetic tone at rest. The parasympathetic influence on the heart rate is mediated by the release of acetylcholine by the vagus nerve; sympathetic stimulation is mediated by the release of epinephrine. HRV has been noted as a noninvasive signature of the interaction and balance between the effects of the PNS and SNS on the heart [20]. See Kamath et al. [21] and Ishaque et al. [22] for reviews of several works on HRV.

Olshansky et al. [19] studied the role of the PNS in heart failure and made the following observations. The resting rate of a normal heart is governed by a parasympathetic mechanism. The resting heart rate, which is an indicator of vagus nerve function, can predict mortality. Increase in the parasympathetic component of HRV and higher vagus nerve activity result in slower heart rate and better outcome in terms of cardiac health. In cases of heart failure, the regulation of heart rate by parasympathetic activation is poor. High resting heart rate typically leads to adverse outcomes. They also noted that the high-frequency components of HRV are associated with the vagus nerve and the parasympathetic effect, that the low-frequency components are due to sympathetic and parasympathetic activation, and, furthermore, that parasympathetic activation and its physiological effects are attenuated in cases of heart failure. Olshansky et al. observed that the electrophysiological benefits of parasympathetic activation include anti-inflammatory effects, reduced heart rate, increased HRV, and direct antiarrhythmic effects. While high levels of sympathetic activity are associated with poor prognosis, a high level of parasympathetic activation could provide protection. Based on these observations, Olshansky et al. indicated that direct or indirect vagus nerve stimulation could have beneficial effects on clinical outcomes.

Reduced HRV in patients following acute myocardial infarction has been observed to be related to poor prognosis [23, 24]. Malik and Camm [23] noted that HRV is the single most important predictor of patients at high risk of sudden death and ventricular arrhythmia.

It is evident that the heart rate is controlled by several systems in addition to the cardiac system itself. Several physiological conditions and variables affect the heart rate. Such control and related variability indicate coupling between multiple systems. It is important to recognize such coupling and analyze the related systems together.

See Malik et al. [25] for recommendations on measurement, physiological interpretation, and clinical use of HRV. See Sections 7.2.2, 7.9, and 8.12 for discussions on methods for the analysis of HRV.

2.2.6 The EMG and VMG

The EMG signal has been studied extensively, and the relationship between EMG signal parameters and muscle contraction level has been established [26, 27]. It is known that the root mean-squared (*RMS*) and mean frequency values of the EMG increase with increasing muscle contraction until fatigue sets in, at which point both values begin to decrease. In this situation, while the muscle output measured is mechanical contraction (using force or strain transducers), the signal analyzed is electrical in character. A direct mechanical signal related to basic muscle-fiber or motor-unit phenomena may be desired in some situations.

Problem: *Obtain a mechanical signal that is a direct indicator of muscle-fiber or motor-unit activity to study muscle contraction and force development.*

Solution: The VMG, as introduced in Section 1.2.15, is a vibration signal measured from a contracting muscle. The signal is a direct manifestation of the contraction of muscle fibers and, as such, is related to mechanical activity at the muscle-fiber or motor-unit level. The VMG is the mechanical counterpart and contemporary of the EMG. Although no direct relationship has been established between the force outputs of individual motor units and the net force output of the muscle, it has been shown that the *RMS* and mean frequency parameters of the VMG signal increase with muscle force output, in patterns that parallel those of the EMG. Thus, the VMG may be used to quantify muscular contraction [28].

Given the simplicity and noninvasive nature of EMG and VMG measurement, simultaneous analysis of the two signals is an attractive and viable application. Such techniques may find use in biofeedback and rehabilitation [29]. Figure 2.4 shows simultaneous EMG – VMG recordings at two levels of contraction of the rectus femoris muscle [29]; see Figures 1.57 and 1.58. Both signals are interference patterns of several active motor units even at low levels of muscular effort and cannot be analyzed visually. However, a general increase in the power levels of the signals from the lower effort to the higher effort case may be observed. Signal processing techniques for simultaneous EMG – VMG studies are described in Section 5.11.

2.2.7 The knee-joint and muscle-vibration signals

We saw in Section 1.2.14 that the vibration (VAG) signals produced by the knee joint during active swinging movement of the leg may bear diagnostic information. However, the VMG associated with the rectus femoris muscle that must necessarily be active during extension of the leg could appear as an interference and corrupt the VAG signal [30].

Problem: *Suggest an approach to remove muscle-contraction interference from the knee-joint vibration signal.*

Solution: The VMG interference signal gets transmitted from the source muscle location to the VAG recording position at the skin surface over the patella (knee cap) through the intervening muscles and bones (see Figure 1.58 and Section 3.3.6). Although the interference signal has been found to be of very low frequency (10 – 100 Hz), the frequency content of the signal varies with muscular effort and knee-joint angle. The rectus femoris muscle and the knee-joint systems are coupled dynamic systems with vibration characteristics that vary with activity level, and hence over time; thus, simple highpass or bandpass filtering of the VAG signal is not an appropriate solution.

An approach to solve the problem would be to record the VMG signal at the rectus femoris at the same time the VAG signal of interest is acquired from the patella position. Adaptive filtering and noise cancellation techniques [30–32] could then be applied, with the VAG signal as the primary input and the VMG signal as the reference input. Assuming that the VMG signal that arrives at the

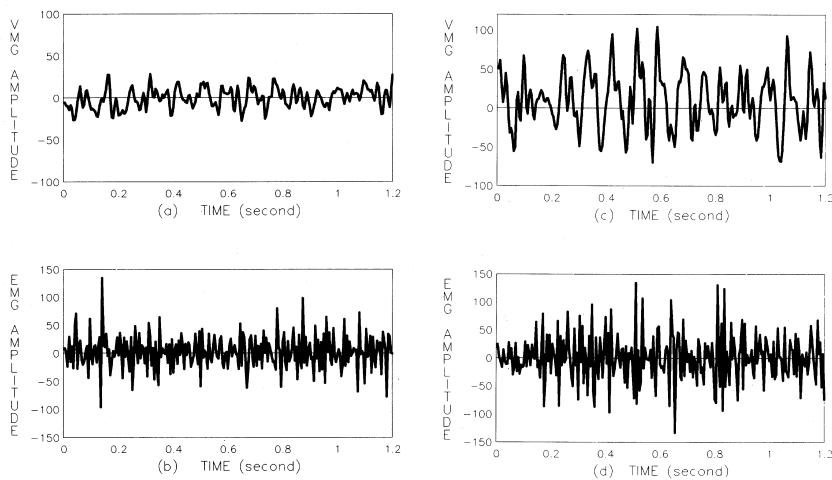


Figure 2.4 Simultaneous EMG – VMG records at two levels of contraction of the rectus femoris muscle. (a) VMG at 40% of the maximal voluntary contraction (MVC) level. (b) EMG at 40% MVC. (c) VMG at 60% MVC. (d) EMG at 60% MVC. See also Figures 1.57 and 1.58. Reproduced with permission from Y.T. Zhang, C.B. Frank, R.M. Rangayyan, and G.D. Bell, Relationships of the vibromyogram to the surface electromyogram of the human rectus femoris muscle during voluntary isometric contraction, *Journal of Rehabilitation Research and Development*, 33(4): 395–403, 1996. ©Department of Veterans Affairs.

patella is strongly correlated with the VMG signal at the rectus femoris and not correlated with the VAG signal of interest, the adaptive filter should remove the interference and estimate the desired VAG signal. Details of adaptive filters are provided in Sections 3.10 and 3.15.

2.3 Application: Segmentation of the PCG into Systolic and Diastolic Parts

Problem: Show how the ECG and carotid pulse signals may be used to break a PCG signal into its systolic and diastolic parts.

Solution: A cardiac cycle may be divided into two important parts based on ventricular activity: systole and diastole. The systolic part starts with S1 and ends at the beginning of S2; it includes any systolic murmur that may be present in the signal. The diastolic part starts with S2, and ends just before the beginning of the S1 of the next cardiac cycle. (The aortic and pulmonary valves close slightly before the A2 and P2 components of S2. Therefore, systole may be considered to have ended just before S2. Although Shaver et al. [3] and Reddy et al. [4] have included S2 in the part of their article on systolic sounds, we shall include S2 in the diastolic part of the PCG.) The diastolic part includes any diastolic murmur that may be present in the signal; it might also include S3 and S4, if present, as well as AV valve opening snaps, if any.

We saw in Section 2.2.1 that the QRS complex in the ECG may be used as a reliable marker of the beginning of S1. We also saw, in Section 2.2.2, that the dicrotic notch in the carotid pulse may be used to locate the beginning of S2. Thus, if we have both the ECG and carotid pulse signals along with the PCG, it becomes possible to break the PCG into its systolic and diastolic parts.

Figure 2.5 shows three-channel PCG – ECG – carotid pulse signals of a subject with systolic murmur due to aortic stenosis (the same as in Figure 1.46), with the systolic and diastolic parts of the PCG marked in relation to the QRS and D events. The demarcation was performed by visual inspection of the signals in this example. Signal processing techniques to detect the QRS and D

waves are presented in Section 4.3. Adaptive filtering techniques to break the PCG into stationary segments without the use of any other reference signal are described in Section 8.11.

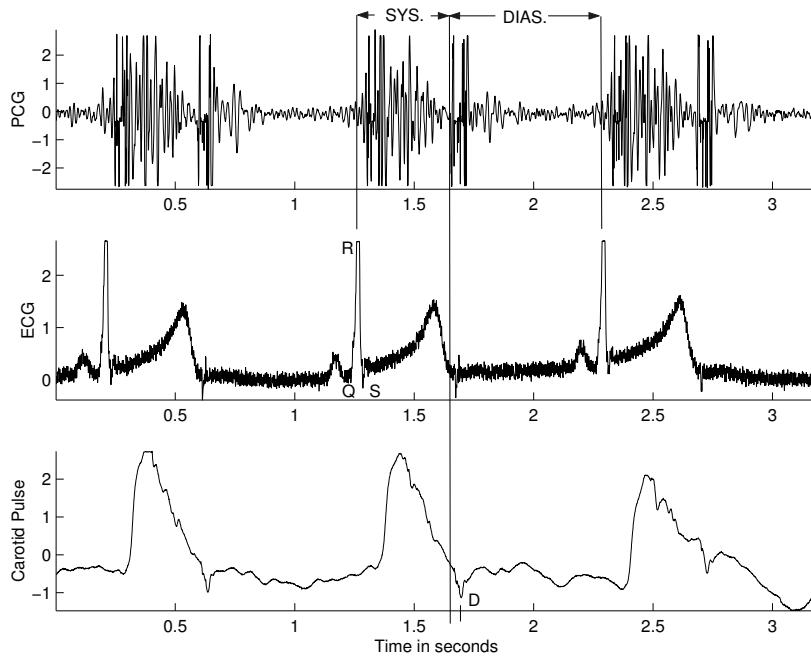


Figure 2.5 Demarcation of the systolic (SYS.) and diastolic (DIAS.) parts of the PCG signal in Figure 1.46 using the ECG and carotid pulse as reference signals. The QRS complex and the dicrotic notch D are marked on the ECG and carotid pulse signals, respectively.

2.4 Application: Diagnosis and Monitoring of Sleep Apnea

Problem: Propose approaches based on biomedical signal analysis to detect sleep apnea. Analyze the problem taking into consideration the various physiological systems that are either a part of the problem or are affected by the resulting condition.

Solution: The term “apnea” indicates a condition in which one stops breathing for several seconds, usually during sleep [33–37]. The term “hypopnea” indicates a condition when airflow is diminished. According to Chesson et al. [38], the general condition of disordered breathing during sleep includes the conditions of apnea, defined as cessation or near cessation of respiration for a minimum period of 10 s; hypopnea, defined as a reduction in airflow for a minimum period of 10 s; and episodes of increased respiratory effort due to partial upper-airway obstruction. See Alshaer et al. [39] and Bradley and Floras [40] for additional conditions that define apnea and hypopnea. Disordered breathing during sleep causes fragmentation of sleep and lack of adequate rest. The total number of episodes of apnea and hypopnea per hour of sleep is defined as the apnea–hypopnea index (AHI). While a patient affected by sleep apnea may stop breathing 10 – 100 times per hour in episodes of duration 10 – 30 s each, the diagnosis of sleep apnea is typically based on an AHI threshold of 10 to 15 [40].

Causes of sleep apnea include the lack of neural input from the CNS to the diaphragm to cause contraction and breathing, known as central sleep apnea (CSA), or collapse of the upper airway, referred to as obstructive sleep apnea (OSA). CSA involves disruption of breathing control where breathing does not match the metabolic requirements. In contrast, OSA is predominantly an anatom-

ical disorder in which the upper airway does not remain open. During sleep, when muscle tone is reduced or absent, the airway collapses, and the individual can no longer breathe sufficiently to maintain normal blood gases; as a result, the level of oxygen drops and that of carbon dioxide rises. OSA is the most common type of sleep apnea.

A common symptom of OSA is snoring, a respiratory noise caused by airflow through a partially obstructed airway. An episode of apnea is typically followed by arousal, allowing breathing to resume and blood gases to return to normal levels; there could be movements of the body or the limbs during such occasions of arousal. Frequent and numerous such arousals reduce the quality of sleep and the amount of rest obtained. Chronic reduction in oxygen levels and arousals increase SNS activity, and could cause numerous complications, including daytime sleepiness, hypertension, heart failure, depression, cardiac arrhythmia, and stroke. An effective treatment for OSA is continuous positive airway pressure (CPAP) applied via a facial or nasal mask; the pressure applied inflates the upper airway and prevents it from collapsing.

Sleep apnea leads to decreased levels of oxyhemoglobin in the blood. Oxygen transported by blood is bound to hemoglobin in red blood cells. The absorption of different wavelengths of light by hemoglobin changes when it is bound to oxygen; thus, oxyhemoglobin appears red, and deoxyhemoglobin appears blue. A fingertip sensor with an LED and a photo sensor, known as a fingertip pulse oximeter, or a *CO*-oximeter is commonly used to estimate the level of oxyhemoglobin. Using such a device, the absorption levels for red and infrared light are compared to estimate the proportion of oxyhemoglobin to deoxyhemoglobin in the blood. The measure known as SpO_2 is defined as the ratio of the amount of oxyhemoglobin to deoxyhemoglobin; SaO_2 is defined as the ratio of the amount of arterial-blood oxyhemoglobin to the sum of all types of hemoglobin.

One approach to detect sleep apnea is to measure oronasal airflow using a pneumotachometer; however, due to the inconvenient nature of this approach, indirect indicators of airflow through the nose and mouth may be obtained using an oronasal thermistor or by nasal pressure measurement [39, 41]. Respiratory inductance plethysmography is an indirect way of assessing airflow via movement of thoracic or abdominal belts; the signal obtained is proportional to volumetric displacement of the lungs [39, 41, 42]. See Section 10.13 for further discussions on sleep apnea.

2.4.1 Monitoring of sleep apnea by polysomnography

Polysomnography (PSG) [38, 43, 44], which involves multichannel recording of several biomedical signals and parameters, is the current standard for the evaluation of sleep-related problems, including sleep apnea. Some of the signals and parameters measured are respiratory effort, airflow, oxygenation, sleep state, EMG of the submental muscle (beneath the chin), EMG of the legs, the electrooculogram (EOG), EEG, ECG, and snoring sound. PSG requires the subject to sleep overnight in a laboratory.

Figure 2.6 shows eight channels selected from a 14-channel PSG dataset of a patient with OSA [39]. Periods of diminished airflow to the lungs, and consequently low SaO_2 levels are evident. Also seen are related intervals of increased submental EMG activity, snoring, and increased thoracic and/or abdominal activity related to recovery of respiration following episodes of apnea. (Artifacts in the submental EMG channel have caused baseline movement; filters may be used to remove such artifacts.) SaO_2 values are seen to increase following periods of increased ventilation, but fall again as OSA causes reduced airflow to the lungs. The leg EMG channel shows activity related to leg movement during periods of recovery from episodes of apnea, but is contaminated with the ECG.

2.4.2 Home monitoring of sleep apnea

Although PSG is a comprehensive and accurate method for the diagnosis of sleep apnea, it is expensive, inconvenient, and often not available. To address these difficulties, practical home-monitoring

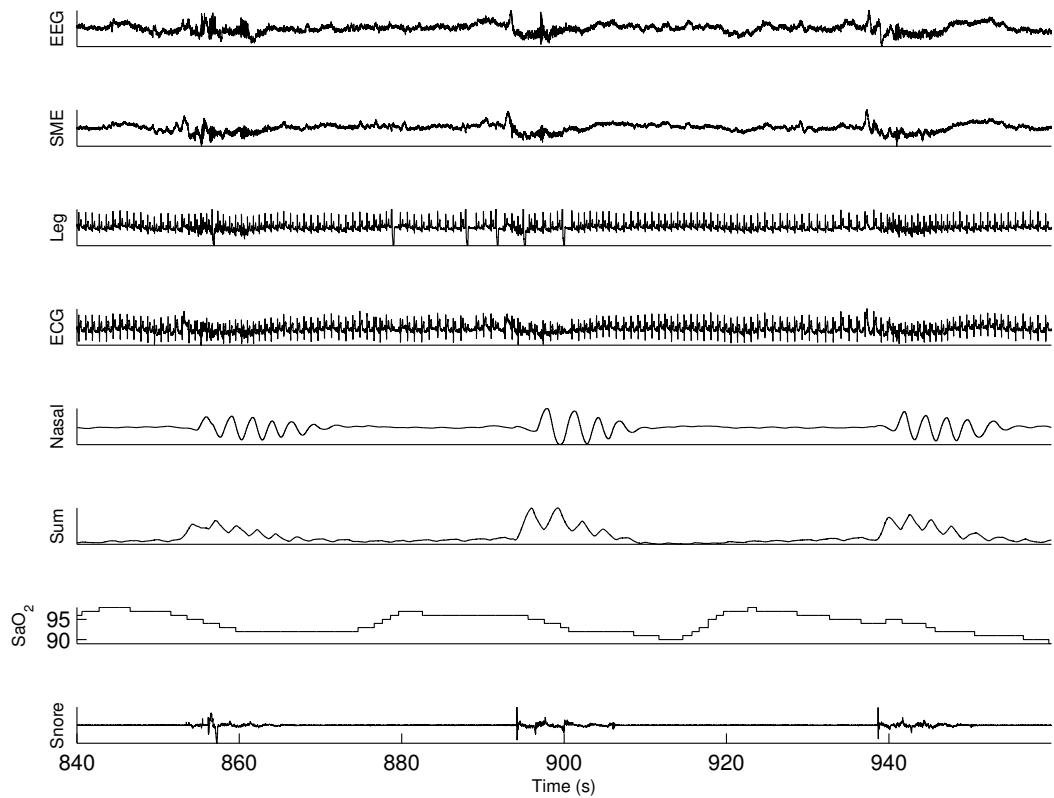


Figure 2.6 Top to bottom: EEG (F4), submental EMG (SME), leg EMG, ECG, airflow (nasal pressure cannula), sum of thoracic and abdominal activity (sum, from respiratory inductance plethysmography), SaO_2 %, and snoring sound signals selected from a 14-channel PSG record of a patient with OSA. Amplitude information has been removed from all channels except SaO_2 % to reduce clutter. Data obtained from Hisham Alshaer and T. Douglas Bradley, Sleep Research Laboratory of the University Health Network Toronto Rehabilitation Institute, Toronto, Ontario, Canada, with permission [39, 41].

systems have been developed and are commercially available for the diagnosis and follow-up of sleep apnea. One such system [34] captures the following signals to diagnose patients suspected of having sleep apnea: SpO_2 using a fingertip pulse oximeter; heart rate, also from the pulse oximeter; pulse amplitude; nasal airflow (through measurement of pressure with a nasal cannula); snoring sound signal using a microphone attached to the throat; and body position using an accelerometer. Additional signals that may be acquired are respiratory airflow (with a pneumotachograph, when CPAP is used); mask pressure (when CPAP is used); respiratory movements; EMG of the legs (for the diagnosis of periodic leg movement); and EMG of the masseter (jaw muscle, for the diagnosis of bruxism).

Figure 2.7 shows a segment of four channels of signals (duration = 4 min) from a home apnea monitor for a subject with mild sleep apnea. Concurrent episodes of increasing heart rate and snoring sound are seen in relation to the two brief episodes of apnea (periods of no or low variations in the nasal pressure signal just before the markers for 32 and 33 min).

Figure 2.8 shows a segment of four channels of signals (of duration 5 min) from an apnea monitor for a subject with severe apnea. Several episodes of substantial hemoglobin desaturation are seen in the record. It is evident that periods of apnea are directly associated with episodes of

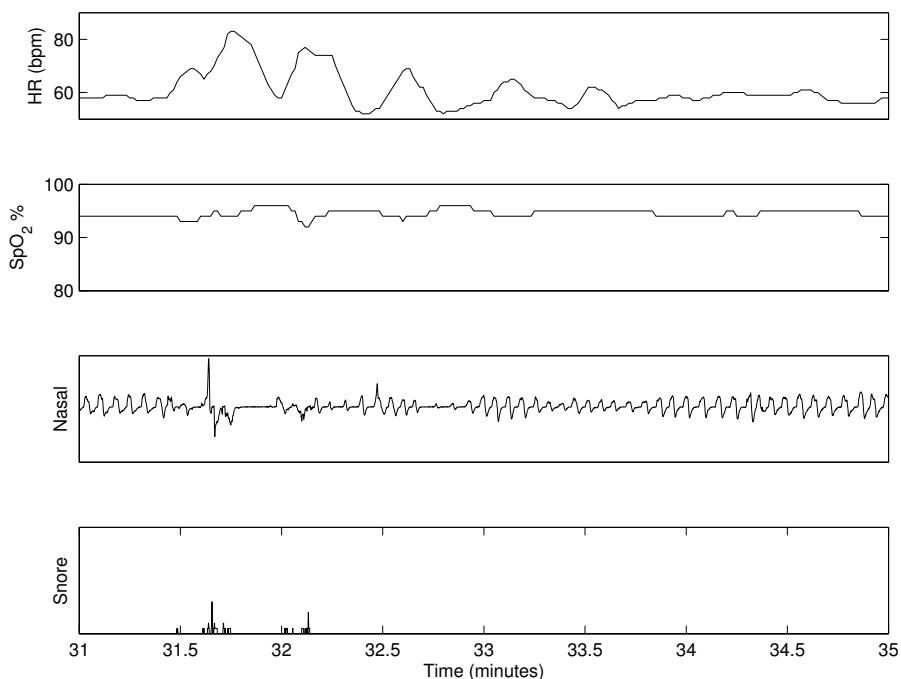


Figure 2.7 Top to bottom: Heart rate, SpO_2 , nasal pressure, and snoring sound signals from a home apnea monitoring record of a moderately symptomatic subject. Amplitude information has been removed from the nasal pressure and snoring sound channels to reduce clutter. Data courtesy of R. Platt, SagaTech Electronics Inc., Calgary, Alberta, Canada, <https://www.sagatech.ca/>.

snoring, reduced SpO_2 , and increased heart rate. The clinical report of the study indicated that the patterns of airflow and oxygen saturation signals seen in the figure are compatible with an obstructive condition.

See Koley and Dey [45] for the description of a system to detect apnea and hypopnea using the SpO_2 signal from a fingertip oximeter. Ongoing research in the field of sleep signal analysis includes exploiting sparse characteristics of wearable sensor signals collected at the convenience of home, including the use of compressive sensing and data compression techniques to suit low-bandwidth and low-power applications [46].

2.4.3 Multivariate and multiorgan analysis

Bianchi et al. [47] proposed a multivariate and multiorgan approach for the analysis of cardiorespiratory variability signals with application to the analysis of sleep and sleep-related disorders. This approach emphasizes that the ANS influences several organs and systems, including the cardiovascular, respiratory, and endocrine-metabolic systems; it also indicates a direct connection to the central and peripheral nervous systems. Bianchi et al. [47] conducted PSG studies including the following signals: three EEG leads (C3–A2, C4–A1, and O2–A1), two EOG leads, three EMG leads (chin, right tibia, and left tibia), one ECG lead, nasal and oral airflow using thermistors, and thoracic and abdominal respiration with piezoelectric belts. An HRV signal was obtained from the ECG as a sequence of RR intervals. The results obtained indicated that respiratory activation and tachycardia precede periodic leg movement, which was taken to correlate with sympathetic preactivation. Bianchi et al. observed that autonomic changes could be the first manifestation of central activity involving other peripheral systems, and precede related changes in the EEG and EMG. They also

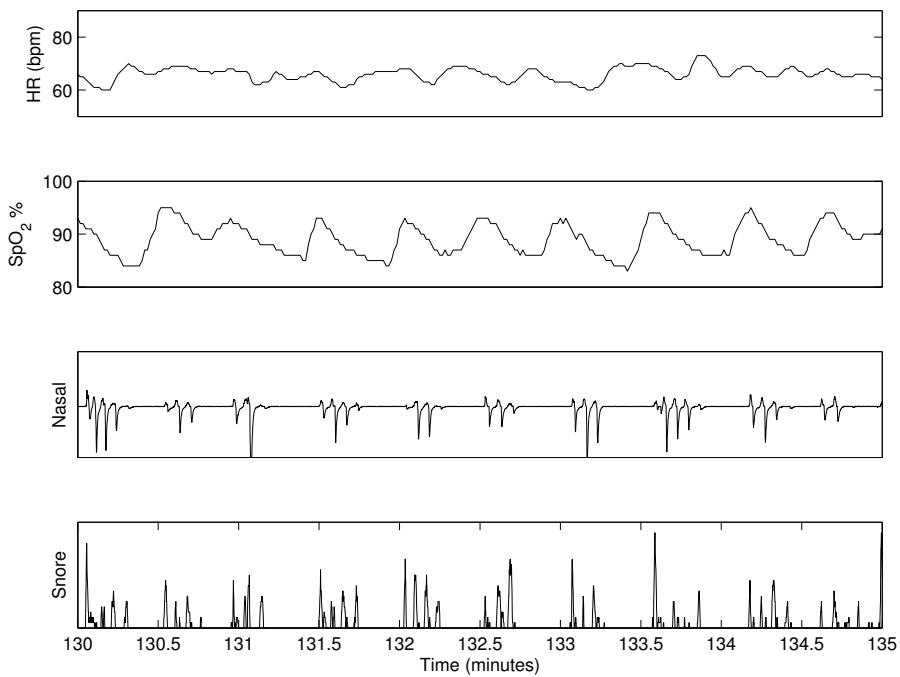


Figure 2.8 Top to bottom: Heart rate, SpO_2 , nasal pressure, and snoring sound signals from a home apnea monitoring record of a subject with severe sleep apnea. Amplitude information has been removed from the nasal pressure and snoring sound channels to reduce clutter. Data courtesy of R. Platt, Sagatech Electronics Inc., Calgary, Alberta, Canada, <https://www.sagatech.ca/>.

observed that HRV may start a few cardiac beats before short arousals, an arousal being considered as a homeostatic mechanism to recover cardiorespiratory functionality. The results were taken to indicate the existence of strong coupling between the ANS and CNS during sleep, and that a mismatch in the coupling strength may lead to pathological situations. It was noted that therapeutic strategies should not be targeted to a single organ or system but should take into consideration the interactions among various systems.

Cerutti [48] presents compelling arguments for an integrated approach for analysis of signals arising from coupled and correlated biological or physiological systems. Figure 2.9 shows four traces of signals including the EEG, EMG from the tibia, RR intervals, and respiration from a patient with myoclonus (involuntary twitching or jerking of a muscle) during sleep. Repeated jerking of the leg, known as the restless leg syndrome, is associated with sleep disruption and apnea, among several other disorders. The signals in Figure 2.9 demonstrate synchronization between arousal events in the EEG, spikes of activity related to myoclonus in the tibial EMG, increased heart rate (diminished RR intervals), and decreased respiratory activity. (See Somers et al. [49] for a discussion on the autonomic and hemodynamic responses to OSA and high sympathetic nerve activity in patients with OSA.)

In order to simplify the process of monitoring sleep apnea, several researchers have studied various secondary markers of apnea and hypopnea. Contrary to the multivariate approach described above, it might be desirable to be able to detect apnea using only one or a small number of signals obtained by minimally intrusive ways. It is well known that the heart rate varies with respiration [11]; see Section 2.2.4 and Figure 2.3. Based on this knowledge, researchers have attempted to derive information related to respiration from HRV signals (see Sections 7.2.2, 7.9, and 8.12 for discussions on HRV). Episodes of apnea have been observed to be accompanied by cyclical variations

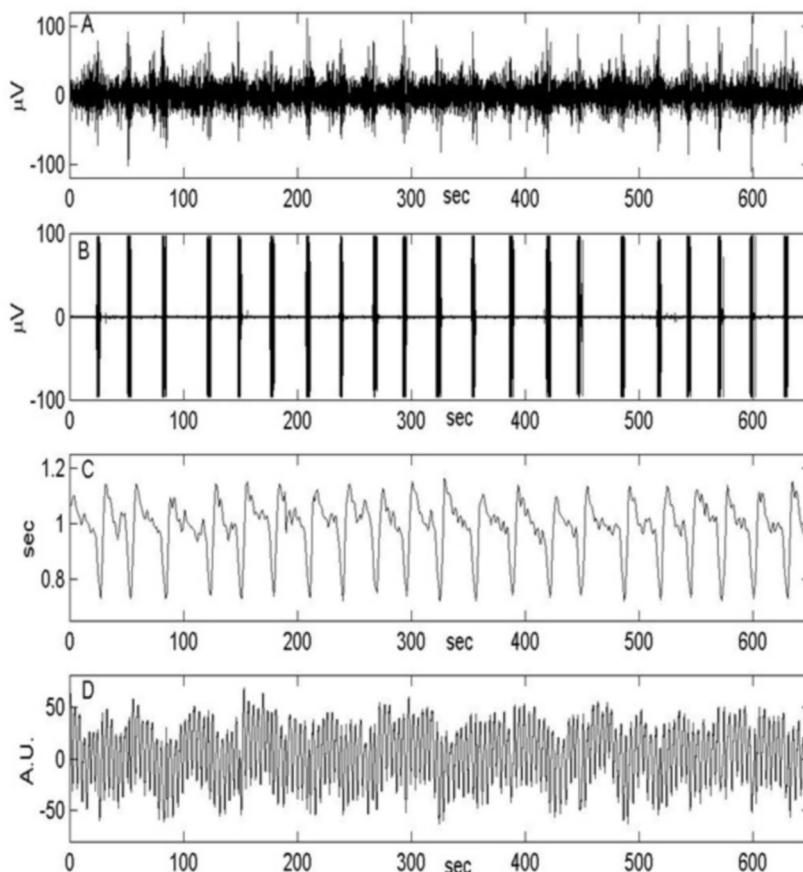


Figure 2.9 Top to bottom: EEG, EMG from the tibia, RR intervals, and respiration signal from a patient with myoclonus (involuntary twitching or jerking of a muscle) during sleep. Reproduced with permission from S. Cerutti, “Methods of biomedical signal processing: multiparametric and multidisciplinary integration toward a better comprehension of pathophysiological mechanisms,” pp 3–31, Chapter 1 in *Advanced Methods of Biomedical Signal Processing*, Edited by S. Cerutti and C. Marchesi, IEEE and Wiley, New York, NY, 2011. ©IEEE.

in RR intervals of ECG signals, demonstrating bradycardia during apnea followed by tachycardia upon return to normal respiration. de Chazal et al. [50] proposed methods to derive a surrogate signal related to respiration from a single-lead ECG signal. Several features were derived from the RR intervals of the ECG signal and the surrogate respiration signal, including the mean, SD , correlation coefficients, measures of variation, and spectral density. Features based only on RR intervals provided a minute-by-minute rate of over 85% in the identification of episodes of apnea. Bsoul et al. [33] developed a real-time sleep apnea monitor using single-lead ECG. Measures were extracted from an HRV signal formulated using RR intervals and from a surrogate respiration signal derived from one-minute segments of the ECG signal for the detection of sleep apnea. High sensitivity of up to 96% was obtained.

Madhav et al. [51] proposed methods to derive respiratory activity from signals such as the ECG and the PPG using modeling techniques; see Section 1.2.11. Arunachalam and Brown [52] proposed a real-time algorithm to estimate and remove baseline wander and obtain a surrogate respiration signal from an ECG signal. The respiration signal was estimated from the amplitude modulation of R waves in the ECG caused by breathing. See Khandoker et al. [53] for related studies.

Mendez et al. [54] proposed a method for the detection of OSA based on the ECG recorded during sleep. Several parameters were derived from the ECG signal using time-variant modeling techniques. One of the features giving good results in the detection of OSA was the coherence between the RR interval data and the area under the QRS wave. This result was considered to suggest that respiration not only causes respiratory sinus arrhythmia, but also modifies the ECG waveshape. The coherence feature was considered to be a direct representation of the relation between the respiratory aspects of OSA and their synchronization with the heart rate.

Patangay et al. [55] proposed methods for the detection of apnea based on features of ECG as well as PCG signals. It was observed that the decrease in intrathoracic pressure during OSA causes a decrease in left-ventricular pressure; this, in turn, results in increasingly stronger left-ventricular contraction. As a result, the PCG signal during OSA, in particular the S1 amplitude, has a crescendo-like change in amplitude. A composite feature vector was prepared using subband decomposition of S1 amplitudes and RR intervals. A sensitivity of 85.5% and a specificity of 92.2% were obtained in the detection of episodes of OSA.

Alshaer et al. [39, 41] developed a specially designed face frame with a microphone to record breath sounds. They proposed single-channel signal processing techniques to detect sleep-disordered breathing (CSA or OSA) via characterization of the envelope and spectral properties of the breath sounds. Using a diagnostic limit of $AHI \geq 10$ based on PSG, the overall accuracy of the breath-sound-based method was about 88%, with an overall correlation of AHI with PSG of 94%.

Notwithstanding the benefits of detecting apnea with a single signal, it should be noted that each such approach may present its own limitations and may not provide the desired degree of suitability, robustness, or dependability in a clinical application. Similar to the acceptance of redundancy in the 12-lead system of ECG in clinical practice for the sake of the accompanying robustness, it may be desirable to use multiple signals from several systems that may demonstrate different effects or manifestations of apnea. Although substantial numbers of research studies are being conducted on various issues related to OSA and CSA, as reviewed briefly in the present section, systems and methods to detect and treat apnea are now commercially available [34, 56].

2.5 Remarks

This chapter has introduced the notion of using multiple channels of biomedical signals to obtain information on concurrent, coupled, and correlated phenomena with the aim of obtaining an improved understanding of physiological systems or obtaining reference signals for various purposes. The main point to note is that physiological systems are complex systems with multiple variables and outputs that should be studied from various approaches in order to gain multifaceted information. See Cerutti [48] and Baselli et al. [57] for discussions on several techniques and parametric models for the analysis of interactions between biomedical signals and systems.

Some of the problems described in the present chapter have been stated in fairly general terms due to the introductory nature of the chapter. The subsequent chapters present more illustrations of specific problems and applications of the notions gained from this chapter. A number of examples are provided in the chapters that follow to illustrate the use of multiple channels of signals to obtain clinically useful information.

2.6 Study Questions and Problems

1. A patient has ventricular bigeminy: Every other pulse from the SA node is replaced by a PVC with a full compensatory pause. (See Figure 10.10 for an illustration of bigeminy.) The firing rate of the SA node is regular at 80 beats a minute, and each ectopic beat precedes the blocked SA-node pulse by 100 ms.

Draw a schematic three-channel representation of the ECG, the atrial electrogram (or the SA node's firing pattern), and the firing pattern of the ectopic focus for 10 beats, marking the time scale in detail. Identify the correspondences and relationships between the activities in the three channels.

2. Draw schematic representations of the ECG, PCG, and carotid pulse signals. Label all waves in the three signals. Identify their common relationships to events in the cardiac cycle.

2.7 Laboratory Exercises and Projects

(Note: The following projects require access to a physiological signal recording laboratory. See also the note at the beginning of Section 1.8.)

1. Using a multichannel biomedical signal acquisition system, obtain simultaneous recordings of an ECG channel and a signal related to respiration (temperature, airflow, or pressure in the nostril). Study the variations in the RR interval with inspiration and expiration. Repeat the experiment with the subject holding his/her breath during the signal acquisition period.
2. Obtain simultaneous recordings of an ECG lead, the PCG, the carotid pulse, and the pulse at the wrist. Study the temporal correspondences (and delays) between events in the various channels.
3. Record an ECG lead and PCG signals from two or three auscultation areas (mitral, aortic, pulmonary, tricuspid, and apex: see Figure 1.33) simultaneously. Study the variations in the intensities and characteristics of S1 and S2 and their components in the PCGs from the various recording sites.

References

- [1] Rushmer RF. *Cardiovascular Dynamics*. WB Saunders, Philadelphia, PA, 4th edition, 1976.
- [2] Akay AM, Semmlow JL, Welkowitz W, Bauer MD, and Kostis JB. Detection of coronary occlusions using autoregressive modeling of diastolic heart sounds. *IEEE Transactions on Biomedical Engineering*, 37(4):366–373, 1990.
- [3] Shaver JA, Salerni R, and Reddy PS. Normal and abnormal heart sounds in cardiac diagnosis, Part I: Systolic sounds. *Current Problems in Cardiology*, 10(3):1–68, 1985.
- [4] Reddy PS, Salerni R, and Shaver JA. Normal and abnormal heart sounds in cardiac diagnosis, Part II: Diastolic sounds. *Current Problems in Cardiology*, 10(4):1–55, 1985.
- [5] Lehner RJ and Rangayyan RM. A three-channel microcomputer system for segmentation and characterization of the phonocardiogram. *IEEE Transactions on Biomedical Engineering*, 34:485–489, 1987.
- [6] Tavel ME. *Clinical Phonocardiography and External Pulse Recording*. Year Book Medical, Chicago, IL, 3rd edition, 1978.
- [7] Jenkins JM, Wu D, and Arzbaecher RC. Computer diagnosis of abnormal cardiac rhythms employing a new P-wave detector for interval measurement. *Computers and Biomedical Research*, 11:17–33, 1978.
- [8] Jenkins JM, Wu D, and Arzbaecher RC. Computer diagnosis of supraventricular and ventricular arrhythmias. *Circulation*, 60(5):977–987, 1979.
- [9] Jenkins JM. Computerized electrocardiography. *CRC Critical Reviews in Bioengineering*, 6:307–350, November 1981.
- [10] Jenkins JM. Automated electrocardiography and arrhythmia monitoring. *Progress in Cardiovascular Disease*, 25(5):367–408, 1983.
- [11] Sayers B.McA. Analysis of heart rate variability. *Ergonomics*, 16(1):17–32, 1973.
- [12] Kobayashi M and Musha T. 1/f fluctuation of heartbeat period. *IEEE Transactions on Biomedical Engineering*, 29(6):456–457, 1982.
- [13] Rompelman O, Snijders JBIM, and van Spronsen CJ. The measurement of heart rate variability spectra with the help of a personal computer. *IEEE Transactions on Biomedical Engineering*, 29(7):503–510, 1982.

- [14] deBoer RW, Karemaker JM, and Strackee J. Comparing spectra of a series of point events particularly for heart rate variability studies. *IEEE Transactions on Biomedical Engineering*, 31(4):384–387, 1984.
- [15] Rosenblum MG, Kurths J, Pikovsky A, Schäfer C, Tass P, and Abel HH. Synchronization in noisy systems and cardiorespiratory interaction. *IEEE Engineering in Medicine and Biology Magazine*, 17(6):46–53, 1998.
- [16] Pompe B, Blidh P, Hoyer D, and Eiselt M. Using mutual information to measure coupling in the cardiorespiratory system. *IEEE Engineering in Medicine and Biology Magazine*, 17(6):32–39, 1998.
- [17] Durand LG, Genest Jr. J, and Guardo R. Modeling of the transfer function of the heart-thorax acoustic system in dogs. *IEEE Transactions on Biomedical Engineering*, 32(8):592–601, 1985.
- [18] Kantz H, Kurtis J, and Mayer-Kress G, editors. *Nonlinear Analysis of Physiological Data*. Springer-Verlag, Berlin, Germany, 1998.
- [19] Olshansky B, Sabbah HN, Hauptman PJ, and Colucci WS. Parasympathetic nervous system and heart failure: Pathophysiology and potential implications for therapy. *Circulation*, 118:863–871, 2008.
- [20] Kamath MV and Fallen EL. Power spectral analysis of heart rate variability: A noninvasive signature of cardiac autonomic function. *Critical Reviews in Biomedical Engineering*, 21(3):245–311, 1993.
- [21] Kamath MV, Watanabe M, and Upton A, editors. *Heart Rate Variability (HRV) Signal Analysis: Clinical Applications*. CRC Press, Boca Raton, FL, 2012.
- [22] Ishaque S, Khan N, and Krishnan S. Trends in heart-rate variability signal analysis. *Frontiers in Digital Health*, 3(639444), 2021.
- [23] Malik M and Camm AJ. Heart rate variability. *Clinical Cardiology*, 13(8):570–576, 1990.
- [24] La Rovere MT. Evaluation of the autonomic nervous system: From algorithms to clinical practice. In Cerutti S and Marchesi C, editors, *Advanced Methods of Biomedical Signal Processing*, pages 83–97. IEEE and Wiley, New York, NY, 2011.
- [25] Malik MJ, Bigger T, Camm AJ, Kleiger RE, Malliani A, Moss AJ, and Schwartz PJ. Guidelines — Heart rate variability: Standards of measurement, physiological interpretation, and clinical use (Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology). *European Heart Journal*, 17(3):354–381, 1996.
- [26] Goodgold J and Eberstein A. *Electrodiagnosis of Neuromuscular Diseases*. Williams and Wilkins, Baltimore, MD, 3rd edition, 1983.
- [27] de Luca CJ. Physiology and mathematics of myoelectric signals. *IEEE Transactions on Biomedical Engineering*, 26:313–325, 1979.
- [28] Zhang YT, Frank CB, Rangayyan RM, and Bell GD. A comparative study of vibromyography and electromyography obtained simultaneously from active human quadriceps. *IEEE Transactions on Biomedical Engineering*, 39(10):1045–1052, 1992.
- [29] Zhang YT, Frank CB, Rangayyan RM, and Bell GD. Relationships of the vibromyogram to the surface electromyogram of the human rectus femoris muscle during voluntary isometric contraction. *Journal of Rehabilitation Research and Development*, 33(4):395–403, 1996.
- [30] Zhang YT, Rangayyan RM, Frank CB, and Bell GD. Adaptive cancellation of muscle contraction interference from knee joint vibration signals. *IEEE Transactions on Biomedical Engineering*, 41(2):181–191, 1994.
- [31] Haykin S. *Adaptive Filter Theory*. Prentice-Hall, Upper Saddle River, NJ, 3rd edition, 1996.
- [32] Widrow B, Glover Jr. JR, McCool JM, Kaunitz J, Williams CS, Hearn RH, Zeidler JR, Dong Jr. E, and Goodlin RC. Adaptive noise cancelling: Principles and applications. *Proceedings of the IEEE*, 63(12):1692–1716, 1975.
- [33] Bsoul M, Minn H, and Tamil L. Apnea MedAssist: Real-time sleep apnea monitor using single-lead ECG. *IEEE Transactions on Information Technology in Biomedicine*, 15(3):416–427, 2011.
- [34] SagaTech Electronics Inc., Calgary, Alberta, Canada, www.sagatech.ca, accessed on 2023-04-26. *Remmers Sleep Recorder*.

- [35] Young T, Palta M, Dempsey J, Skatrud J, Weber S, and Badr S. The occurrence of sleep-disordered breathing among middle-aged adults. *New England Journal of Medicine*, 328:1230–1235, 1993.
- [36] Yen FC, Behbehani H, Lucas E, Burk J, and Axe J. A noninvasive technique for detecting obstructive and central sleep apnea. *IEEE Transactions on Biomedical Engineering*, 44(12):1262–1268, 1997.
- [37] Penzel T, Moody GB, Mark RG, Goldberger AL, and Peter JH. The Apnea-ECG database. In *Proceedings of IEEE Computers in Cardiology*, pages 255–258, <https://www.physionet.org/content/apnea-ecg/1.0.0/>, 2000.
- [38] Chesson, Jr. AL, Ferber RA, Fry JM, Grigg-Damberger M, Hartse KM, Hurwitz TD, Johnson S, Littner M, Kader GA, Rosen G, Sangal RB, Schmidt-Nowara W, and Sher A. Practice parameters for the indications for polysomnography and related procedures. *Sleep*, 20:406–422, 1997.
- [39] Alshaer H, Fernie GR, Maki E, and Bradley TD. Validation of an automated algorithm for detecting apneas and hypopneas by acoustic analysis of breath sounds. *Sleep Medicine*, 14(6):562–571, 2013.
- [40] Bradley TD and Floras JS. Sleep apnea and heart failure, Part I: Obstructive sleep apnea. *Circulation*, 107(12):1671–1678, 2003.
- [41] Alshaer H, Fernie GR, and Bradley TD. Monitoring of breathing phases using a bioacoustic method in healthy awake subjects. *Journal of Clinical Monitoring and Computing*, 25(5):285–294, 2011.
- [42] Yasuda Y, Umezawa A, Horihata S, Yamamoto K, Miki R, and Koike S. Modified thoracic impedance plethysmography to monitor sleep apnea syndromes. *Sleep Medicine*, 6(3):215–224, 2005.
- [43] Kushida CA, Littner MR, Morgenthaler T, Alessi CA, Bailey D, Coleman, Jr. J, Friedman L, Hirshkowitz M, Kapen S, Kramer M, Lee-Chiong T, Loube DL, Owens J, Pancer JP, and Wise M. Practice parameters for the indications for polysomnography and related procedures: An update for 2005. *Sleep*, 28(4):499–521, 2005.
- [44] Martin RJ, Block AJ, Cohn MA, Conway WA, Hudgel DW, Powles ACP, Sanders MH, and Smith PL. Indications and standards for cardiopulmonary sleep studies. *Sleep*, 8:371–379, 1985.
- [45] Koley BL and Dey D. On-line detection of apnea/hypopnea events using SpO₂ signal: A rule-based approach employing binary classifier models. *IEEE Journal of Biomedical and Health Informatics*, 18(1):231–239, 2014.
- [46] Krishnan S. *Biomedical Signal Analysis for Connected Healthcare*. Academic Press, New York, NY, 2021.
- [47] Bianchi AM, Ferini-Strambi L, Castronovo V, and Cerutti S. Multivariate and multiorgan analysis of cardiorespiratory variability signals: The CAP sleep case. *Biomedical Technology*, 51:167–173, 2006.
- [48] Cerutti S. Methods of biomedical signal processing: Multiparametric and multidisciplinary integration toward a better comprehension of pathophysiological mechanisms. In Cerutti S and Marchesi C, editors, *Advanced Methods of Biomedical Signal Processing*, pages 3–31. IEEE and Wiley, New York, NY, 2011.
- [49] Somers VK, Dyken ME, Clary MP, and Abboud FM. Sympathetic neural mechanisms in obstructive sleep apnea. *Journal of Clinical Investigation*, 96(4):1897–1904, 1995.
- [50] de Chazal P, Heneghan C, Sheridan E, Reilly R, Nolan P, and O’Malley M. Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnea. *IEEE Transactions on Biomedical Engineering*, 50(6):686–696, 2003.
- [51] Madhav KV, Raghuram M, Krishna EH, Komalla NR, and Reddy KA. Extraction of respiratory activity from ECG and PPG signals using vector autoregressive model. In *Proceedings of the 2012 IEEE International Symposium on Medical Measurements and Applications Proceedings*, pages 132–135, Budapest, Hungary, June 2012.
- [52] Arunachalam SP and Brown LF. Real-time estimation of the ECG-derived respiration (EDR) signal using a new algorithm for baseline wander noise removal. In *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5681–5684, Minneapolis, MN, September 2009.
- [53] Khandoker AH, Palaniswami M, and Karmakar C. Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings. *IEEE Transactions on Information Technology in Biomedicine*, 13(1):37–48, 2009.

- [54] Mendez MO, Bianchi AM, Matteucci M, Cerutti S, and Penzel T. Sleep apnea screening by autoregressive models from a single ECG lead. *IEEE Transactions on Biomedical Engineering*, 56(12):2838–2850, 2009.
- [55] Patangay A, Vemuri P, and Tewfik A. Monitoring of obstructive sleep apnea in heart failure patients. In *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1043–1046, Lyon, France, August 2007.
- [56] ResMed. www.resmed.com, accessed on 2023-06-26.
- [57] Baselli G, Porta A, and Bolzern P. Parametric models for the analysis of interactions in biomedical signals. In Cerutti S and Marchesi C, editors, *Advanced Methods of Biomedical Signal Processing*, pages 101–125. IEEE and Wiley, New York, NY, 2011.

CHAPTER 3

FILTERING FOR REMOVAL OF ARTIFACTS

Most biomedical signals appear as weak signals in an environment that is teeming with many other signals of various origins. Any signal other than that of interest could be termed as an interference, artifact, or simply *noise*. The sources of noise could be physiological, the instrumentation used, or the environment of the experiment.

This chapter starts with an introduction to the nature of the artifacts that are commonly encountered in biomedical signals. Several illustrations of signals corrupted by various types of artifacts are provided. Details of the design of filters, spanning a broad range of approaches, from linear time-domain and frequency-domain fixed filters to the optimal Wiener filter to adaptive filters, are then presented. The chapter concludes with demonstrations of application of the filters described to ECG, ERP, and VAG signals.

(*Note:* A good background in signal and system analysis [1–4] as well as in probability, random variables, and stochastic processes [5–10] is required to follow the procedures and analysis described in this chapter. Familiarity with systems theory and transforms, such as the Laplace and Fourier transforms in both the continuous and discrete forms as well as the z -transform, is assumed. A few related topics are briefly described where needed.)

3.1 Problem Statement

Noise is omnipresent! The problems caused by artifacts in biomedical signals are vast in scope and variety; their potential to degrade the performance of the most highly trained experts and sophisticated signal processing algorithms is high. The enormity of the problem of noise removal and its importance are reflected by the size of this chapter and its placement as the first one on sig-

nal processing techniques. Let us start with a generic statement of the problem and investigate its nature:

Analyze the various types of artifacts that corrupt biomedical signals and explore filtering techniques to remove them without degrading the signal of interest.

During a procedure to acquire the ECG signal, if the subject coughs or squirms, the EMG associated with such activity will create an interference or artifact. In adult patients, such physiological interference may be minimized by strict instructions and self-control; this solution may, however, not be applicable to infants, children, and severely ill patients. An intriguing example of physiological interference is that of the expectant mother's ECG appearing along with that of the fetus, with the latter being of interest. No external control is feasible or desirable in this case, and the investigator is forced to develop innovative solutions to extract the signal of interest.

Due to the weak levels of most biomedical signals at their source, high amplification factors of several hundred to several thousand may be required. Electronic noise in the instrumentation amplifiers also gets amplified along with the desired signal. While it is possible to reduce the thermal component of the noise by cooling the devices to low temperatures, this step may not be practical in most applications; the cost could also be prohibitive. Low-noise power supplies and specialized electronic amplifiers with high input impedance, high common-mode rejection ratio, and high power-supply rejection ratio are desirable for the acquisition of biomedical signals [11].

Our environment is filled with EM waves, both natural and man-made. EM waves broadcast by radio and television (TV) stations and those radiated by fluorescent lighting devices, computer monitors, and other systems used in the laboratory or work environment are picked up by cables, devices, and connectors. The 50 Hz or 60 Hz power-supply waveform is notorious for the many ways in which it can get mixed with and corrupt a signal of interest. Such interference may be termed as being due to the environment of the experiment. Simple EM shielding of cables and grounding of the chassis of equipment reduce EM and power-supply interference in most cases. Experiments dealing with weak signals such as ERPs and EEGs may require a wire-mesh-shielded (Faraday) cage to contain the subject and the instruments.

The ECG is a relatively strong signal with a readily identifiable waveform. Most types of interference that affect ECG signals may be removed by bandpass filters. Other signals of less-recognizable waveforms and broader bandwidths may not be amenable to simple filtering procedures. In the case of signals such as ERPs or SEPs, the noise levels could be much higher than the signal levels, rendering the latter unrecognizable in a single recording. It is important to gain a good understanding of the noise processes involved before one attempts to filter or preprocess a signal.

3.2 Random, Structured, and Physiological Noise

Several types of artifacts or noise could corrupt biomedical signals in various ways. The following sections provide descriptions of various types of noise and methods to characterize them.

3.2.1 Random noise

A *deterministic signal* is one whose value at a given instant of time may be computed using a closed-form mathematical function of time, or predicted from a knowledge of a few past values of the signal. A signal that does not meet this condition may be labeled as a *nondeterministic signal* or a random signal. The term *random noise* refers to an interference that arises from a random process, such as thermal noise in electronic devices.

Test for randomness: Random signals are generally expected to display more excursions about a certain reference level within a specified interval than signals that are predictable. Kendall [12], and Challis and Kitney [13] recommend a test for randomness based on the number of peaks or troughs

in the signal. A peak or a trough is defined by a set of three consecutive samples of the signal, with the central sample being either the maximum or minimum, respectively. As the direction of excursion of the signal changes at peaks and troughs, such points are collectively known as *turning points*. At a turning point, the sign of the first-order difference (derivative) at the current sample of the signal is not equal to that at the preceding sample. Given a signal of N samples, the signal may be labeled as being random if the number of turning points is greater than the threshold $\frac{2}{3}(N - 2)$ [12, 13]. In the case of a signal of varying characteristics, that is, a nonstationary signal, the test would have to be conducted using a running window of N samples. The width of the window should be chosen by taking into consideration the shortest duration over which the signal may remain in a given state. The method as above was used by Mintchev et al. [14] to study the dynamics of the level of randomness in EGG signals.

Figure 3.1 illustrates the variation in the number of turning points in a moving window of 50 ms (400 samples with the sampling frequency $f_s = 8 \text{ kHz}$) for the speech signal of the word “safety.” The limit on the number of turning points for randomness for $N = 400$ according to the threshold mentioned above is 265. It is seen from the figure that the test indicates that the signal is random for the fricatives /S/ (over the interval of 0.2 – 0.4 s, approximately) and /F/ (0.7 – 0.9 s), and not random for the remaining portions, as expected. (See also Section 1.2.13 and Figures 1.54 and 1.55.)

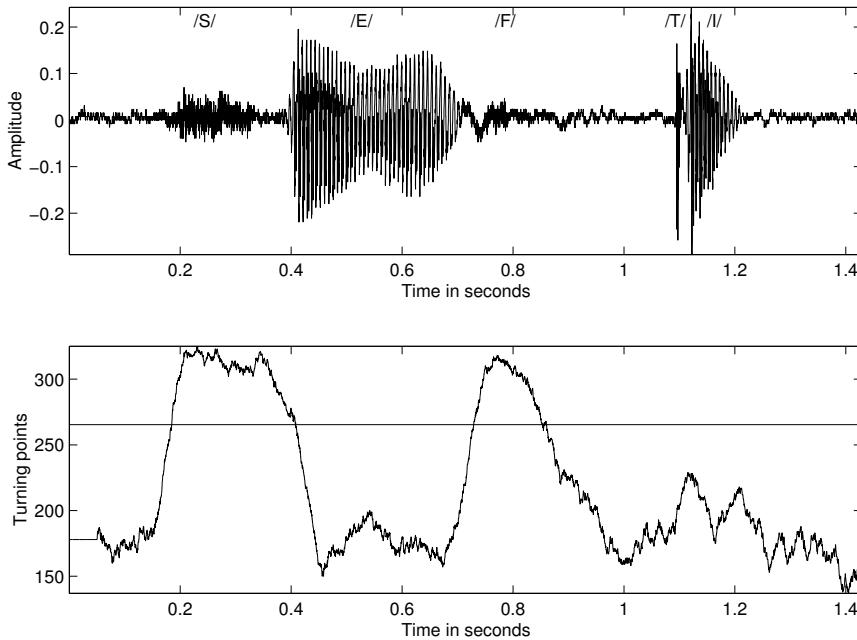


Figure 3.1 Top: Speech signal of the word “safety” uttered by a male speaker. Bottom: Count of turning points in a moving window of 50 ms (400 samples with $f_s = 8 \text{ kHz}$) shifted one sample at a time. The threshold for randomness for $N = 400$ is 265.

Statistical analysis of random processes: A random process is characterized by its probability density function (PDF) representing the probabilities of occurrence of all possible values of the related random variable. (See Papoulis [5] and Bendat and Piersol [6] for background material on probability, random variables, and stochastic processes.) Consider a random process η that is characterized by the PDF $p_\eta(\eta)$. While the PDF provides a complete representation of the random process, a few well-known statistical measures are useful in characterizing the random variable for practical purposes. The mode is the value of the random variable corresponding to the peak or the highest value of the PDF (assuming a unimodal or single-peaked PDF). The mean μ_η of the random

process η is given by the first-order moment of the PDF, defined as

$$\mu_\eta = E[\eta] = \int_{-\infty}^{\infty} \eta p_\eta(\eta) d\eta, \quad (3.1)$$

where $E[]$ represents the *statistical expectation operator*. It is common to assume the mean of a random noise process to be zero.

The mean-squared (*MS*) value of the random process η is given by the second-order moment of the PDF, defined as

$$E[\eta^2] = \int_{-\infty}^{\infty} \eta^2 p_\eta(\eta) d\eta. \quad (3.2)$$

The variance σ_η^2 of the process is defined as the second central moment:

$$\sigma_\eta^2 = E[(\eta - \mu_\eta)^2] = \int_{-\infty}^{\infty} (\eta - \mu_\eta)^2 p_\eta(\eta) d\eta. \quad (3.3)$$

The square root of the variance gives the *SD* σ_η of the process. Note that $\sigma_\eta^2 = E[\eta^2] - \mu_\eta^2$. If the mean is zero, it follows that $\sigma_\eta^2 = E[\eta^2]$, that is, the variance and the *MS* values are the same.

The coefficient of variation (*CV*) of a process is defined as the ratio σ/μ . This dimensionless measure places the variability of a process in the context of its mean and is useful in the analysis of the variability of processes with widely different means; however, as $\mu \rightarrow 0$, $CV \rightarrow \infty$. *CV* is applicable in the analysis of nonnegative quantities.

A few other measures used to characterize random processes are skewness, kurtosis, and entropy. Skewness is defined as a normalized version of the third central moment, given by

$$S_\eta = \frac{1}{\sigma_\eta^3} \int_{-\infty}^{\infty} (\eta - \mu_\eta)^3 p_\eta(\eta) d\eta. \quad (3.4)$$

Skewness characterizes the lack of symmetry of a PDF. Symmetric PDFs, such as the Gaussian, have a skewness value of zero. A random process with negative skewness is indicated by a longer or thicker “tail” of the PDF to the left or lower side of the random variable; the mean value is lower than or to the left of the mode. On the other hand, a process with positive skewness has a longer or thicker tail of the PDF to the right or higher side of the random variable; the mean value is higher than or to the right of the mode.

Kurtosis is defined as a normalized version of the fourth central moment, given by

$$K_\eta = \frac{1}{\sigma_\eta^4} \int_{-\infty}^{\infty} (\eta - \mu_\eta)^4 p_\eta(\eta) d\eta. \quad (3.5)$$

Kurtosis characterizes the presence of a long tail in the PDF. The Gaussian PDF has a kurtosis value of 3. The value $K' = K - 3$ (referred to as kurtosis excess) is used to represent the difference in the kurtosis of a PDF with respect to that of a Gaussian PDF. A positive K' indicates a PDF with a strong peak near the mean that declines with a heavy tail. A PDF that is nearly flat or uniform has a negative K' value.

The entropy of a random process is a statistical measure of the information conveyed by the process [15–17]. It is also a measure of the extent of disorder in the process. The most commonly used measure of entropy of a PDF is defined as

$$H_\eta = - \int_{-\infty}^{\infty} p_\eta(\eta) \log_2[p_\eta(\eta)] d\eta. \quad (3.6)$$

The unit of entropy, defined as above, is *bits (b)*. The entropy of a process is at its maximum when all associated values or events occur with equal probability, that is, the PDF is uniform.

The definitions of the moments of a random process given above refer to a continuous variable and the situation when its PDF is known. In a practical situation, when we have only a certain number, N , of the values of the process observed as $\eta(n)$, $n = 0, 1, 2, \dots, N-1$, and no knowledge of its PDF, we could estimate the statistical measures from the given samples as

$$\mu_\eta = \frac{1}{N} \sum_{n=0}^{N-1} \eta(n), \quad (3.7)$$

$$MS_\eta = \frac{1}{N} \sum_{n=0}^{N-1} [\eta(n)]^2, \quad (3.8)$$

$$RMS_\eta = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} [\eta(n)]^2}, \quad (3.9)$$

and

$$\sigma_\eta = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} [\eta(n) - \mu_\eta]^2}. \quad (3.10)$$

[Some authors use the notation (nT) , $T = 1/f_s$ being the sampling interval, where f_s is the sampling frequency, to denote the index of a sampled signal; in this book we use just (n) , the sample number.] Estimates of statistical measures as above, computed from a small number of observed values of a random variable, are referred to as sample statistics or ensemble averages.

In the case of a process with a finite number L of discrete (or quantized) values η_l , $l = 0, 1, 2, \dots, L-1$, the entropy is defined as

$$H_\eta = - \sum_{l=0}^{L-1} p_\eta(\eta_l) \log_2[p_\eta(\eta_l)], \quad (3.11)$$

where $p_\eta(\eta_l)$ is the probability of occurrence of the l^{th} quantized value of η . Each probability value in the equation given above may be estimated from a large number of observations of the values of the process. This form of entropy is known as Shannon entropy [15, 16].

When the values of a random process η form a time series or a function of time, such as an electrical or electronic signal, we have a random signal (or a stochastic process) $\eta(t)$. The statistical measures described above then have physical meanings: The mean represents the DC component, the MS value represents the average power, and the RMS value gives the average magnitude or level of the signal. The measures are useful in calculating the SNR , which is commonly defined as the ratio of the peak-to-peak amplitude range of the signal of interest to the RMS value of the noise, or as the ratio of the average power of the signal to that of the noise.

Observe the use of the same symbol η to represent the random variable, the random process, and the random signal as a function of time. The subscript of the PDF or the statistical parameter derived indicates the random process of concern. The context of the discussion or expression should make the meaning of the symbol clear.

A biomedical signal of interest $x(t)$ may also, for the sake of generality, be considered to be a realization or observation of a random process x . For example, although a normal heart sound signal is heard as the same comforting *lub-dub* sound over every cycle, the corresponding PCG vibration waveforms are not precisely the same from one cycle to another. The PCG signal may be represented as a random process exhibiting certain characteristics *on the average*.

When a (random) signal $x(t)$ is observed in an environment with random noise, the measured signal $y(t)$ may be treated as a realization of another random process y . In most cases, the noise is

additive, and the observed signal is expressed as

$$y(t) = x(t) + \eta(t). \quad (3.12)$$

Each of the random processes x and y is characterized by its own PDF $p_x(x)$ and $p_y(y)$, respectively.

The means of the signals $y(t)$, $x(t)$, and $\eta(t)$ in Equation 3.12 are related as

$$E[y] = \mu_y = \mu_x + \mu_\eta; \quad (3.13)$$

if $\mu_\eta = 0$, then $\mu_y = \mu_x$. Here, μ represents the mean of the random process indicated by the subscript.

In most practical applications, the random processes representing a signal of interest and the noise affecting the signal may be assumed to be *statistically independent processes*. Two random processes x and η are said to be statistically independent if their joint PDF $p_{x,\eta}(x, \eta)$ is equal to the product of their individual PDFs given as $p_x(x) p_\eta(\eta)$.

The processes x and η are uncorrelated if $E[x\eta] = E[x]E[\eta]$; they are (mutually) orthogonal if $E[x\eta] = 0$. If x and η are independent, they are also uncorrelated. If x and η are uncorrelated and the mean of at least one of them is zero, they are also orthogonal [5].

If x and η are uncorrelated, it follows that

$$E[(y - \mu_y)^2] = \sigma_y^2 = \sigma_x^2 + \sigma_\eta^2; \quad (3.14)$$

furthermore, their covariance and correlation coefficient are zero [5].

Ensemble averages: When the PDFs of the random processes of concern are not known, it is common to approximate the statistical expectation operation by averages computed using a collection or *ensemble* of sample observations of the random process. Such averages are known as *ensemble averages*. Suppose we have M observations of the random process x as functions of time: $x_1(t), x_2(t), \dots, x_M(t)$. We may estimate the mean of the process at a particular instant of time t_1 as

$$\mu_x(t_1) = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{k=1}^M x_k(t_1). \quad (3.15)$$

Figure 3.2 illustrates 10 sample acquisitions of flash visual ERPs (see also Figure 3.43). The vertical lines at $t = t_1$ and $t = t_2 = t_1 + \tau$ represent the ensemble averaging process at two different instants of time.

The autocorrelation function (ACF) $\phi_{xx}(t_1, t_1 + \tau)$ of a random process x that is a time series is given by

$$\begin{aligned} \phi_{xx}(t_1, t_1 + \tau) &= E[x(t_1) x(t_1 + \tau)] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(t_1) x(t_1 + \tau) p_{x_1, x_2}(x_1, x_2) dx_1 dx_2, \end{aligned} \quad (3.16)$$

where x_1 and x_2 represent the random variables corresponding to the processes $x(t_1)$ and $x(t_1 + \tau)$, respectively, and $p_{x_1, x_2}(x_1, x_2)$ is the joint PDF of the two processes. The ACF may be estimated as

$$\phi_{xx}(t_1, t_1 + \tau) = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{k=1}^M x_k(t_1) x_k(t_1 + \tau), \quad (3.17)$$

where τ is the delay parameter. If the signals are complex, one of the functions in the expression above should be conjugated; in this book, we deal with physiological signals that are always real. The two vertical lines at $t = t_1$ and $t = t_2 = t_1 + \tau$ in Figure 3.2 represent the ensemble averaging process to compute $\phi_{xx}(t_1, t_2)$. The ACF indicates how the values of a signal at a particular instant

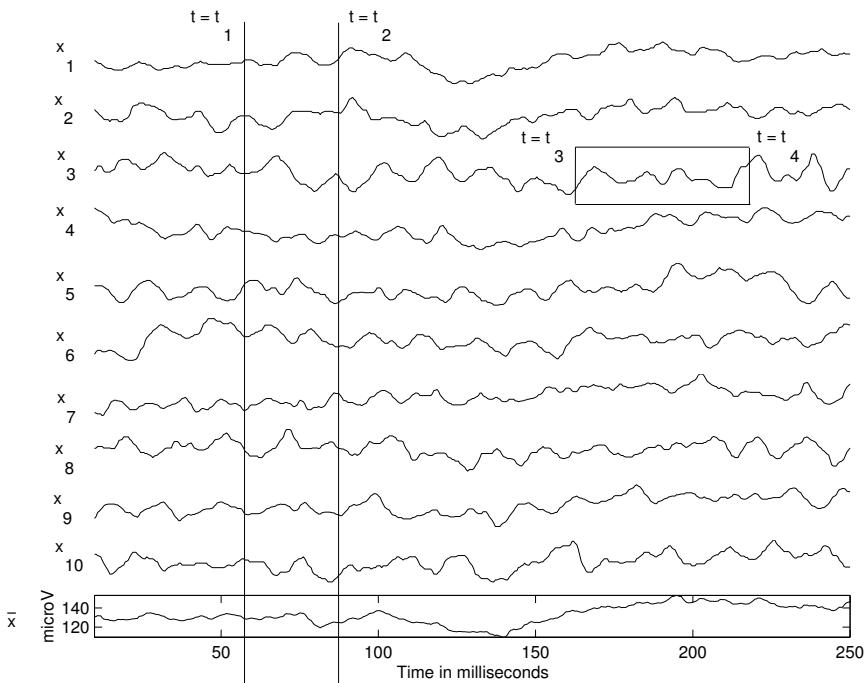


Figure 3.2 Ten sample acquisitions (x_1 to x_{10}) of individual flash visual ERPs from the occipital midline (oz) position of a normal adult male. The earlobes were used to form the reference lead (a1a2), and the left forehead was used as the reference (see Figure 1.39). The signals may be treated as ten realizations of a random process in the form of time series or signals. The vertical lines at $t = t_1$ and $t = t_2 = t_1 + \tau$ represent the ensemble averaging process at two different instants of time. The last plot (framed) gives the ensemble average or prototype $\bar{x}(t)$ of the 10 individual signals. The horizontal box superimposed on the third trace represents the process of computing temporal statistics over the duration $t = t_3$ to $t = t_4$ of the sample ERP $x_3(t)$. See also Figure 3.43. Data courtesy of L. Alfaro and H. Darwish, Alberta Children's Hospital, Calgary.

of time are statistically related to (or have characteristics in common with) the values of the same signal at another instant of time.

When dealing with random processes that are observed as functions of time (or stochastic processes), it becomes possible to compute ensemble averages at every point of time. Then, we obtain an averaged function of time $\bar{x}(t)$ as

$$\bar{x}(t) = \mu_x(t) = \frac{1}{M} \sum_{k=1}^M x_k(t) \quad (3.18)$$

for all time t . The signal $\bar{x}(t)$ may be used to represent the random process x as a prototype; see the last trace (framed) in Figure 3.2. The signal $\bar{x}(t)$ is also a filtered version of the 10 observations $x_1(t)$ to $x_{10}(t)$, with diminished noise due to averaging.

Time averages: When we have a sample observation of a random process $x_k(t)$ as a function of time, it is possible to compute *time averages* or *temporal statistics* by integrating along the time axis, such as

$$\mu_x(k) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x_k(t) dt. \quad (3.19)$$

The integral would be replaced by a summation in the case of sampled or discrete-time signals. The time-averaged ACF $\phi_{xx}(\tau, k)$ is given by

$$\phi_{xx}(\tau, k) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} x_k(t) x_k(t + \tau) dt. \quad (3.20)$$

(See Section 6.3 for details on estimation of the ACF of finite-length data sequences.) The horizontal box superimposed on the third trace in Figure 3.2 represents the process of computing temporal statistics over the duration $t = t_3$ to $t = t_4$ of the sample ERP $x_3(t)$ selected from the ensemble of ERPs illustrated in the figure.

Random noise may be characterized in terms of ensemble and/or temporal statistics. The mean does not play an important role: it is usually assumed to be zero, or may be subtracted out if it is not zero. The ACF plays an important role in the characterization of random processes. The Fourier transform of the ACF is the power spectral density (PSD) function, which is useful in spectral analysis and filter design.

Covariance and cross-correlation: When two random processes x and y need to be compared, we could compute the covariance between them as

$$C_{xy} = E[(x - \mu_x)(y - \mu_y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) p_{x,y}(x, y) dx dy, \quad (3.21)$$

where $p_{x,y}(x, y)$ is the joint PDF of the two processes. The covariance parameter may be normalized to get the correlation coefficient, defined as

$$\rho_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}, \quad (3.22)$$

with $-1 \leq \rho_{xy} \leq +1$. A high covariance indicates that the two processes have similar statistical variability or behavior. The processes x and y are uncorrelated if $\rho_{xy} = 0$. Two processes that are statistically independent are also uncorrelated; the converse of this property is, in general, not true.

When dealing with random processes x and y that are functions of time, the cross-correlation function (CCF) between them is defined as

$$\theta_{xy}(t_1, t_1 + \tau) = E[x(t_1)y(t_1 + \tau)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(t_1) y(t_1 + \tau) p_{x,y}(x, y) dx dy. \quad (3.23)$$

Correlation functions are useful in analyzing the nature of variability and spectral bandwidth of signals, as well as for detection of events by template matching.

3.2.2 Structured noise

Power-line interference at 50 Hz or 60 Hz is an example of structured noise: The typical waveform of the interference is known in advance. It should, however, be noted that the phase of the interfering waveform will not usually be known. Furthermore, the interfering waveform may not be an exact sinusoid; this is indicated by the presence of harmonics of the fundamental 50 Hz or 60 Hz component. Analysis of the power spectrum of a given noisy signal can reveal the presence of periodic noise in the form of peaks or spikes at the fundamental frequency and its harmonics.

3.2.3 Physiological interference

The human body is a complex conglomeration of several systems and processes. Several physiological processes could be active at a given instant of time, each one producing many signals of

different types. A patient or experimental subject may not be able to exercise control on all physiological processes and systems. The appearance of signals from systems or processes other than those of interest may be termed as physiological interference; several examples are listed below.

- EMG related to coughing, breathing, or squirming affecting the ECG
- EGG interfering with precordial ECG
- Maternal ECG getting added to the fetal ECG of interest
- ECG interfering with the EEG
- Ongoing EEG in ERPs and SEPs
- Breath, lung, or bowel sounds contaminating heart sounds (PCG)
- Heart sounds getting mixed with breath or lung sounds
- Muscle sound (VMG) interference in joint sounds (VAG)
- Needle-insertion activity appearing at the beginning of a needle-EMG recording

Physiological interference may not be characterized by any specific waveform or spectral content, and is typically dynamic and nonstationary (varying with the level of the activity of relevance and hence with time; see the following section for a discussion on stationarity). Thus, simple linear bandpass filters will usually not be effective in removing physiological interference.

3.2.4 Stationary, nonstationary, and cyclostationary processes

We have seen in Section 3.2.1 that random processes may be characterized in terms of their ensemble and/or temporal statistics. A random process is said to be *stationary in the strict sense* or *strongly stationary* if its statistics are not affected by a shift in the origin of time. In practice, only first-order and second-order averages are used. A random process is said to be *weakly stationary* or *stationary in the wide sense* if its mean is a constant and its ACF depends only on the difference (or shift) in time. Then, from Equations 3.15 and 3.17, we have $\mu_x(t_1) = \mu_x$ and $\phi_{xx}(t_1, t_1 + \tau) = \phi_{xx}(\tau)$. The ACF is now a function of the delay or shift parameter τ only; the PSD of the process does not vary with time.

A stationary process is said to be *ergodic* if the temporal statistics computed are independent of the sample observed; that is, the same result is obtained for any sample observation $x_k(t)$. The time averages in Equations 3.19 and 3.20 are then independent of k : $\mu_x(k) = \mu_x$ and $\phi_{xx}(\tau, k) = \phi_{xx}(\tau)$. Ensemble statistics may be replaced by the corresponding temporal statistics when analyzing ergodic processes. Ergodic processes are an important type of stationary random processes because their statistics may be computed from a single observation as a function of time. The use of ensemble and temporal averages for noise filtering is illustrated in Sections 3.5 and 3.6.1, respectively.

Signals or processes that do not meet the conditions described above may be, in general, called *nonstationary processes*. A nonstationary process possesses statistics that vary with time. It is readily seen in Figure 1.30 (see also Figure 3.6) that the mean level (baseline) of the signal is varying over the duration of the signal. Therefore, the signal is nonstationary in the mean, a first-order statistical measure. Figure 3.3 illustrates the variance of the speech signal of the word “safety” computed in a moving window of 50 ms (400 samples with $f_s = 8 \text{ kHz}$). Because the variance changes substantially from one portion of the signal to another, it should be concluded that the signal is nonstationary in its second-order statistics (variance, *SD*, or *RMS*). While the speech signal is stationary in the mean, this is not an important characteristic as the mean is typically zero, having

been removed by a highpass filter with cutoff at around 20 Hz. (A DC signal bears no information related to vibration or sound.)

Note that variance displays a behavior that is almost the opposite of that of the turning points in Figure 3.1. Variance is sensitive to changes in amplitude, with large swings about the local mean leading to large variance values. The procedure to detect turning points examines the presence of peaks and troughs with no consideration of their relative amplitudes; the low-amplitude ranges of the fricatives in the signal have resulted in low variance values, even though their counts of the turning points are high.

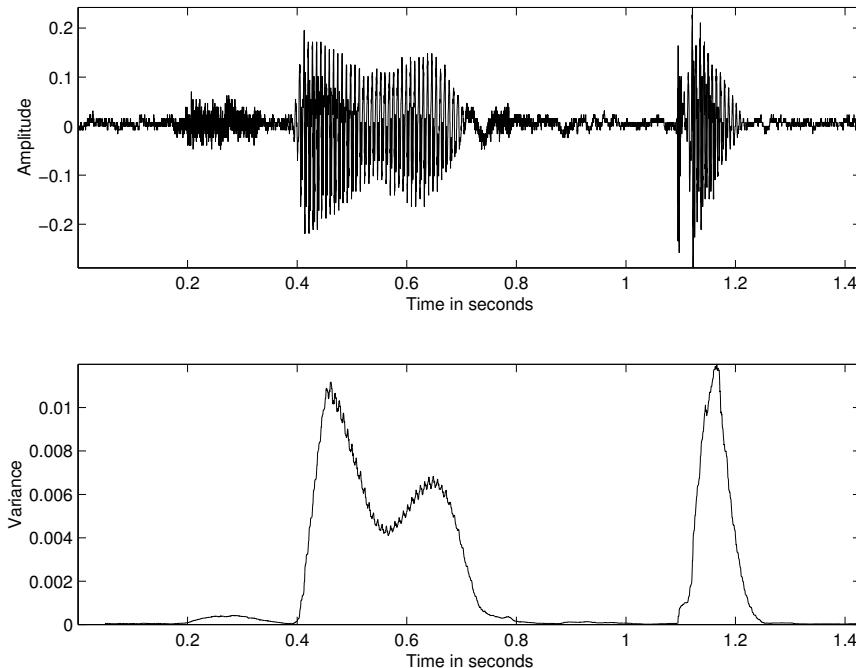


Figure 3.3 Top: Speech signal of the word “safety” uttered by a male speaker. Bottom: Variance computed in a moving window of 50 ms (400 samples with $f_s = 8 \text{ kHz}$) shifted one sample at a time.

Most biomedical systems are dynamic and produce nonstationary signals (for example, EMG, EEG, VMG, PCG, VAG, and speech signals). However, a physical or physiological system has limitations in the rate at which it can change its characteristics. This limitation facilitates breaking a signal into segments of short duration (typically a few tens of milliseconds), over which the statistics of interest are not varying, or may be assumed to remain the same. The signal is then referred to as a *quasistationary process*; the approach is known as *short-time analysis*. Figure 3.4 illustrates the spectrogram of the speech signal of the word “safety.” The spectrogram was composed by computing an array of Fourier magnitude spectra of segments of the signal of duration 64 ms; an overlap of 32 ms was permitted between successive segments. It is evident that the spectral characteristics of the signal vary over its duration: The fricatives demonstrate more high-frequency content than the vowels and also lack formant (resonance) structure. The signal is, therefore, nonstationary in terms of its PSD; because the PSD is related to the ACF, the signal is also nonstationary in the second-order statistical measure of the ACF. However, it can be observed that the PSD shows the same or similar spectral content over the duration of each of the phonemes present in the signal. Thus, over such intervals, the signal may be considered to be quasistationary.

Further discussion and examples of techniques of this nature are presented in Sections 8.4.1 and 8.5. Adaptive signal processing techniques may be designed to detect changes in certain statis-

tical measures of an observed signal; the signal may then be broken into quasistationary segments of variable duration that meet the specified conditions of stationarity. Methods for analysis of non-stationary signals are discussed in Chapter 8. Adaptive segmentation of the EEG, VAG, and PCG signals is discussed in Sections 8.5, 8.6, 8.10, and 8.11.

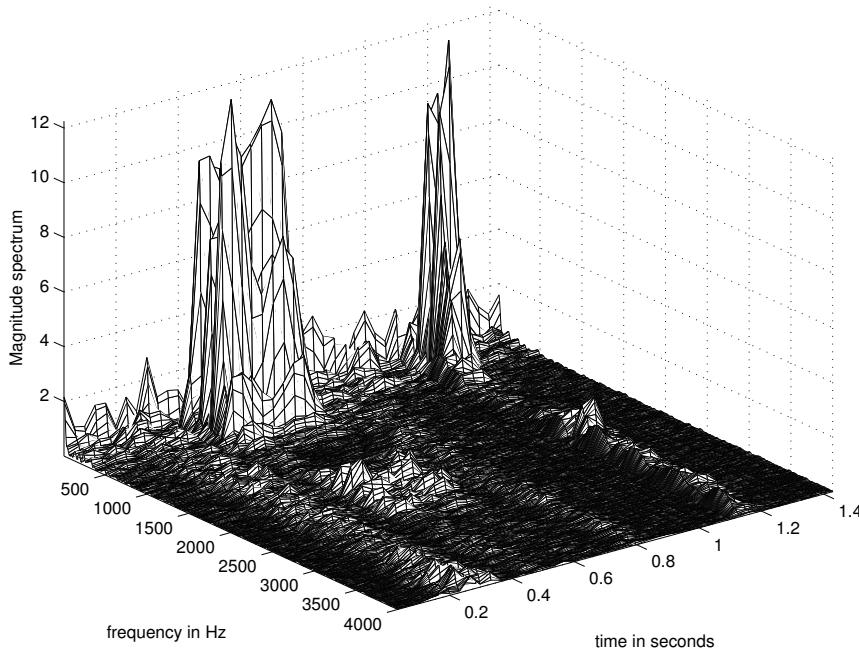


Figure 3.4 Spectrogram of the speech signal of the word “safety” uttered by a male speaker. (The signal is also illustrated in Figures 1.54, 3.1, and 3.3.) Each function or plot along a line parallel to the frequency axis represents the Fourier magnitude spectrum of the signal in a moving window of duration 64 ms (512 samples with $f_s = 8 \text{ kHz}$), with the window advance interval being 32 ms. The spectrogram is plotted on a linear scale to display better the major differences between the voiced and unvoiced sounds.

Certain systems, such as the cardiac system, normally perform rhythmic operations. The resulting signal, such as the ECG, PCG, or carotid pulse, is then almost periodic and may be referred to as a *cyclostationary signal*. The statistics of the PCG signal vary within the duration of a cardiac cycle, especially when murmurs are present, but repeat themselves at regular intervals related to the heart rate or the cardiac cycle. The cyclic repetition of a process facilitates ensemble averaging, using epochs or events extracted from an observation of the signal over many cycles (which is, strictly speaking, a single function of time). Exploitation of the cyclic nature of the ECG signal for synchronized averaging to reduce noise is illustrated in Section 3.5. Application of the same concept to estimate the envelopes of PCG signals is described in Section 5.5.2. Further extensions of the approach to extract A2 from S2 in PCG signals are demonstrated in Section 4.10. Procedures to estimate the PSDs of PCG segments in systole and diastole are presented in Section 6.3.6.

3.3 Illustration of the Problem with Case Studies

The following case studies present several examples of various types of interference in biomedical signals of different origins. The aim of this section is to gain familiarity with the various possibilities of interference and their general characteristics. Filtering techniques to remove various types of interference are described in sections to follow.

3.3.1 Noise in event-related potentials

An ERP is a signal obtained in response to a stimulus. The response is usually of small amplitude (of the order of $10 \mu V$), and it is submerged in ambient EEG activity and noise that could be larger than the ERP. The waveform of a single response may be barely recognizable against the background activity. Figure 3.2 shows 10 individual flash visual ERP signals. The signals were recorded at the occipital midline position, with the left and right earlobes combined to form the reference lead. The left forehead was used as the reference. The ERP signals are buried in ongoing EEG and power-line ($60 Hz$) interference, and they cannot be analyzed using the individual acquisitions shown in the figure.

3.3.2 High-frequency noise in the ECG

Figure 3.5 shows a segment of an ECG signal with high-frequency noise. The noise could be due to the instrumentation amplifiers, the recording system, pickup of ambient EM signals by the cables, and other sources. The signal illustrated has also been corrupted by power-line interference at $60 Hz$ and its harmonics, which may also be considered as a part of high-frequency noise relative to the low-frequency nature of the ECG signal.

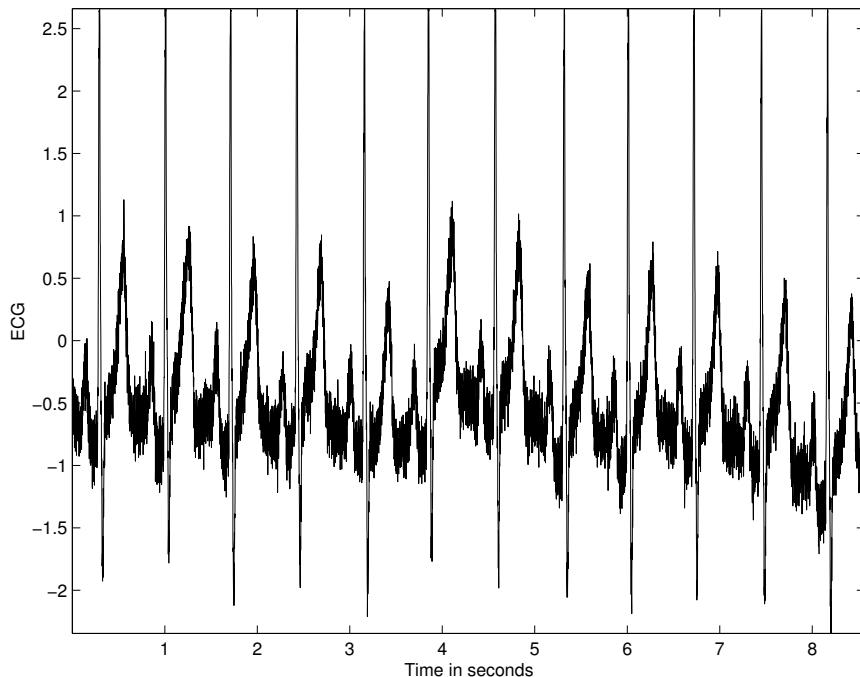


Figure 3.5 ECG signal with high-frequency noise.

3.3.3 Motion artifact in the ECG

Low-frequency artifacts and baseline drift may be caused in chest-lead ECG signals by coughing or breathing with large movement of the chest, or when an arm or leg is moved in the case of limb-lead ECG acquisition. The EGG is a common source of artifact in chest-lead ECG. Poor contact and polarization of the electrodes may also cause low-frequency artifacts. Baseline drift may sometimes be caused by variations in temperature and bias in the instrumentation and amplifiers as

well. Figure 3.6 shows an ECG signal with low-frequency artifact. Baseline drift makes analysis of isoelectricity of the ST segment difficult. A large baseline drift may cause the positive or negative peaks in the ECG to be clipped by the amplifiers or the ADC.

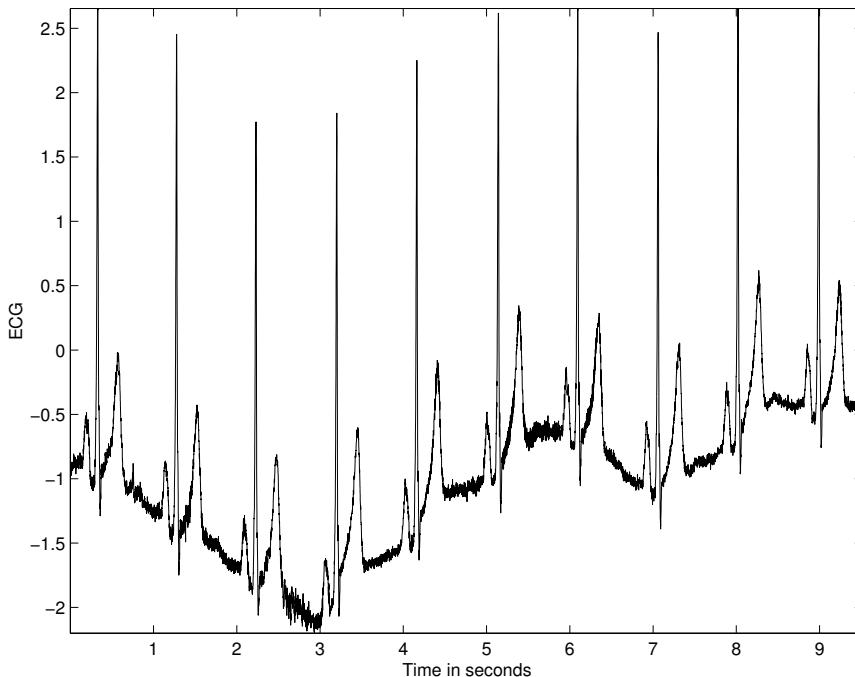


Figure 3.6 ECG signal with low-frequency artifact.

3.3.4 Power-line interference in ECG signals

The most commonly encountered periodic artifact in biomedical signals is the power-line interference at 50 Hz or 60 Hz . If the power-line waveform is not a pure sinusoid due to distortions, harmonics of the fundamental frequency could also appear. Harmonics will also appear if the interference is a periodic waveform that is not a sinusoid (such as rectangular pulses).

Power-line interference may be difficult to detect visually in signals having nonspecific waveforms such as the PCG or EMG; however, the interference is easily visible if present on well-defined signal waveforms such as the ECG or carotid pulse signals. In either case, the power spectrum of the signal should provide a clear indication of the presence of power-line interference as an impulse or spike at 50 Hz or 60 Hz ; harmonics, if present, will appear as additional spikes at integral multiples of the fundamental frequency.

Figure 3.7 shows a segment of an ECG signal with 60 Hz interference. Observe the regular or periodic structure of the interference, which rides on top of the ECG waves and the baseline. Figure 3.8 shows the Fourier power spectrum of the signal. The periodic interference is clearly displayed not only as a spike at its fundamental frequency of 60 Hz , but also as spikes at 180 Hz and 300 Hz , which represent the third and fifth harmonics, respectively. (The recommended sampling rate for ECG signals is 500 Hz ; the higher rate of 1,000 Hz was used in this case because the ECG was recorded as a reference signal with the PCG. The larger bandwidth also permits better illustration of artifacts and filtering.)

The bandwidth of interest of the ECG signal, which is usually in the range 0.05 – 100 Hz , includes the 60– Hz component; hence, simple lowpass filtering will not be appropriate for removal

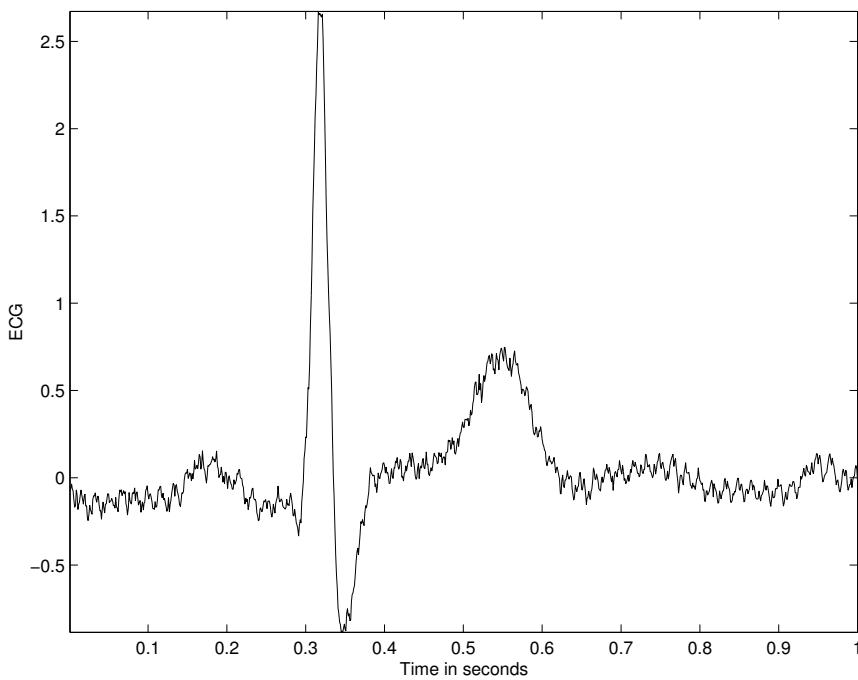


Figure 3.7 ECG signal with power-line (60 Hz) interference.

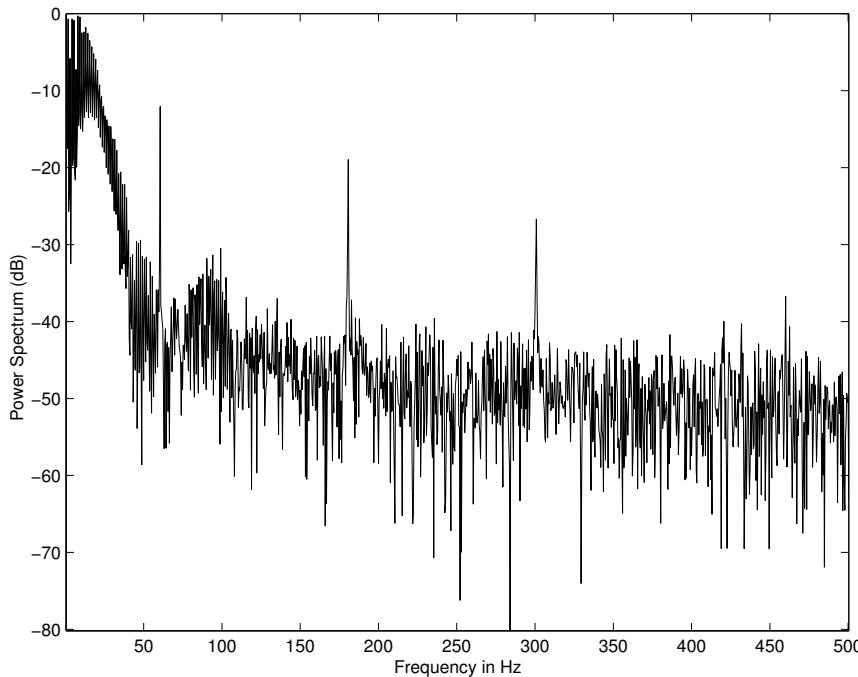


Figure 3.8 Fourier power spectrum of the ECG signal in Figure 3.7 with power-line interference. The spectrum illustrates peaks at the fundamental frequency of 60 Hz as well as the third and fifth harmonics at 180 Hz and 300 Hz , respectively. Only one half of the spectrum is shown for positive frequencies from 0 to 500 Hz , with the sampling rate being $f_s = 1,000\text{ Hz}$. The other half of the spectrum from -500 Hz to 0 is even-symmetric with respect to the part shown in the figure. See Section 3.4.5 for details.

of power-line interference. Lowpass filtering of the ECG to a bandwidth lower than 60 Hz could smooth and blur the QRS complex as well as affect the PQ and ST segments. The ideal solution would be to remove the $60 - \text{Hz}$ component without sacrificing any other frequency component.

3.3.5 Maternal ECG interference in fetal ECG

Figure 3.9 shows an ECG signal recorded from the abdomen of a pregnant woman. Shown also is a simultaneously recorded ECG from the woman's chest. Comparing the two, we see that the abdominal ECG demonstrates multiple peaks (QRS complexes) corresponding to the maternal ECG (occurring at the same time instants as the QRS complexes in the chest lead) as well as several others at weaker levels and a higher repetition rate. The QRS complexes that are not of the expectant mother represent the ECG of the fetus. Observe that the QRS complex shapes of the maternal ECG from the chest and abdominal leads have different shapes due to the projection of the cardiac electrical vector on to different axes. Given that the two signals being combined have almost the same bandwidth individually, how would we be able to separate them and obtain the fetal ECG of interest?

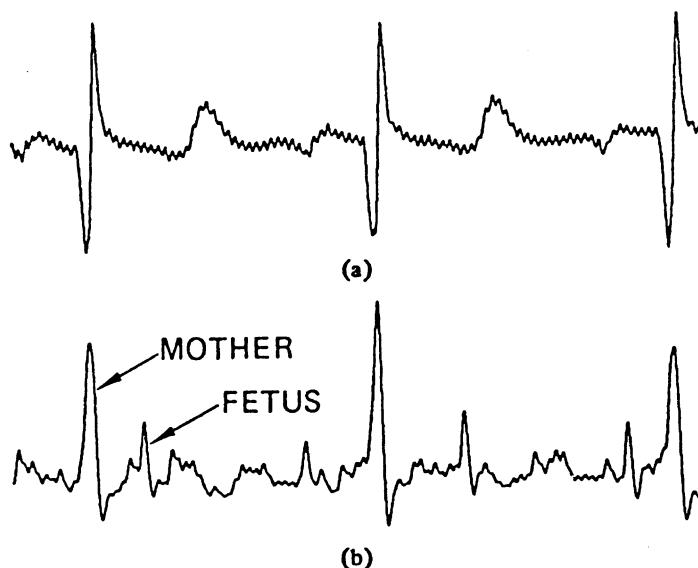


Figure 3.9 ECG signals of a pregnant woman from abdominal and chest leads: (a) chest-lead ECG and (b) abdominal-lead ECG; the former presents the maternal ECG whereas the latter is a combination of the maternal and fetal ECG signals. Observe the presence of power-line artifact in the upper plot. (See also Figure 3.104.) Reproduced with permission from B. Widrow, J.R. Glover, Jr., J.M. McCool, J. Kaunitz, C.S. Williams, R.H. Hearn, J.R. Zeidler, E. Dong, Jr., and R.C. Goodlin, Adaptive noise cancelling: Principles and applications, *Proceedings of the IEEE*, 63(12):1692–1716, 1975. ©IEEE.

3.3.6 Muscle-contraction interference in VAG signals

VAG signals are difficult to analyze because they have no predefined or recognizable waveforms; it is even more difficult to identify any noise or interference that may be present in VAG signals. The signals shown in Figure 1.58 indicate that a transformed version of the VMG could get added to the VAG, especially during extension of the leg when the rectus femoris muscle is active [the second halves of the VAG signals in parts (b)–(d) of the right-hand column]. The left-hand column

of VMG signals in Figure 1.58 illustrates that the VMG generated at the distal rectus femoris gets transmitted well down the leg and appears at the other recording positions. It may be observed from the VAG signals in the right-hand column that vibration signals comparable to the VMG are present in the VAG channels (b)–(d) during extension (second halves) but are not as prominent in flexion (first halves). Interestingly, the knee-joint grinding (crepitus) and click signals that appear in the first half of the VAG signal at the midpatella position [right (b)] have been transmitted downward along the leg to the tibial tuberosity [right (c)] and midtibial shaft [right (d)] positions farther down the leg, presumably along the tibia, but not upwards to the distal rectus femoris position [right (a)].

It should also be noted that the VAG signal cannot be expected to be the same during the extension and flexion parts of a swing cycle: Extension causes more stress or force per unit area on the patellofemoral joint than flexion. Furthermore, the VAG and VMG signals are nonstationary: Characteristics of the VAG vary with the quality of the cartilage surfaces that come into contact at different joint angles, while the VMG varies in accordance with the level of contraction of the muscles involved. To make the problem even more difficult, the bandwidths of the two signals overlap in the range of about 0 – 100 Hz. These factors make removal of the VMG or muscle-contraction interference from VAG signals a challenge.

3.3.7 Potential solutions to the problem

Now that we have gained an understanding of a few sources of artifacts in biomedical signals and their nature, we are prepared to look at specific problems and develop effective filtering techniques to solve them. The following sections present studies of artifacts of various types and demonstrate increasingly complex signal processing techniques to remove them. The problem statement at the beginning of each section defines the nature of the problem in as general terms as possible, sets the terms and conditions, and defines the scope of the investigation to follow. The solution proposed provides the details of an appropriate filtering technique. Each solution is demonstrated with an illustration of its application. Further examples of application of the techniques studied are provided at the end of the chapter. Comparative evaluation of filtering techniques is also provided where applicable. Examples of both success and failure of filtering methods are presented to facilitate learning.

A practical problem encountered by an investigator in the field may not precisely match a specific problem considered in this chapter. However, it is expected that the knowledge of several techniques and an appreciation of the results of their application gained from this chapter will help in designing innovative and appropriate solutions to new problems.

3.4 Fundamental Concepts of Filtering

A filter is a signal processing system, algorithm, or method, realized in hardware or software, that is used to modify a given signal in a particular manner. Quite often, a filtering operation is performed on a signal to remove undesired components that are referred to as noise or artifacts. Regardless of the nature of the operations performed or their effects, filters may be categorized as

- linear or nonlinear,
- fixed (time invariant) or adaptive (time variant),
- active or passive, or
- statistical or deterministic.

In this chapter, we study several filters that span almost all of the categories mentioned above. Before studying specific filters, it would be beneficial to remind ourselves of a few fundamental notions associated with filters and their characteristics, as presented in the following sections.

3.4.1 Linear shift-invariant filters and convolution

Linear time-invariant (LTI) or shift-invariant (LSI) filters form an important category of filters that have a long and well-established history in the field of signal processing [1–3, 18]. The following presentation summarizes important notions and characteristics of LSI filters and filtering operations. Both continuous-time and discrete-time notations are used, as appropriate or convenient. The proofs of the various properties stated are left as exercises for the reader. Readers unfamiliar with the material are referred to Lathi [1, 2], Oppenheim et al. [3], and Oppenheim and Schafer [18].

A fundamental characteristic of an LSI system is its impulse response, which is the output of the system when the input is a Dirac delta or impulse function. Before looking at the details of the impulse response, let us review a few important definitions related to the delta or impulse function. In continuous time, the Dirac delta function is defined as [1–3, 18]

$$\delta(t) = \begin{cases} \text{undefined} & \text{at } t = 0, \\ 0 & \text{otherwise,} \end{cases} \quad (3.24)$$

and

$$\int_{t=-\infty}^{\infty} \delta(t) dt = 1. \quad (3.25)$$

The delta function may be visualized as the limit of a function whose integral over the full extent of time or an independent variable is maintained equal to unity while its duration or extent is compressed toward zero. Figure 3.10 demonstrates this definition graphically using a rectangular pulse function. Another example of this phenomenon is given by

$$\delta(t) = \frac{1}{2} \lim_{a \rightarrow 0} a|t|^{(a-1)}. \quad (3.26)$$

Figure 3.11 illustrates three plots of the function defined above for $a = 0.8, 0.4$, and 0.2 . The emergence of the delta function as $a \rightarrow 0$ is evident.

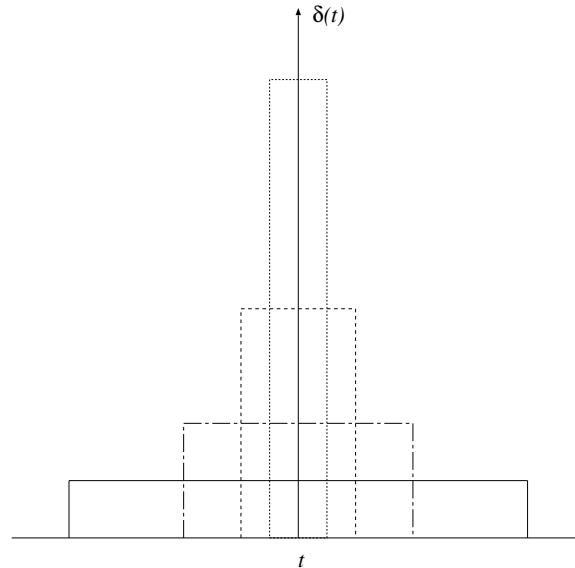


Figure 3.10 Schematic representation of the delta function as a limit of a rectangular pulse function. The time duration of the pulse is reduced while maintaining unit area under the function.

The delta function is the derivative of the unit step function $u(t)$, which is defined as

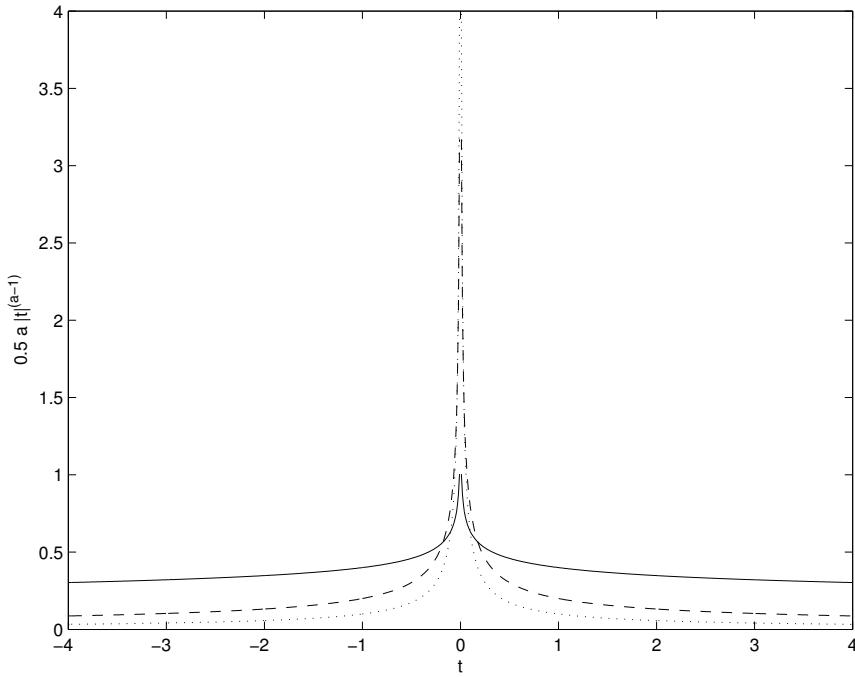


Figure 3.11 The delta function as the limit of $0.5a|t|^{(a-1)}$ as $a \rightarrow 0$. The function is plotted for $a = 0.8$ (solid line), $a = 0.4$ (dashed line), and $a = 0.2$ (dotted line).

$$u(t) = \begin{cases} 1 & \text{for } t > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (3.27)$$

The delta function is also defined in terms of its action within an integral over a certain interval $[T_1, T_2]$:

$$\int_{T_1}^{T_2} x(t) \delta(t - t_o) dt = \begin{cases} x(t_o) & \text{if } T_1 < t_o < T_2, \\ 0 & \text{otherwise,} \end{cases} \quad (3.28)$$

where $x(t)$ is a function that is continuous at t_o . This is known as the *sifting property* of the delta function, because the value of the function $x(t)$ at the location t_o of the delta function is sifted or selected from all of its values. The expression may be extended to all t as

$$x(t) = \int_{\alpha=-\infty}^{\infty} x(\alpha) \delta(t - \alpha) d\alpha, \quad (3.29)$$

where α is a temporary variable, which may also be interpreted as resolving an arbitrary signal $x(t)$ into a weighted combination of mutually orthogonal delta functions.

Let us consider a continuous-time signal, $x(t)$, that is being processed by an LSI system, as shown schematically in Figure 3.12. An LSI system is completely characterized or specified by its impulse response, $h(t)$, which is the output of the system when the input is a delta function. The output of the system, $y(t)$, is given by the convolution of the input, $x(t)$, with the impulse response, $h(t)$, defined as

$$y(t) = \int_{\tau=-\infty}^{\infty} x(\tau) h(t - \tau) d\tau, \quad (3.30)$$

where τ is a temporary variable of integration. An equivalent result is given by

$$y(t) = \int_{\tau=-\infty}^{\infty} h(\tau) x(t - \tau) d\tau. \quad (3.31)$$

The convolution operation given above is *linear convolution*; another version of convolution known as periodic or circular convolution is defined in Equation 3.90. When a causal system or causality is considered, the lower limit of the integrals given above for convolution may be changed to zero and the upper limit changed to t , the current instant of time, resulting in

$$y(t) = \int_{\tau=0}^t x(\tau) h(t - \tau) d\tau \quad (3.32)$$

or

$$y(t) = \int_{\tau=0}^t h(\tau) x(t - \tau) d\tau. \quad (3.33)$$

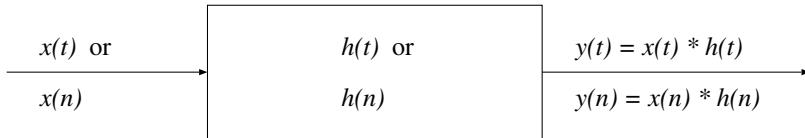


Figure 3.12 A schematic representation of a continuous-time or discrete-time LTI/LSI filter.

The discrete-time unit impulse function or delta function is defined as [1–3]

$$\delta(n) = \begin{cases} 1 & \text{if } n = 0, \\ 0 & \text{otherwise.} \end{cases} \quad (3.34)$$

Figure 3.13 illustrates a few versions of the discrete-time delta function.

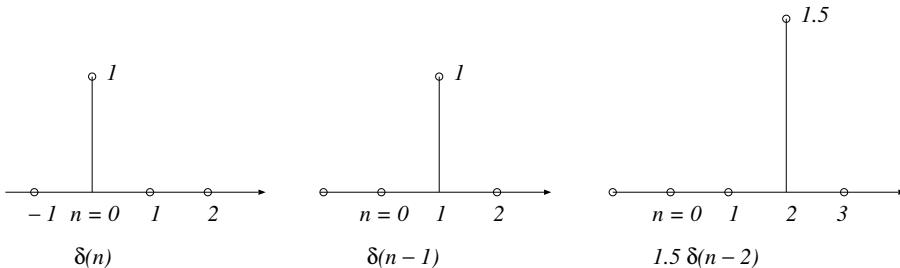


Figure 3.13 A schematic representation of the discrete-time unit impulse function or delta function as well as shifted and scaled versions of the same.

The discrete-time unit step function is defined as

$$u(n) = \begin{cases} 1 & \text{for } n \geq 0, \\ 0 & \text{otherwise.} \end{cases} \quad (3.35)$$

Figure 3.14 illustrates a few versions of the discrete-time unit step function.

A discrete-time LSI system is shown schematically in Figure 3.15, displaying its impulse response, $h(n)$. The output, $y(n)$, of the system is given by the linear convolution of the input, $x(n)$, with the impulse response, $h(n)$, as

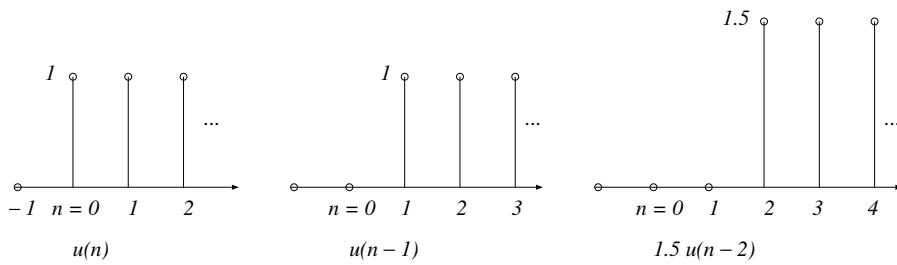


Figure 3.14 A schematic representation of the discrete-time unit step function as well as shifted and scaled versions of the same.

$$y(n) = \sum_{k=0}^n x(k) h(n-k), \quad (3.36)$$

where k is a temporary variable of summation. An equivalent result is given by

$$y(n) = \sum_{k=0}^n h(k) x(n-k). \quad (3.37)$$

In the two equations given above, causality has been assumed.

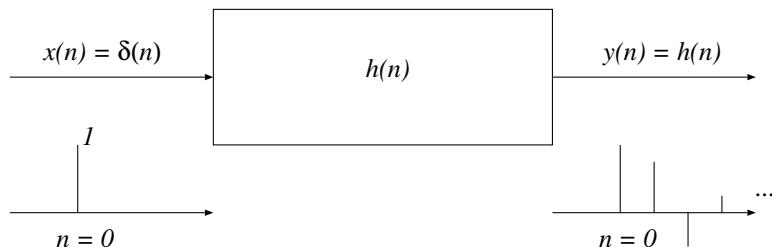


Figure 3.15 A schematic representation of the impulse response, $h(n)$, of a discrete-time LSI filter.

To understand convolution, let us consider the discrete-time version in Equation 3.36. In addition to the change of the independent (time) variable from n to k , there are two important points to note:

- $h(-k)$ represents a reversal in time of the function or signal $h(k)$; and
- $h(n - k)$ represents a shift of the reversed signal $h(-k)$ by n samples.

Multiplication of $h(n - k)$ by $x(k)$ can be viewed as scaling. The summation represents accumulation of the results or integration of $x(k) h(n - k)$ over the interval $k = 0$, the origin of time, to n , the present instant of time. A simple illustration of shifting or adding a delay in time, as well as scaling, is shown in Figure 3.13. A numerical illustration of shifting a signal is shown in Figure 3.16. An illustration of reversing as well as shifting a signal is shown in Figure 3.17.

The linear convolution of two discrete-time signals is illustrated numerically in Figure 3.18. To facilitate understanding of the procedure, we could expand and modify Equation 3.36 as follows:

<i>n:</i>	0	1	2	3	4	5	6	7
<i>x(n):</i>	4	5	3	1				
<i>x(n - 1):</i>		4	5	3	1			
<i>x(n - 2):</i>			4	5	3	1		
<i>x(n - 3):</i>				4	5	3	1	

Figure 3.16 A numerical illustration of shifted versions of signal. Blank spaces indicate samples of the signal that are undefined or zero in value. This type of shifting is known as linear shifting. See Figure 3.39 for an example of circular or periodic shifting.

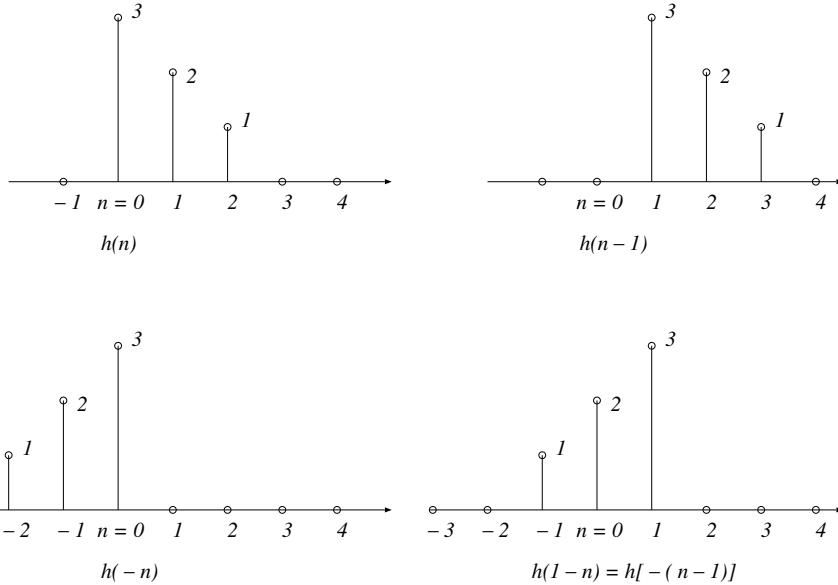


Figure 3.17 A schematic illustration of reversing and shifting of a signal.

$$\begin{aligned}
 y(n) &= \sum_{k=0}^n x(k) h(n-k), \\
 y(0) &= \sum_{k=0}^0 x(k) h(0-k) \\
 &= x(0)h(0), \\
 y(1) &= \sum_{k=0}^1 x(k) h(1-k) \\
 &= x(0)h(1) + x(1)h(0), \\
 y(2) &= \sum_{k=0}^2 x(k) h(2-k) \\
 &= x(0)h(2) + x(1)h(1) + x(2)h(0), \\
 y(3) &= \sum_{k=0}^3 x(k) h(3-k) \\
 &= x(0)h(3) + x(1)h(2) + x(2)h(1) + x(3)h(0).
 \end{aligned} \tag{3.38}$$

The procedure is stopped, and the output ceases to exist (or has only zero values) when the shift exceeds a certain amount such that $x(k)$ and $h(n - k)$ do not overlap in time any more. It is evident that the linear convolution of two discrete-time signals with durations of N_1 and N_2 samples leads to a result with the duration of $N_1 + N_2 - 1$ samples.

$$\begin{array}{cccccccc}
 n: & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
 x(n): & 4 & 1 & 3 & 1 & & & & \\
 h(n): & 3 & 2 & 1 & & & & & \\
 k: & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
 x(k): & 4 & 1 & 3 & 1 & 0 & 0 & 0 & 0 \\
 \\
 h(0 - k): & 1 & 2 & 3 & & & & & \\
 h(1 - k): & & 1 & 2 & 3 & & & & \\
 h(2 - k): & & & 1 & 2 & 3 & & & \\
 h(3 - k): & & & & 1 & 2 & 3 & & \\
 h(4 - k): & & & & & 1 & 2 & 3 & \\
 h(5 - k): & & & & & & 1 & 2 & 3 \\
 h(6 - k): & & & & & & & 1 & 2 & 3 \\
 \\
 y(n): & 12 & 11 & 15 & 10 & 5 & 1 & 0 & 0 \\
 n: & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7
 \end{array}$$

Figure 3.18 A numerical illustration of the convolution of two discrete-time signals. The notations used in the figure agree with Equation 3.36. For each shift, the corresponding output sample value is obtained by multiplying each pair of samples of $x(k)$ and $h(n - k)$ that overlap (exist at the same instant of time k) and then adding the results.

Taking a different approach, we could expand and modify Equation 3.36 as follows:

$$\begin{aligned}
 y(n) &= \sum_{k=0}^n x(k) h(n - k), \\
 &= x(0) h(n) + x(1) h(n - 1) + x(2) h(n - 2) + x(3) h(n - 3) + \dots
 \end{aligned} \tag{3.39}$$

This result may be interpreted as the sum of several delayed and weighted versions of the impulse response of the system, the weights being provided by the samples of the input signal. Just as the system produces the impulse response, $h(n)$, when the input is $\delta(n)$, when the system is triggered by the input $x(0)$ at $n = 0$, it produces the output $x(0)h(n)$, which lasts for the duration of $h(n)$. When the input is $x(1)$, occurring at $n = 1$, the output is $x(1)h(n - 1)$, recognizing the fact that the system is causal, and the corresponding output can only start from $n = 1$. Note that the part of the output signal that started at $n = 0$ will, in general, continue past $n = 1$, and hence the parts of the output that start at $n = 0$ and $n = 1$ will overlap. The process continues, as illustrated in Figure 3.19.

Illustrations of application: To illustrate the effects of random noise being added to a deterministic signal, the following signal was created:

$$x(t) = 5 \sin(2\pi 2t) + 2 \cos(2\pi 3t), \tag{3.40}$$

<i>n:</i>	0	1	2	3	4	5	6	7
<i>x(n):</i>	4	1	3	1				
<i>h(n):</i>	3	2	1					
<i>x(0) h(n - 0):</i>	12	8	4	0	0	0	0	0
<i>x(1) h(n - 1):</i>	0	3	2	1	0	0	0	0
<i>x(2) h(n - 2):</i>	0	0	9	6	3	0	0	0
<i>x(3) h(n - 3):</i>	0	0	0	3	2	1	0	0
<i>y(n):</i>	12	11	15	10	5	1	0	0
<i>n:</i>	0	1	2	3	4	5	6	7

Figure 3.19 A numerical illustration of the convolution of two discrete-time signals. The output, $y(n)$, is obtained by adding the corresponding values in the four rows labeled as $x(0)h(n-0)$ through $x(3)h(n-3)$. The notations used in the figure agree with Equation 3.39. Compare this procedure with that shown in Figure 3.18. Although the view is different, the end result is the same.

using a sampling frequency of 2 kHz . Random noise with a Gaussian PDF was simulated and added to the signal x for an effective SNR of 10 dB . The original signal, the noise samples generated, and the noisy signal are shown in Figure 3.20. The distinction between the deterministic nature of the original signal and the random nature of the noise is clearly evident. The histogram of a realization of the noise process used is shown in Figure 3.21, which approximates a Gaussian PDF with zero mean.

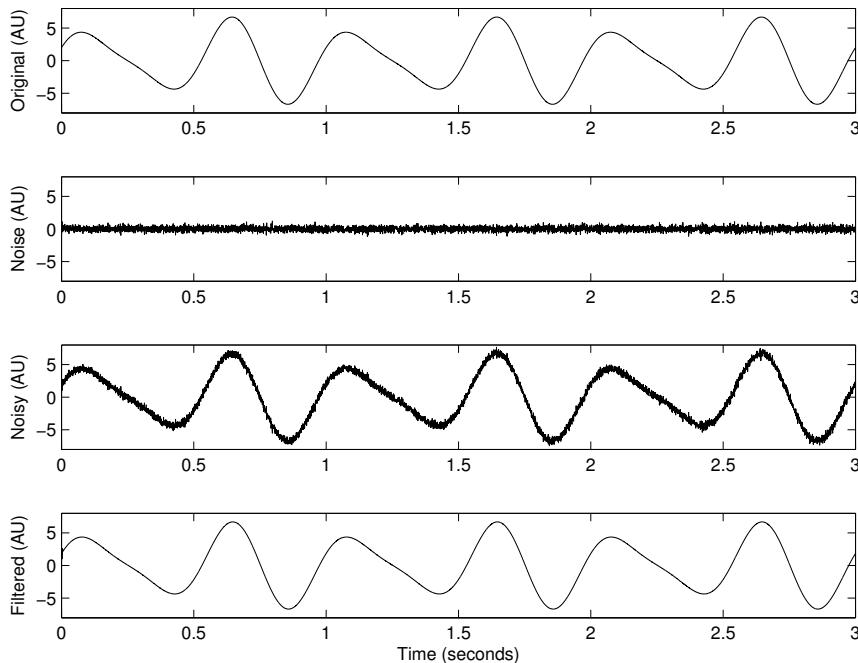


Figure 3.20 Top to bottom: original signal $x(t)$ as in Equation 3.40; Gaussian-distributed random noise; noisy signal; result of filtering using the mean of the signal values in a sliding window of width 11 samples or 5.5 ms . The window was moved one sample at a time. See Figure 3.21 for the histogram of the noise process.

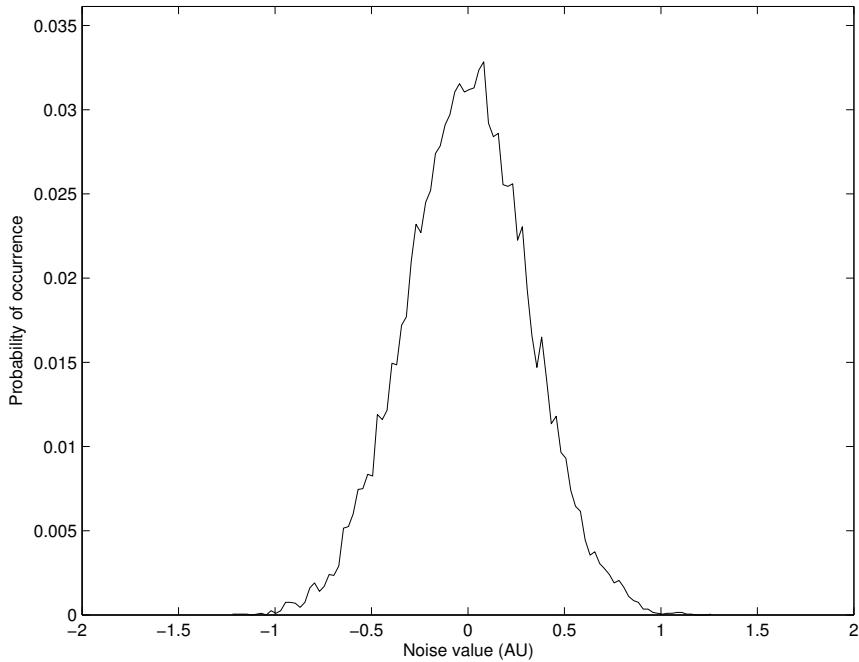


Figure 3.21 Histogram of a realization of the noise process used in the example in Figure 3.20.

The noisy signal was filtered by computing the mean of each sample and the preceding 10 samples. The equation of the corresponding filter could be expressed as

$$y(n) = \frac{1}{11} \sum_{k=0}^{10} x(n-k), \quad (3.41)$$

for $n = 10, 11, \dots, N-1$, where N is the number of samples in the signal. A filter as above is known as a moving-average (MA) filter: average values of the input signal are computed in a moving temporal window and used to define the output signal. (See Figure 3.2 for a related illustration.) Note that the operation as defined in Equation 3.41 cannot be started until the first 11 samples are available in the input data stream. The operation may also be viewed as a convolution of $x(n)$ with the impulse response of the filter, as in Equation 3.37, with $h(k) = 1/11$, $k = 0, 1, \dots, 10$. Due to the random nature of the noise process (with the mean of the process being zero), the mean of a number of samples of the process tends to zero as the number of samples used increases. Therefore, the values of the filtered result, also shown in Figure 3.20, approach the corresponding values in the original signal. While a small number of samples used in computing the mean would not suppress much of the noise, a large number of samples could overly smooth the signal and remove fine details.

Figure 3.22 shows another version of a noisy signal generated using the same processes as in the preceding example. The noisy signal was filtered with a filter having its impulse response as a linearly decreasing function or a ramp of duration 0.25 s, defined in the continuous-time notation as

$$h(t) = 10(0.25 - t), \quad 0 \leq t \leq 0.25 \text{ s}, \quad (3.42)$$

with the sampling frequency of 2 kHz. The output was divided by the sum of all of the values of $h(n)$. The result is a weighted average of the corresponding values of the input signal. Figure 3.23 shows the filter function in Equation 3.42 superimposed on the noisy signal being filtered. The impulse response of the filter has been reversed in time, as required in Equation 3.36 for convolution,

and placed at $t = 2.3\text{ s}$. The output at $t = 2.3\text{ s}$ is given by the area under the product of the signal and the impulse response; equivalently, it is given by the sum of the products of all of the overlapping samples of the signal and the impulse response. It is seen that the filtered output is smooth and free of noise; however, some of the minor details in the original signal have been suppressed by the lengthy filter. Furthermore, the delay introduced by the filter is clearly seen by comparing the peaks in the original signal with the corresponding peaks in the filtered output.

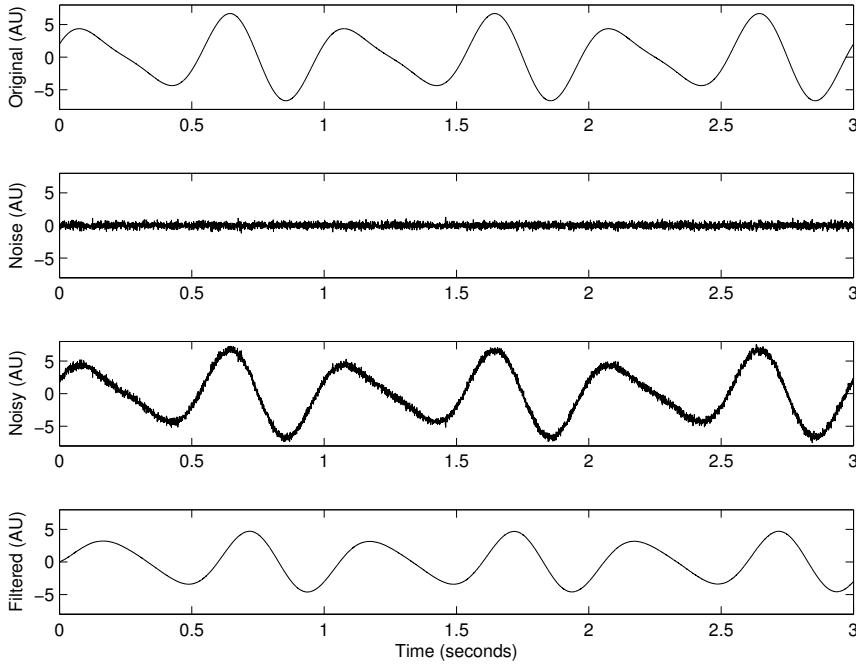


Figure 3.22 Top to bottom: original signal $x(t)$ as in Equation 3.40; Gaussian-distributed random noise; noisy signal; result of filtering using the ramp function in Equation 3.42. See Figure 3.21 for the histogram of the noise process. See also Figure 3.23.

LSI systems in series or parallel: When a number of LSI systems are used in series or in parallel, their effects may be combined into a single equivalent system or operation. Figure 3.24 shows two LSI systems in series (cascade). The first system, with the impulse response $h_1(n)$, operates upon the input, $x(n)$, to give the output as

$$s(n) = x(n) * h_1(n). \quad (3.43)$$

The second system, with the impulse response $h_2(n)$, operates upon $s(n)$ to produce the output

$$\begin{aligned} y(n) &= s(n) * h_2(n) \\ &= x(n) * h_1(n) * h_2(n) \\ &= x(n) * h(n), \end{aligned} \quad (3.44)$$

where

$$h(n) = h_1(n) * h_2(n) \quad (3.45)$$

is the impulse response of the combination of the two LSI systems in series.

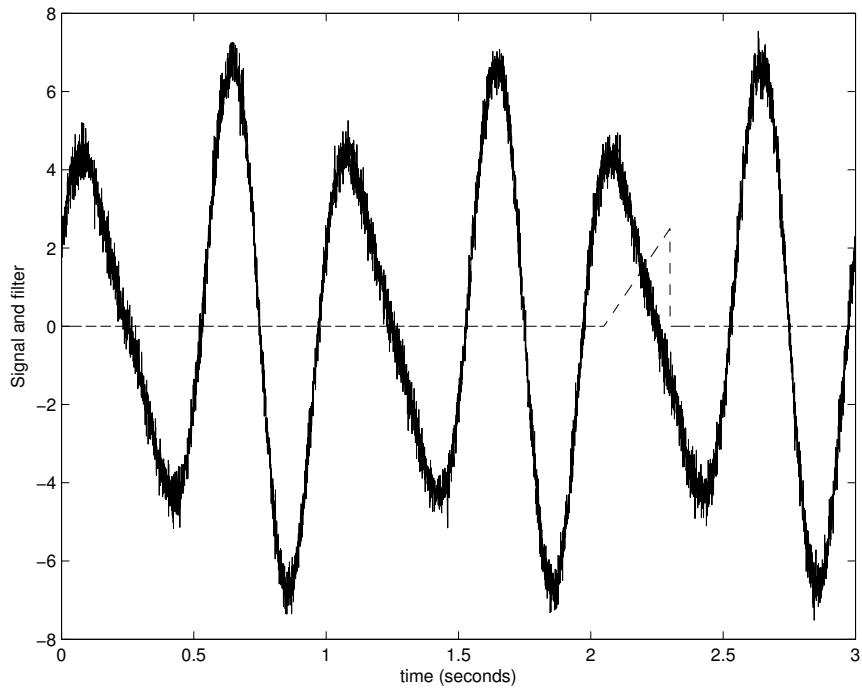


Figure 3.23 The linear ramp filter in Equation 3.42 is shown superimposed (in dashed line) on the noisy signal in the example shown in Figure 3.22. The impulse response of the filter has been reversed in time and placed at $t = 2.3$ s. The output at $t = 2.3$ s is given by the area under the product of the signal and the impulse response; equivalently, it is given by the sum of the products of all samples of the signal and the impulse response that overlap.

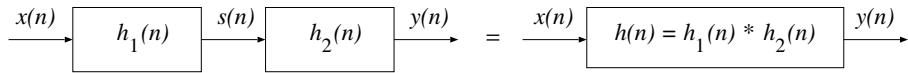


Figure 3.24 Two LSI systems in series and the equivalent system.

Figure 3.25 shows two LSI systems in parallel. For the first system, we have the output as

$$s_1(n) = x(n) * h_1(n). \quad (3.46)$$

Similarly, the second system produces the output

$$s_2(n) = x(n) * h_2(n). \quad (3.47)$$

The combined result is

$$\begin{aligned} y(n) &= s_1(n) + s_2(n) \\ &= x(n) * h_1(n) + x(n) * h_2(n) \\ &= x(n) * [h_1(n) + h_2(n)] \\ &= x(n) * h(n), \end{aligned} \quad (3.48)$$

where

$$h(n) = h_1(n) + h_2(n) \quad (3.49)$$

is the impulse response of the combination of the two LSI systems in parallel. Relationships as given above are useful in the analysis of sophisticated signal processing systems designed by combining several LSI systems.

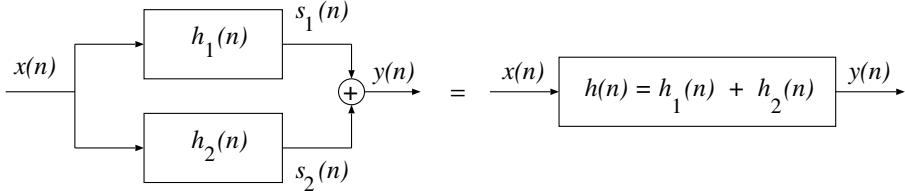


Figure 3.25 Two LSI systems in parallel and the equivalent system.

3.4.2 Transform-domain analysis of signals and systems

Quite often, it is convenient to analyze the behavior, characteristics, and performance of an LTI or LSI system in a transform domain. The commonly used transforms for the analysis of continuous-time systems and signals are the Laplace and Fourier transforms [1–3]. The Laplace transform, $H(s)$, of the impulse response, $h(t)$, of an LTI system is defined as

$$H(s) = \int_{-\infty}^{\infty} h(t) \exp(-st) dt, \quad (3.50)$$

where $s = \sigma + j\omega$ is a complex variable representing the transform domain. Here, $\omega = 2\pi f$ represents the frequency variable in radians per second (rad/s), with f being the frequency variable in Hz ; the unit for σ is neper or Napier. If the signal $h(t)$ is causal and of finite duration, existing only over the interval of time $[0, T]$, the limits of the integral could be changed as

$$H(s) = \int_0^T h(t) \exp(-st) dt. \quad (3.51)$$

The function $H(s)$ is known as the system transfer function, or simply the transfer function of the filter.

Figure 3.26 shows a schematic representation of the s -plane. When $H(s)$ is evaluated on the imaginary axis in the s -plane, with $s = j\omega$, we get the frequency response of the system as

$$H(\omega) = H(s)|_{s=j\omega} = \int_0^T h(t) \exp(-j\omega t) dt, \quad (3.52)$$

which is the Fourier transform of the impulse response, $h(t)$. [Note: Some authors express the function above as $H(e^{j\omega})$ or $H(j\omega)$.] In general, $H(\omega)$ is a complex quantity even when $h(t)$ is a real-valued signal. The magnitude response of the system or filter is given by $|H(\omega)|$ and the phase response is given by $\angle H(\omega)$.

Figure 3.27 shows the relationships among a signal $x(t)$, its Laplace and Fourier transforms, and other related entities. We will be using all of the transformations and representations indicated in the figure in the analysis of signals and related functions either in the continuous or discrete form.

It is common to express the transfer function, $H(s)$, of an LTI system as a ratio of two polynomials in s . The roots of the polynomial in the numerator give the zeros of the system, whereas the roots of the polynomial in the denominator give the poles of the system. For a stable LTI system,

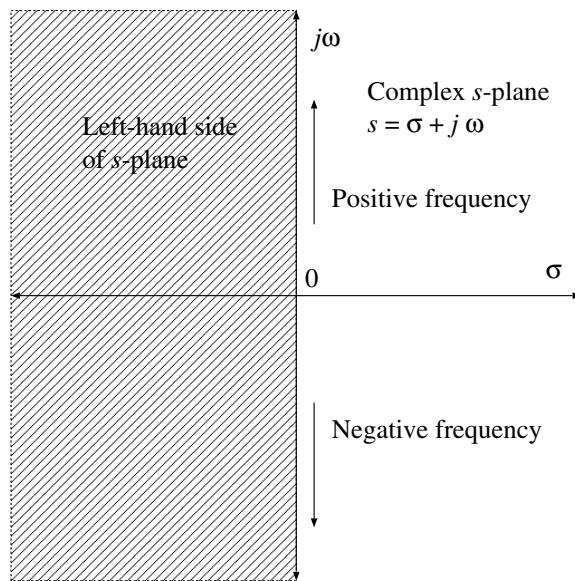


Figure 3.26 Schematic representation of the s -plane or Laplace transform domain.

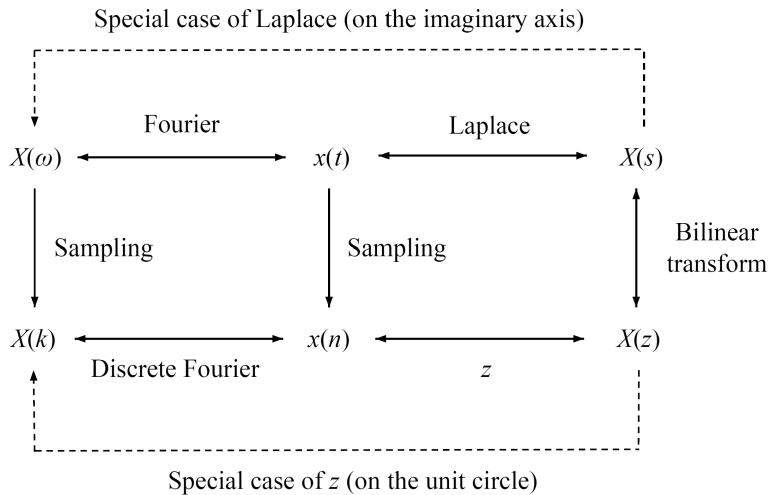


Figure 3.27 Relationship among a signal $x(t)$ and its sampled discrete-time form $x(n)$ with their Laplace transform, Fourier transform, and/or z -transform. It should be noted that a sampled signal or spectrum can be transformed back to its continuous version by interpolation.

all of the poles should be located on the left-hand side (LHS) of the s -plane with $\sigma < 0$. For a real-valued impulse response or signal, poles and zeros should occur in complex-conjugate pairs or with real values. The pole-zero plot of a system is adequate to derive its impulse response, transfer function, and output (for a given input signal) except for a scale factor or gain.

The Laplace transform is a linear and reversible transform. An important property of the Laplace transform when dealing with LTI systems is that the convolution of two signals in the time domain is converted to the product of their individual Laplace transforms in the s -domain, expressed as

$$\begin{aligned}
 & \text{if } y(t) = x(t) * h(t), \\
 & \text{then } Y(s) = X(s) H(s), \\
 & Y(\omega) = X(\omega) H(\omega).
 \end{aligned} \tag{3.53}$$

Figure 3.28 shows the input–output relationship for an LTI system in the s -domain.

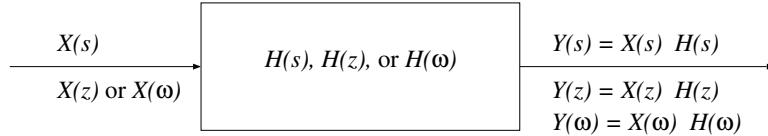


Figure 3.28 Input–output relationship for an LTI system in the s -domain or an LSI system in the z -domain. The various types of input and output shown indicate that the same system may be analyzed in multiple domains to gain different perspectives for analysis.

It follows that, when two LTI systems are in parallel, as in Figure 3.25, we have the output as $Y(s) = [H_1(s) + H_2(s)]X(s)$, with the equivalent transfer function of $H(s) = H_1(s) + H_2(s)$. When two LTI systems are in series, as in Figure 3.24, we have the output as $Y(s) = H_1(s) H_2(s) X(s)$, with the equivalent transfer function of $H(s) = H_1(s) H_2(s)$. For further details regarding the properties of the Laplace transform, the inverse Laplace transform, and examples of signals and their transforms, refer to Lathi [1, 2] and Oppenheim et al. [3].

The discrete-time counterpart of the Laplace transform is the z -transform, which is useful for the analysis of discrete-time LSI systems and signals. The z -transform of a signal $x(n)$ is defined as

$$X(z) = \sum_{n=-\infty}^{\infty} x(n) z^{-n}, \tag{3.54}$$

where z is a complex variable. Lathi [1, 2] and Oppenheim et al. [3] provide detailed discussions on the properties of the z -transform, its region of convergence (ROC), and the inverse z -transform.

If we consider the z -transform of the impulse response, $h(n)$, of a causal, LSI, finite-impulse response (FIR) system, with $h(n)$ existing only for $n = 0, 1, 2, \dots, N - 1$, then we have

$$H(z) = \sum_{n=0}^{N-1} h(n) z^{-n}, \tag{3.55}$$

which represents the transfer function of the system.

An important property of the z -transform when dealing with LSI systems is that the convolution of two signals in the time domain is converted to the product of their individual z -transforms, expressed as

$$\begin{aligned}
 & \text{if } y(n) = x(n) * h(n), \\
 & \text{then } Y(z) = X(z) H(z).
 \end{aligned} \tag{3.56}$$

Figure 3.28 shows the input–output relationship for an LSI system in the z -domain. The relationship expressed above may be derived in detail as follows:

$$\begin{aligned}
Y(z) &= \sum_{n=-\infty}^{\infty} y(n) z^{-n} \\
&= \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} x(k) h(n-k) z^{-n} \\
&= \sum_{k=-\infty}^{\infty} x(k) \sum_{n=-\infty}^{\infty} h(n-k) z^{-n}.
\end{aligned} \tag{3.57}$$

Here, the definition of the convolution relationship $y(n) = x(n) * h(n)$ has been expanded as in Equation 3.36. The range of summation has been indicated as $(-\infty, \infty)$ for the sake of generality and ease of manipulation of the equations. In the last step above, the order of the two summation operations has been switched, and the terms have been rearranged. Now, let $m = n - k$. Then, $n = m + k$, and we have

$$\begin{aligned}
Y(z) &= \sum_{k=-\infty}^{\infty} x(k) \sum_{m=-\infty}^{\infty} h(m) z^{-(m+k)} \\
&= \sum_{k=-\infty}^{\infty} x(k) \sum_{m=-\infty}^{\infty} h(m) z^{-m} z^{-k} \\
&= \sum_{k=-\infty}^{\infty} x(k) z^{-k} \sum_{m=-\infty}^{\infty} h(m) z^{-m} \\
&= X(z) H(z).
\end{aligned} \tag{3.58}$$

The effects of combining two LSI systems in series or in parallel, as seen in the z -domain, are shown in Figures 3.29 and 3.30. For the combination in series, we have

$$S(z) = X(z) H_1(z) \tag{3.59}$$

and

$$\begin{aligned}
Y(z) &= S(z) H_2(z) \\
&= X(z) H_1(z) H_2(z) \\
&= X(z) H(z),
\end{aligned} \tag{3.60}$$

where

$$H(z) = H_1(z) H_2(z). \tag{3.61}$$

For the combination in parallel, we have

$$S_1(z) = X(z) H_1(z), \tag{3.62}$$

$$S_2(z) = X(z) H_2(z), \tag{3.63}$$

and

$$\begin{aligned}
 Y(z) &= S_1(z) + S_2(z) \\
 &= X(z) H_1(z) + X(z) H_2(z) \\
 &= X(z) [H_1(z) + H_2(z)] \\
 &= X(z) H(z),
 \end{aligned} \tag{3.64}$$

where

$$H(z) = H_1(z) + H_2(z). \tag{3.65}$$

Relationships such as these are useful in the analysis of sophisticated signal processing systems that are designed by combining several LSI systems.

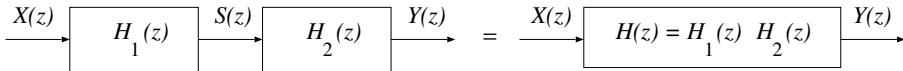


Figure 3.29 Two LSI systems in series and the equivalent system in the z -domain.

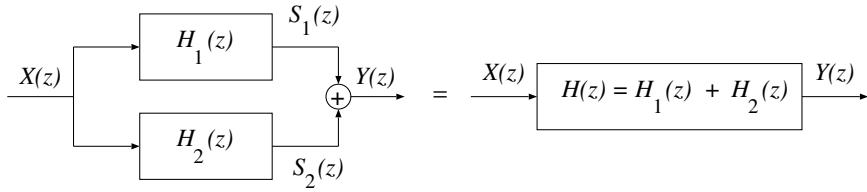


Figure 3.30 Two LSI systems in parallel and the equivalent system in the z -domain.

A few of the important properties of and examples related to the z -transform are as follows [1–3]:

- The z -transform of $\delta(n)$ is 1. The ROC is all z .
- The z -transform of the unit step function, $u(n)$, is $\frac{1}{1-z^{-1}}$. The ROC is $|z| > 1$.
- The z -transform of $a^n u(n)$ is $\frac{1}{1-az^{-1}}$. The ROC is $|z| > |a|$.
- The z -transform of a causal signal $x(n)$ shifted by k samples, that is, $x(n - k)$, is $z^{-k} X(z)$, where $X(z)$ is the z -transform of $x(n)$.
- If $y(n) = x(n) * h(n)$, then $Y(z) = X(z) H(z)$.

Figure 3.31 illustrates the relationship between the Laplace domain and the z -domain with $z = \exp(sT)$, where $T = 1/f_s$ is the sampling interval, and f_s is the sampling frequency. (The maximum frequency permitted in the signal without aliasing errors is $f_m = f_s/2$; this is also referred to as the folding frequency or the Nyquist frequency.) The entire LHS of the s -plane is mapped to the region within the unit circle in the z -plane. Thus, all of the poles of a stable system should be located within the unit circle in the z -plane. See Oppenheim and Schafer [18, 19] for a detailed discussion on this relationship. The bilinear transformation, also known as the Tustin approximation [20], is another method of mapping from the s -plane to the z -plane and vice versa [1–3, 18, 19].

Figure 3.32 provides interpretation of the frequency variable around the unit circle in the z -domain for two different sampling frequencies, $f_s = 2,000 \text{ Hz}$ and $f_s = 500 \text{ Hz}$. The unit circle

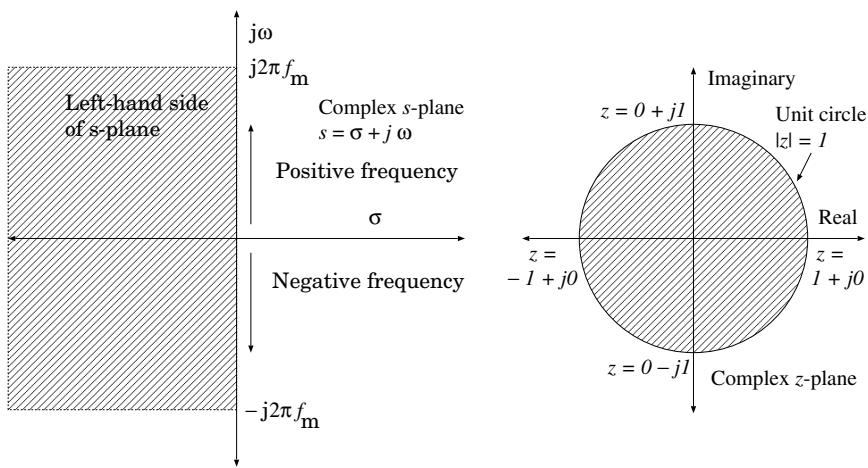


Figure 3.31 Transformation from the Laplace domain to the z -domain with $z = \exp(j\omega T)$. $f_m = f_s/2$.

is given by $z = \exp(j\omega T)$, and represents the frequency axis in a circular and periodic manner instead of the linear vertical axis in the s -plane. (The same symbol ω is being used in the present discussion for the frequency variable in both cases of continuous-time and discrete-time analysis. Different symbols will be used for the two cases when the two variables need to be distinguished.) The notions of a limited bandwidth of $[-f_s/2, +f_s/2]$ or $[0, f_s]$ and the introduction of aliasing errors if the Nyquist sampling rate [1–3] is not satisfied are implied by the circular and periodic nature of the frequency axis.

Just as the Fourier transform may be seen as the Laplace transform evaluated on the imaginary axis with $s = j\omega$, evaluation of the z -transform on the unit circle in the z -plane gives us the Fourier transform, as

$$\begin{aligned} X(\omega) &= \sum_{n=0}^{N-1} x(n) z^{-n}|_{z=\exp(j\omega T)} \\ &= \sum_{n=0}^{N-1} x(n) \exp(-j\omega nT). \end{aligned} \quad (3.66)$$

Using $\omega = 2\pi f$ and $T = 1/f_s$, we have the argument of the \exp function above as $-j\omega nT = -j2\pi n f / f_s$. We may now consider the ratio f/f_s to represent a normalized frequency variable in the range $[0, 1]$, with zero corresponding to DC and unity corresponding to f_s . The result of multiplication of this ratio with 2π may be seen as division of the range $[0, 2\pi]$ into the frequency axis spanning the range $[0, f_s]$. Thus, the variable T may be dropped. (Note: MATLAB® normalizes one-half of the sampling frequency to unity; the maximum normalized frequency present in the sampled signal is then unity, that is, $f_m = 1$.)

Figure 3.27 shows the relationships among a signal $x(n)$, its z -transform and Fourier transform, and a few related functions. It is important to note that while the signal has been sampled and discretized in time, the frequency variable ω in its Fourier transform in Equation 3.66 is a continuous variable.

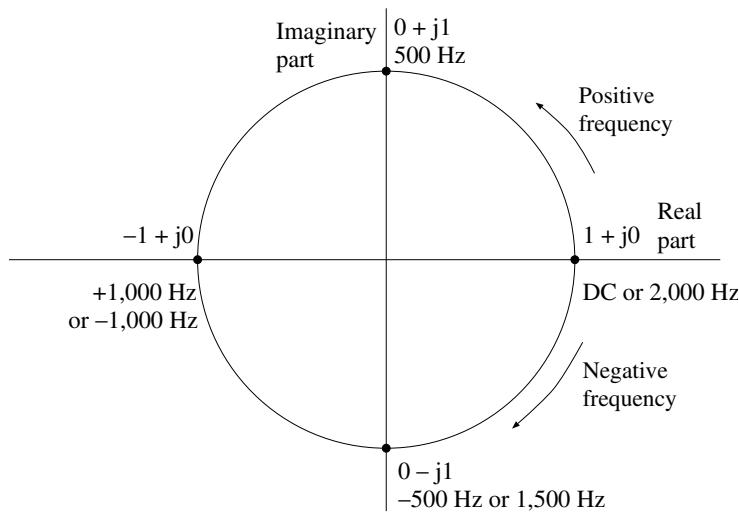
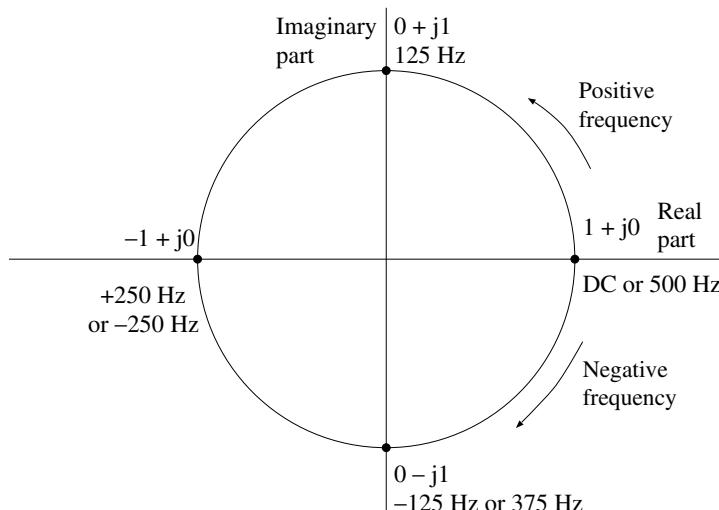
(a) Unit circle in the z -plane with $f_s = 2,000$ Hz(b) Unit circle in the z -plane with $f_s = 500$ Hz

Figure 3.32 Interpretation of the frequency variable with uniform sampling along the unit circle in the z -domain for two different sampling frequencies.

3.4.3 The pole–zero plot

Consider a signal processing system or filter with the transfer function specified as a ratio of two polynomials in z or a rational function of z :

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{k=0}^N b_k z^{-k}}{1 + \sum_{k=1}^M a_k z^{-k}}. \quad (3.67)$$

The equivalent time-domain difference equation of the filter is

$$y(n) = \sum_{k=0}^N b_k x(n-k) - \sum_{k=1}^M a_k y(n-k). \quad (3.68)$$

The equations given above do not include a gain factor that could be specified separately.

The polynomials in the numerator and denominator of Equation 3.67 may be solved for their roots. Let the roots of the numerator be z_k , $k = 1, 2, \dots, N$, which are the zeros of the transfer function $H(z)$. Let the roots of the denominator be p_k , $k = 1, 2, \dots, M$, which are the poles of the transfer function $H(z)$. A plot of the roots of $H(z)$ in the z -plane is the pole-zero plot of the system; see Figure 3.33. The transfer function may be expressed in terms of the poles and zeros as

$$H(z) = \frac{\prod_{k=1}^N (1 - z_k z^{-1})}{\prod_{k=1}^M (1 - p_k z^{-1})}, \quad (3.69)$$

which may be modified to

$$H(z) = z^{(M-N)} \frac{\prod_{k=1}^N (z - z_k)}{\prod_{k=1}^M (z - p_k)}. \quad (3.70)$$

The term $(z - z_k)$ represents the vector from an arbitrary point z to the zero z_k . Similarly, the term $(z - p_k)$ represents the vector from an arbitrary point z to the pole p_k .

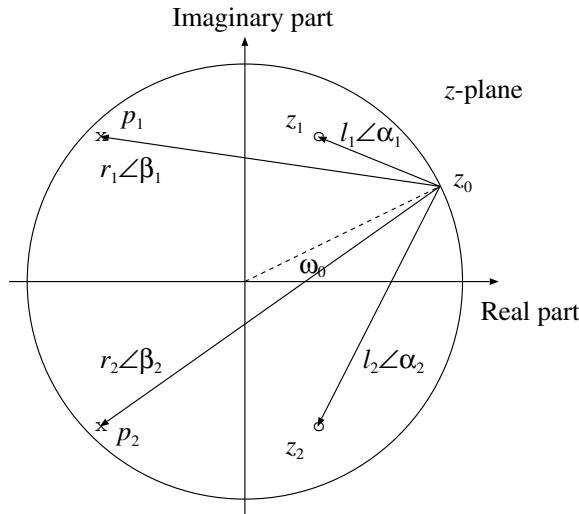


Figure 3.33 Derivation of the frequency response of a system from its pole–zero plot.

Now, consider the evaluation of $H(z)$ for a point z_0 on the unit circle in the z -plane, as shown in Figure 3.33. This gives the frequency response of the system for the corresponding radian frequency ω_0 , as

$$H(\omega_0)|_{z=z_0} = z_0^{(M-N)} \frac{\prod_{k=1}^N (z_0 - z_k)}{\prod_{k=1}^M (z_0 - p_k)}. \quad (3.71)$$

Following the illustration in Figure 3.33, let us represent the vector from z_0 to the zero z_1 as $(z_0 - z_1) = l_1 \angle \alpha_1$, the vector from z_0 to the pole p_1 as $(z_0 - p_1) = r_1 \angle \beta_1$, and apply the same procedure to the remaining poles and zeros. Then, we have

$$|H(\omega_0)| = \frac{\prod_{k=1}^N l_k}{\prod_{k=1}^M r_k} \quad (3.72)$$

and

$$\angle H(\omega_0) = \angle z_0^{(M-N)} + \sum_{k=1}^N \alpha_k - \sum_{k=1}^M \beta_k. \quad (3.73)$$

Thus, the magnitude of the response of a system at a particular frequency is given by the ratio of the product of the distances from the corresponding frequency point in the z -plane to all of the zeros of the system to the product of the distances from the same frequency point to all of the poles of the system (except for a gain factor). Similarly, the phase response is given by the difference between the sum of the angles of the vectors from the frequency point to all of the zeros of the system and the sum of the angles of the vectors from the same frequency point to all of the poles of the system (except for any additional linear phase factor).

It is evident from the equations and the related discussions given above that, as the point z_0 approaches a zero of the system, one of the distances l_k will become small, leading to a low value of the response. If a zero is present on the unit circle, the corresponding l_k will be zero, resulting in a zero-valued magnitude response at the related frequency. Thus, a zero on the unit circle is associated with a spectral null of the system. As the frequency variable crosses from one side of a zero on the unit circle to the other side, there will be an associated 180° change in the phase response. On the contrary, as the point z_0 approaches a pole of the system that is close to the unit circle, the corresponding distance r_k will become small, leading to a high value of the magnitude response at the related frequency. Thus, a pole close to the unit circle is associated with a spectral peak or resonance of the system. In this manner, the pole-zero plot of a system may be inspected to elicit information related to its frequency response in an immediate, albeit qualitative, manner.

3.4.4 The Fourier transform

The Fourier transform is the most commonly used transform to study the frequency-domain characteristics of signals [1–3, 18, 19]. This is mainly because the Fourier transform uses sinusoidal functions as its basis functions. Projections are computed of the given signal $x(t)$ on to the complex exponential basis function of frequency ω rad/s, given by

$$\exp(j\omega t) = \cos(\omega t) + j \sin(\omega t), \quad (3.74)$$

as

$$X(\omega) = \int_{-\infty}^{\infty} x(t) \exp(-j\omega t) dt, \quad (3.75)$$

or in the frequency variable f in Hz as

$$X(f) = \int_{-\infty}^{\infty} x(t) \exp(-j2\pi ft) dt. \quad (3.76)$$

[The complex exponential function is conjugated in computing the projection. In some fields, the forward Fourier transform is defined with $\exp(+j\omega t)$ in the integral.] Equations 3.75 and 3.76 represent *analysis* of the signal $x(t)$ with reference to the complex exponential basis functions. The Fourier transform could also be viewed as decomposition of the given arbitrary signal into sinusoids of all frequencies, the weight or strength of each sinusoid being the corresponding transform coefficient. The lower limit of the integral will be 0 if the signal is causal; the upper limit will be equal to the duration of the signal in the case of a finite-duration signal. The value of $X(\omega)$ or $X(f)$ at

each frequency $\omega = 2\pi f$ represents the amount or strength of the corresponding cosine and sine functions present in the signal $x(t)$. Note that, in general, $X(\omega)$ is complex for real signals, and includes the magnitude and phase of the corresponding complex exponential.

The inverse transformation, representing *synthesis* of the signal $x(t)$ as a weighted combination of the complex exponential basis functions, is given as

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) \exp(j\omega t) d\omega = \int_{-\infty}^{\infty} X(f) \exp(j2\pi ft) df. \quad (3.77)$$

The second version of the equation given above with the frequency variable f in Hz may be more convenient in some situations than the first one with ω in rad/s, due to the absence of the $\frac{1}{2\pi}$ factor. [If the forward Fourier transform is defined with $\exp(+j\omega t)$, the inverse Fourier transform will have $\exp(-j\omega t)$ in the integral.]

In the case of a discrete-time signal $x(n)$, we may still compute the Fourier transform with a continuous frequency variable ω as

$$X(\omega) = \sum_{n=-\infty}^{\infty} x(n) \exp(-j\omega n), \quad (3.78)$$

with the normalized-frequency range $0 \leq \omega \leq 2\pi$ (equivalent to $0 \leq f \leq 1$). The lower limit of the summation will be 0 if the signal is causal. The upper limit of the summation will be equal to the index $(N - 1)$ of the last sample in the case of a finite-duration signal with N samples.

3.4.5 The discrete Fourier transform

If we consider evaluation of the Fourier transform for only certain sampled values of the frequency, we could span the range $[0, 2\pi]$ with K samples, with even steps of $2\pi/K$. This is equivalent to spanning the range $[0, 1]$ of the normalized frequency with K samples. The discrete Fourier transform (DFT) could then be expressed as

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp\left(-j\frac{2\pi}{K}nk\right), \quad (3.79)$$

with $k = 0, 1, 2, \dots, K - 1$. It can be shown that, when the given signal has only N nonzero samples, we need only N samples of the DFT evenly spaced over the unit circle in the z -plane [18] for exact recovery of the original signal from the DFT. Then, we have

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp\left(-j\frac{2\pi}{N}nk\right), \quad (3.80)$$

for $k = 0, 1, 2, \dots, N - 1$, and

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) \exp\left(+j\frac{2\pi}{N}nk\right), \quad (3.81)$$

for $n = 0, 1, 2, \dots, N - 1$, as the forward and inverse DFT relationships, respectively.

Figure 3.27 shows the relationships among a signal $x(n)$, its DFT $X(k)$, and a number of related items. Now, both the time and frequency variables have been sampled or discretized in Equation 3.80. It is important to recognize all of the interrelationships indicated in the figure between a given signal and its transforms in various domains; we use all of these representations in the analysis of signals and related entities either in the continuous form (as functions of t, s, ω, f , and z) or discrete form (as functions of n and k).

If we define a complex variable

$$W_N = \exp\left(-j\frac{2\pi}{N}\right), \quad (3.82)$$

which can also be seen as a vector or a phasor, we can write the DFT relationship as

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{nk}, \quad (3.83)$$

for $k = 0, 1, 2, \dots, N - 1$. Considering a simple case with $N = 8$, Figure 3.34 shows the vectors (or phasors) representing the $N = 8$ roots of unity, with W_8^k , $k = 0, 1, 2, \dots, 7$, where $W_8 = \exp\left(-j\frac{2\pi}{8}\right)$. Given the relationship

$$W_8^{nk} = \exp\left(-j\frac{2\pi}{N}nk\right) = \cos\left(\frac{2\pi}{N}nk\right) - j \sin\left(\frac{2\pi}{N}nk\right), \quad (3.84)$$

the sinusoidal nature of the basis functions used in the DFT becomes clear. Figures 3.35 and 3.36 show stem plots of the cos and sin functions as above, for $k = 0, 1, 2, \dots, 7$ and $N = 8$. Figure 3.37 shows stem plots of the sin functions as above, for $k = 0, 1, 2, \dots, 7$; for improved illustration of the sinusoidal variations at increasing frequencies, each signal was created with $N = 64$ samples, with $n = 0, 1, 2, \dots, 63$.

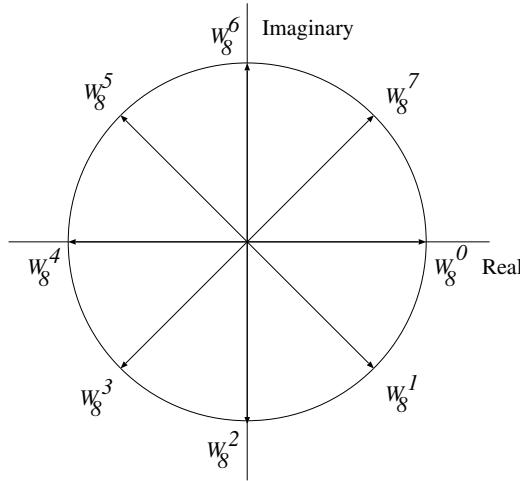


Figure 3.34 Vectors (or phasors) representing the $N = 8$ roots of unity, with W_8^k , $k = 0, 1, 2, \dots, 7$, where $W_8 = \exp\left(-j\frac{2\pi}{8}\right)$. Based on a similar figure by Hall [21].

Let us consider the DFT relationship again, with the basis functions written in terms of the sin and cos functions as follows:

$$\begin{aligned} X(k) &= \sum_{n=0}^{N-1} x(n) \exp\left(-j\frac{2\pi}{N}nk\right) \\ &= \sum_{n=0}^{N-1} x(n) \left\{ \cos\left(\frac{2\pi}{N}nk\right) - j \sin\left(\frac{2\pi}{N}nk\right) \right\} \\ &= \sum_{n=0}^{N-1} x(n) \cos\left(\frac{2\pi}{N}nk\right) - j \sum_{n=0}^{N-1} x(n) \sin\left(\frac{2\pi}{N}nk\right). \end{aligned} \quad (3.85)$$

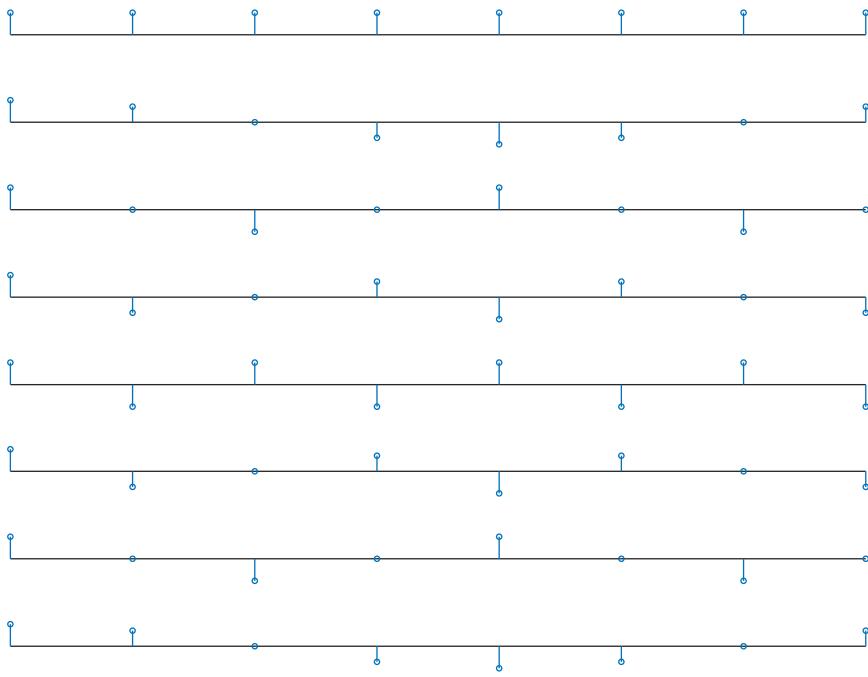


Figure 3.35 Basis functions used in the DFT. The eight plots (top to bottom) show the function $\cos\left(\frac{2\pi}{N}nk\right)$, for $k = 0, 1, 2, \dots, 7$, with $N = 8$. The functions are shown for $n = 0, 1, 2, \dots, 7$, and vary over the range $[-1, 1]$, except for $k = 0$ for which the values of the function are all equal to unity. The axis labels have been suppressed for improved visualization.

It is evident that the real part of $X(k)$ is given by the dot product of the given signal, $x(n)$, with the k^{th} cos basis function, $\cos\left(\frac{2\pi}{N}nk\right)$. Similarly, the imaginary part of $X(k)$ is given by the dot product with the corresponding sin function. The dot product represents the projection of one vector or series of values on to another, and the resulting quantity indicates the extent of commonality between them. Thus, it is clear that the DFT coefficients indicate the amount or strength of each sinusoid present in the given signal. This is the essence of the analysis of a signal in the frequency domain.

Along similar lines, we can view the inverse DFT relationship as a regeneration or synthesis of the original signal as a weighted combination of sinusoids of various frequencies, the weights being the corresponding DFT coefficients, as follows:

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) \cos\left(\frac{2\pi}{N}nk\right) + j \frac{1}{N} \sum_{k=0}^{N-1} X(k) \sin\left(\frac{2\pi}{N}nk\right), \quad (3.86)$$

for $n = 0, 1, 2, \dots, N - 1$.

The periodic nature of the frequency variable and the range of the basis functions used in the DFT relationships leads to an important property: The results of the DFT and inverse DFT operations are periodic signals. Even when we start with a discrete-time signal of finite duration that is not periodic, its spectrum obtained via the DFT is a periodic signal; furthermore, the result of application of the inverse DFT to the spectrum so obtained (and further manipulated as desired) is a periodic discrete-time signal. Unless proper attention is paid to this property, errors comparable to aliasing due to inadequate sampling of a continuous-time signal arise in the final result. Of particular interest are implications of the following property:

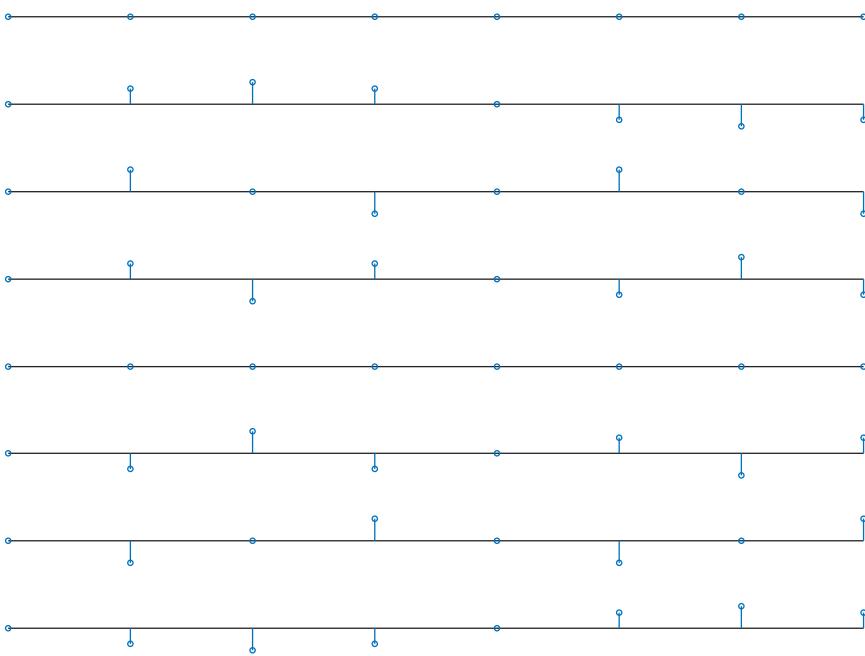


Figure 3.36 Basis functions used in the DFT. The eight plots (top to bottom) show the function $\sin(\frac{2\pi}{N} nk)$, for $k = 0, 1, 2, \dots, 7$, with $N = 8$. The functions are shown for $n = 0, 1, 2, \dots, 7$, and vary over the range $[-1, 1]$, except for $k = 0$ and $k = 4$ for which the values of the functions are all equal to zero. The axis labels have been suppressed for improved visualization.

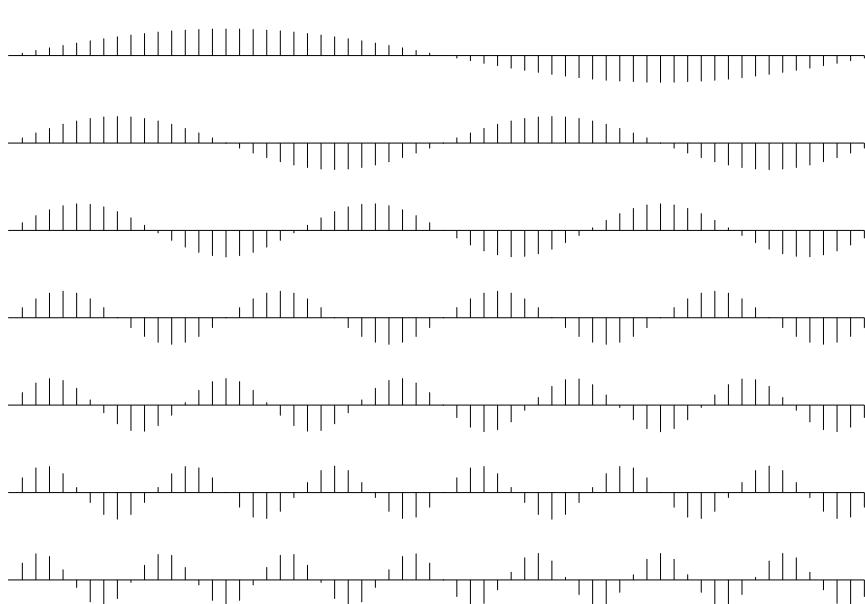


Figure 3.37 Basis functions used in the DFT. The eight plots (top to bottom) show the function $\sin(\frac{2\pi}{N} nk)$, for $k = 0, 1, 2, \dots, 7$, with $N = 64$. The functions are shown for $n = 0, 1, 2, \dots, 63$, and vary over the range $[-1, 1]$, except for $k = 0$ for which the values of the function are equal to zero. The axis labels have been suppressed for improved visualization.

$$\begin{aligned} \text{if } y(n) &= x(n) * h(n), \\ \text{then } Y(k) &= X(k) H(k), \end{aligned} \quad (3.87)$$

which are elaborated upon in Section 3.4.6.

The symmetry and periodicity of the basis functions used in the DFT can be exploited to reduce the computational requirements for the DFT, which has led to the creation of several versions of the fast Fourier transform (FFT) [18, 19, 22]. The two properties are expressed as follows:

$$W_N^{-nk} = (W_N^{nk})^* \quad (3.88)$$

and

$$W_N^{nk} = W_N^{n(k+N)} = W_N^{(n+N)k}. \quad (3.89)$$

Note: In order to indicate the relationships between DFT samples and frequency in Hz, Figure 3.38 illustrates samples around the unit circle in the z -plane for $N = 8$ samples and $f_s = 200$ Hz. Numbering of the DFT samples is indicated two ways: as $k = 0, 1, 2, \dots, N-1$, and $k = 1, 2, \dots, N$, with the latter as in most computer programming languages, including MATLAB®. There are two unique samples at $z = 1$ representing the DC component or $f = 0$ and $z = -1$ representing the folding-frequency component at $f_m = f_s/2$. There are $N-2$ samples in $(N-2)/2$ complex-conjugate pairs on the upper and lower halves of the unit circle. In general, the DFT samples are complex-valued for a real-valued signal; however, the components at DC and f_m possess real values. This discussion assumes that an even number of samples are placed around the unit circle starting from $z = 1$. The distinction between the two ways of numbering the DFT samples is important to note to avoid errors in programming applications.

The discrete cosine transform (DCT) uses only cosine functions as basis functions and offers advantages over the DFT; see Ahmed and Rao [23] and Ahmed et al. [24] for details.

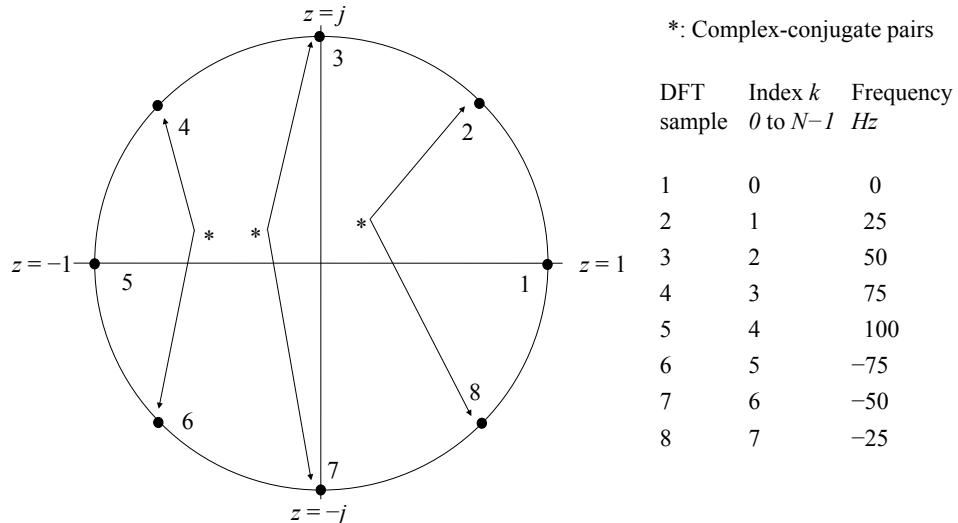


Figure 3.38 Samples of the DFT around the unit circle in the z -plane with $N = 8$ and $f_s = 200$ Hz.

3.4.6 Convolution using the DFT

The advantages provided by FFT algorithms are so substantial that it is usually computationally less expensive to use the second line of Equation 3.87 to realize convolution than the original definition of convolution itself. The algorithm to achieve convolution via the DFT or FFT is as follows:

1. Compute the DFTs, $X(k)$ and $H(k)$, of the two signals given, $x(n)$ and $h(n)$.
2. Multiply the two DFTs, sample by sample, to get the DFT of the output, $Y(k) = X(k) H(k)$.
3. Compute the inverse DFT of $Y(k)$. The output signal, $y(n)$, is given by the real part of the result.

The procedure given above requires the two signals and their transforms to have the same number of samples. However, it should be noted that the convolution achieved in this manner is circular or periodic convolution; the results, in general, will not be the same as those provided by linear convolution. This arises due to the fact that the results of DFT and inverse DFT operations are periodic entities. When we are interested in computing the output of an LSI system, we need to perform linear convolution. To understand the differences between the two types of convolution, let us examine the following examples in detail.

Figure 3.39 shows a numerical illustration of periodic or circular shift applied to a signal. Comparing this figure with Figure 3.16 facilitates understanding the differences between circular and linear shifting.

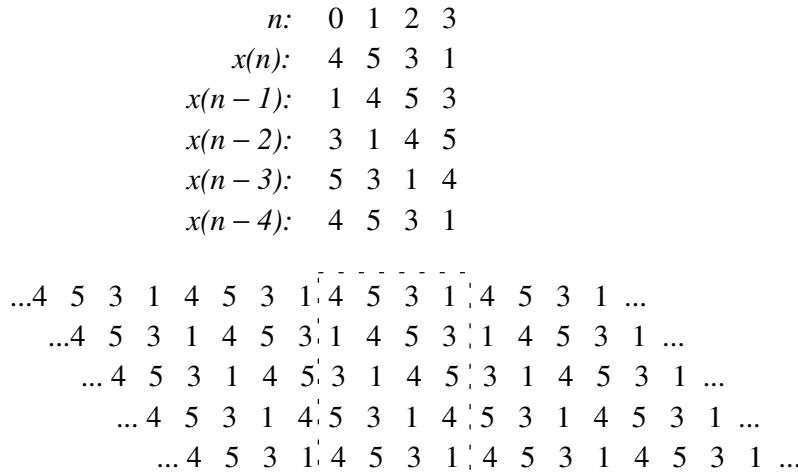


Figure 3.39 A numerical illustration of circularly shifted versions of a periodic signal. The window shown in dashed lines on the lower set of signals over four cycles represents one period. Shifted versions of the signal are shown for one period in the upper part of the figure. The circular nature of the shift is evident in the upper illustration, even if the shift in the lower illustration is considered to be linear. This type of shifting is known as circular or periodic shifting. A sample that gets shifted out of the observation window on the right reenters the window from the left. See Figure 3.16 for an example of linear shifting.

In view of circular shifting of periodic signals, we can express the convolution of two periodic signals, having the same period of N samples, as

$$y_p(n) = \sum_{k=0}^{N-1} x_p(k) h_p[(n - k) \bmod N], \quad (3.90)$$

where the subscript p indicates that the signals are periodic, and $[(n - k) \bmod N]$ returns values within the range $[0, N - 1]$, by adding or subtracting N , as required. The result is also periodic, having the same period of N samples. Figure 3.40 shows a numerical example of circular or periodic convolution of two periodic discrete-time signals.

$n:$	0	1	2	3
$x(n):$	4	1	3	1
$h(n):$	4	3	2	1
$k:$	0	1	2	3
$x(k):$	4	1	3	1
$h(0 - k):$	4	1	2	3
$h(1 - k):$	3	4	1	2
$h(2 - k):$	2	3	4	1
$h(3 - k):$	1	2	3	4
$y(n):$	26	21	24	19
$n:$	0	1	2	3

Figure 3.40 A numerical illustration of circular or periodic convolution of two periodic discrete-time signals.

Given two signals, $x(n)$, $n = 0, 1, 2, \dots, N_1$, and $h(n)$, $n = 0, 1, 2, \dots, N_2$, we know that the result of linear convolution of the two signals should have $N_1 + N_2 - 1$ samples. The application of periodic convolution in place of linear convolution leads to erroneous results; compare the numerical example in Figure 3.41 with that in Figure 3.18.

$n:$	0	1	2	3
$x(n):$	4	1	3	1
$h(n):$	3	2	1	0
$k:$	0	1	2	3
$x(k):$	4	1	3	1
$h(0 - k):$	3	0	1	2
$h(1 - k):$	2	3	0	1
$h(2 - k):$	1	2	3	0
$h(3 - k):$	0	1	2	3
$y(n):$	17	12	15	10
$n:$	0	1	2	3

Figure 3.41 A numerical illustration of the undesired effect of periodic convolution of two discrete-time signals when linear convolution is desired. Compare this with the illustration of linear convolution of the same signals in Figure 3.18.

In order to gain the computational benefits provided by the FFT, the DFT approach to convolution may be modified by padding two given nonperiodic signals of unequal duration with zeros so as to have the same extended duration as that of the desired result, and then considering them to represent

one period each of their periodic extensions. The algorithm to achieve linear convolution via the DFT is as follows:

1. Extend the two signals given, $x(n)$ and $h(n)$, to have N samples each, with $N \geq N_1 + N_2 - 1$, by appending or padding with zeros.
2. Compute the DFTs, $X(k)$ and $H(k)$, of the two extended signals, for $k = 0, 1, 2, \dots, N - 1$.
3. Multiply the two DFTs, sample by sample, to get the DFT of the output, $Y(k) = X(k) H(k)$, $k = 0, 1, 2, \dots, N - 1$.
4. Compute the inverse DFT of $Y(k)$. The output signal, $y(n)$, for $n = 0, 1, 2, \dots, N_1 + N_2 - 1$, is given by the real part of the result.

Note that the sample-by-sample product given as $Y(k) = X(k) H(k)$ cannot be computed if $X(k)$ and $H(k)$ are of different lengths (or periods). The final result $y(n)$ will, in general be complex, and may have extra values for $N_1 + N_2 - 1 < n < N$ that are not of interest. Figure 3.42 shows a numerical example of circular or periodic convolution of two nonperiodic discrete-time signals. By comparing the illustration in Figure 3.42 with the related case of linear convolution of the same signals in Figure 3.18, it is evident that the final results are the same.

$n:$	0	1	2	3	4	5
$x(n):$	4	1	3	1	0	0
$h(n):$	3	2	1	0	0	0
$k:$	0	1	2	3	4	5
$x(k):$	4	1	3	1	0	0
$h(0 - k):$	3	0	0	0	1	2
$h(1 - k):$	2	3	0	0	0	1
$h(2 - k):$	1	2	3	0	0	0
$h(3 - k):$	0	1	2	3	0	0
$h(4 - k):$	0	0	1	2	3	0
$h(5 - k):$	0	0	0	1	2	3
$y(n):$	12	11	15	10	5	1
$n:$	0	1	2	3	4	5

Figure 3.42 A numerical illustration of the use of periodic convolution of two discrete-time and nonperiodic signals to achieve linear convolution by adequate zero padding. Compare this with the illustration of linear convolution of the same signals in Figure 3.18. With the expected number of samples in the result being six, the two input signals, with four and three samples each, have been padded with zeros to the same length of six samples. The effective period used in periodic convolution is six samples.

3.4.7 Properties of the Fourier transform

Some of the important properties of the DFT and their implications are summarized in the following paragraphs [1–3, 18, 19].

- A signal $x(n)$ and its DFT $X(k)$ are both periodic sequences.

- If a signal $x(n)$ has N samples, its DFT $X(k)$ must be computed with at least N samples equally spaced over the normalized frequency range $0 \leq \omega \leq 2\pi$ (or, equivalently, around the unit circle in the z -plane) for complete representation and determination of $X(\omega)$, and hence, exact reconstruction of $x(n)$ via the inverse DFT of $X(k)$. One may use more than N samples to compute $X(k)$ in order to employ an FFT algorithm with $L = 2^M \geq N$ samples, where M is an integer, or to obtain $X(\omega)$ with finer frequency sampling than $2\pi/N$.
- The DFT is linear: the DFT of $ax(n) + by(n)$ is $aX(k) + bY(k)$, where $X(k)$ and $Y(k)$ are the DFTs of $x(n)$ and $y(n)$, respectively.
- The DFT of $x(n - n_o)$ is $\exp(-j\frac{2\pi}{N}kn_o)X(k)$, where $X(k)$ is the DFT of $x(n)$. A time shift leads to a linear component being added to the phase of the original signal. As all sequences in DFT relationships are periodic, the shift operation should be defined as a circular or periodic shift. If at least n_o zeros are present or are padded at the end of the signal before the shift operation, a circular shift of less than or equal to n_o samples will be equivalent to a linear shift.
- The DFT of $x(n) * h(n)$ is $X(k)H(k)$, where $X(k)$ and $H(k)$ are the DFTs of $x(n)$ and $h(n)$, respectively. The inverse DFT of $X(k)H(k)$ is $x(n) * h(n)$. Similarly, $x(n)h(n)$ and $X(k) * \bar{H}(k)$ form a DFT pair. Convolution in one domain is equivalent to multiplication in the other. It is necessary for all of the signals in the relationships mentioned here to have the same period or number of samples N .

As all sequences in DFT relationships are periodic, the convolution operations in the relationships mentioned here are *periodic convolution* and not linear convolution. Note that circular or periodic convolution is defined for periodic signals having the same period, and that the result will also be periodic with the same period as that of the individual input signals.

The result of linear convolution of two signals $x(n)$ and $h(n)$ with different durations N_x and N_h samples, respectively, will have a duration of $N_x + N_h - 1$ samples. If linear convolution is desired via the inverse DFT of $X(k)H(k)$, the DFTs must be computed with $L \geq N_x + N_h - 1$ samples. The individual signals should be padded with zeros at the end to make their effective durations equal for the sake of DFT computation and multiplication. All signals and their DFTs are then periodic with the augmented period of L samples.

- The DFT of a real signal $x(n)$ possesses conjugate-symmetry, that is, $X(-k) = X^*(k)$. As a consequence, the real part and the magnitude of $X(k)$ are even-symmetric sequences, whereas the imaginary part and the phase of $X(k)$ are odd-symmetric sequences.
- According to Parseval's theorem, the total energy of the signal must remain the same before and after Fourier transformation. We then have the following equalities:

$$\begin{aligned} \int_{-\infty}^{\infty} |x(t)|^2 dt &= \frac{1}{2\pi} \int_{-\infty}^{\infty} |X(\omega)|^2 d\omega, \\ \sum_{n=0}^{N-1} |x(n)|^2 &= \frac{1}{N} \sum_{k=0}^{N-1} |X(k)|^2. \end{aligned} \quad (3.91)$$

Because the integral of $|X(\omega)|^2$ over all ω or the sum of $|X(k)|^2$ over all k represents the total energy of the signal (or average power, if the quantity is divided by the duration of the signal), $|X(\omega)|^2$ and $|X(k)|^2$ represent the spread or *density* of the power of the signal along the frequency axis; hence the name power spectral *density* or PSD.

Note: A signal or function $x(n)$ possesses even-symmetry if $x(-n) = x(n)$ or odd-symmetry if $x(-n) = -x(n)$. An arbitrary signal $x(n)$ may be expressed as a combination of a part with even

symmetry, $x_e(n)$, and a part with odd-symmetry, $x_o(n)$, with the even part given by

$$x_e(n) = \frac{1}{2}[x(n) + x(-n)], \quad (3.92)$$

and the odd part given by

$$x_o(n) = \frac{1}{2}[x(n) - x(-n)]. \quad (3.93)$$

Then, we have

$$x(n) = x_e(n) + x_o(n). \quad (3.94)$$

See Figure 3.8 for an illustration of the PSD of the ECG signal in Figure 3.7 computed via the FFT. Several examples of Fourier spectra and PSDs are given in the sections that follow.

The remaining sections in the present chapter provide details of application of the various concepts described in the present section to filtering of biomedical signals for several purposes.

3.5 Synchronized Averaging

Problem: Propose a technique to remove random noise given the possibility of acquiring multiple realizations of the signal or event of interest.

Solution: Linear filters fail to perform well or are not applicable when the signal and noise spectra overlap. Synchronized signal averaging can separate a repetitive signal from noise without distorting the signal [13, 25]. ERP or SEP epochs may be obtained a number of times by repeated application of the stimulus; they may then be averaged by using the stimulus as a trigger to align the epochs. ECG signals may be filtered by detecting the QRS complexes and using their positions to align the waveforms for synchronized averaging. If the noise is random with zero mean and is uncorrelated with the signal, averaging will improve the *SNR*.

Let $y_k(n)$ represent one realization of a signal, with $k = 1, 2, \dots, M$ representing the ensemble index, and $n = 0, 1, 2, \dots, N - 1$ representing the time-sample index. Here, M is the number of observations (events, epochs, or realizations) of the signal available, and N is the number of time samples in each observation of the signal (event). We may express the observed signal as

$$y_k(n) = x_k(n) + \eta_k(n), \quad (3.95)$$

where $x_k(n)$ represents the original uncorrupted signal, and $\eta_k(n)$ represents the noise in the k^{th} observed signal. Now, if for each instant of time n we add the M observations of the signal, we get

$$\sum_{k=1}^M y_k(n) = \sum_{k=1}^M x_k(n) + \sum_{k=1}^M \eta_k(n); \quad n = 0, 1, 2, \dots, N - 1. \quad (3.96)$$

If the repetitions of the signal are identical and aligned, $\sum_{k=1}^M x_k(n) = Mx(n)$. If the noise is random and has zero mean and variance σ_η^2 , $\sum_{k=1}^M \eta_k(n)$ will tend to zero as M increases, with a variance of $M\sigma_\eta^2$. The *RMS* value of the noise in the summed signal is $\sqrt{M}\sigma_\eta$. Thus, the *SNR* of the signal will increase by a factor of $\frac{M}{\sqrt{M}}$ or \sqrt{M} . The larger the number of epochs or realizations that are averaged, the better will be the *SNR* of the result. Note that synchronized averaging is a type of ensemble averaging.

An algorithmic description of synchronized averaging is as follows:

1. Obtain a number of realizations of the signal or event of interest.
2. Determine a reference point for each realization of the signal. This is directly given by the trigger if the signal is obtained by external stimulation (such as ERPs or SEPs), or may be

obtained by detecting the repetitive events in the signal if it is quasiperiodic (such as the QRS complex in the ECG or S1 and S2 in the PCG).

3. Extract parts of the signal corresponding to the events and add them to a buffer. Note that it is possible for the various parts to be of different durations. Alignment of the signals to be averaged at the trigger point is important; their tail ends need not be aligned.
4. Divide the result in the buffer by the number of events or realizations added.

Illustrations of application: Figure 3.43 illustrates two single-flash ERPs in the upper two traces. The results of averaging the ERPs over 10 and 20 flashes are shown in the third and fourth plots, respectively, in the same figure. The averaging process has facilitated identification of the first positivity and the preceding and succeeding troughs (marked on the fourth trace) with certainty; the corresponding features are not reliably seen in the single acquisitions (see also the single-flash ERPs in Figure 3.2). Visual ERPs are analyzed in terms of the latencies of the first major peak or positivity, labeled as P120 due to the fact that the normal expected latency for adults is 120 ms; the trough or negativity before P120, labeled as N80; and the trough following P120, labeled as N145. The N80, P120, and N145 latencies measured from the averaged signal in Trace 4 of Figure 3.43 are 85.7, 100.7, and 117 ms, respectively, which are considered to be within the normal range for adults.

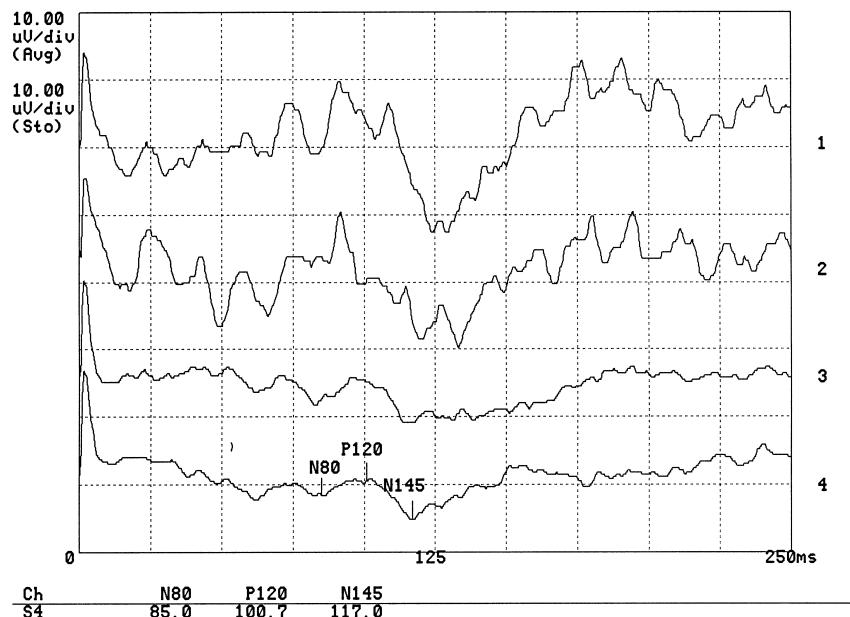


Figure 3.43 Traces 1 and 2: Two sample acquisitions of individual flash visual ERPs from the occipital midline (oz) position of a normal adult male. The earlobes were used to form the ground lead (a1a2), and the left forehead was used as the reference (see Figure 1.39). Trace 3: Average of 10 ERPs. Trace 4: Average of 20 ERPs. The latencies of interest have been labeled on Trace 4 by an EEG technologist as N80, 85.0 ms; P120, 100.7 ms; and N145, 117.0 ms. The height of each grid division (box) is 10 μ V, and the width is 25 ms. See also Figure 3.2. Data courtesy of L. Alfaro and H. Darwish, Alberta Children's Hospital, Calgary.

The upper trace in Figure 3.44 illustrates a noisy ECG signal over several beats. In order to obtain trigger points, a sample QRS complex of 86 ms duration (86 samples at a sampling rate of 1,000 Hz) was extracted from the first beat in the signal and used as a template. Template matching

was performed using a normalized correlation coefficient defined as [13]

$$\gamma_{xy}(k) = \frac{\sum_{n=0}^{N-1}[x(n) - \bar{x}][y(k - N + 1 + n) - \bar{y}_k]}{\sqrt{\sum_{n=0}^{N-1}[x(n) - \bar{x}]^2 \sum_{n=0}^{N-1}[y(k - N + 1 + n) - \bar{y}_k]^2}}, \quad (3.97)$$

where

$$\bar{x} = \frac{1}{N} \sum_{n=0}^{N-1} x(n)$$

and

$$\bar{y}_k = \frac{1}{N} \sum_{n=0}^{N-1} y(k - N + 1 + n).$$

\bar{y}_k is the average of the part of the signal y being used in the template matching procedure. The operation given above is causal, and it is valid from $k = N$ to the last sample of the signal y . x is the template, and \bar{x} is the average value of x . k is the time index of the signal y at which the template is placed. {Jenkins et al. [26] used a measure similar to $\gamma_{xy}(k)$ but without subtraction of the mean and without the shift parameter k to match segmented ECG cycles with a template.} The lower trace in Figure 3.44 shows $\gamma_{xy}(k)$, where it is seen that the cross-correlation result peaks to values near unity at the locations of the QRS complexes in the signal. Averaging inherent in the cross-correlation formula (over N samples) has reduced the effect of noise on template matching.

By choosing an appropriate threshold, it becomes possible to obtain a trigger point to extract the QRS complex locations in the ECG signal. (*Note:* The QRS template matches with the P and T waves with cross-correlation values of about 0.5; wide QRS complexes may yield high cross-correlation values with tall P and T waves. The threshold has to be chosen so as to detect only the QRS complexes.) A threshold of 0.9 was applied to $\gamma_{xy}(k)$, and the QRS positions of all of the 12 beats in the signal were detected.

Figure 3.45 illustrates two ECG cycles extracted using the trigger points obtained by thresholding the cross-correlation function, as well as the result of averaging the first 11 cycles in the signal. It is seen that the noise has been effectively suppressed by synchronized averaging. The low-level baseline variation and power-line interference present in the signal have caused minor artifacts in the result, which are negligible in this illustration.

The most important requirement in synchronized averaging is indicated by the first word in the name of the process: The realizations of the signal that are added for averaging *must be synchronized or aligned* such that the repetitive part of the signal appears at exactly the same instant in each realization of the signal. If this condition is not met, the waveform of the event in the signal will be blurred or smudged along the time axis.

A major advantage of synchronized averaging is that no frequency-domain filtering is performed — either explicitly or implicitly. No spectral content of the signal is lost as is the case with frequency-domain (lowpass) filters or time-domain filters such as moving-window averaging filters. Although the averaging operation is performed with the signal in the time domain, it is a statistical operation applied to values of the signal across an ensemble at a given instant of time, as indicated by the vertical lines in Figure 3.2, and not to signal values along the time axis.

Structured noise such as power-line interference may also be suppressed by synchronized averaging if the phase of the interference in each realization is different. To facilitate this application, the repetition rate of the stimulus should be set so that it is not directly related to the power-line frequency (for example, the flashes used to acquire the averaged ERPs in Figure 3.43 were delivered at 2.1 pps). Physiological interference such as background EEG in ERPs and SEPs may also be suppressed by synchronized averaging, as such activity may bear no interrelationship from one epoch

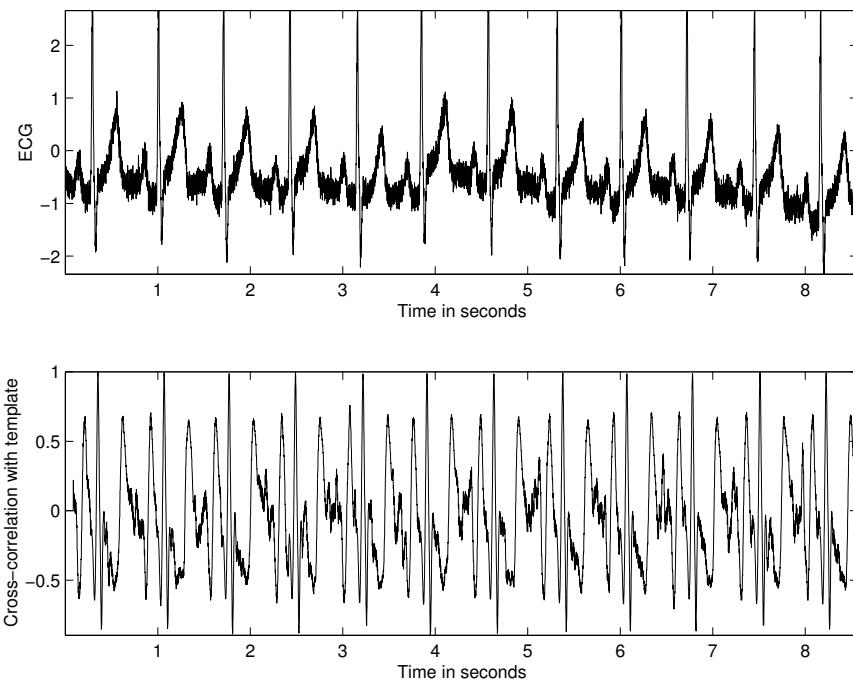


Figure 3.44 An ECG signal with noise (upper trace) and the result of cross-correlation (lower trace) with the QRS template selected from the first cycle. The cross-correlation coefficient is normalized to the range $[-1, 1]$.

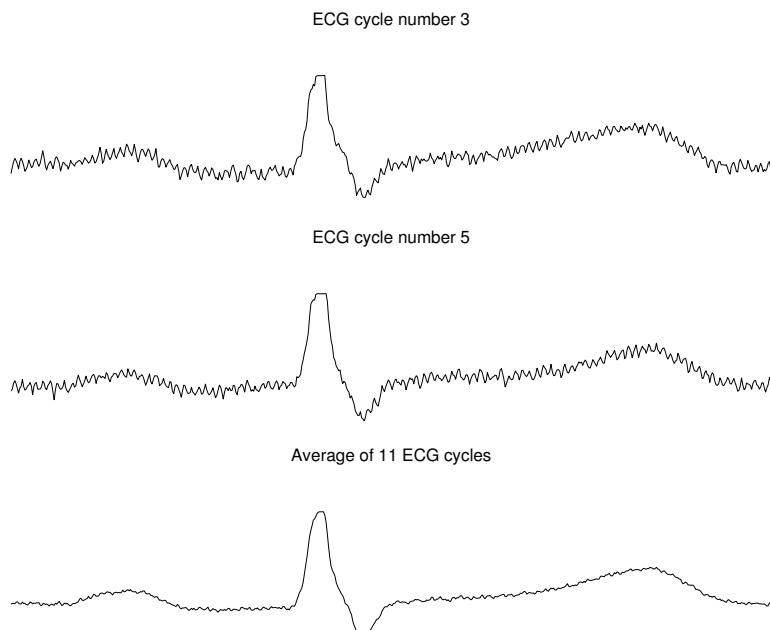


Figure 3.45 Upper two traces: two cycles of the ECG extracted from the signal in Figure 3.44. Bottom trace: the result of synchronized averaging of 11 cycles from the same ECG signal.

of the desired signal to another. See Sections 3.12, 4.10, 5.5.2, and 6.3.6 for further applications and discussions on synchronized averaging.

3.6 Time-domain Filters

Certain types of noise may be filtered directly in the time domain using signal processing techniques or digital filters. An advantage of time-domain filtering is that spectral characterization of the signal and noise may not be required (at least in a direct manner).

3.6.1 Moving-average filters

Problem: Propose a time-domain technique to remove random noise given only one realization of the signal or event of interest.

Solution: When an ensemble of several realizations of an event is not available, synchronized averaging will not be possible. We are then constrained to consider temporal averaging for noise removal, with the assumption that the processes involved are ergodic, that is, temporal statistics may be used instead of ensemble statistics. As temporal statistics are computed using a few samples of the signal along the time axis, and the temporal window of samples is moved to obtain the output at various points of time, such a filtering procedure is called a moving-window averaging filter in general; the term moving-average or MA filter is commonly used. See Figures 3.20, 3.22, and 3.23 as well as the related discussions.

The general form of the difference equation of an MA filter is

$$\begin{aligned} y(n) &= \sum_{k=0}^N b_k x(n-k) \\ &= b_0 x(n) + b_1 x(n-1) + b_2 x(n-2) + \cdots + b_N x(n-N), \end{aligned} \quad (3.98)$$

where x and y are the input and output of the filter, respectively. The b_k values are the filter coefficients or tap weights, $k = 0, 1, 2, \dots, N$, where N is the order of the filter. The effect of division by the number of samples used ($N + 1$) is included in the values of the filter coefficients; typically $\sum_{k=0}^N b_k = 1$, with any desired gain or amplification factor specified separately. The signal-flow diagram of a generic MA filter is shown in Figure 3.46.

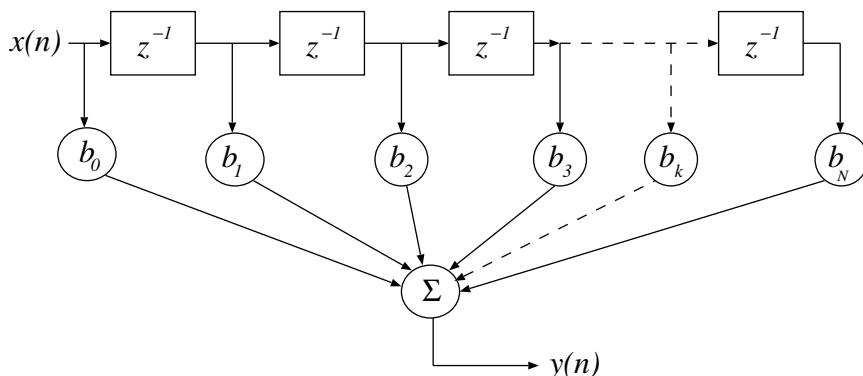


Figure 3.46 Signal-flow diagram of an MA filter of order N . Each block with the symbol z^{-1} represents a delay of one sample and serves as a memory unit for the corresponding signal sample value.

Applying the z -transform, we get the transfer function $H(z)$ of the filter as

$$H(z) = \frac{Y(z)}{X(z)} = \sum_{k=0}^N b_k z^{-k} = b_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_N z^{-N}, \quad (3.99)$$

where $X(z)$ and $Y(z)$ are the z -transforms of $x(n)$ and $y(n)$, respectively.

A simple MA filter to filter noise is the von Hann (also known as Hann or Hanning) filter [25], given by

$$y(n) = \frac{1}{4}[x(n) + 2x(n-1) + x(n-2)]. \quad (3.100)$$

The signal-flow diagram of the Hann filter is shown in Figure 3.47. The impulse response of the filter is obtained by letting $x(n) = \delta(n)$, resulting in

$$h(n) = \frac{1}{4}[\delta(n) + 2\delta(n-1) + \delta(n-2)]; \quad (3.101)$$

see Figure 3.48.

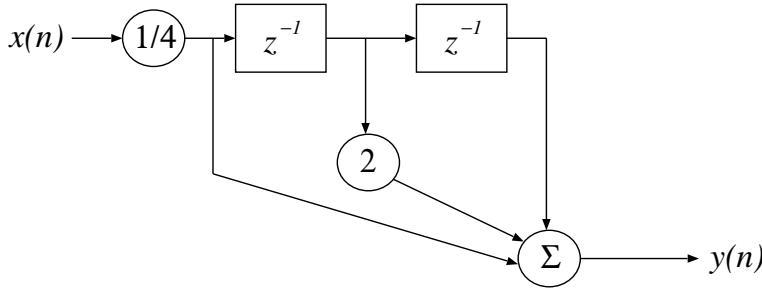


Figure 3.47 Signal-flow diagram of the Hann filter.

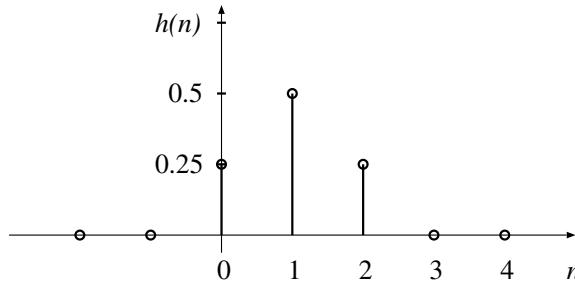


Figure 3.48 Impulse response of the Hann filter.

Applying the z -transform to Equation 3.100, we get

$$Y(z) = \frac{1}{4}[X(z) + 2z^{-1}X(z) + z^{-2}X(z)], \quad (3.102)$$

which leads to the derivation of the transfer function of the Hann filter as

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1}{4}[1 + 2z^{-1} + z^{-2}]. \quad (3.103)$$

The transfer function has a double-zero at $z = -1$.

An MA filter is an FIR filter with the following attributes and advantages:

- The impulse response $h(k)$ has a finite number of terms: $h(k) = b_k$, $k = 0, 1, 2, \dots, N$.
- An FIR filter may be realized nonrecursively (with no feedback).
- The output depends only on the present input sample and a few past input samples.
- The filter is defined by a set of tap weights of the delay stages, as illustrated in Figure 3.46.
- The filter transfer function has no poles except at $z = 0$: the filter is inherently stable.
- The filter has linear phase if the series of tap weights is symmetric or antisymmetric.

(Note: A recursive filter uses previous values of the output to compute the current output value.)

The frequency response of a filter is obtained by substituting $z = e^{j\omega T}$ in the expression for $H(z)$, where T is the sampling interval in seconds, and ω is the radian frequency ($\omega = 2\pi f$, where f is the frequency in Hz). Note that we may set $T = 1$ and deal with normalized frequency in the range $0 \leq \omega \leq 2\pi$ or $0 \leq f \leq 1$; then, $f = 1$ or $\omega = 2\pi$ represents the sampling frequency, with lower frequency values being represented as a normalized fraction of the sampling frequency.

The frequency response of the Hann filter is obtained as

$$H(\omega) = H(z)|_{z=\exp(j\omega)} = \frac{1}{4}[1 + 2e^{-j\omega} + e^{-j2\omega}]. \quad (3.104)$$

This expression can be simplified as follows:

$$\begin{aligned} H(\omega) &= \frac{1}{4}[1 + 2e^{-j\omega} + e^{-j2\omega}] \\ &= \frac{1}{4}e^{-j\omega}[e^{+j\omega} + 2 + e^{-j\omega}] \\ &= \frac{1}{4}e^{-j\omega}[2\cos(\omega) + 2] \\ &= \frac{1}{2}[1 + \cos(\omega)]e^{-j\omega}. \end{aligned} \quad (3.105)$$

Note that $e^{\pm j\omega} = \cos(\omega) \pm j\sin(\omega)$, $e^{+j\omega} + e^{-j\omega} = 2\cos(\omega)$, and $e^{+j\omega} - e^{-j\omega} = 2j\sin(\omega)$. The magnitude and phase of the frequency response are

$$|H(\omega)| = \left| \frac{1}{2}[1 + \cos(\omega)] \right| \quad (3.106)$$

and

$$\angle H(\omega) = -\omega. \quad (3.107)$$

The magnitude and phase responses of the Hann filter are plotted in Figure 3.49. It is clear that the filter is a lowpass filter with linear phase.

Note that, although we started with a description of the Hann filter in the time domain, subsequent analysis of the filter was performed in the frequency domain using the z -transform and the frequency response. System analysis is easier to perform in the z -domain in terms of the poles and zeros of the transfer function and in the frequency domain in terms of the magnitude and phase responses. The magnitude and phase responses assist in understanding the effect of the filter on the frequency components of the signal and noise.

It is seen from the magnitude response of the Hann filter (Figure 3.49) that components beyond about 20% of the sampling frequency of 1,000 Hz are reduced in amplitude by more than 3 dB, that is, to less than one-half of their levels in the input. High-frequency components beyond 40%

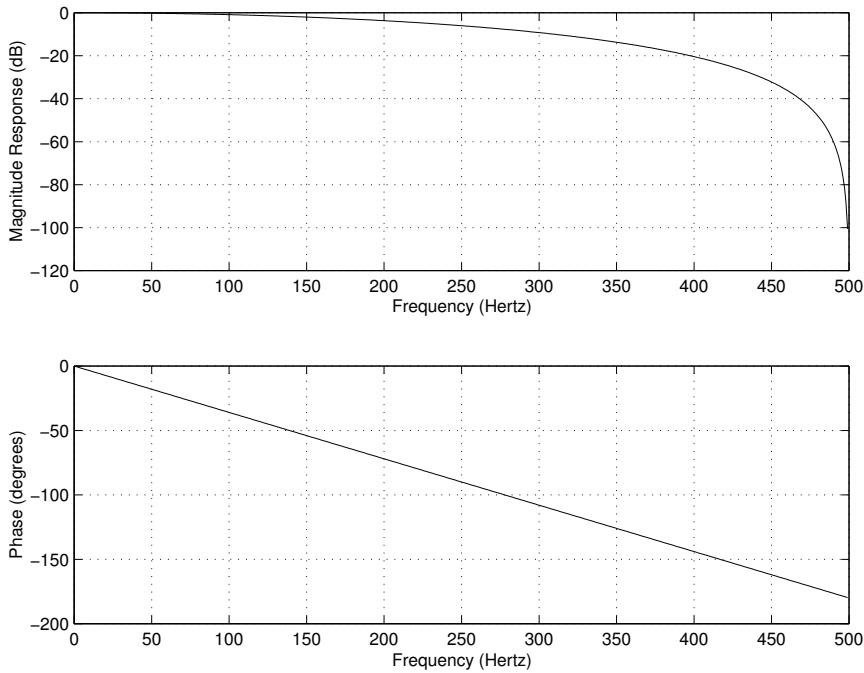


Figure 3.49 Magnitude and phase responses of the Hann (smoothing) filter.

of the sampling frequency are suppressed to less than 20 dB below their input levels. The filter will perform adequate filtering of ECG signals sampled at 200 Hz , with the gain being lower than $-20\ dB$ beyond 80 Hz . However, if the signal is sampled at 1,000 Hz (as in the present example), the gain remains above $-20\ dB$ for frequencies up to 400 Hz ; such a lowpass filter may not be adequate to filter ECG signals, but may be appropriate for other signals, such as the PCG and the EMG.

Increased smoothing may be achieved by averaging signal samples over longer time windows, at the expense of increased filter delay. If the signal samples over a window of eight samples are averaged, we get the 8-point MA filter with the output defined as

$$\begin{aligned} y(n) &= \frac{1}{8} \sum_{k=0}^7 x(n-k) \\ &= \frac{1}{8} [x(n) + x(n-1) + x(n-2) + x(n-3) + x(n-4) \\ &\quad + x(n-5) + x(n-6) + x(n-7)]. \end{aligned} \quad (3.108)$$

The impulse response of the filter is

$$\begin{aligned} h(n) &= \frac{1}{8} [\delta(n) + \delta(n-1) + \delta(n-2) + \delta(n-3) + \delta(n-4) \\ &\quad + \delta(n-5) + \delta(n-6) + \delta(n-7)]. \end{aligned} \quad (3.109)$$

The transfer function of the filter is

$$H(z) = \frac{1}{8} \sum_{k=0}^7 z^{-k}, \quad (3.110)$$

and the frequency response is given by

$$\begin{aligned} H(\omega) &= \frac{1}{8} \sum_{k=0}^7 \exp(-j\omega k) \\ &= \frac{1}{8} \{1 + \exp(-j4\omega) \times [1 + 2 \cos(\omega) + 2 \cos(2\omega) + 2 \cos(3\omega)]\}. \end{aligned} \quad (3.111)$$

The frequency response of the 8-point MA filter is shown in Figure 3.50; the pole-zero plot of the filter is depicted in Figure 3.51. It is seen that the filter has zeros at $\frac{f_s}{8} = 125$ Hz, $\frac{f_s}{4} = 250$ Hz, $\frac{3f_s}{8} = 375$ Hz, and $\frac{f_s}{2} = 500$ Hz. Zeros are also present at -125 , -250 , and -375 Hz to form complex-conjugate pairs with the zeros at the corresponding positive frequencies. Comparing the frequency response of the 8-point MA filter with that of the Hann filter in Figure 3.49, we see that the former provides increased attenuation in the range $90 - 400$ Hz over the latter. Note that the attenuation provided by the filter after about 100 Hz is nonuniform, which may not be desirable in certain applications. Furthermore, the phase response of the filter is not linear, although it is piecewise linear. (Note that the impulse response of the 8-point MA filter corresponds to a rectangle function in the continuous-time domain, for which the Fourier transform is a *sinc* function.)

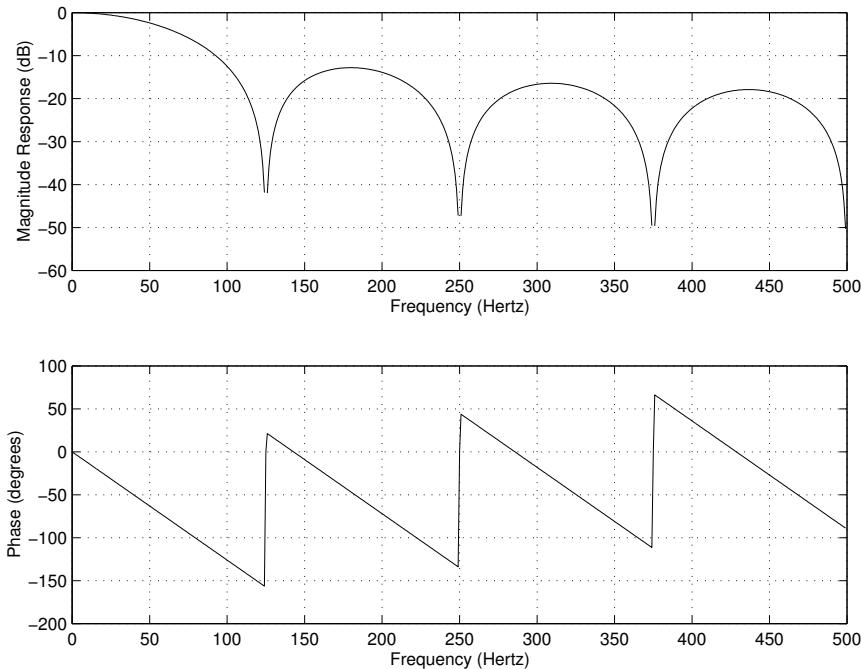


Figure 3.50 Magnitude and phase responses of the 8-point MA (smoothing) filter; $f_s = 1,000$ Hz.

Relationship of MA filtering to integration: Disregarding the $\frac{1}{8}$ scale factor for a moment, the operation in Equation 3.108 may be interpreted as the summation or integration of the signal over the duration $n - 7$ to n . A comparable integration of a continuous-time signal $x(t)$ over the interval $t - \tau$ to t is expressed as

$$y(t) = \int_{t-\tau}^t x(t) dt. \quad (3.112)$$

The general definition of the integral of a signal is

$$y(t) = \int_{-\infty}^t x(t) dt, \quad (3.113)$$

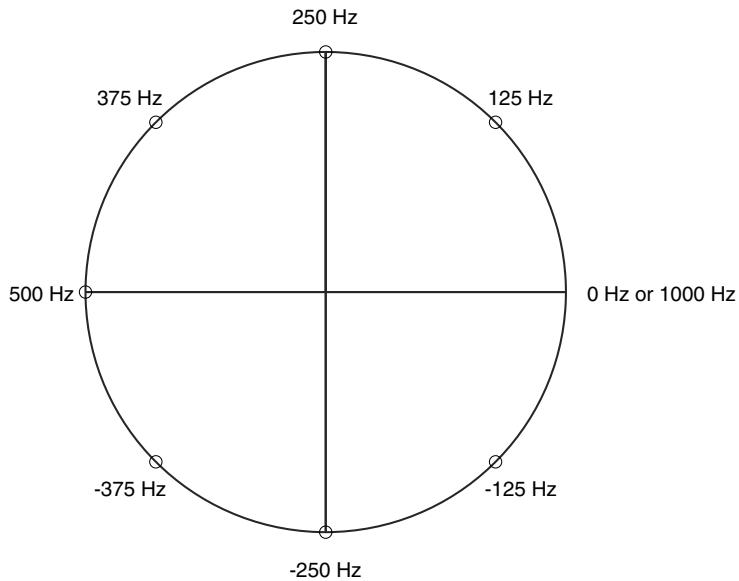


Figure 3.51 Pole–zero plot of the 8-point MA (smoothing) filter; $f_s = 1,000 \text{ Hz}$.

or, if the signal is causal,

$$y(t) = \int_0^t x(t) dt. \quad (3.114)$$

The Fourier transforms of the signals in the relationship given above are related as [1–3]

$$Y(\omega) = \frac{1}{j\omega} X(\omega) + \pi X(0)\delta(\omega). \quad (3.115)$$

Keeping aside the second term related to DC, the frequency response of the integration operator is

$$H(\omega) = \frac{1}{j\omega}, \quad (3.116)$$

with the magnitude response

$$|H(\omega)| = \left| \frac{1}{\omega} \right| \quad (3.117)$$

and phase response

$$\angle H(\omega) = -\frac{\pi}{2}. \quad (3.118)$$

It is seen from the frequency response that the gain of the filter reduces (nonlinearly) as the frequency is increased; therefore, the corresponding filter has lowpass characteristics.

Integration or accumulation of a discrete-time signal for all samples up to the present sample results in the transfer function $H(z) = \frac{1}{1-z^{-1}}$ [1–3]. Such an operation is seldom used in filtering. Instead, a moving-window sum is computed over a duration much shorter than the duration of the signal, such as the 8-point MA filter in Equation 3.108. It follows from Equation 3.108 that

$$\begin{aligned} y(n-1) &= \frac{1}{8}[x(n-1) + x(n-2) + x(n-3) + x(n-4) + x(n-5) \\ &\quad + x(n-6) + x(n-7) + x(n-8)]. \end{aligned} \quad (3.119)$$

By combining Equations 3.108 and 3.119, we get the relationship

$$y(n) = y(n - 1) + \frac{1}{8}x(n) - \frac{1}{8}x(n - 8). \quad (3.120)$$

The recursive form as above clearly depicts the integration aspect of the filter. The transfer function of this expression is easily derived to be

$$H(z) = \frac{1}{8} \left[\frac{1 - z^{-8}}{1 - z^{-1}} \right]. \quad (3.121)$$

The frequency response of the filter is given by

$$H(\omega) = \frac{1}{8} \left[\frac{1 - e^{-j8\omega}}{1 - e^{-j\omega}} \right] = \frac{1}{8} e^{-j\frac{\pi}{2}\omega} \left[\frac{\sin(4\omega)}{\sin(\frac{\omega}{2})} \right], \quad (3.122)$$

which is equivalent to that in Equation 3.111. Summation over a limited discrete-time window results in a frequency response having sinc-type characteristics, as illustrated in Figure 3.50. See Tompkins [25] for a discussion on other types of integrators.

Illustration of application: Figure 3.52 shows a segment of an ECG signal with high-frequency noise. Figure 3.53 shows the result of filtering the signal with the 8-point MA filter described above. Although the noise level has been reduced, some noise is still present in the result. This is due to the fact that the attenuation of the simple 8-point MA filter is not more than -20 dB at most frequencies (except near the zeros of the filter).

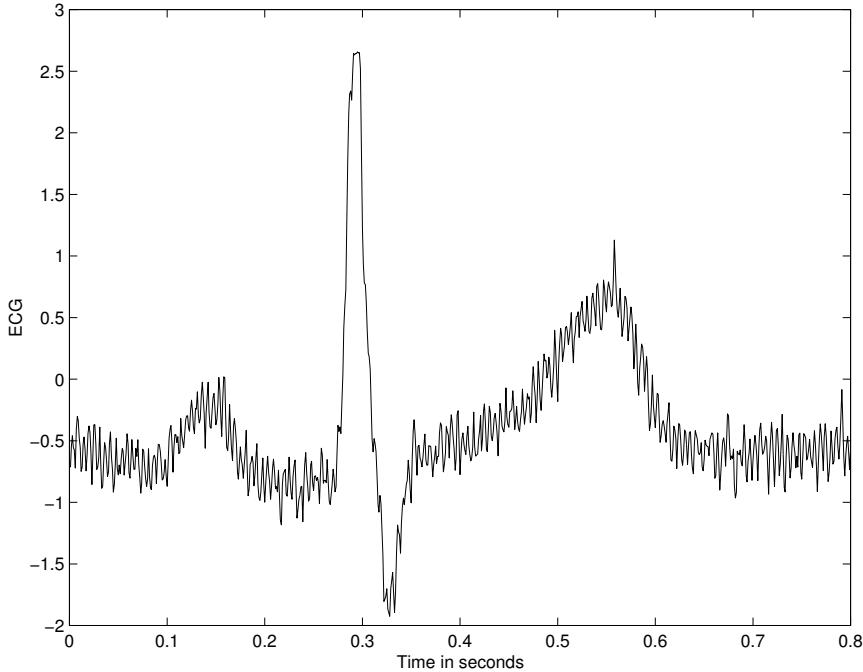


Figure 3.52 ECG signal with high-frequency noise; $f_s = 1,000 \text{ Hz}$.

3.6.2 Derivative-based operators to remove low-frequency artifacts

Problem: Develop a time-domain technique to remove baseline drift in the ECG signal.

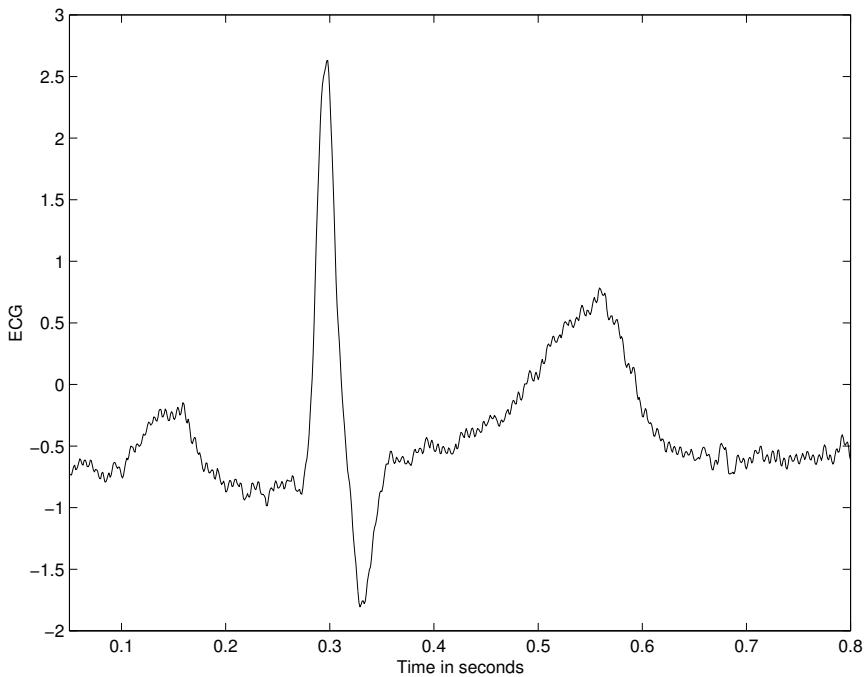


Figure 3.53 The ECG signal with high-frequency noise in Figure 3.52 after filtering by the 8-point MA filter shown in Figure 3.50 and Equation 3.108.

Solution: The derivative operator in the time domain removes the parts of the input that are constant (the output is zero). Large changes in the input lead to high values in the output of the derivative operator. Improved understanding of the derivative operation may be obtained by studying its transform in the frequency domain.

The ideal $\frac{d}{dt}$ operator in the time domain results in multiplication of the Fourier transform of the original signal by $j\omega = j2\pi f$ in the frequency domain. If $X(f)$ represents the Fourier transform of the signal $x(t)$, then the Fourier transform of $\frac{dx}{dt}$ is $j2\pi f X(f)$ or $j\omega X(\omega)$. The frequency response of the operation is $H(\omega) = j\omega$. It is seen that the magnitude of the frequency response increases linearly with frequency, starting with $H(\omega) = 0$ at $\omega = 0$. Thus, the DC component is removed by the derivative operator, and higher frequencies receive linearly increasing gain: the operation represents a highpass filter. The derivative operator may be used to remove DC and suppress low-frequency components (and boost high-frequency components).

It follows readily that the second-order derivative operator $\frac{d^2}{dt^2}$ has the frequency response $H(\omega) = (j\omega)(j\omega) = -\omega^2$, with a quadratic increase in gain for higher frequencies. The second-order derivative operator may be used to obtain even higher gain for higher frequencies than the first-order derivative operator; the former may be realized as a cascade of two of the latter.

In DSP, the basic derivative is given by the first-order difference operator [25]

$$y(n) = \frac{1}{T} [x(n) - x(n-1)]. \quad (3.123)$$

The scale factor including the sampling interval T is required in order to obtain the rate of change of the signal with respect to the true time. The transfer function of the operator is

$$H(z) = \frac{1}{T} (1 - z^{-1}). \quad (3.124)$$

The filter has a zero at $z = 1$, the DC point.

The frequency response of the operator is

$$H(\omega) = \frac{1}{T} [1 - \exp(-j\omega)] = \frac{1}{T} \exp\left(-j\frac{\omega}{2}\right) \left[2j \sin\left(\frac{\omega}{2}\right)\right], \quad (3.125)$$

which leads to

$$|H(\omega)| = \frac{2}{T} \left|\sin\left(\frac{\omega}{2}\right)\right| \quad (3.126)$$

and

$$\angle H(\omega) = \frac{\pi}{2} - \frac{\omega}{2}. \quad (3.127)$$

The magnitude and phase responses of the first-order difference operator are plotted in Figure 3.54. The gain of the filter increases for higher frequencies up to the folding frequency $f_s/2$. The gain may be taken to approximate that of the ideal derivative operator, that is, $|\omega|$, for low values of ω . Any high-frequency noise present in the signal will be amplified significantly: therefore, the result could be noisy.

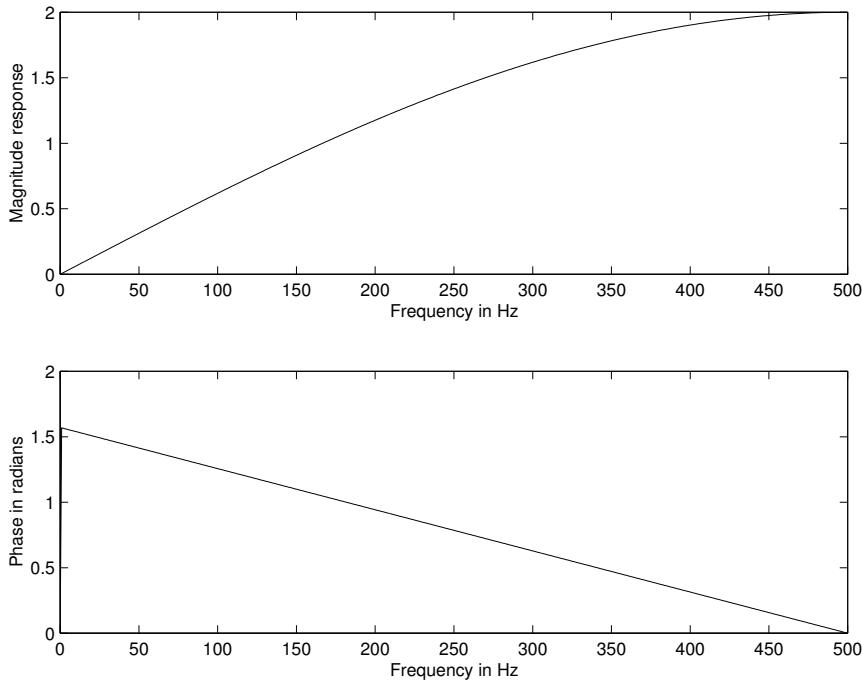


Figure 3.54 Magnitude and phase responses of the first-order difference operator in Equation 3.123. The magnitude response is shown on a linear scale in order to illustrate better its proportionality to frequency.

The noise-amplification problem with the first-order difference operator in Equation 3.123 may be controlled by taking the average of two successive output values:

$$\begin{aligned} y_3(n) &= \frac{1}{2} [y(n) + y(n-1)] \\ &= \frac{1}{2T} \{[x(n) - x(n-1)] + [x(n-1) - x(n-2)]\} \\ &= \frac{1}{2T} [x(n) - x(n-2)]. \end{aligned} \quad (3.128)$$

The transfer function of the operator above, known as the three-point central difference [25], is

$$H(z) = \frac{1}{2T} (1 - z^{-2}) = \left[\frac{1}{T} (1 - z^{-1}) \right] \left[\frac{1}{2} (1 + z^{-1}) \right]. \quad (3.129)$$

Observe that the transfer function of the three-point central-difference operator is the product of the transfer functions of the simple first-order difference operator and a two-point MA filter. The three-point central-difference operation may, therefore, be performed by the simple first-order difference operator and a two-point MA filter in series (cascade).

The magnitude and phase responses of the three-point central-difference operator are given by

$$|H(\omega)| = \frac{1}{T} |\sin(\omega)| \quad (3.130)$$

and

$$\angle H(\omega) = \frac{\pi}{2} - \omega, \quad (3.131)$$

and are plotted in Figure 3.55. The transfer function has zeros at $z = 1$ and $z = -1$, with the latter pulling the gain at the folding frequency ($f_s/2$) to zero: The operator is a bandpass filter, including a lowpass filter and a highpass filter in series. Although the operator does not have the noise-amplification problem of the first-order difference operator, the approximation of the $\frac{d}{dt}$ operation is poor after about $f_s/10$ [25].

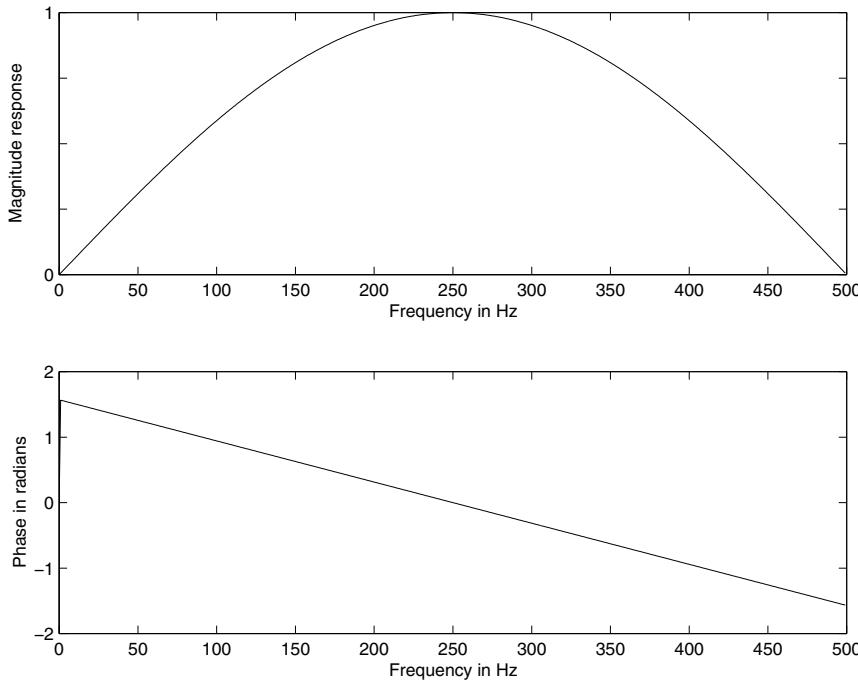


Figure 3.55 Magnitude and phase responses of the three-point central-difference operator in Equation 3.128. The magnitude response is shown on a linear scale.

Illustration of application: Figures 3.56 and 3.57 show the results of filtering the ECG signal with low-frequency noise shown in Figure 3.6, using the first-order difference and three-point central-difference operators, respectively. It is seen that the baseline drift has been removed, with

the latter being less noisy than the former. However, it is obvious that the highpass and high-frequency emphasis effects inherent in both operators have removed the slow P and T waves, and have altered the QRS complexes to such an extent as to make the resulting waveforms look unlike ECG signals. (We will see in Section 4.3 that, although the derivative operators are not useful in the present application, they are useful in detecting the QRS complex and the dicrotic notch.)

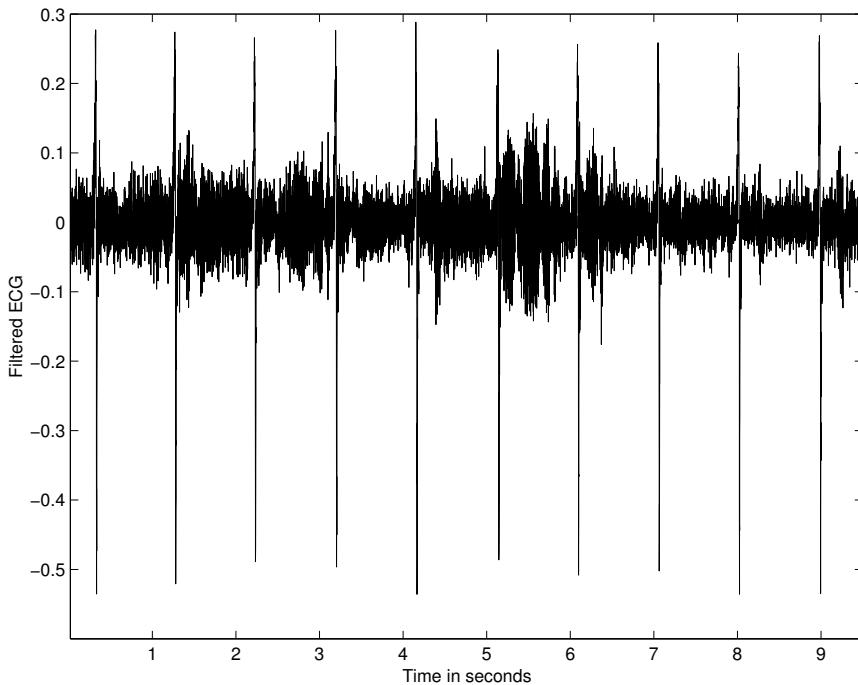


Figure 3.56 Result of filtering the ECG signal with low-frequency noise shown in Figure 3.6 using the first-order difference operator in Equation 3.123.

Problem: How can we improve the performance of the basic first-order difference operator as a filter to remove low-frequency noise or baseline wander without distorting the QRS complex?

Solution: A drawback of the first-order difference and the three-point central-difference operators lies in the fact that their magnitude responses remain low for a significant range of frequencies well beyond the band related to baseline wander. The zero of the first-order difference operator at $z = 1$ is desired in order to reject the DC component and low frequencies. However, we would like to maintain the levels of the components present in the signal beyond about 0.5 Hz, that is, we would like the gain of the filter to be close to unity after about 0.5 Hz.

The gain of a filter at specific frequencies may be boosted by placing poles at the related locations around the unit circle in the z -plane. For the sake of stability of the filter, the poles should be placed within the unit circle. Since we are interested in maintaining a high gain at low frequencies, we could place a pole on the real axis (zero frequency), for example, at $z = 0.995$ [27]. The transfer function of the modified first-order difference filter is then

$$H(z) = \frac{1}{T} \left[\frac{1 - z^{-1}}{1 - 0.995 z^{-1}} \right] \quad (3.132)$$

or, equivalently,

$$H(z) = \frac{1}{T} \left[\frac{z - 1}{z - 0.995} \right]. \quad (3.133)$$

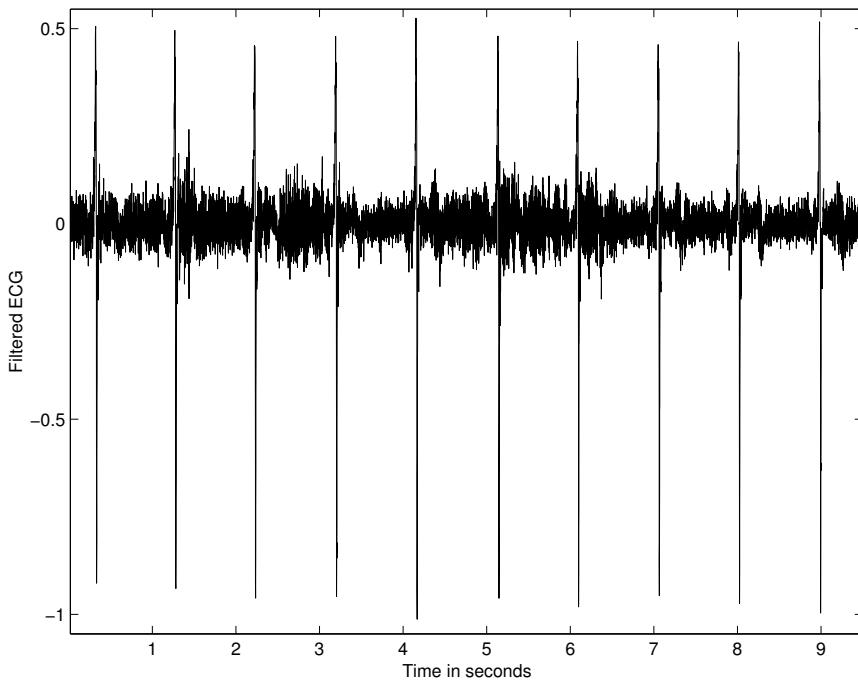


Figure 3.57 Result of filtering the ECG signal with low-frequency noise shown in Figure 3.6 using the three-point central-difference operator in Equation 3.128.

The time-domain input–output relationship is given as

$$y(n) = \frac{1}{T} [x(n) - x(n-1)] + 0.995 y(n-1). \quad (3.134)$$

Two equivalent signal-flow diagrams of the filter are shown in Figure 3.58. (Note: The filter is no longer an FIR filter; details on infinite impulse response or IIR filters are presented in Section 3.7.1.)

The form of $H(z)$ in Equation 3.133 in terms of z helps in understanding the graphical method for the evaluation of the frequency response of discrete-time filters [1–3, 25] shown in Figure 3.33; see also Figure 3.59. The frequency response of a system is obtained by evaluating its transfer function at various points on the unit circle in the z -plane, that is, by letting $z = \exp(j\omega)$ and evaluating $H(z)$ for various values of the frequency variable ω of interest. The numerator in Equation 3.133 expresses the vectorial distance between a specified point in the z -plane and the zero at $z = 1$; the denominator gives the distance to the pole at $z = 0.995$. For ease of illustration, the position of the pole is shown farther from $z = 1$ in Figure 3.59. The magnitude transfer function of a system for a particular value of z is given by the product of the distances from the corresponding point in the z -plane to all of the zeros of the system's transfer function, divided by the product of the distances to all of its poles. The phase response is given by the sum of the angles of the vectors joining the point to all of the zeros, minus the sum of the angles to all of the poles [1–3, 25]. It is obvious that the magnitude response of the filter in Equations 3.132 and 3.133 is zero at $z = 1$, due to the presence of a zero at that point. Furthermore, for values of z away from $z = 1$, the distances to the zero at $z = 1$ and the pole at $z = 0.995$ will be almost equal; therefore, the gain of the filter will be close to unity for frequencies greater than about 5 Hz; see Figure 3.59. The magnitude and phase responses of the filter shown in Figure 3.60 confirm these observations: The filter is a highpass filter with nonlinear phase.

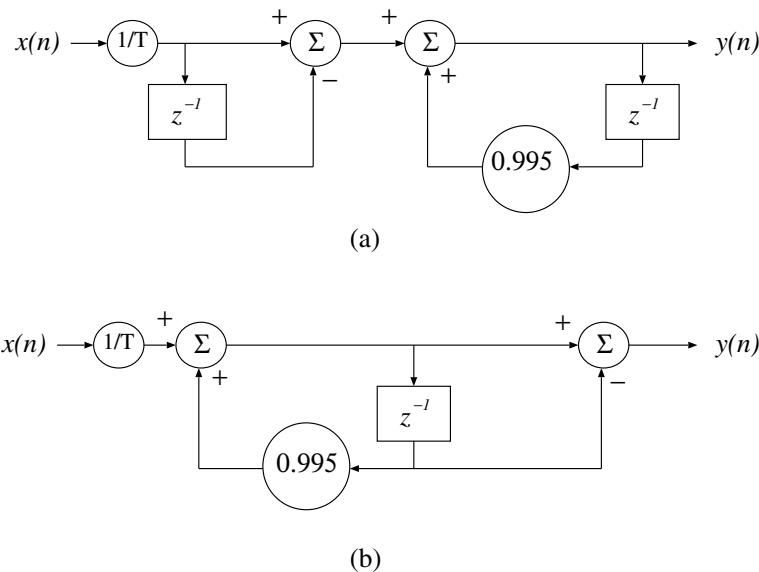


Figure 3.58 Two equivalent signal-flow diagrams of the filter to remove low-frequency noise or baseline wander. The form in (a) uses two delays, whereas that in (b) uses only one delay.

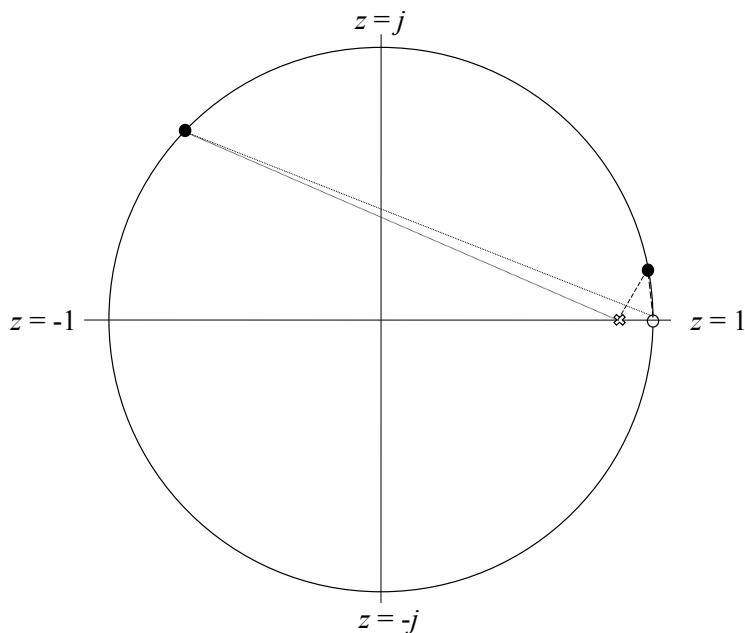


Figure 3.59 Graphical evaluation of the frequency response of a highpass filter.

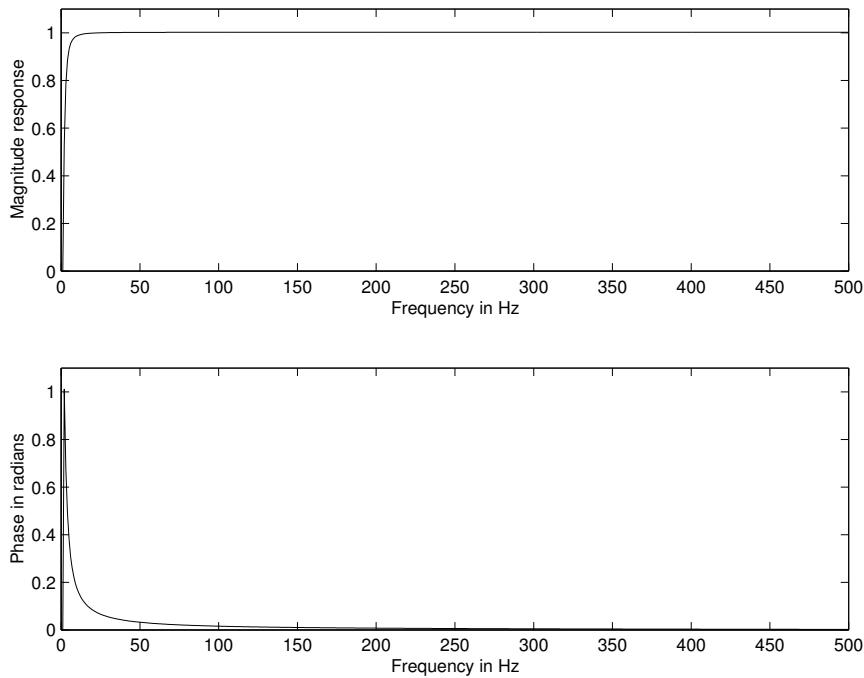


Figure 3.60 Normalized magnitude and phase responses of the filter to remove baseline wander as in Equation 3.132. The magnitude response is shown on a linear scale.

The result of application of the filter in Equation 3.134 to the ECG signal with low-frequency noise shown in Figure 3.6 is displayed in Figure 3.61. It is evident that the low-frequency baseline artifact has been removed without any significant distortion of the ECG waveforms, as compared with the results of differentiation in Figures 3.56 and 3.57. Close inspection, however, reveals that the S wave has been enhanced (made deeper), and that negative undershoots have been introduced after the P and T waves. Removal of the low-frequency baseline artifact has been achieved at the cost of distortion of the ECG waves due to the use of a derivative-based filter and its nonlinear phase response.

3.6.3 Various specifications of a filter

Although we started the present section with the design of filters in the time domain, we used several ways to specify the parameters and analyze the characteristics of filters that are not limited to the time domain. Some of the several possible specifications of a filter are listed below.

- Difference equation.
- Signal-flow diagram.
- Tap-weight or filter coefficients.
- Impulse response, $h(n)$.
- Transfer function, $H(z)$.
- Frequency response, $H(\omega)$, including its magnitude and phase parts.

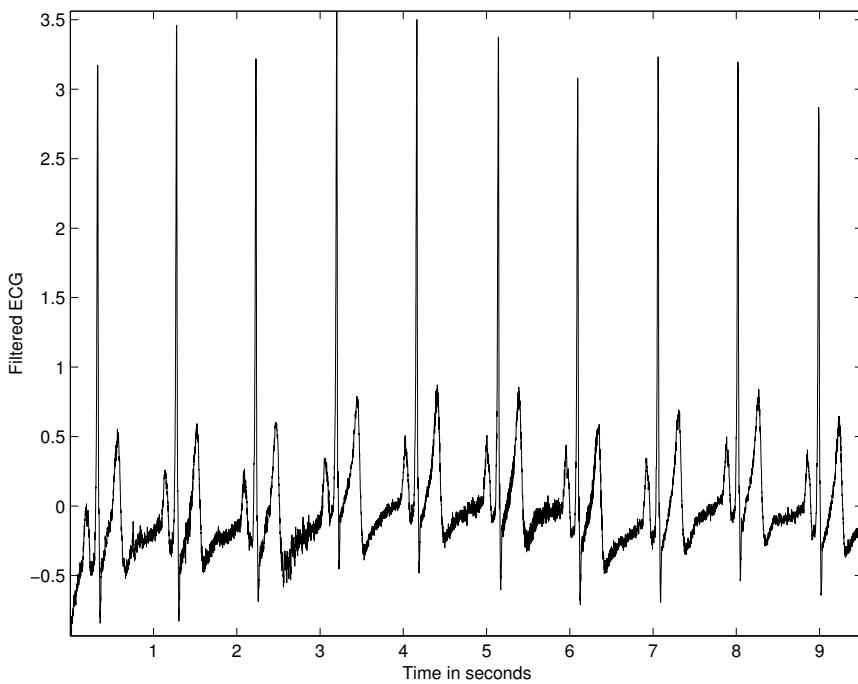


Figure 3.61 Result of processing the ECG signal with low-frequency noise shown in Figure 3.6 using the filter to remove baseline wander as in Equation 3.134. (Compare with the results in Figures 3.56 and 3.57.)

- Pole–zero diagram and a gain factor.

The items listed above are interrelated, and any one of them may be used to derive any of the others, as seen in the several illustrations provided in the present and previous sections; the following sections provide more examples.

3.7 Frequency-domain Filters

The filters described in the previous section perform relatively simple operations in the time domain; although their frequency-domain characteristics were explored, the operators were not specifically designed to possess any particular frequency response at the outset. The frequency response of the 8-point MA filter in Figure 3.50, in particular, is not attractive: the attenuation in the stopband is not high and is not uniform, with the gain falling below -20 dB only around the zeros of the transfer function.

Filters may be designed in the frequency domain to provide specific lowpass, highpass, bandpass, or band-reject (notch) characteristics. Frequency-domain filters may be implemented in software after obtaining the Fourier transform of the input signal, or converted into equivalent time-domain filters and applied directly upon the signal samples.

Many design procedures are available in the literature to design various types of filters: The most-commonly used designs are the Butterworth, Chebyshev, elliptic, and Bessel filters [18, 28–33]. Since these filters have been well established in the analog-filter domain, it is common to commence with an analog design and apply the bilinear transformation to obtain a digital filter in the z -domain. It is also common to design a lowpass filter with the desired passband, transition, and stopband characteristics on a normalized-frequency axis, and then transform it to the desired lowpass, highpass, bandpass, or band-reject characteristics [18, 28]. Frequency-domain filters may also

be specified directly in terms of the values of the desired frequency response at certain frequency samples only, and then transformed into the equivalent time-domain filter coefficients via the inverse Fourier transform.

3.7.1 Removal of high-frequency noise: Butterworth lowpass filters

Problem: Design a frequency-domain filter to remove high-frequency noise with minimal loss of signal components in the specified passband.

Solution: The Butterworth filter is a commonly used frequency-domain filter due to its simplicity and the property of a maximally flat magnitude response in the passband. For a Butterworth lowpass filter of order N , the first $2N - 1$ derivatives of the squared magnitude response are zero at $\Omega = 0$, where Ω represents the analog radian frequency. The Butterworth filter response is monotonic in the passband as well as in the stopband.

The basic Butterworth lowpass filter function is given as [18, 19]

$$|H_a(j\Omega)|^2 = \frac{1}{1 + \left(\frac{j\Omega}{j\Omega_c}\right)^{2N}}, \quad (3.135)$$

where H_a is the frequency response of the analog filter, and Ω_c is the cutoff frequency (in rad/s). A Butterworth filter is completely specified by its cutoff frequency Ω_c and order N . As the order N increases, the filter response becomes more flat in the passband, and the transition to the stopband becomes faster or sharper. $|H_a(j\Omega_c)|^2 = \frac{1}{2}$ for all N .

Changing to the Laplace variable s , we get

$$H_a(s)H_a(-s) = \frac{1}{1 + \left(\frac{s}{j\Omega_c}\right)^{2N}}. \quad (3.136)$$

The poles of the squared transfer function are located with equal spacing around a circle of radius Ω_c in the s -plane, distributed symmetrically on either side of the imaginary axis $s = j\Omega$. No pole will lie on the imaginary axis itself; poles will appear on the real axis for odd N . The angular spacing between the poles is $\frac{\pi}{N}$. If $H_a(s)H_a(-s)$ has a pole at $s = s_p$, it will have a pole at $s = -s_p$ as well. Furthermore, for the filter coefficients to be real, complex poles must appear in conjugate pairs. In order to obtain a stable and causal filter, we need to form $H_a(s)$ with only the N poles on the LHS of the s -plane. The pole positions in the s -plane are given by

$$s_k = \Omega_c \exp \left[j\pi \left(\frac{1}{2} + \frac{(2k-1)}{2N} \right) \right], \quad (3.137)$$

$k = 1, 2, \dots, 2N$ [28].

Once the pole positions are obtained in the s -plane, they may be combined to obtain the transfer function in the analog Laplace domain as

$$H_a(s) = \frac{G}{(s - p_1)(s - p_2)(s - p_3) \cdots (s - p_N)}, \quad (3.138)$$

where p_k , $k = 1, 2, \dots, N$, are the N poles of the transfer function in the left-half of the s -plane, and G is a gain factor specified as needed or calculated to normalize the gain at DC ($s = 0$) to be unity.

If we use the bilinear transformation

$$s = \frac{2}{T} \left[\frac{1 - z^{-1}}{1 + z^{-1}} \right], \quad (3.139)$$

the Butterworth circle in the s -plane maps to a circle in the z -plane with its real-axis intercepts at $z = \frac{2-\Omega_c T}{2+\Omega_c T}$ and $z = \frac{2+\Omega_c T}{2-\Omega_c T}$. The poles at $s = s_p$ and $s = -s_p$ in the s -plane map to the locations $z = z_p$ and $z = 1/z_p$, respectively. The poles in the z -plane are not uniformly spaced around the transformed Butterworth circle. For stability, all poles of $H(z)$ must lie within the unit circle in the z -plane.

Consider the unit circle in the z -plane given by $z = e^{j\omega}$. For points on the unit circle, we have

$$s = \sigma + j\Omega = \frac{2}{T} \left(\frac{1 - e^{-j\omega}}{1 + e^{-j\omega}} \right) = \frac{2j}{T} \tan\left(\frac{\omega}{2}\right), \quad (3.140)$$

including the bilinear transformation in Equation 3.139. For the unit circle, $\sigma = 0$; therefore, we get the relationships between the continuous-time (s -domain) frequency variable Ω and the discrete-time (z -domain) frequency variable ω as

$$\Omega = \frac{2}{T} \tan\left(\frac{\omega}{2}\right) \quad (3.141)$$

and

$$\omega = 2 \tan^{-1}\left(\frac{\Omega T}{2}\right). \quad (3.142)$$

This is a nonlinear relationship that warps the frequency values as they are mapped from the imaginary (vertical) axis in the s -plane to the unit circle in the z -plane (or vice versa), and it should be taken into account in specifying cutoff frequencies.

The transfer function $H_a(s)$ may be mapped to the z -domain by applying the bilinear transformation, that is, by substituting $s = \frac{2}{T} \frac{1-z^{-1}}{1+z^{-1}}$ in Equation 3.138. The transfer function $H(z)$ may then be simplified to the form

$$H(z) = \frac{G' (1 + z^{-1})^N}{\sum_{k=0}^N a_k z^{-k}}, \quad (3.143)$$

where a_k , $k = 0, 1, 2, \dots, N$, are the filter coefficients or tap weights (with $a_0 = 1$), and G' is the gain factor [usually calculated so as to obtain $|H(z)| = 1$ at DC, that is, at $z = 1$]. Observe that the filter has N zeros at $z = -1$ due to the use of the bilinear transformation. The filter is now in the familiar form of an IIR filter. Two forms of realization of a generic IIR filter are illustrated as signal-flow diagrams in Figures 3.62 and 3.63: The former represents a direct realization using $2N$ delays and $2N + 1$ multipliers (with $a_0 = 1$), whereas the latter uses only N delays and $2N + 1$ multipliers.

A time-domain representation of the filter will be required if the filter is to be applied to data samples directly in the time domain. From the filter transfer function $H(z)$ in Equation 3.143, it becomes easy to represent the filter in the time domain with the difference equation

$$y(n) = \sum_{k=0}^N b_k x(n-k) - \sum_{k=1}^N a_k y(n-k). \quad (3.144)$$

The coefficients b_k are given by the coefficients of the expansion of $G'(1+z^{-1})^N$. (The MATLAB® command *butter* provides Butterworth filters [34].)

It is also possible to directly specify the Butterworth lowpass filter as

$$|H(\omega)|^2 = \frac{1}{1 + \left(\frac{\omega}{\omega_c}\right)^{2N}}, \quad (3.145)$$

with ω normalized to the range $[0, 2\pi]$ for sampled or discrete-time signals; in such a case, the equation is valid only for the range $[0, \pi]$, with the function in the range $[\pi, 2\pi]$ being a reflection of that over $[0, \pi]$. The cutoff frequency ω_c should be specified in the range $[0, \pi]$.

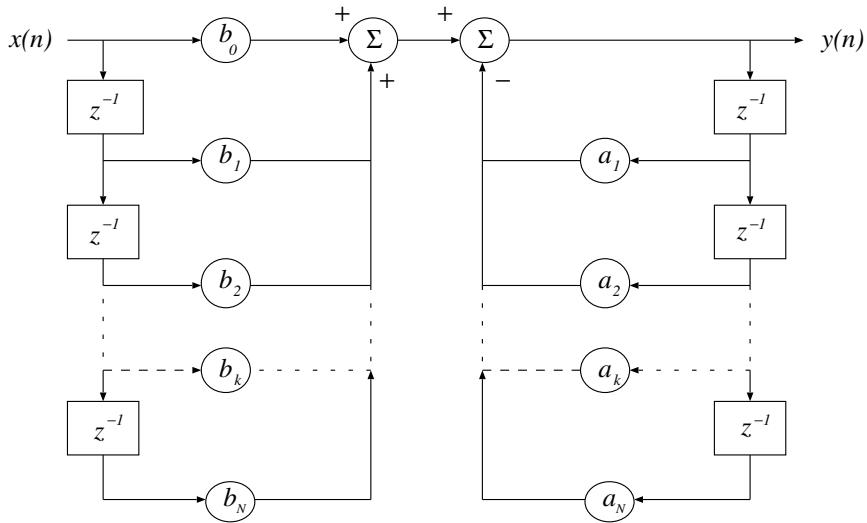


Figure 3.62 Signal-flow diagram of a direct realization of a generic IIR filter. This form uses $2N$ delays and $2N + 1$ multipliers for a filter of order N .

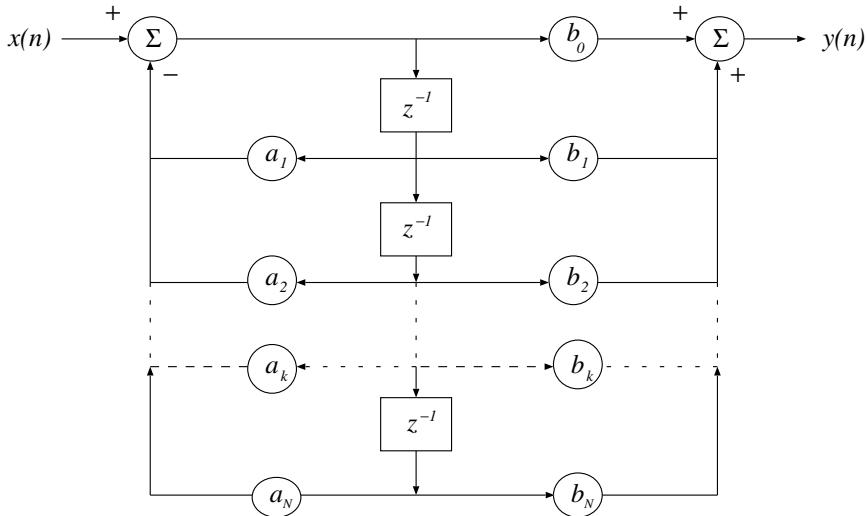


Figure 3.63 Signal-flow diagram of a realization of an IIR filter that uses only N delays and $(2N + 1)$ multipliers for a filter of order N .

If the DFT is used to compute the Fourier transforms of the signals being filtered, Equation 3.145 may be modified to

$$|H(k)|^2 = \frac{1}{1 + \left(\frac{k}{k_c}\right)^{2N}}, \quad (3.146)$$

where k is the index of the DFT array standing for discretized frequency. With K being the number of points in the DFT array, k_c is the array index corresponding to the cutoff frequency ω_c (that is, $k_c = \lceil K \frac{\omega_c}{\omega_s} \rceil$). The equation given above is valid for $k = 0, 1, 2, \dots, \frac{K}{2}$, with the second half over $(\frac{K}{2} + 1, K - 1)$ being a reflection of the first half [that is, $H(k) = H(K - k)$, $k =$

$\frac{K}{2} + 1, \dots, K - 1$. Note that the DFT includes two unique values: the DC component in $H(0)$ and the folding-frequency component in $H(\frac{K}{2})$; see Figure 3.38. The variable k in the filter equation could also be used to represent normalized frequency in the range $[0, 1]$, with unity standing for the sampling frequency, 0.5 standing for the maximum frequency present in the sampled signal (that is, the folding frequency), and k_c being specified in the range $[0, 0.5]$. (Note: MATLAB® normalizes one-half of the sampling frequency to unity; the maximum normalized frequency present in the sampled signal is then unity, that is, $f_m = 1$. MATLAB® and a few programming languages do not allow an array index to be zero: In such a case, the indices mentioned above must be incremented by 1.)

One could compute the DFT of the given signal, multiply the result by $|H(k)|$, and compute the inverse DFT to obtain the filtered signal. The advantage of this procedure is that no phase change is involved: The filter is a strictly magnitude-only transfer function. The time-domain implementation described earlier will include a phase response which may not be desired. However, time-domain implementation is often required in on-line, real-time signal processing applications.

Butterworth lowpass filter design example: In order to design a Butterworth lowpass filter, we need to specify two parameters: ω_c and N . The two parameters may be specified based on a knowledge of the characteristics of the filter as well as those of the signal and noise. It is also possible to specify the required minimum gain at a certain frequency in the passband and the required minimum attenuation at another frequency in the stopband. The two values may then be used with Equation 3.135 to obtain two equations in the two unknowns ω_c and N , which may be solved to derive the filter parameters [19].

Given the 3 dB cutoff frequency f_c and order N , the procedure to design a Butterworth lowpass filter is as follows:

1. Convert the specified 3 dB cutoff frequency f_c to radians in the normalized range $[0, 2\pi]$ as $\omega_c = \frac{f_c}{f_s} 2\pi$. Then, $T = 1$. Prewarp the cutoff frequency ω_c using Equation 3.141 and obtain Ω_c .
2. Derive the positions of the poles of the filter in the s -plane as given by Equation 3.137.
3. Form the transfer function $H_a(s)$ of the Butterworth lowpass filter in the Laplace domain using the poles in the left-half plane only as given by Equation 3.138.
4. Apply the bilinear transformation as per Equation 3.139 and obtain the transfer function of the filter $H(z)$ in the z -domain as in Equation 3.143.
5. Convert the filter to the series of coefficients b_k and a_k as in Equation 3.144.

Let us now design a Butterworth lowpass filter with $f_c = 40\text{ Hz}$, $f_s = 200\text{ Hz}$, and $N = 4$. We have $\omega_c = \frac{40}{200} 2\pi = 0.4\pi\text{ rad/s}$. The prewarped s -domain cutoff frequency is $\Omega_c = \frac{\pi}{T} \tan\left(\frac{\omega_c}{2}\right) = 1.453085\text{ rad/s}$.

The poles of $H_a(s)H_a(-s)$ are placed around a circle of radius 1.453085 rad/s with an angular separation of $\frac{\pi}{N} = \frac{\pi}{4}\text{ rad}$. The poles of interest are located at angles $\frac{5}{8}\pi$ and $\frac{7}{8}\pi$ and the corresponding conjugate positions. Figure 3.64 shows the positions of the poles of $H_a(s)H_a(-s)$ in the Laplace plane. The coordinates of the poles of interest are $(-0.556072 \pm j 1.342475)$ and $(-1.342475 \pm j 0.556072)$. The transfer function of the filter is

$$H_a(s) = \frac{4.458247}{(s^2 + 1.112143s + 2.111456)(s^2 + 2.684951s + 2.111456)}. \quad (3.147)$$

Applying the bilinear transformation, we get

$$H(z) = \frac{0.046583(1+z^{-1})^4}{(1-0.447765z^{-1}+0.460815z^{-2})(1-0.328976z^{-1}+0.064588z^{-2})}. \quad (3.148)$$

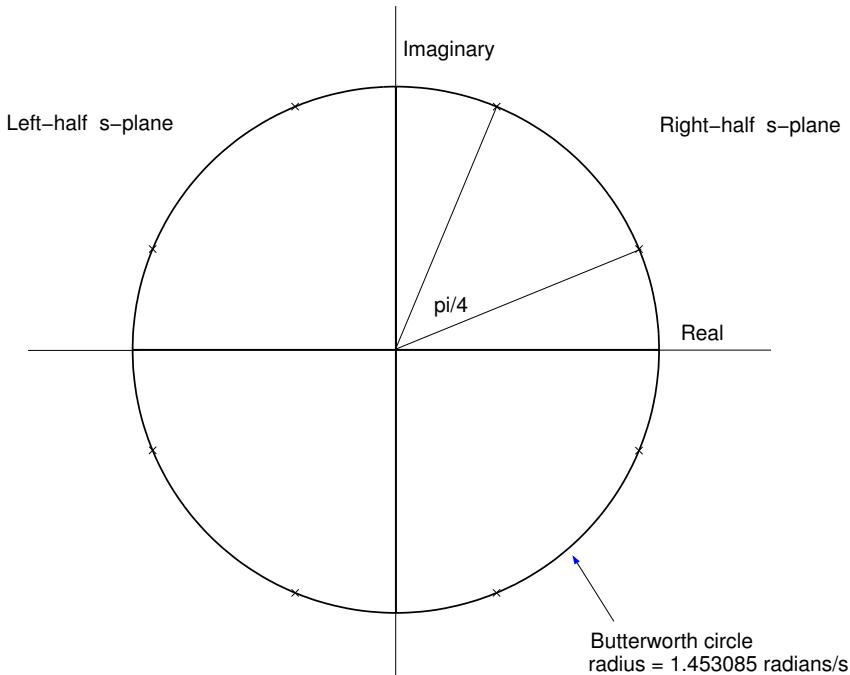


Figure 3.64 Pole positions on the Butterworth circle in the s -plane of the squared magnitude response of the Butterworth lowpass filter with $f_c = 40 \text{ Hz}$, $f_s = 200 \text{ Hz}$, and $N = 4$.

This filter has four poles at $(0.223882 \pm j 0.640852)$ and $(0.164488 \pm j 0.193730)$, and four zeros at $-1 + j 0$. The b_k coefficients of the filter as in Equation 3.144 are $\{0.0465829, 0.186332, 0.279497, 0.186332, 0.046583\}$, and the a_k coefficients are $\{1, -0.776740, 0.672706, -0.180517, 0.029763\}$. The pole-zero plot and the frequency response of the filter are given in Figures 3.65 and 3.66, respectively. The frequency response displays the expected monotonic decrease in gain and -3 dB power point or 0.707 gain at 40 Hz .

Figure 3.67 compares the magnitude responses of three Butterworth lowpass filters with $f_c = 40 \text{ Hz}$ and $f_s = 200 \text{ Hz}$, with the order increasing from $N = 4$ (dotted) to $N = 8$ (dashed) to $N = 12$ (solid). All three filters have their half-power points (gain = 0.707) at 40 Hz , but the transition band becomes sharper as the order N is increased.

The Butterworth design is popular because of its simplicity, a monotonically decreasing magnitude response, and a maximally flat magnitude response in the passband. Its main disadvantages are a slow (or wide) transition from the passband to the stopband, and a nonlinear phase response. The nonlinear phase may be corrected by passing the filter output again through the same filter but after a reversal in time [29]. This process, however, leads to a magnitude response that is the square of that provided by the initial filter design. The squaring effect may be compensated for in the initial design; however, the approach cannot be applied in real time. The elliptic filter design provides a sharp transition band at the expense of ripples in the passband and the stopband. The Bessel design provides a group delay that is maximally flat at DC, and a phase response that approximates a linear response. Details on the design of Bessel, Chebyshev, elliptic, and other filters may be found in other sources on filter design [18, 28–33].

Illustrations of application: The upper trace in Figure 3.68 illustrates a carotid pulse signal with high-frequency noise. The lower trace in the same figure shows the result of processing with a Butterworth lowpass filter ($f_c = 40 \text{ Hz}$, $f_s = 200 \text{ Hz}$, and $N = 12$). The high-frequency noise has been effectively reduced.

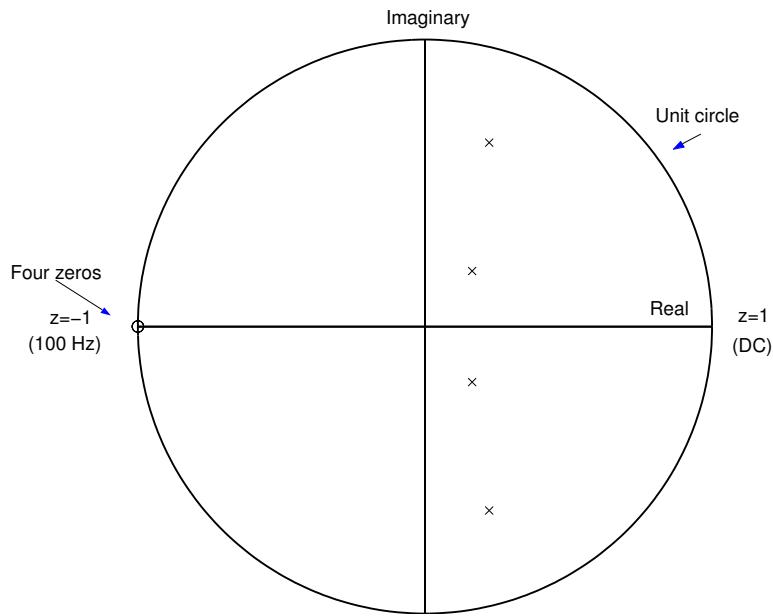


Figure 3.65 Positions of the poles and zeros with reference to the unit circle in the z -plane of the Butterworth lowpass filter with $f_c = 40$ Hz, $f_s = 200$ Hz, and $N = 4$.

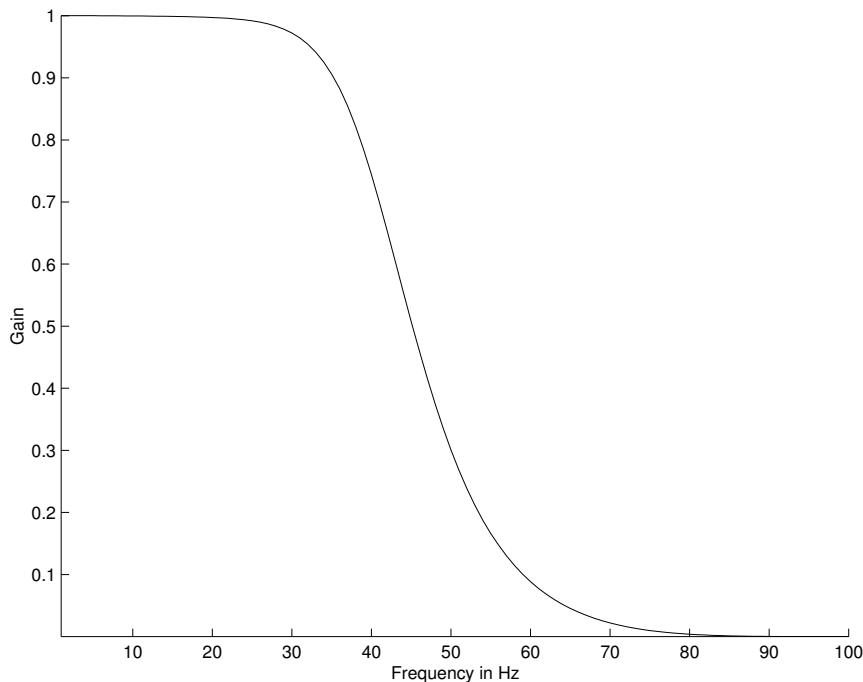


Figure 3.66 Magnitude response of the Butterworth lowpass filter with $f_c = 40$ Hz, $f_s = 200$ Hz, and $N = 4$.

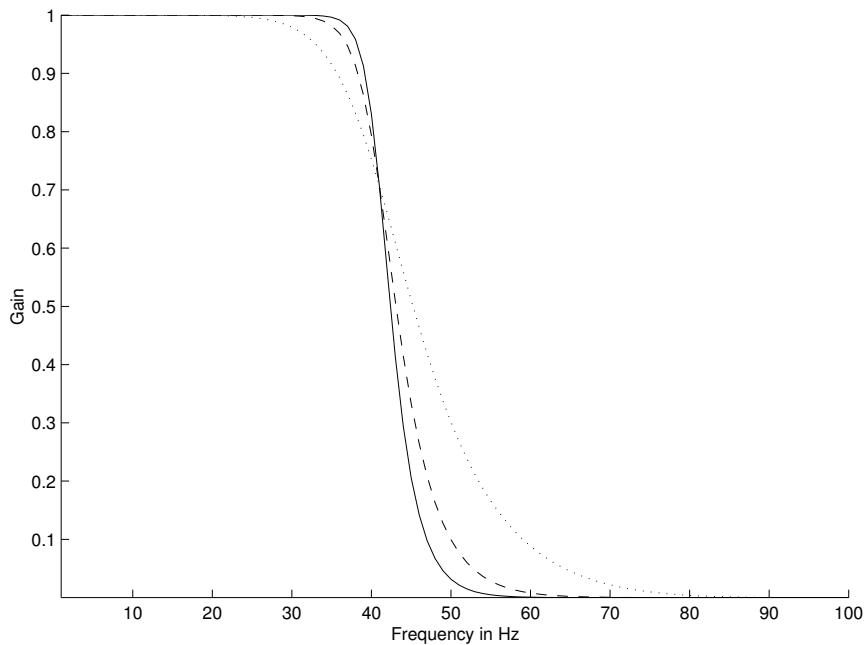


Figure 3.67 Magnitude responses of three Butterworth lowpass filters with $f_c = 40 \text{ Hz}$, $f_s = 200 \text{ Hz}$, and variable order: $N = 4$ (dotted), $N = 8$ (dashed), and $N = 12$ (solid).

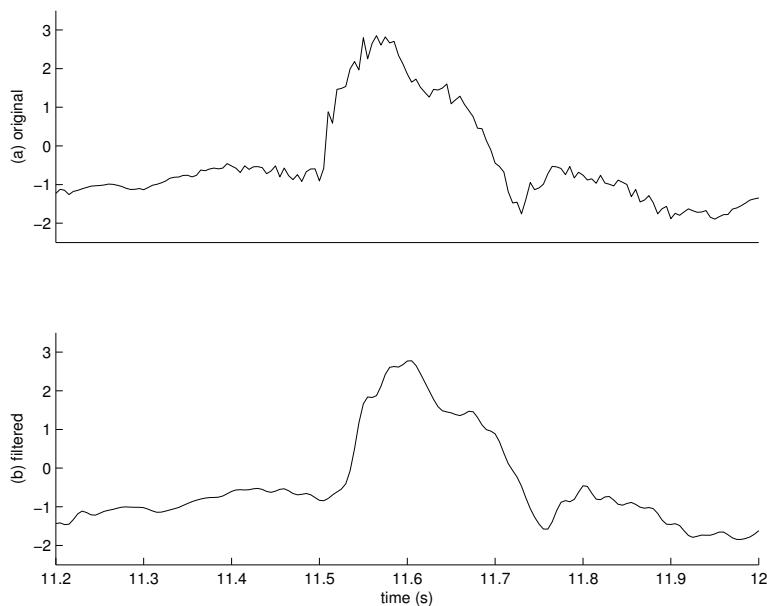


Figure 3.68 Upper trace: a carotid pulse signal with high-frequency noise. Lower trace: result of filtering using a Butterworth lowpass filter with $f_c = 40 \text{ Hz}$, $f_s = 200 \text{ Hz}$, and $N = 12$. The filtering operation was performed in the time domain using the MATLAB® *filter* command.

Figure 3.69 shows the result of filtering the noisy ECG signal shown in Figure 3.52 with an eighth-order Butterworth lowpass filter as in Equations 3.145 and 3.146; the cutoff frequency is 70 Hz and $f_s = 1,000$ Hz. The frequency response $|H(\omega)|$ of the filter is shown in Figure 3.70. It is evident that the high-frequency noise has been suppressed by the filter.

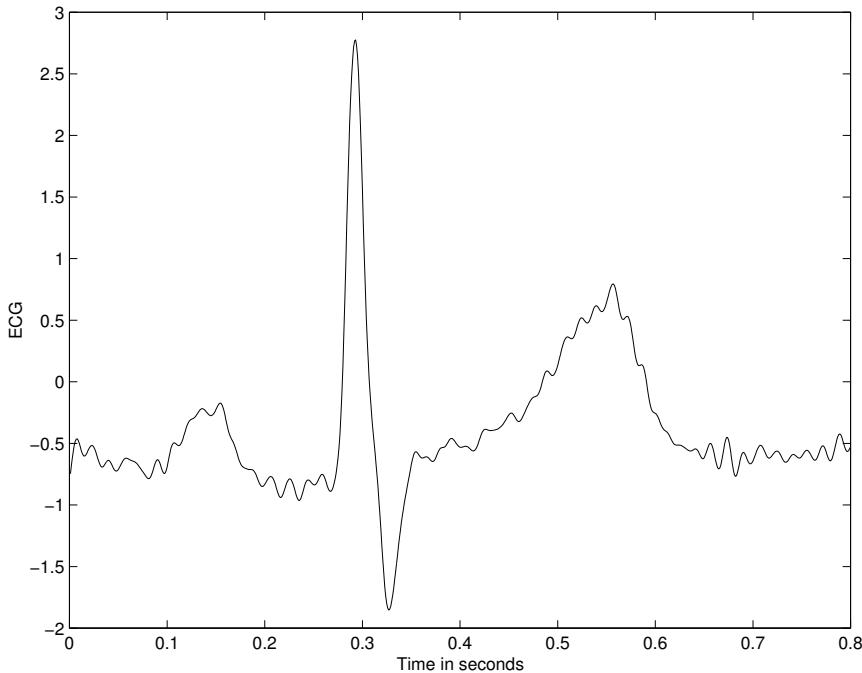


Figure 3.69 Result of frequency-domain filtering of the noisy ECG signal in Figure 3.52 with an eighth-order Butterworth lowpass filter with cutoff frequency = 70 Hz.

3.7.2 Removal of low-frequency noise: Butterworth highpass filters

Problem: Design a frequency-domain filter to remove low-frequency noise with minimal loss of signal components in the passband.

Solution: Highpass filters may be designed on their own, or obtained by transforming a normalized prototype lowpass filter [19, 28]. The latter approach is easier since lowpass filter prototypes with various characteristics are readily available, as are the transformations required to derive high-pass, bandpass, and bandstop filters [19, 28]. [MATLAB® provides Butterworth highpass filters with the command `butter(N, fc, 'high')`.]

As in the case of the Butterworth lowpass filter in Equation 3.146, the Butterworth highpass filter may be specified directly in the discrete-frequency domain as

$$|H(k)|^2 = \frac{1}{1 + \left(\frac{k_c}{k}\right)^{2N}}. \quad (3.149)$$

Illustration of application: Figure 3.6 shows a segment of an ECG signal with low-frequency noise appearing in the form of a wandering baseline (baseline drift). Figure 3.71 shows the result of filtering the signal with an eighth-order Butterworth highpass filter as in Equation 3.149 and a cutoff frequency of 2 Hz. The frequency response of the filter is shown in Figure 3.72. While the low-frequency artifact has been removed by the filter, it should be noted that the high-frequency noise present in the signal has not been affected.

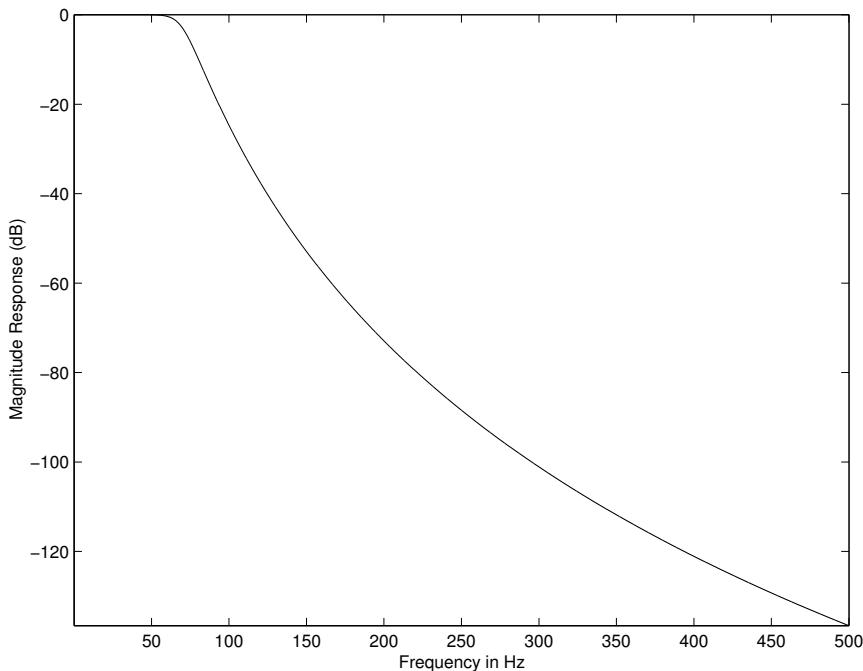


Figure 3.70 Frequency response of the eighth-order Butterworth lowpass filter with cutoff frequency $f_c = 70 \text{ Hz}$ and $f_s = 1,000 \text{ Hz}$.

Observe that the filtered result retains the characteristics of the QRS complex, unlike the case with the derivative-based time-domain filters (compare Figure 3.71 with Figures 3.56 and 3.57.) This advantage is due to the fact that the Butterworth highpass filter that was used has a gain of almost unity over the frequency range of $3 - 100 \text{ Hz}$; the derivative-based filters severely attenuate these components and hence distort the QRS complex. However, it should be observed that the Butterworth highpass filter has distorted the P and T waves to some extent. The result in Figure 3.71 compares well with that in Figure 3.61, obtained using the simpler IIR filter in Equation 3.134. (Compare the frequency responses in Figures 3.72, 3.54, 3.55, and 3.60.)

3.7.3 Removal of periodic artifacts: Notch and comb filters

Problem: Design a frequency-domain filter to remove periodic artifacts, such as power-line interference.

Solution: The simplest method to remove periodic artifacts is to compute the Fourier transform of the signal, delete the undesired component(s) from the spectrum, and then compute the inverse Fourier transform. The undesired components could be set to zero, or better, to the average level of the signal components over a few frequency samples around the component that is to be removed; the former method will remove the noise components as well as the signal components at the frequencies of concern, whereas the latter assumes that the signal spectrum is smooth in the affected regions.

Periodic interference may also be removed by notch filters with zeros on the unit circle in the z -domain at the specific frequencies to be rejected. If f_o is the interference frequency, the angles of the (complex conjugate) zeros required will be $\pm \frac{f_o}{f_s}(2\pi)$; the radius of the zeros will be unity. If harmonics are also present, multiple zeros will be required at $\pm \frac{n f_o}{f_s}(2\pi)$, n representing the orders of all of the harmonics present. The angles of the zeros are limited to the range $[-\pi, \pi]$. The filter

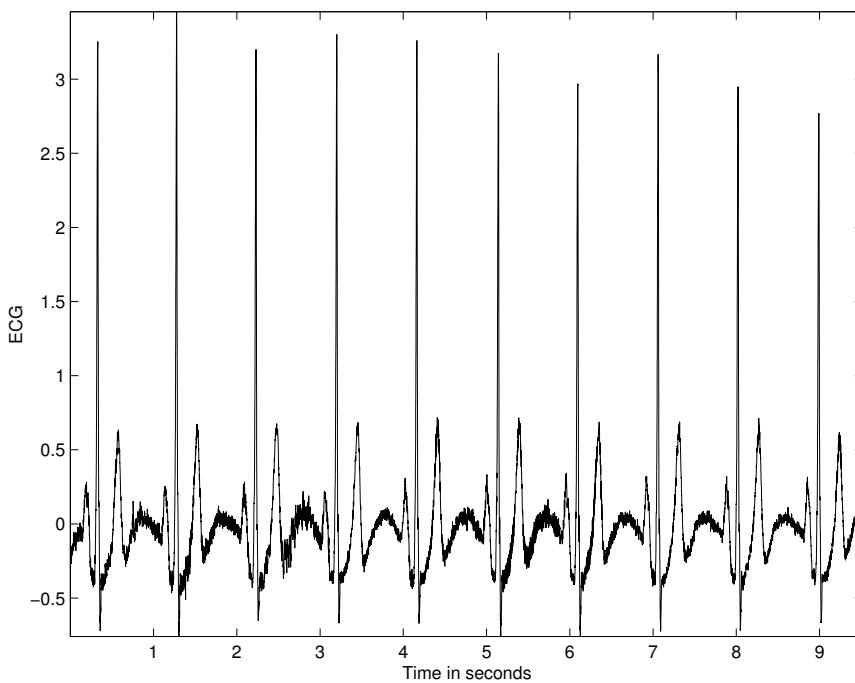


Figure 3.71 Result of frequency-domain filtering of the ECG signal with low-frequency noise in Figure 3.6 with an eighth-order Butterworth highpass filter with cutoff frequency = 2 Hz. (Compare with the results in Figures 3.56, 3.57, and 3.61.)

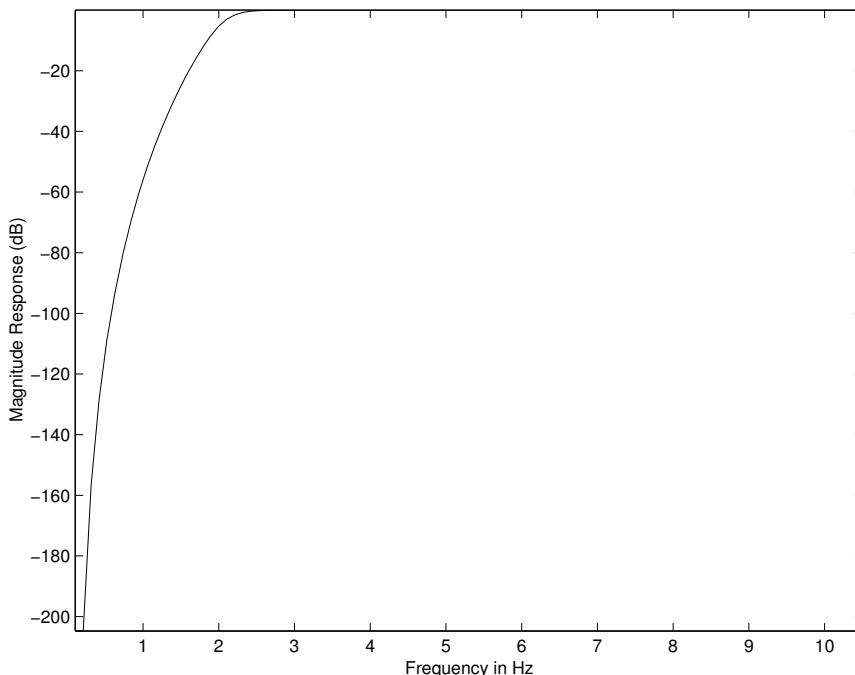


Figure 3.72 Frequency response of an eighth-order Butterworth highpass filter with cutoff frequency = 2 Hz. f_s = 1,000 Hz. The frequency response is shown on an expanded scale for the range 0 – 10 Hz only.

is then called a “comb” filter. In some situations, higher-order harmonics beyond $\frac{f_s}{2}$ may appear at aliased locations (see Figures 3.8 and 3.103); zeros may then be placed at such frequencies as well.

Notch filter design example: Consider a signal with power-line interference at $f_o = 60 \text{ Hz}$ and sampling rate of $f_s = 1,000 \text{ Hz}$ (see Figures 3.7 and 3.8). The notch filter is then required to have zeros at $\omega_o = \pm \frac{f_o}{f_s}(2\pi) = \pm 0.377 \text{ rad} = \pm 21.6^\circ$. The locations of the zeros are given by $\cos(\omega_o) \pm j \sin(\omega_o)$ or $z_1 = 0.92977 + j0.36812$ and $z_2 = 0.92977 - j0.36812$. The transfer function is

$$H(z) = (1 - z^{-1} z_1)(1 - z^{-1} z_2) = 1 - 1.85955z^{-1} + z^{-2}. \quad (3.150)$$

Substituting $z = 1$ in the expression above, we get the DC response or gain as $H(1) = 1 - 1.85955 + 1 = 0.14045$. Therefore, if the gain at DC is required to be unity, $H(z)$ as above should be divided by 0.14045.

Figure 3.73 shows a plot of the zeros of the notch filter as above in the z -plane. Figure 3.74 shows the magnitude and phase responses of the notch filter. Observe that the filter attenuates not only the 60 Hz component but also a band of frequencies around 60 Hz . The sharpness of the notch may be improved by placing a few poles near or symmetrically around the zeros and inside the unit circle [1, 2, 27]. Note also that the gain of the filter is at its maximum at $f_s/2$; additional lowpass filtering in the case of application to ECG signals could be used to reduce the gain at frequencies beyond about 80 Hz .

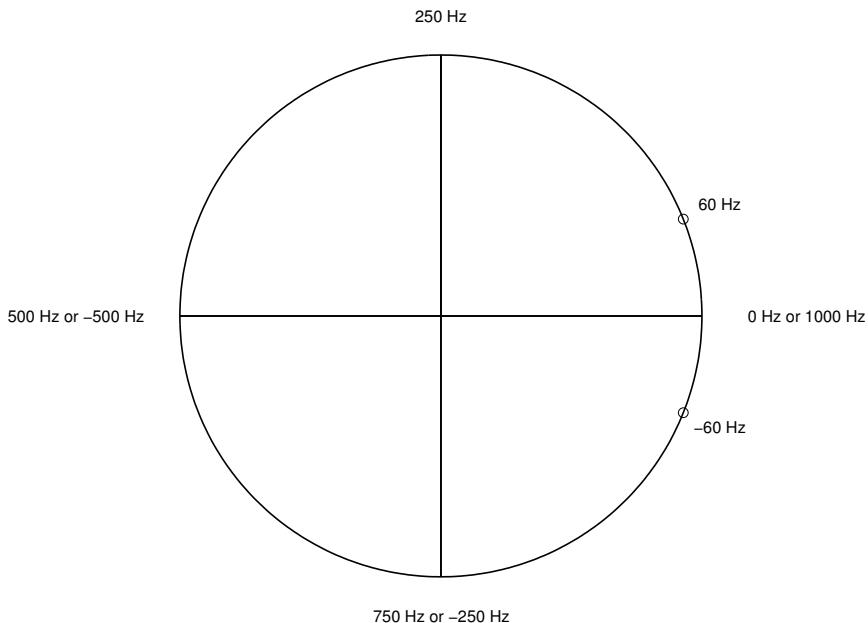


Figure 3.73 Zeros of the notch filter on the unit circle in the z -domain to remove 60 Hz interference, the sampling frequency being 1,000 Hz .

Comb filter design example: Let us consider the presence of a periodic artifact with the fundamental frequency of 60 Hz and odd harmonics at 180 Hz , 300 Hz , and 420 Hz ; see Figures 3.7 and 3.8. Let $f_s = 1,000 \text{ Hz}$, and assume the absence of any aliasing error. Zeros are then desired at 60 Hz , 180 Hz , 300 Hz , and 420 Hz , which translate to $\pm 21.6^\circ$, $\pm 64.8^\circ$, $\pm 108^\circ$, and $\pm 151.2^\circ$, with 360° corresponding to 1,000 Hz . The coordinates of the zeros are $0.92977 \pm j0.36812$, $0.42578 \pm j0.90483$, $-0.30902 \pm j0.95106$, and $-0.87631 \pm j0.48175$. The transfer function of

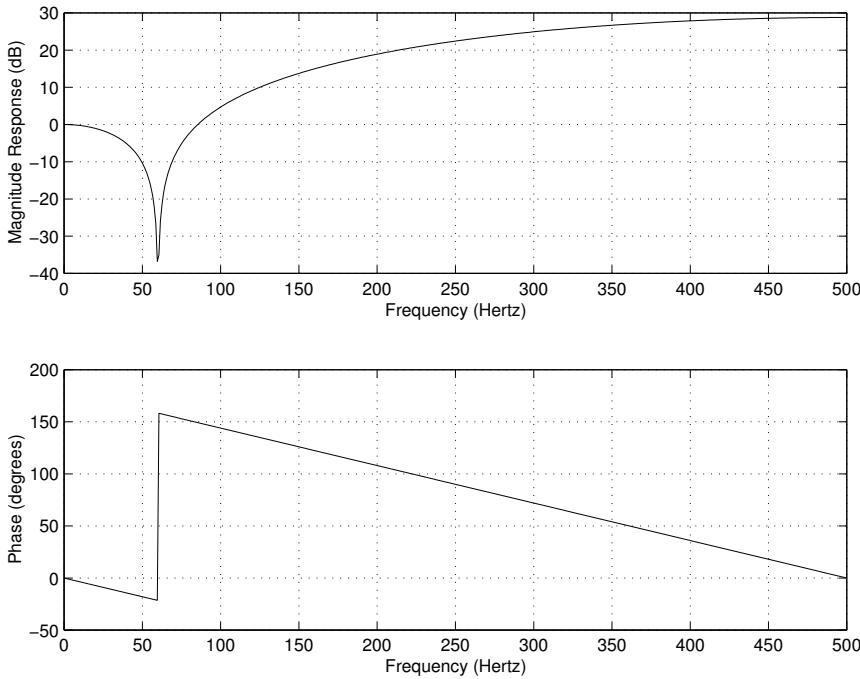


Figure 3.74 Magnitude and phase responses of the 60 Hz notch filter with zeros as shown in Figure 3.73. $f_s = 1,000$ Hz. Observe the 180° change in phase at 60 Hz due to the presence of the zero in the z -plane at the same frequency.

the filter is

$$\begin{aligned} H(z) &= G (1 - 1.85955z^{-1} + z^{-2})(1 - 0.85156z^{-1} + z^{-2}) \\ &\quad \times (1 + 0.61803z^{-1} + z^{-2})(1 + 1.75261z^{-1} + z^{-2}), \end{aligned} \quad (3.151)$$

where G is the desired gain or scaling factor. With G computed so as to set the gain at DC to be unity, the filter transfer function becomes

$$\begin{aligned} H(z) &= 0.6310 - 0.2149z^{-1} + 0.1512z^{-2} - 0.1288z^{-3} + 0.1227z^{-4} \\ &\quad - 0.1288z^{-5} + 0.1512z^{-6} - 0.2149z^{-7} + 0.6310z^{-8}. \end{aligned} \quad (3.152)$$

A plot of the locations of the zeros in the z -plane is shown in Figure 3.75. The frequency response of the comb filter is shown in Figure 3.76. Observe the low gain at not only the notch frequencies but also in the adjacent regions.

Illustrations of application: Figure 3.77 shows an ECG signal with power-line interference at $f_o = 60$ Hz. Figure 3.78 shows the result of applying the notch filter in Equation 3.150 to the signal. The 60 Hz interference has been effectively removed, with no perceptible distortion of the ECG waveform.

Figure 3.79 shows an ECG signal with noise including power-line interference at 60 Hz and $f_s = 200$ Hz. A notch filter was designed with two zeros at ± 60 Hz with radius equal to unity. The coordinates of the zeros in the z -plane are $-0.3090 \pm j 0.9511$. The frequency response of the notch filter is shown in Figure 3.80 (dashed line). The result of filtering the noisy signal in Figure 3.79 is shown in Figure 3.81, which indicates effective removal of the interference.

In order to improve the notch filter, specifically, to make the notch in the frequency response narrower or sharper, four poles were incorporated in the filter. The frequencies of the poles were set

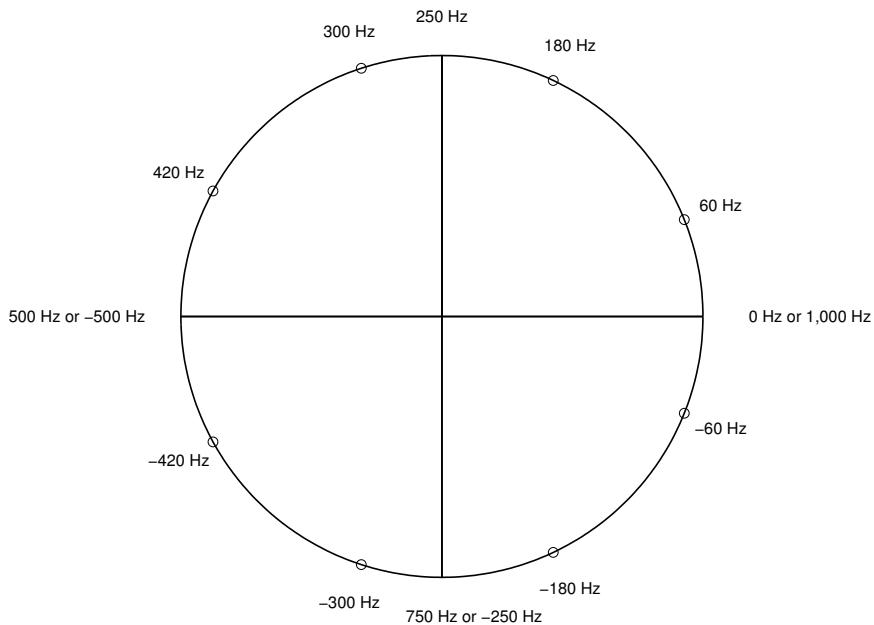


Figure 3.75 Zeros of the comb filter on the unit circle in the z -domain to remove 60 Hz interference with odd harmonics; the sampling frequency is 1,000 Hz .

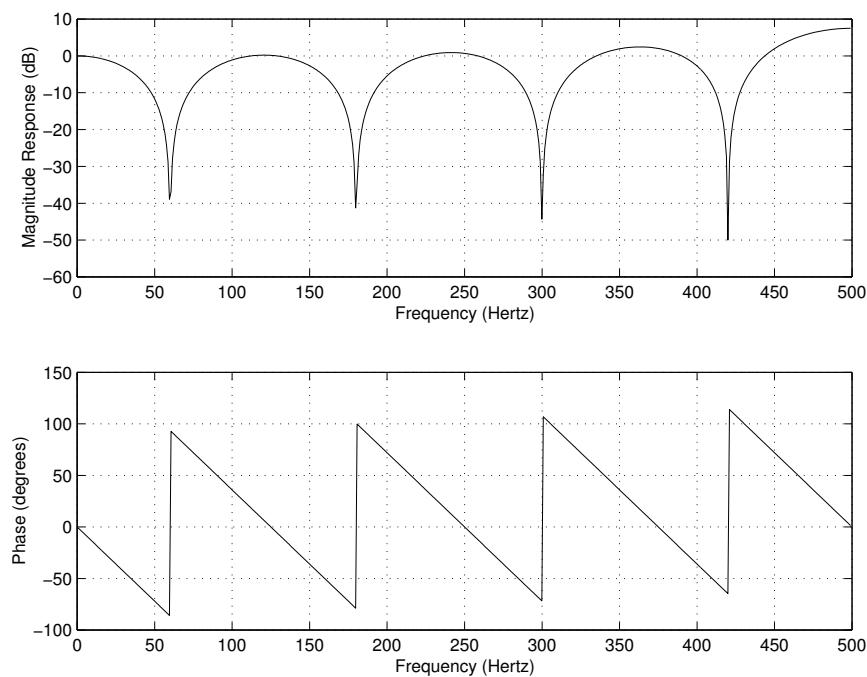


Figure 3.76 Magnitude and phase responses of the comb filter with zeros as shown in Figure 3.75.

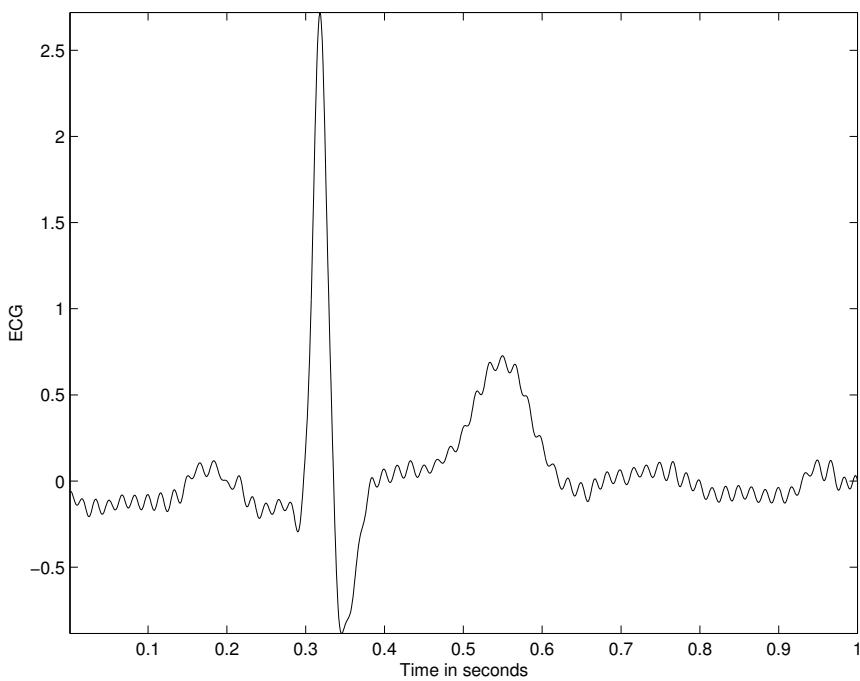


Figure 3.77 ECG signal with 60 Hz interference.

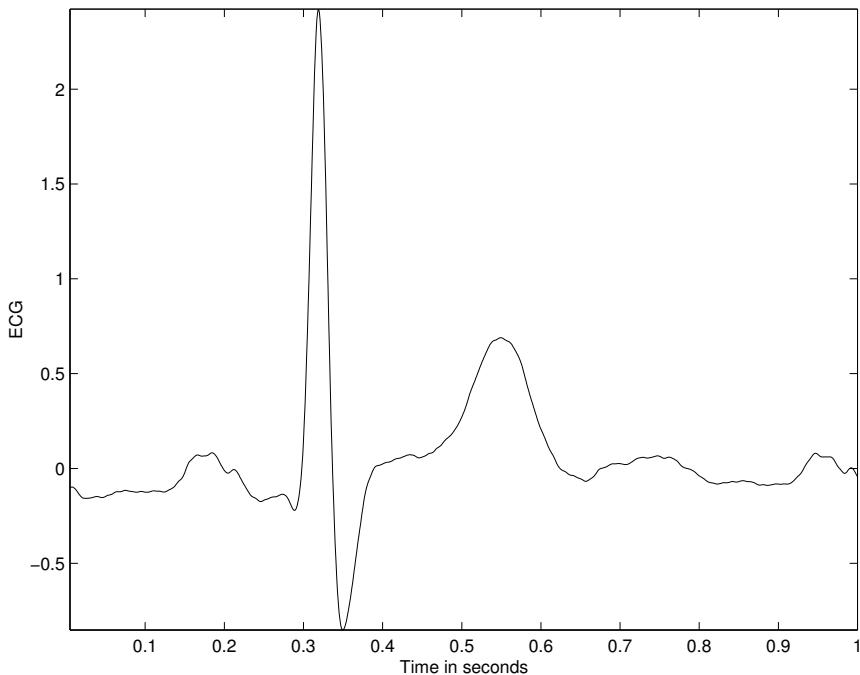


Figure 3.78 The ECG signal in Figure 3.77 after filtering with the $60 - \text{Hz}$ notch filter shown in Figures 3.73 and 3.74.

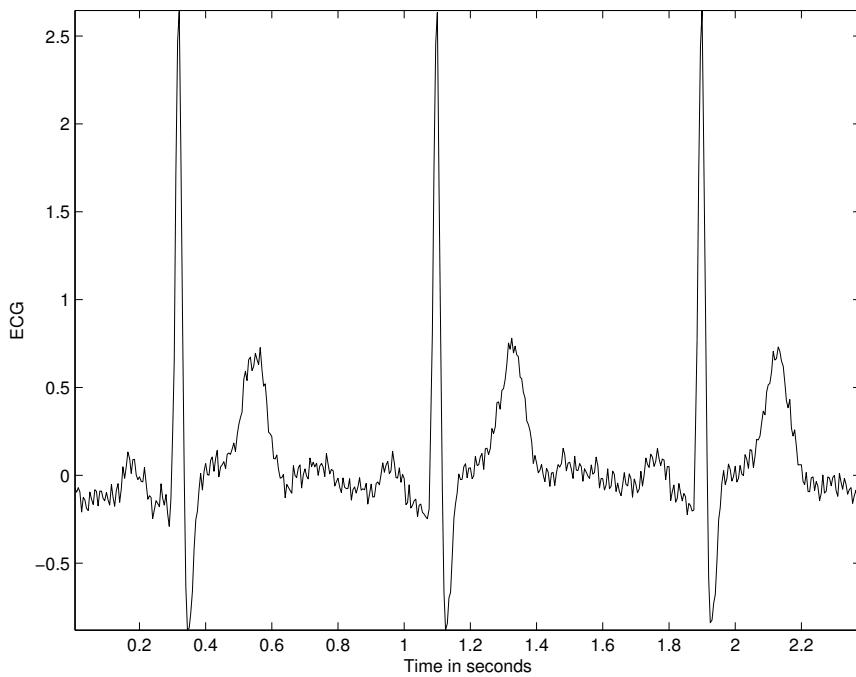


Figure 3.79 An ECG signal with power-line interference at 60 Hz .

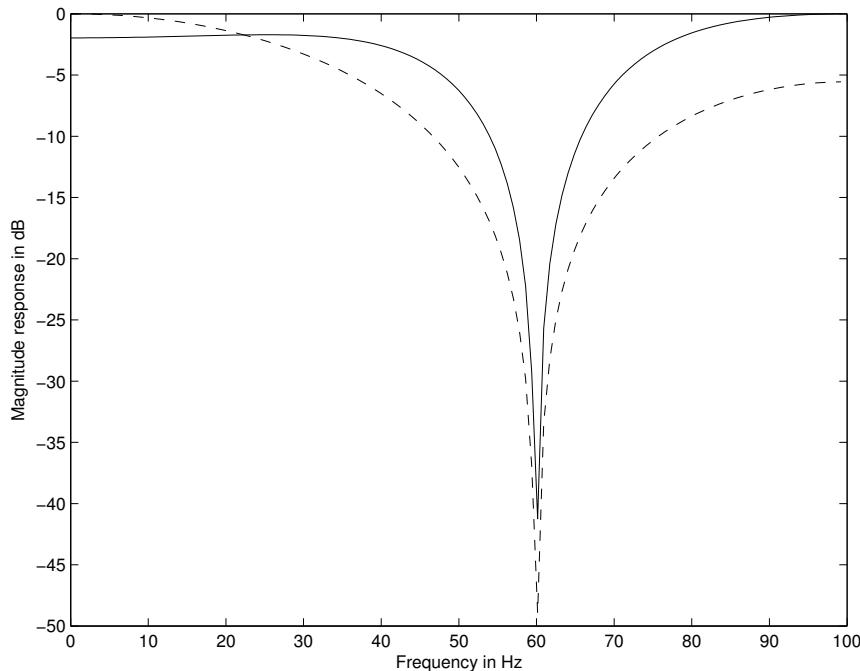


Figure 3.80 Dashed line: frequency response (magnitude, in dB) of a notch filter with two zeros at $\pm 60\text{ Hz}$; $f_s = 200\text{ Hz}$. Solid line: response of the filter with the inclusion of four poles; see the text for details.

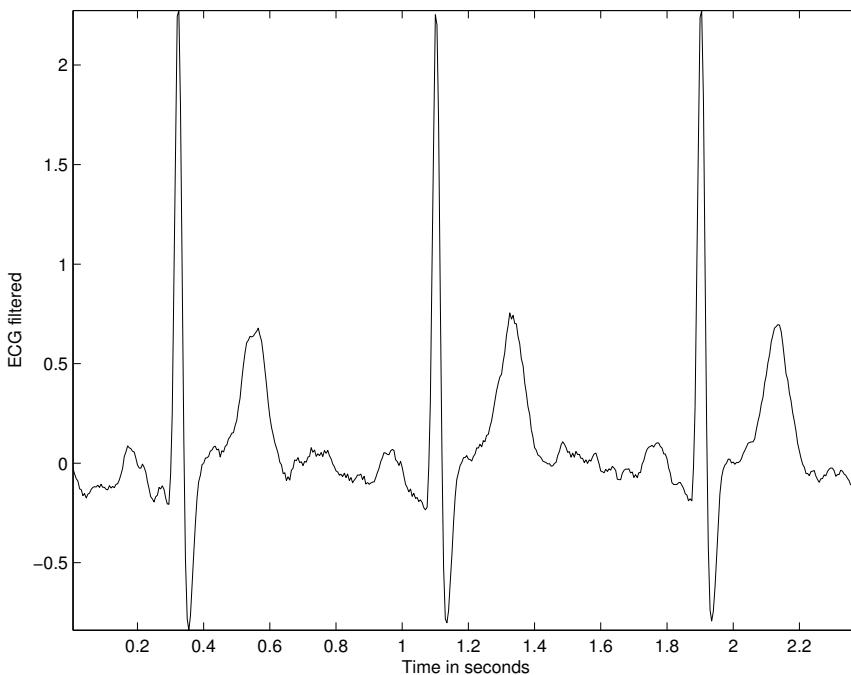


Figure 3.81 The result of filtering the noisy signal in Figure 3.79 using the notch filter with two zeros at $\pm 60\text{ Hz}$.

at $\pm 59.95\text{ Hz}$ and $\pm 60.05\text{ Hz}$; the radii of the poles were set at 0.4 to place them well within the unit circle in the z -plane. The coordinates of the poles in the z -plane are $-0.1230 \pm j 0.3806$ and $-0.1242 \pm j 0.3802$. The frequency response of the modified notch filter is shown in Figure 3.80 (solid line). It is evident that the notch is sharper or narrower for the modified filter as compared to the response of the filter with only two zeros; however, the gain increases for higher frequencies. As a result, the output of the filter, shown in Figure 3.82, has more noise than the previous result in Figure 3.81.

An illustration of the application of the comb filter is provided in Section 3.13. See Chen et al. [35] and Luo and Johnson [36] for discussions on recent developments in filtering ECG signals.

3.8 Order-statistic Filters

The class of filters based on order statistics includes several nonlinear filters that are useful in filtering different types of noise in signals [17, 37, 38]. The first step in order-statistic filtering is to arrange in rank order, usually from the minimum to the maximum, the values of the signal in a moving window positioned at the current sample being processed. Then, the i^{th} entry in the list is the output of the i^{th} order-statistic filter. A few commonly used order-statistic filters are defined in the following list:

- *Min filter*: the first entry in the rank-ordered list, useful in removing high-valued impulsive noise.
- *Max filter*: the last entry in the rank-ordered list, useful in removing low-valued impulsive noise.

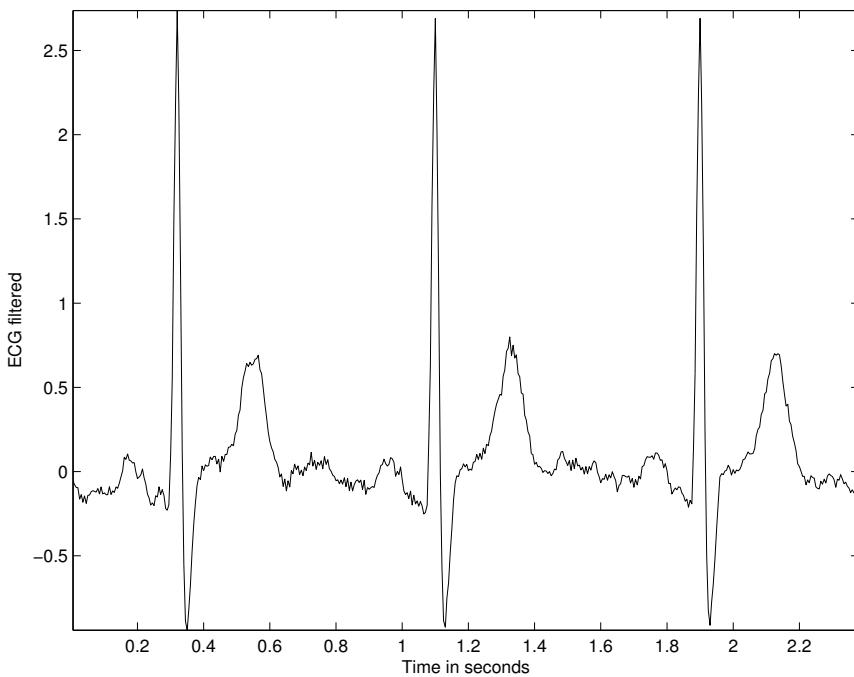


Figure 3.82 The result of filtering the noisy signal in Figure 3.79 using the notch filter with two zeros and four poles. Compare with the result in Figure 3.81.

- *Min/Max filter*: sequential application of the Min and Max filters, useful in removing impulsive noise of both of the types mentioned above.
- *Median filter*: the entry in the middle of the list. The median filter is the most popular and commonly used filter among the order-statistic filters.
- *α -trimmed mean filter*: the mean of a reduced or trimmed list, where the first $\alpha \times 100\%$ and the last $\alpha \times 100\%$ of the entries in the original list are removed, with $0 \leq \alpha < 0.5$. Outliers, which are samples with values substantially different from the rest of the samples in the list, are rejected by the trimming process. A value close to but less than 0.5 for α leads to the rejection of the entire list except the median or a few values close to the median; the output is then close to or equal to that of the median filter. The mean of the trimmed list provides a compromise between the generic mean and median filters.
- *L-filters*: a weighted combination of all of the elements in the rank-ordered list. Appropriate weights can provide outputs equivalent to those of all of the filters listed above, and facilitate the design of several nonlinear filters based on order statistics.

The methods described above may be used to realize several types of linear, nonlinear, nonstationary, and adaptive filters for the removal of various kinds of noise. It should be noted that nonlinear filters are not amenable to analysis using the Fourier transform.

Illustrations of application: Figure 3.83 (a) shows a synthesized test signal with a rectangular pulse; part (b) of the same figure shows the test signal contaminated with simulated impulsive (shot) noise. The results of filtering the noisy signal using the mean and median with filter length $M = 3$ samples are shown in parts (c) and (d) of Figure 3.83, respectively. The mean filter has blurred the edges of the pulse, which is undesirable; it has also created variations or details within the pulse that are artifacts. Furthermore, the result of the mean filter has retained the noise spikes present near

the beginning and end of the signal, albeit reduced in amplitude. The median filter has completely eliminated the noise without distorting the pulse.

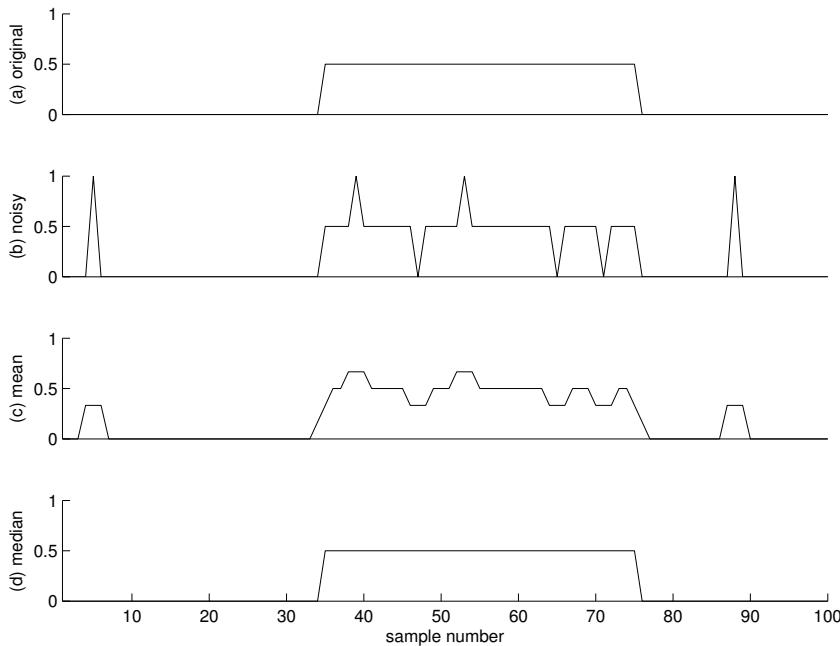


Figure 3.83 (a) A synthesized test signal with a rectangular pulse. (b) Test signal contaminated with simulated impulsive or shot noise. Result of filtering the noisy signal using (c) the mean and (d) the median with a sliding window having $M = 3$ samples.

Figure 3.84 (a) shows another synthesized test signal with two rectangular pulses, the first one having a width of only two samples. Part (b) of the same figure shows the test signal degraded with uniformly distributed noise. The results of filtering the noisy signal using the mean and median with filter length $M = 5$ samples are shown in parts (c) and (d) of Figure 3.84, respectively. The mean filter has reduced the noise, but has also blurred the edges of the pulses; furthermore, the amplitude of the first short pulse has been reduced. While the median filter has removed the noise to some extent without distorting the edges of the long pulse, it has obliterated the short pulse. The examples demonstrate the need to choose an appropriate type and length of the filter in accordance with the nature and strength of the noise as well as those of the signal.

Figure 3.85 shows a noisy ECG signal ($f_s = 1,000 \text{ Hz}$) and the results of filtering with the mean and median using a causal sliding window including nine samples. Both of the filters have substantially reduced the noise without causing any noticeable distortion.

3.9 The Wiener Filter

The filters described in the preceding sections can take into account only limited information about the temporal or spectral characteristics of the signal and noise processes. They are often labeled as *ad hoc* filters: One may have to try several filter parameters and settle upon the filter that appears to provide a usable result. The output is not guaranteed to be the best achievable result: It is not optimized in any sense.

Problem: Design an optimal filter to remove noise from a signal, given that the signal and noise processes are statistically independent, stationary, and random processes. You may assume that the

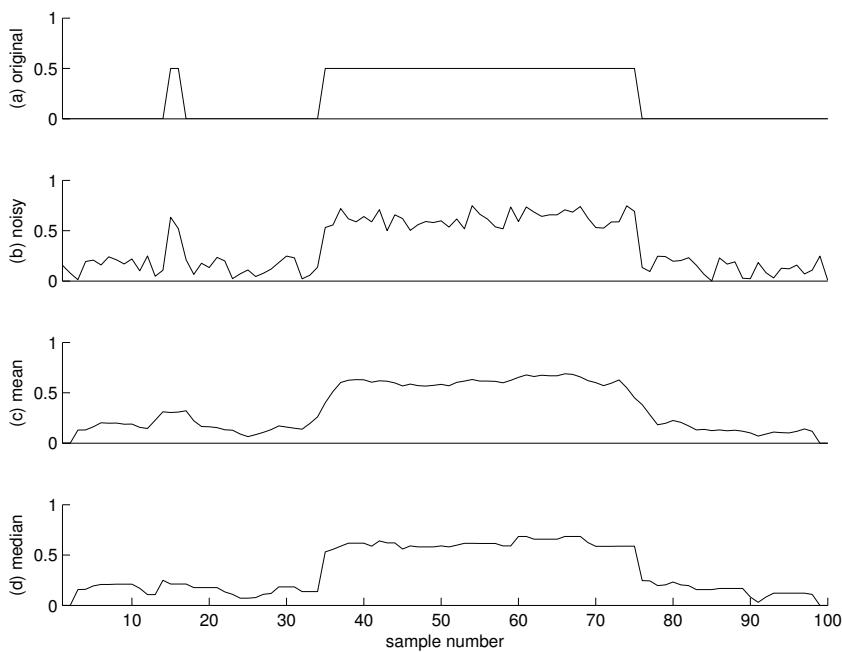


Figure 3.84 (a) A synthesized test signal with two rectangular pulses. (b) Degraded signal with uniformly distributed noise. Result of filtering the degraded signal using (c) the mean and (d) the median operation with a sliding window having $M = 5$ samples.

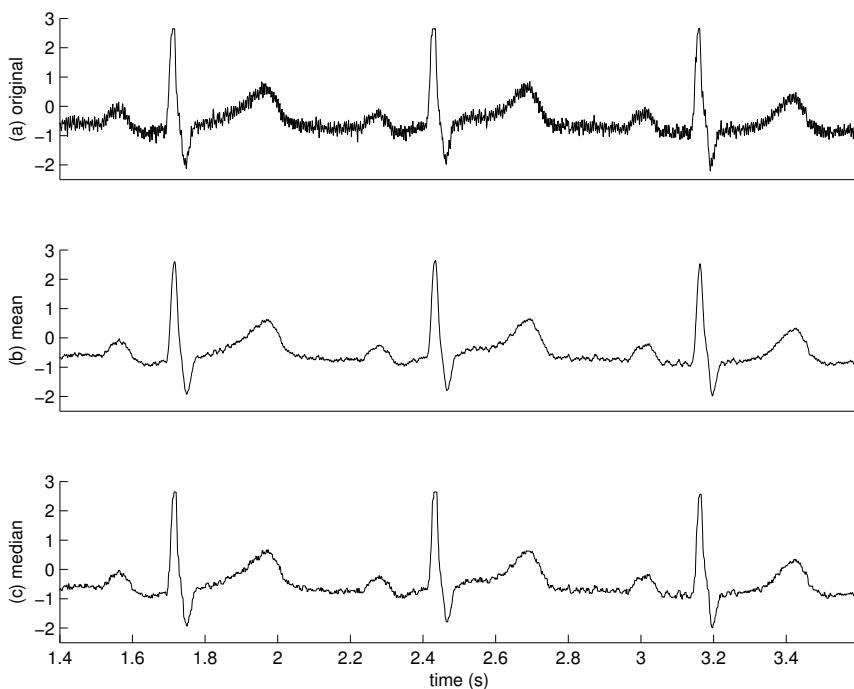


Figure 3.85 (a) An ECG signal with noise. Result of filtering the ECG signal using (b) the mean and (c) the median operation with a sliding window having $M = 9$ samples.

“desired” or ideal characteristics of the uncorrupted signal are known. The noise characteristics may also be assumed to be known.

Solution: Wiener filter theory [39] provides for *optimal* filtering by taking into account the statistical characteristics of the signal and noise processes. The filter parameters are *optimized* with reference to a *performance criterion*. The output is guaranteed to be the best achievable result under the conditions imposed and the information provided. The Wiener filter is a powerful tool that changed traditional approaches to signal processing.

Considering the application of filtering a biomedical signal to remove noise, let us limit ourselves to a single-input, single-output, FIR filter with real input signal values and real coefficients. Figure 3.86 shows the general signal-flow diagram of a transversal filter with coefficients or tap weights $w_i, i = 0, 1, 2, \dots, M - 1$, input $x(n)$, and output $\tilde{d}(n)$ [40]. The output is usually considered to be an estimate of some “desired” signal $d(n)$ that represents the ideal, uncorrupted signal, and is, therefore, indicated as $\tilde{d}(n)$. If we assume, for the moment, that the desired signal is available, we could compute the *estimation error* between the output and the desired signal as

$$e(n) = d(n) - \tilde{d}(n). \quad (3.153)$$

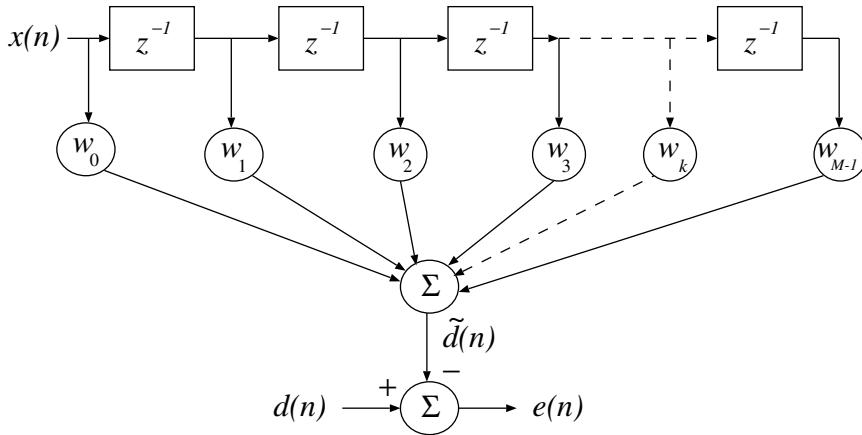


Figure 3.86 Signal-flow diagram of the Wiener filter.

Since $\tilde{d}(n)$ is the output of a linear FIR filter, it can be expressed as the convolution of the input $x(n)$ with the tap-weight sequence w_i (which is also the impulse response of the filter) as

$$\tilde{d}(n) = \sum_{k=0}^{M-1} w_k x(n-k). \quad (3.154)$$

For easier handling of the optimization procedures, the tap-weight sequence may be written as an $M \times 1$ *tap-weight vector*

$$\mathbf{w} = [w_0, w_1, w_2, \dots, w_{M-1}]^T, \quad (3.155)$$

where the boldfaced character \mathbf{w} represents a vector, and the superscript T indicates vector transposition. As the tap weights are combined with M values of the input in the convolution expression, we could also write the M input values as an $M \times 1$ vector:

$$\mathbf{x}(n) = [x(n), x(n-1), \dots, x(n-M+1)]^T. \quad (3.156)$$

Note that the vector $\mathbf{x}(n)$ varies with time: At a given instant n , the vector contains the current input sample $x(n)$ and the preceding $(M-1)$ input samples from $x(n-1)$ to $x(n-M+1)$. The

convolution expression in Equation 3.154 may now be written in a simpler form as the inner or dot product of the vectors \mathbf{w} and $\mathbf{x}(n)$:

$$\tilde{d}(n) = \mathbf{w}^T \mathbf{x}(n) = \mathbf{x}^T(n) \mathbf{w} = \langle \mathbf{x}, \mathbf{w} \rangle. \quad (3.157)$$

The estimation error is then given by

$$e(n) = d(n) - \mathbf{w}^T \mathbf{x}(n). \quad (3.158)$$

Wiener filter theory estimates the tap-weight sequence that minimizes the *MS* value of the estimation error; the output could then be called the *minimum mean-squared error* (MMSE) estimate of the desired response, the filter then being an *optimal filter*. The mean-squared error (MSE) is used to define the performance criterion as

$$\begin{aligned} J(\mathbf{w}) &= E[e^2(n)] \\ &= E[\{d(n) - \mathbf{w}^T \mathbf{x}(n)\}\{d(n) - \mathbf{x}^T(n) \mathbf{w}\}] \\ &= E[d^2(n)] - \mathbf{w}^T E[\mathbf{x}(n)d(n)] - E[d(n)\mathbf{x}^T(n)]\mathbf{w} + \mathbf{w}^T E[\mathbf{x}(n)\mathbf{x}^T(n)]\mathbf{w}. \end{aligned} \quad (3.159)$$

Note that the statistical expectation operator $E[\cdot]$ is not applicable to \mathbf{w} as it is not a random variable.

Under the assumption that the input vector $\mathbf{x}(n)$ and the desired response $d(n)$ are jointly stationary, the statistical expectation expressions in the equation given above have the following interpretations [40]:

- $E[d^2(n)]$ is the variance of $d(n)$, written as σ_d^2 , with the further assumption that the mean of $d(n)$ is zero.
- $E[\mathbf{x}(n)d(n)]$ is the cross-correlation between the input vector $\mathbf{x}(n)$ and the desired response $d(n)$, which is an $M \times 1$ vector:

$$\Theta = E[\mathbf{x}(n)d(n)]. \quad (3.160)$$

Note that $\Theta = [\theta(0), \theta(-1), \dots, \theta(1-M)]^T$, where

$$\theta(-k) = E[x(n-k)d(n)], \quad k = 0, 1, 2, \dots, M-1. \quad (3.161)$$

- $E[d(n)\mathbf{x}^T(n)]$ is the transpose of $E[\mathbf{x}(n)d(n)]$; therefore,

$$\Theta^T = E[d(n)\mathbf{x}^T(n)]. \quad (3.162)$$

- $E[\mathbf{x}(n)\mathbf{x}^T(n)]$ represents the autocorrelation of the input vector $\mathbf{x}(n)$ computed as the outer product of the vector with itself, written as

$$\Phi = E[\mathbf{x}(n)\mathbf{x}^T(n)] \quad (3.163)$$

or in its full $M \times M$ matrix form as

$$\Phi = \begin{bmatrix} \phi(0) & \phi(1) & \cdots & \phi(M-1) \\ \phi(-1) & \phi(0) & \cdots & \phi(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(-M+1) & \phi(-M+2) & \cdots & \phi(0) \end{bmatrix}, \quad (3.164)$$

with the element in row k and column i given by

$$\phi(i - k) = E[x(n - k)x(n - i)], \quad (3.165)$$

with the property that $\phi(i - k) = \phi(k - i)$. (Note: $\phi = \phi_{xx}$.) With the assumption of wide-sense stationarity, the $M \times M$ matrix Φ is completely specified by M values of the autocorrelation $\phi(0), \phi(1), \dots, \phi(M - 1)$ for lags $0, 1, \dots, M - 1$.

With the interpretations as listed above, the MSE expression in Equation 3.160 is simplified to

$$J(\mathbf{w}) = \sigma_d^2 - \mathbf{w}^T \Theta - \Theta^T \mathbf{w} + \mathbf{w}^T \Phi \mathbf{w}. \quad (3.166)$$

This expression indicates that the MSE is a second-order function of the tap-weight vector \mathbf{w} . To determine the optimal tap-weight vector, denoted by \mathbf{w}_o , we could differentiate $J(\mathbf{w})$ with respect to \mathbf{w} , set it to zero, and solve the resulting equation. To perform this differentiation, we should note the following derivatives:

$$\begin{aligned} \frac{d}{d\mathbf{w}}(\Theta^T \mathbf{w}) &= \Theta, \\ \frac{d}{d\mathbf{w}}(\mathbf{w}^T \Theta) &= \Theta, \\ \frac{d}{d\mathbf{w}}(\mathbf{w}^T \Phi \mathbf{w}) &= 2\Phi \mathbf{w}. \end{aligned}$$

Now, we obtain the derivative of $J(\mathbf{w})$ with respect to \mathbf{w} as

$$\frac{dJ(\mathbf{w})}{d\mathbf{w}} = -2\Theta + 2\Phi \mathbf{w}. \quad (3.167)$$

Setting this expression to zero, we obtain the condition for the optimal filter as

$$\Phi \mathbf{w}_o = \Theta. \quad (3.168)$$

This equation is known as the *Wiener–Hopf* equation. It is also known as the *normal equation* as it can be shown that [40], for the optimal filter, each element of the input vector $\mathbf{x}(n)$ and the estimation error $e(n)$ are mutually orthogonal, and, furthermore, that the filter output $\tilde{d}(n)$ and the error $e(n)$ are mutually orthogonal (that is, the expectation of their products is zero). The optimal filter is obtained as

$$\mathbf{w}_o = \Phi^{-1} \Theta. \quad (3.169)$$

In expanded form, we have the Wiener–Hopf equation as

$$\begin{bmatrix} \phi(0) & \phi(1) & \cdots & \phi(M - 1) \\ \phi(-1) & \phi(0) & \cdots & \phi(M - 2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(-M + 1) & \phi(-M + 2) & \cdots & \phi(0) \end{bmatrix} \begin{bmatrix} w_{o0} \\ w_{o1} \\ \vdots \\ w_{o(M-1)} \end{bmatrix} = \begin{bmatrix} \theta(0) \\ \theta(-1) \\ \vdots \\ \theta(1 - M) \end{bmatrix} \quad (3.170)$$

or as

$$\sum_{i=0}^{M-1} w_{oi} \phi(i - k) = \theta(-k), \quad k = 0, 1, 2, \dots, M - 1. \quad (3.171)$$

The minimum MSE is given by

$$J_{\min} = \sigma_d^2 - \Theta^T \Phi^{-1} \Theta. \quad (3.172)$$

Given the condition that the signals involved are stationary, we have $\phi(i - k) = \phi(k - i)$ and $\theta(-k) = \theta(k)$; that is, the functions ϕ and θ are even-symmetric. Then, we may write Equation 3.171 as

$$\sum_{i=0}^{M-1} w_{oi} \phi(k - i) = \theta(k), \quad k = 0, 1, 2, \dots, M - 1. \quad (3.173)$$

Thus, we have the convolution relationship

$$w_{ok} * \phi(k) = \theta(k). \quad (3.174)$$

Applying the Fourier transform to the equation given above, we get

$$W(\omega)S_{xx}(\omega) = S_{xd}(\omega), \quad (3.175)$$

which may be modified to obtain the Wiener filter frequency response $W(\omega)$ as

$$W(\omega) = \frac{S_{xd}(\omega)}{S_{xx}(\omega)}, \quad (3.176)$$

where $S_{xx}(\omega)$ is the PSD of the input signal, and $S_{xd}(\omega)$ is the cross-spectral density (CSD) between the input signal and the desired signal.

Note that derivation of the optimal filter requires specific knowledge about the input and the desired response in the form of the autocorrelation Φ of the input $x(n)$ and the cross-correlation Θ between the input $x(n)$ and the desired response $d(n)$. In practice, although the desired response $d(n)$ may not be known, it should be possible to obtain an estimate of its temporal or spectral statistics, which may be used to estimate Θ . Proper estimation of the statistical entities mentioned above requires a large number of samples of the corresponding signals.

{Note: Haykin [40] allows all the entities involved to be complex. Vector transposition T is then Hermitian or complex-conjugate transposition H . Products of two entities require one to be conjugated: for example, $e^2(n)$ is obtained as $e(n)e^*(n)$, and Equation 3.154 will have w_k^* in place of w_k . Furthermore, $\frac{d}{dw}(\Theta^H w) = 0$ and $\frac{d}{dw}(w^H \Theta) = 2\Theta$. The final Wiener–Hopf equation, however, simplifies to the same as above in Equation 3.171.}

Let us now consider the problem of removing noise from a corrupted input signal. For this case, let the input $x(n)$ contain a mixture of the desired (original) signal $d(n)$ and noise $\eta(n)$, that is,

$$x(n) = d(n) + \eta(n). \quad (3.177)$$

Using the vector notation as before, we have

$$\mathbf{x}(n) = \mathbf{d}(n) + \boldsymbol{\eta}(n), \quad (3.178)$$

where $\boldsymbol{\eta}(n)$ is the vectorial representation of the noise function $\eta(n)$. The autocorrelation matrix of the input is given by

$$\Phi = E[\mathbf{x}(n)\mathbf{x}^T(n)] = E[\{\mathbf{d}(n) + \boldsymbol{\eta}(n)\}\{\mathbf{d}(n) + \boldsymbol{\eta}(n)\}^T]. \quad (3.179)$$

If we assume that the noise process is statistically independent of the signal process, and that at least one of the processes has a mean value of zero, we have

$$E[\mathbf{d}(n)\boldsymbol{\eta}^T(n)] = E[\boldsymbol{\eta}^T(n)\mathbf{d}(n)] = \mathbf{0}. \quad (3.180)$$

(This result is valid also if the two processes are mutually orthogonal.) Then,

$$\Phi = E[\mathbf{d}(n)\mathbf{d}^T(n)] + E[\boldsymbol{\eta}(n)\boldsymbol{\eta}^T(n)] = \Phi_d + \Phi_\eta, \quad (3.181)$$

where Φ_d and Φ_η are the $M \times M$ autocorrelation matrices of the signal and noise, respectively. Furthermore,

$$\Theta = E[\mathbf{x}(n)d(n)] = E[\{\mathbf{d}(n) + \boldsymbol{\eta}(n)\}d(n)] = E[\mathbf{d}(n)d(n)] = \Phi_{1d}, \quad (3.182)$$

where Φ_{1d} is an $M \times 1$ autocorrelation vector of the desired signal. The optimal Wiener filter is then given by

$$\mathbf{w}_o = (\Phi_d + \Phi_\eta)^{-1}\Phi_{1d}. \quad (3.183)$$

The frequency response of the Wiener filter may be obtained by modifying Equation 3.176 by taking into account the spectral relationships

$$S_{xx}(\omega) = S_d(\omega) + S_\eta(\omega) \quad (3.184)$$

and

$$S_{xd}(\omega) = S_d(\omega), \quad (3.185)$$

which leads to

$$W(\omega) = \frac{S_d(\omega)}{S_d(\omega) + S_\eta(\omega)} = \frac{1}{1 + \frac{S_\eta(\omega)}{S_d(\omega)}}, \quad (3.186)$$

where $S_d(\omega)$ and $S_\eta(\omega)$ are the PSDs of the desired signal and the noise process, respectively. Note that designing the optimal filter requires knowledge of the PSDs of the desired signal and the noise process (or models thereof).

From Equation 3.186, the following important properties of the Wiener filter are evident: $W(\omega) = 0$ wherever $S_d(\omega) = 0$, $W(\omega) = 1$ wherever $S_\eta(\omega) = 0$, and $W(\omega)$ decreases as $S_\eta(\omega)$ increases. Therefore, the Wiener filter has the following characteristics:

- frequency components that are not present in the input (that is, have a value of zero) will not be modified or restored,
- the input signal's frequency components are passed with unit gain at frequencies where noise does not exist, and
- the gain of the Wiener filter decreases as the *SNR*, expressed as a function of frequency, decreases.

Illustrations of application: The upper trace in Figure 3.87 shows one ECG cycle extracted from the signal with noise in Figure 3.5. A piecewise linear model of the desired version of the signal was created by concatenating linear segments to provide P, QRS, and T waves with amplitudes, durations, and intervals similar to those in the given noisy signal. The baseline of the model was set to zero. The noise-free model is shown in the middle trace of Figure 3.87. The log PSDs of the given noisy signal and the noise-free model, the latter being $S_d(\omega)$ in Equation 3.186, are shown in the upper two plots of Figure 3.88.

The T-P intervals between successive cardiac cycles in an ECG (the interbeat intervals) may be taken to represent the isoelectric baseline. Then, any activity present in these intervals constitutes noise. Four T-P intervals were selected from the noisy signal in Figure 3.5, and their Fourier power spectra were averaged to derive the noise PSD $S_\eta(\omega)$ required in the Wiener filter (Equation 3.186). The estimated log PSD of the noise is shown in the third trace of Figure 3.88. Observe the relatively high levels of power in the noise PSD beyond 100 Hz as compared to the PSDs of the original noisy signal and the model. Observe also the peaks in the original and noise PSDs near 180 Hz, 300 Hz, and 420 Hz, representing the third, fifth, and seventh harmonics of 60 Hz, respectively; the peak at 460 Hz is an aliased version of the ninth harmonic at 540 Hz. The 60 – Hz component itself appears to have been suppressed by a notch filter in the signal acquisition system. (See Sections 3.3.4 and 3.7.3 for details.)

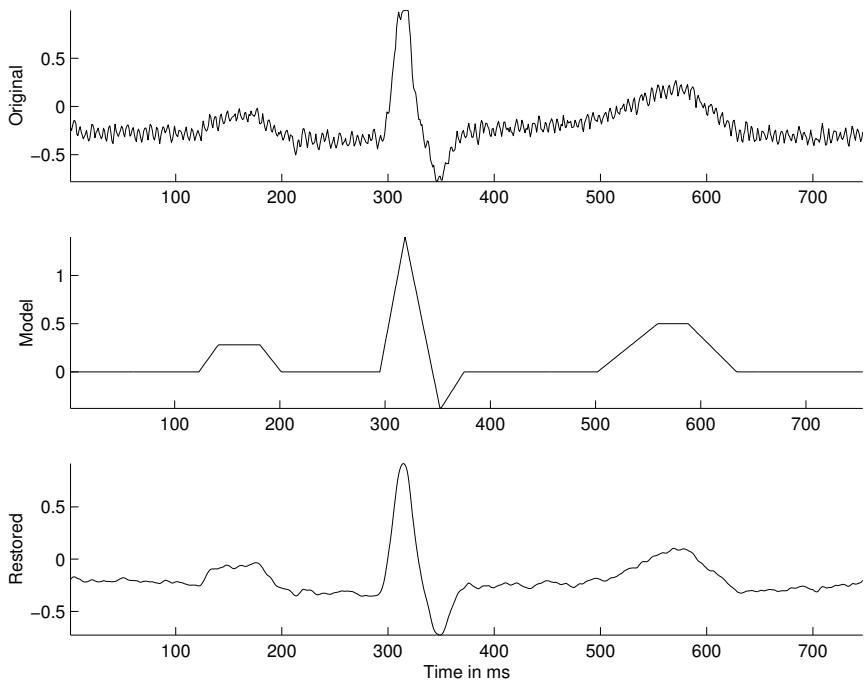


Figure 3.87 From top to bottom: one cycle of the noisy ECG signal in Figure 3.5 (labeled as Original); a piecewise linear model of the desired noise-free signal (Model); and the output of the Wiener filter (Restored).

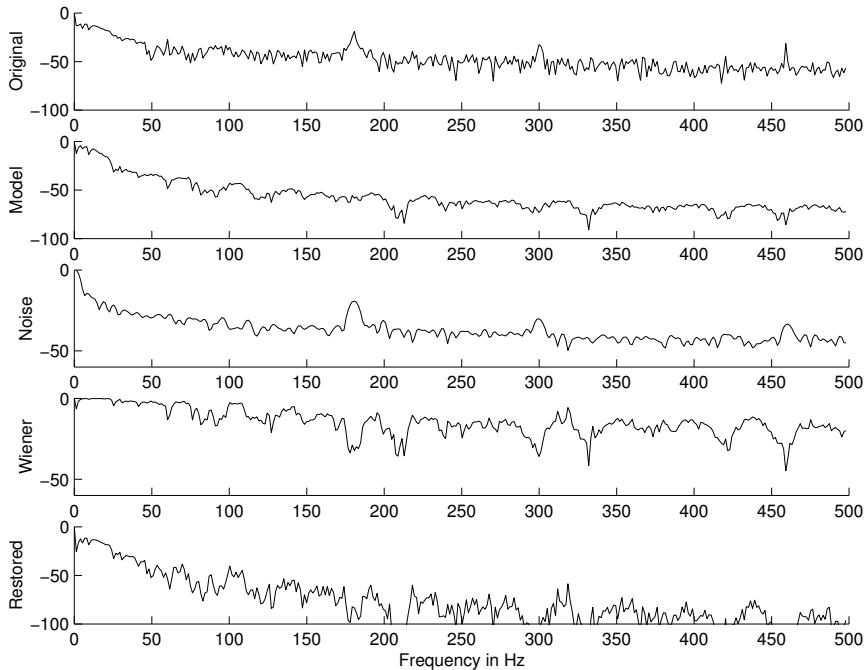


Figure 3.88 From top to bottom: log PSD (in dB) of the given noisy signal (labeled as Original); log PSD of the noise-free signal model (Model); estimated log PSD of the noise process (Noise); log frequency response of the Wiener filter (Wiener); and log PSD of the filter output (Restored). The scaling of the vertical axis is not the same for all PSDs in the figure.

The Wiener filter frequency response was derived as in Equation 3.186, and is shown in the fourth plot in Figure 3.88. Observe the low gain of the filter near 180 Hz, 300 Hz, 420 Hz, and 460 Hz corresponding to the peaks in the noise spectrum. As indicated by Equation 3.186, the Wiener filter gain is inversely related to the noise PSD and directly related to the signal PSD. The result of application of the Wiener filter to the given signal is shown in the third trace of Figure 3.87. It is evident that almost all of the noise has been effectively removed by the filter. This is also confirmed by comparing the first and last plots in Figure 3.88: The spectrum of the restored signal has values below -50 dB for frequencies greater than 100 Hz.

The most important point to observe here is that the filter was derived with models of the noise and signal processes (PSDs), which were obtained from the given signal itself in the present application. No cutoff frequency was required to be specified in designing the Wiener filter, whereas the Butterworth filter requires the specification of a cutoff frequency and a filter order.

In another experiment, a noise-free ECG signal was obtained from a subject under controlled conditions; a one-second-long segment of this signal was selected to derive the desired signal PSD (model) for use in the design of the Wiener filter. Subsequently, a noisy ECG was recorded from the same subject by including artifacts related to motion and contraction of the limbs. The sampling frequency used was 200 Hz. Ten segments of noise were selected from the noisy ECG in the T-P intervals, and their average PSD was computed. The impulse response of the Wiener filter was derived by taking the inverse Fourier transform of the frequency response derived according to Equation 3.186.

Figure 3.89 shows one-second-long traces of the model ECG, the noisy ECG, and the filtered result. Also shown in the same figure are the ACFs of the model ECG and noise, as well as the impulse response of the Wiener filter. The impulse response of the Wiener filter, which possesses even-symmetry due to the real-valued nature of the related frequency response, was truncated by selecting only 10 samples on either side of its maximal value (which occurs at zero time but appears at 0.5 s in Figure 3.89 due to the shift applied for causality), resulting in the filter order of $M = 21$. The coefficients obtained as above were used to filter the noisy signal in the time domain (over a duration of about two minutes). It is evident from the last two traces of Figure 3.89 that the noise has been effectively suppressed without any noticeable degradation of the ECG wave components. Figure 3.90 shows longer traces of the noisy and filtered ECG signals. Close inspection of the results indicates that, while relatively high-frequency noise has been suppressed, noise within the bandwidth of the ECG signal remains in the output. Furthermore, low-frequency artifacts remain for the same reason, along with the fact that the short segments of noise used to model the noise PSD do not effectively incorporate the low-frequency artifacts in the model. The PSDs and frequency response of the Wiener filter shown in Figure 3.91 confirm the observations made above for the present example as well as those in the previous illustration.

Most signal acquisition systems should permit the measurement of at least the variance or power level of the noise present. A uniform (white) PSD model may then be easily derived. Models of the ideal signal and the noise processes may also be created using parametric Gaussian or Laplacian models either in the time domain (ACF) or directly in the frequency domain (PSD). However, it should be observed that the noise PSDs in Figures 3.88 and 3.91 are not flat or white, and that the noise ACF in Figure 3.89 is not an impulse. These characteristics indicate that the noise process has some underlying structure.

The Wiener filter is an optimal filter under the conditions imposed and the assumptions made. The optimality of the filter does not hold if the conditions are violated or if the information provided is not valid. Other types of optimal filters may be derived under other conditions and assumptions. Additional examples of optimal filters are provided in the sections to follow and in Chapter 8.

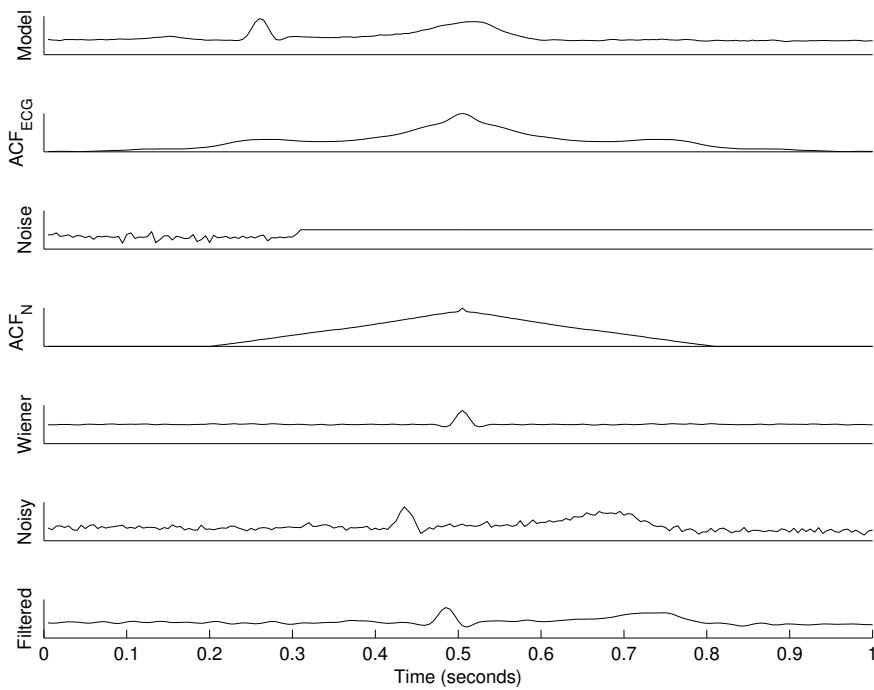


Figure 3.89 From top to bottom: one cycle of the noise-free ECG of a subject (labeled as Model); the ACF of the noise-free ECG; a sample segment of noise from a noisy ECG of the same subject (the actual noise segment has a duration of 0.3 s but has been padded with zeros to the same length as the ECG, which is 1 s); the ACF of the noise (ACF_N) obtained using 10 segments; the impulse response of the Wiener filter (shifted for causality); a segment of the noisy ECG to be filtered; and the corresponding filtered result. The amplitude labels have been suppressed to prevent clutter. ECG signal data courtesy of Emily Marasco and Matthew LaRocque, University of Calgary. See also Figure 3.90.

3.10 Adaptive Filters for Removal of Interference

Filters with fixed characteristics (tap weights or coefficients), as seen in the preceding sections, are suitable when the characteristics of the signal and noise (random or structured) are stationary and known. Design of frequency-domain filters requires detailed knowledge of the spectral contents of the signal and noise. Such filters are not applicable when the characteristics of the signal and/or noise vary with time, that is, when they are nonstationary. They are also not suitable when the spectral contents of the signal and the interference overlap significantly.

Consider the situation when two ECG signals, such as those of a fetus and the expectant mother, or two vibration signals, such as the VAG and the VMG, arrive at the recording site and get added in some proportion. The spectra of the signals in the mixture span the same or similar frequency ranges, and hence fixed filtering cannot separate them. In the case of the VAG/VMG mixture, it is also possible for the spectra of the signals to vary from one point in time to another, due to changes in the characteristics of the cartilage surfaces causing the VAG signal and the effect of variations in the recruitment of the motor units of the muscles involved on the VMG signal. Such a situation calls for the use of a filter that can learn and adapt to the characteristics of the interference, estimate the interfering signal, and remove it from the mixture to obtain the desired signal. This requires the filter to adjust automatically its impulse response (and hence its frequency response) as the characteristics of the signal and/or noise vary. (Several methods for the analysis of nonstationary and multicomponent signals are presented in Chapters 8 and 9.)

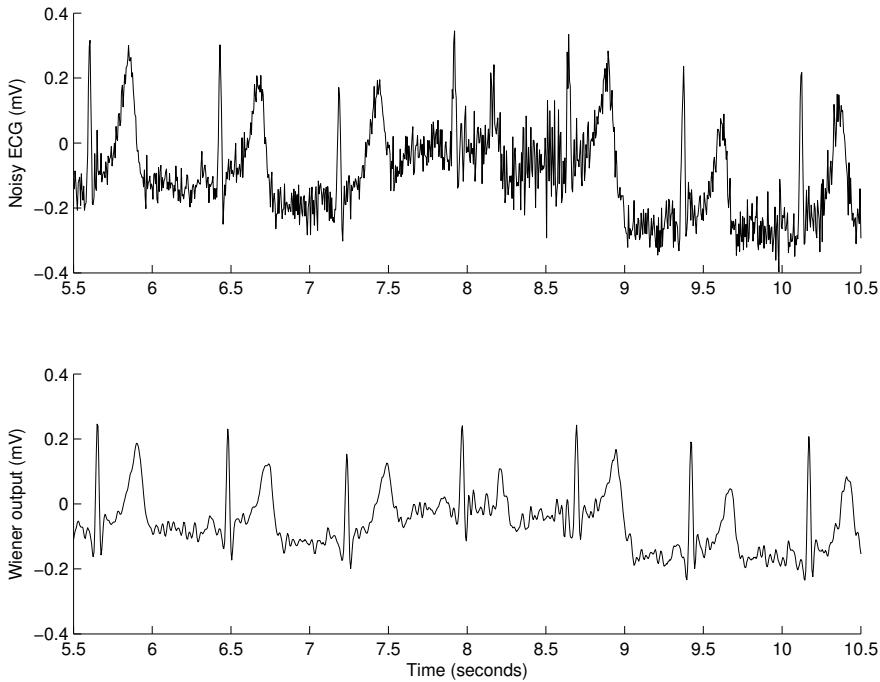


Figure 3.90 From top to bottom: a segment of the noisy ECG to be filtered and the corresponding filtered result. ECG signal data courtesy of Emily Marasco and Matthew LaRocque, University of Calgary. See also Figure 3.89.

Problem: Design an optimal filter to remove a nonstationary interference from a nonstationary signal. An additional channel of information related to the interference is available for use. The filter should continuously adapt to the changing characteristics of the signal and interference.

Solution: We need to address two different concerns in this problem:

1. The filter should be *adaptive*; the tap-weight vector of the filter will then vary with time. The principles of the adaptive filter, also known as the adaptive noise canceler (ANC), are explained in Section 3.10.1.
2. The filter should be *optimal*. Two well-established methods for optimization of the adaptive filter are presented in Sections 3.10.2 and 3.10.3.

Illustrations of the application of the methods are presented at the end of Sections 3.10.2 and 3.10.3 as well as at the end of the chapter, in Sections 3.14 and 3.15.

3.10.1 The adaptive noise canceler

Figure 3.92 shows a generic block diagram of an adaptive filter or ANC [41, 42]. The “primary input” to the filter $x(n)$ is a mixture of the signal of interest $v(n)$ and the “primary noise” $m(n)$:

$$x(n) = v(n) + m(n). \quad (3.187)$$

$x(n)$ is the primary observed signal; it is desired that the interference or noise $m(n)$ be estimated and removed from $x(n)$ in order to obtain the signal of interest $v(n)$. (The notation and terminology used in the present section, while being different from those in the preceding sections, are commonly used

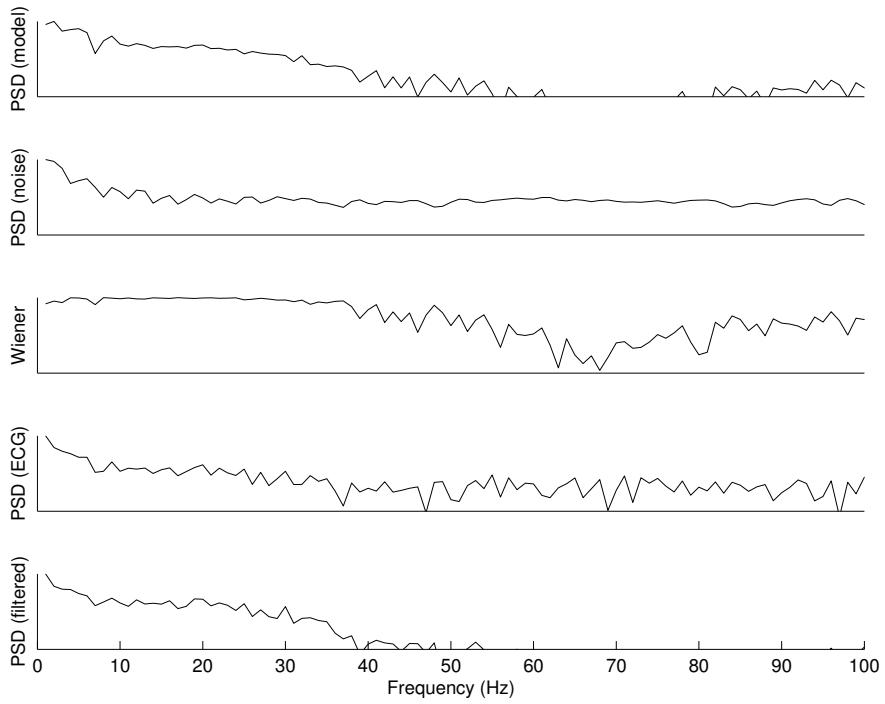


Figure 3.91 From top to bottom: log PSD (in dB) of the noise-free ECG (model); log PSD of the noise obtained using 10 segments; log frequency response of the Wiener filter; log PSD of the noisy ECG; and log PSD of the filtered result. The display of all PSDs is limited to the range $[-50, 0]$ dB. $f_s = 200$ Hz. The PSDs of ECG signals were obtained using only one cardiac cycle of each as shown in Figure 3.89.

in the literature on adaptive filters.) It is assumed that $v(n)$ and $m(n)$ are uncorrelated. Adaptive filtering requires a second input, known as the “reference input” $r(n)$, that is uncorrelated with the signal of interest $v(n)$ but closely related to or correlated with the interference or noise $m(n)$ in some manner that need not be known. The ANC filters or modifies the reference input $r(n)$ to obtain a signal $y(n)$ that is as close to the noise $m(n)$ as possible. $y(n)$ is then subtracted from the primary input to estimate the desired signal:

$$\tilde{v}(n) = e(n) = x(n) - y(n). \quad (3.188)$$

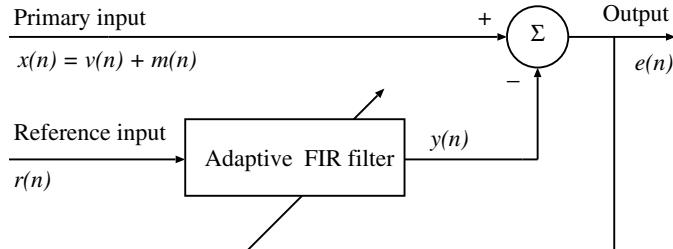


Figure 3.92 Block diagram of a generic ANC.

Let us now analyze the function of the filter. Let us assume that the signal of interest $v(n)$, the primary noise $m(n)$, the reference input $r(n)$, and the primary noise estimate $y(n)$ are statistically

stationary and have zero means. (*Note:* The requirement of stationarity is removed later when the expectations are computed in moving windows.) We have already stated that $v(n)$ is uncorrelated with $m(n)$ and $r(n)$, and that $r(n)$ is correlated with $m(n)$. The output of the ANC is

$$\begin{aligned} e(n) &= x(n) - y(n) \\ &= v(n) + m(n) - y(n), \end{aligned} \quad (3.189)$$

where $y(n) = \tilde{m}(n)$ is the estimate of the primary noise obtained at the output of the adaptive FIR filter. By taking the square and expectation (statistical average) of both sides of Equation 3.189, we obtain

$$E[e^2(n)] = E[v^2(n)] + E[\{m(n) - y(n)\}^2] + 2E[v(n)\{m(n) - y(n)\}]. \quad (3.190)$$

Since $v(n)$ is uncorrelated with $m(n)$ and $y(n)$, and all of them have zero means, we have

$$E[v(n)\{m(n) - y(n)\}] = E[v(n)]E[m(n) - y(n)] = 0. \quad (3.191)$$

Equation 3.190 can be rewritten as

$$E[e^2(n)] = E[v^2(n)] + E[\{m(n) - y(n)\}^2]. \quad (3.192)$$

With reference to Figure 3.92, note that the output $e(n)$ is used (fed back) to control the adaptive FIR filter. In ANC applications, the objective is to obtain an output $e(n)$ that is a least-squares fit to the desired signal $v(n)$. This is achieved by feeding the output back to the adaptive FIR filter and adjusting the filter to minimize the total system output power. The system output serves as the error signal for the adaptive process.

The signal power $E[v^2(n)]$ will be unaffected as the FIR filter is adjusted to minimize $E[e^2(n)]$; accordingly, the minimum output power is

$$\min E[e^2(n)] = E[v^2(n)] + \min E[\{m(n) - y(n)\}^2]. \quad (3.193)$$

As the FIR filter is adjusted so that $E[e^2(n)]$ is minimized, $E[\{m(n) - y(n)\}^2]$ is also minimized. Thus, the FIR filter's output $y(n)$ is the MMSE estimate of the primary noise $m(n)$. Moreover, when $E[\{m(n) - y(n)\}^2]$ is minimized, $E[\{e(n) - v(n)\}^2]$ is also minimized, since from Equation 3.189, we have

$$e(n) - v(n) = m(n) - y(n). \quad (3.194)$$

Adjusting or adapting the FIR filter to minimize the total output power is, therefore, equivalent to causing the ANC's output $e(n)$ to be the MMSE estimate of the signal of interest $v(n)$ for the given structure and adjustability of the adaptive FIR filter and for the given reference input.

The output $e(n)$ will contain the signal of interest $v(n)$ and some noise. From Equation 3.194, the output noise is given by $e(n) - v(n) = \tilde{v}(n) - v(n) = m(n) - y(n)$. Because minimizing $E[e^2(n)]$ minimizes $E[\{m(n) - y(n)\}^2]$, minimizing the total output power minimizes the output noise power. Since the signal component $v(n)$ in the ANC's output remains unaffected, minimizing the total output power maximizes the output SNR.

From Equation 3.192, we have the condition that the output power is minimum when $E[e^2(n)] = E[v^2(n)]$. When this condition is achieved, $E[\{m(n) - y(n)\}^2] = 0$. We then have $y(n) = m(n)$ and $e(n) = v(n)$; that is, the ANC's output is a perfect and noise-free estimate of the desired signal.

Optimization of the FIR filter may be performed by expressing the error in terms of the tap-weight vector and applying the procedure of choice. The output $y(n)$ of the adaptive FIR filter (see Figure 3.92) in response to its input $r(n)$ is given by

$$y(n) = \sum_{k=0}^{M-1} w_k r(n-k), \quad (3.195)$$

where w_k , $k = 0, 1, 2, \dots, M - 1$, are the tap weights, and M is the order of the filter. The estimation error $e(n)$ or the output of the ANC system is

$$e(n) = x(n) - y(n). \quad (3.196)$$

For the sake of notational simplicity, let us define the tap-weight vector at time n as

$$\mathbf{w}(n) = [w_0(n), w_1(n), \dots, w_{M-1}(n)]^T. \quad (3.197)$$

Similarly, the input vector at each time instant n may be defined as the M -dimensional vector

$$\mathbf{r}(n) = [r(n), r(n - 1), \dots, r(n - M + 1)]^T. \quad (3.198)$$

Then, the estimation error $e(n)$ given in Equation 3.196 may be rewritten as

$$e(n) = x(n) - \mathbf{w}^T(n)\mathbf{r}(n). \quad (3.199)$$

It is worth noting that the derivations made above required no knowledge about the processes behind $v(n)$, $m(n)$, and $r(n)$ or their interrelationships, other than the assumptions of statistical independence between $v(n)$ and $m(n)$ and some form of correlation between $m(n)$ and $r(n)$. The arguments can be extended to situations where the primary and reference inputs contain additive random noise processes that are mutually uncorrelated and also uncorrelated with $v(n)$, $m(n)$, and $r(n)$. The procedures may also be extended to cases where $m(n)$ and $r(n)$ are deterministic or structured rather than stochastic processes, such as power-line interference, an ECG, or a VMG signal [41].

Several methods are available to maximize the output SNR ; two such methods based on the least-mean-squares (LMS) and the recursive least-squares (RLS) approaches are described in the following sections.

3.10.2 The least-mean-squares adaptive filter

The purpose of adaptive filtering algorithms is to adjust the tap-weight vector to minimize the MSE. By squaring the expression for the estimation error $e(n)$ given in Equation 3.199, we get

$$e^2(n) = x^2(n) - 2x(n)\mathbf{r}^T(n)\mathbf{w}(n) + \mathbf{w}^T(n)\mathbf{r}(n)\mathbf{r}^T(n)\mathbf{w}(n). \quad (3.200)$$

The squared error is a second-order (quadratic) function of the tap-weight vector (and the inputs), and may be depicted as a concave hyperparaboloidal (bowl-like) surface that is never negative. The aim of the filter optimization procedure would be to reach the bottom of the bowl-like function. Gradient-based methods may be used for this purpose.

By taking the expected values of the entities in Equation 3.200 and taking the derivative with respect to the tap-weight vector, we may derive the Wiener–Hopf equation for the present application. The LMS algorithm takes a simpler approach by assuming the square of the instantaneous error as in Equation 3.200 to stand for an estimate of the MSE [41]. The iterative LMS algorithm is based on the method of steepest gradient descent, where the new tap-weight vector $\mathbf{w}(n + 1)$ is given by the present tap-weight vector $\mathbf{w}(n)$ plus a correction proportional to the negative of the gradient $\nabla e^2(n)$ of the squared error:

$$\mathbf{w}(n + 1) = \mathbf{w}(n) - \mu \nabla e^2(n). \quad (3.201)$$

The parameter μ controls the stability and rate of convergence of the algorithm: the larger the value of μ , the larger the gradient of the error that is introduced and the faster the convergence of the algorithm, and vice versa.

The LMS algorithm approximates $\nabla e^2(n)$ by the derivative of the squared error in Equation 3.200 with respect to the tap-weight vector as

$$\widetilde{\nabla e^2}(n) = -2x(n)\mathbf{r}(n) + 2\{\mathbf{w}^T(n)\mathbf{r}(n)\}\mathbf{r}(n) = -2e(n)\mathbf{r}(n). \quad (3.202)$$

Using this estimate of the gradient in Equation 3.201, we get

$$\mathbf{w}(n+1) = \mathbf{w}(n) + 2\mu e(n) \mathbf{r}(n). \quad (3.203)$$

This expression is known as the Widrow–Hoff LMS algorithm.

The advantages of the LMS algorithm lie in its simplicity and ease of implementation: Although the method is based on the MSE and gradient-based optimization, the filter expression itself is free of differentiation, squaring, or averaging. It has been shown that the expected value of the tap-weight vector provided by the LMS algorithm converges to the optimal Wiener solution when the input vectors are uncorrelated over time [41, 43]. The iterative procedure may be started with an arbitrary tap-weight vector; it will converge in the mean and remain stable as long as μ is greater than zero but less than the reciprocal of the largest eigenvalue of the autocorrelation matrix of the reference input [41].

Illustration of application: Zhang et al. [44] used a two-stage adaptive LMS filter to cancel muscle-contraction interference from VAG signals. The first stage was used to remove the measurement noise in the accelerometers and associated amplifiers, and the second stage was designed to cancel the muscle signal.

Zhang et al. [44] also proposed a procedure for optimization of the step size μ using an *RMS*-error-based misadjustment factor and a time-varying estimate of the input signal power, among other entities. The LMS algorithm was implemented as

$$\mathbf{w}(n+1) = \mathbf{w}(n) + 2\mu(n) e(n) \mathbf{r}(n). \quad (3.204)$$

The step size μ was treated as a variable, its value being determined dynamically as

$$\mu(n) = \frac{\mu}{(M+1) \bar{x}^2(n) [\alpha, r(n), \bar{x}^2(n-1)]}, \quad (3.205)$$

with $0 < \mu < 1$. The notation $\bar{x}^2(n) [\alpha, r(n), \bar{x}^2(n-1)]$ indicates that the updated value of $\bar{x}^2(n)$ is a function of $\alpha, r(n)$, and $\bar{x}^2(n-1)$. A forgetting factor α was introduced in the adaptation process, with $0 \leq \alpha \ll 1$; this feature was expected to overcome problems caused by high levels of nonstationarity in the signal. $\bar{x}^2(n)$ is a time-varying estimate of the input signal power, computed as $\bar{x}^2(n) = \alpha r^2(n) + (1-\alpha) \bar{x}^2(n-1)$.

The filtered versions of the VAG signals recorded from the midpatella and the tibial tuberosity positions, as shown in Figure 1.58 [traces (b) and (c), right-hand column], are shown in Figure 3.93. The muscle-contraction signal recorded at the distal rectus femoris position was used as the reference input [Figure 1.58, right-hand column, trace (a)]. It is seen that the low-frequency muscle-contraction artifact has been successfully removed from the VAG signals (compare the second half of each signal in Figure 3.93 with the corresponding part in Figure 1.58).

3.10.3 The RLS adaptive filter

When the input process of an adaptive system is quasistationary, the best steady-state performance results from slow adaptation. However, when the input statistics are time-variant (nonstationary), the best performance is obtained by a compromise between fast adaptation (necessary to track variations in the input process) and slow adaptation (necessary to limit the noise in the adaptive process). The LMS adaptation algorithm is a simple and efficient approach for ANC; however, it is not appropriate for fast-varying signals due to its slow convergence and due to the difficulty in selecting

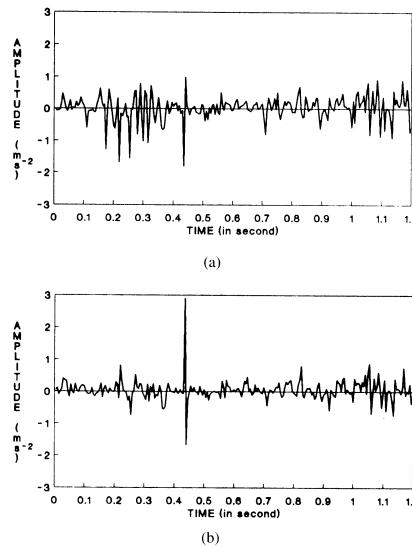


Figure 3.93 LMS-filtered versions of the VAG signals recorded from the midpatella and the tibial tuberosity positions, as shown in Figure 1.58 [traces (b) and (c), right-hand column]. The muscle-contraction signal recorded at the distal rectus femoris position was used as the reference input [Figure 1.58, right-hand column, trace (a)]. The recording setup is shown in Figure 1.57. Reproduced with permission from Y.T Zhang, R.M. Rangayyan, C.B. Frank, and G.D. Bell, Adaptive cancellation of muscle-contraction interference from knee joint vibration signals, *IEEE Transactions on Biomedical Engineering*, 41(2):181–191, 1994. ©IEEE.

the correct value for the step size μ . An alternative approach based on the minimization of the least-squares criterion is the RLS method [40, 42, 45]. The RLS algorithm has been widely used in real-time system identification and noise cancellation because of its fast convergence, which is about an order of magnitude higher than that of the LMS method. (The derivation of the RLS filter in this section has been adapted, with permission, from Sesay [45].)

An important feature of the RLS algorithm is that it utilizes information contained in the input data and extends it back to the instant of time when the algorithm was initiated [40]. Given the least-squares estimate of the tap-weight vector of the filter at time $n - 1$, the updated estimate of the vector at time n is computed upon the arrival of new data.

In the derivation of the RLS algorithm, the *performance criterion* or *objective function* $\xi(n)$ to be minimized in the sense of least squares is defined as

$$\xi(n) = \sum_{i=1}^n \lambda^{n-i} |e(i)|^2, \quad (3.206)$$

where $1 \leq i \leq n$ is the observation interval, $e(i)$ is the estimation error as defined in Equation 3.199, and λ is a weighting factor (also known as the *forgetting factor*) with $0 < \lambda \leq 1$. The values of $\lambda^{n-i} < 1$ give more “weight” to the recent error values compared to past values. Such weighting is desired in the case of nonstationary signals, where changes in the signal statistics make the inclusion of past values less appropriate than recent values. The inverse of $(1 - \lambda)$ is a measure of the memory of the algorithm.

The Wiener–Hopf equation is the necessary and sufficient condition [40] for minimization of the performance index in the least-squares sense and to obtain the optimal values of the tap weights, and may be derived in a manner similar to that presented in Section 3.9 for the Wiener filter. The

normal equation to be solved in the RLS procedure is

$$\Phi(n)\tilde{\mathbf{w}}(n) = \Theta(n), \quad (3.207)$$

where $\tilde{\mathbf{w}}(n)$ is the optimal tap-weight vector for which the performance index is at its minimum, $\Phi(n)$ is an $M \times M$ time-averaged (and weighted) autocorrelation matrix of the reference input $\mathbf{r}(i)$, defined as

$$\Phi(n) = \sum_{i=1}^n \lambda^{n-i} \mathbf{r}(i) \mathbf{r}^T(i), \quad (3.208)$$

and $\Theta(n)$ is an $M \times 1$ time-averaged (and weighted) cross-correlation matrix between the reference input $\mathbf{r}(i)$ and the primary input $x(i)$, defined as

$$\Theta(n) = \sum_{i=1}^n \lambda^{n-i} \mathbf{r}(i) x(i). \quad (3.209)$$

The general scheme of the RLS filter is illustrated in Figure 3.94.

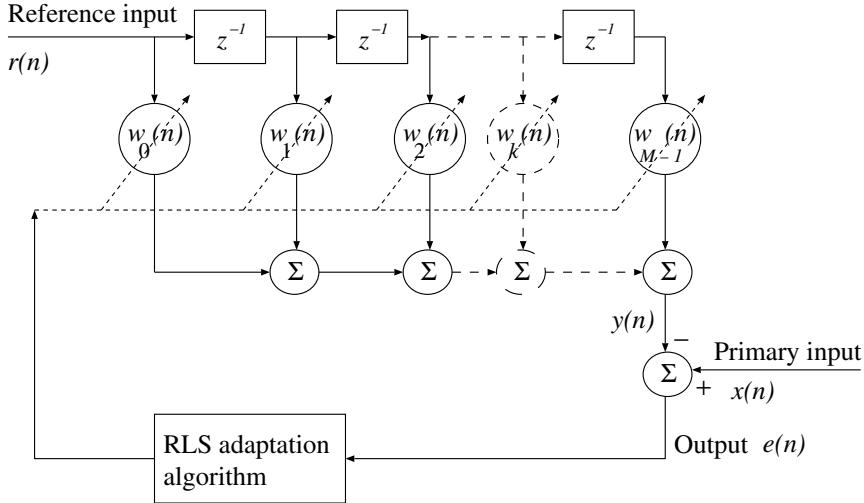


Figure 3.94 General structure of the adaptive RLS filter.

Because of the difficulty in solving the normal equation for the optimal tap-weight vector, recursive techniques need to be considered. In order to obtain a recursive solution, we could isolate the term corresponding to $i = n$ from the rest of the summation on the right-hand side (RHS) of Equation 3.208 and obtain

$$\Phi(n) = \lambda \left[\sum_{i=1}^{n-1} \lambda^{n-i-1} \mathbf{r}(i) \mathbf{r}^T(i) \right] + \mathbf{r}(n) \mathbf{r}^T(n). \quad (3.210)$$

According to the definition in Equation 3.208, the expression inside the square brackets on the RHS of Equation 3.210 equals the time-averaged and weighted autocorrelation matrix $\Phi(n-1)$. Hence, Equation 3.210 can be rewritten as a recursive expression, given by

$$\Phi(n) = \lambda \Phi(n-1) + \mathbf{r}(n) \mathbf{r}^T(n). \quad (3.211)$$

Similarly, Equation 3.209 can be written as the recursive equation

$$\Theta(n) = \lambda \Theta(n-1) + \mathbf{r}(n) x(n). \quad (3.212)$$

To compute the least-squares estimate $\tilde{\mathbf{w}}(n)$ for the tap-weight vector in accordance with Equation 3.207, we have to determine the inverse of the correlation matrix $\Phi(n)$. In practice, such an operation is time consuming (particularly if M is large). To reduce the computational requirements, a matrix inversion lemma known as the “ABCD lemma” could be used (a similar form of the lemma can be found in Haykin [40]). According to the ABCD lemma, given matrices \mathbf{A} , \mathbf{B} , \mathbf{C} , and \mathbf{D} ,

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{DA}^{-1}\mathbf{B} + \mathbf{C}^{-1})^{-1}\mathbf{DA}^{-1}. \quad (3.213)$$

The matrices \mathbf{A} , \mathbf{C} , $(\mathbf{A} + \mathbf{BCD})$, and $(\mathbf{DA}^{-1}\mathbf{B} + \mathbf{C}^{-1})$ are assumed to be invertible. With the correlation matrix $\Phi(n)$ assumed to be positive definite and therefore nonsingular, we may apply the matrix inversion lemma to Equation 3.211 by assigning

$$\begin{aligned} \mathbf{A} &= \lambda\Phi(n-1), \\ \mathbf{B} &= \mathbf{r}(n), \\ \mathbf{C} &= 1, \\ \mathbf{D} &= \mathbf{r}^T(n). \end{aligned}$$

We then have

$$\begin{aligned} \Phi^{-1}(n) &= \lambda^{-1}\Phi^{-1}(n-1) \\ &- \lambda^{-1}\Phi^{-1}(n-1)\mathbf{r}(n) [\lambda^{-1}\mathbf{r}^T(n)\Phi^{-1}(n-1)\mathbf{r}(n) + 1]^{-1} \\ &\times \lambda^{-1}\mathbf{r}^T(n)\Phi^{-1}(n-1). \end{aligned} \quad (3.214)$$

Since the expression inside the brackets of the equation given above is a scalar, the equation can be rewritten as

$$\Phi^{-1}(n) = \lambda^{-1}\Phi^{-1}(n-1) - \frac{\lambda^{-2}\Phi^{-1}(n-1)\mathbf{r}(n)\mathbf{r}^T(n)\Phi^{-1}(n-1)}{1 + \lambda^{-1}\mathbf{r}^T(n)\Phi^{-1}(n-1)\mathbf{r}(n)}. \quad (3.215)$$

For convenience of notation, let

$$\mathbf{P}(n) = \Phi^{-1}(n), \quad (3.216)$$

with $\mathbf{P}(0) = \delta^{-1}\mathbf{I}$, where δ is a small constant, and \mathbf{I} is the identity matrix. Furthermore, let

$$\mathbf{k}(n) = \frac{\lambda^{-1}\mathbf{P}(n-1)\mathbf{r}(n)}{1 + \lambda^{-1}\mathbf{r}^T(n)\mathbf{P}(n-1)\mathbf{r}(n)}. \quad (3.217)$$

$\mathbf{k}(n)$ is analogous to the *Kalman gain vector* in Kalman filter theory [40]; see Section 8.7. Equation 3.215 may then be rewritten in a simpler form as

$$\mathbf{P}(n) = \lambda^{-1}\mathbf{P}(n-1) - \lambda^{-1}\mathbf{k}(n)\mathbf{r}^T(n)\mathbf{P}(n-1). \quad (3.218)$$

By multiplying both sides of Equation 3.217 by the denominator on its RHS, we get

$$\mathbf{k}(n) [1 + \lambda^{-1}\mathbf{r}^T(n)\mathbf{P}(n-1)\mathbf{r}(n)] = \lambda^{-1}\mathbf{P}(n-1)\mathbf{r}(n), \quad (3.219)$$

or

$$\mathbf{k}(n) = [\lambda^{-1}\mathbf{P}(n-1) - \lambda^{-1}\mathbf{k}(n)\mathbf{r}^T(n)\mathbf{P}(n-1)] \mathbf{r}(n). \quad (3.220)$$

Comparing the expression inside the brackets on the RHS of the equation given above with Equation 3.218, we have

$$\mathbf{k}(n) = \mathbf{P}(n)\mathbf{r}(n). \quad (3.221)$$

$\mathbf{P}(n)$ and $\mathbf{k}(n)$ have the dimensions $M \times M$ and $M \times 1$, respectively.

Using Equations 3.207, 3.212, and 3.216, a recursive equation to update the least-squares estimate $\tilde{\mathbf{w}}(n)$ of the tap-weight vector can be obtained as

$$\begin{aligned}\tilde{\mathbf{w}}(n) &= \Phi^{-1}(n)\Theta(n) \\ &= \mathbf{P}(n)\Theta(n) \\ &= \lambda\mathbf{P}(n)\Theta(n-1) + \mathbf{P}(n)\mathbf{r}(n)x(n).\end{aligned}\quad (3.222)$$

Substituting Equation 3.218 for $\mathbf{P}(n)$ in the first term of Equation 3.222, we get

$$\begin{aligned}\tilde{\mathbf{w}}(n) &= \mathbf{P}(n-1)\Theta(n-1) - \mathbf{k}(n)\mathbf{r}^T(n)\mathbf{P}(n-1)\Theta(n-1) + \mathbf{P}(n)\mathbf{r}(n)x(n) \\ &= \Phi^{-1}(n-1)\Theta(n-1) - \mathbf{k}(n)\mathbf{r}^T(n)\Phi^{-1}(n-1)\Theta(n-1) + \mathbf{P}(n)\mathbf{r}(n)x(n) \\ &= \tilde{\mathbf{w}}(n-1) - \mathbf{k}(n)\mathbf{r}^T(n)\tilde{\mathbf{w}}(n-1) + \mathbf{P}(n)\mathbf{r}(n)x(n).\end{aligned}\quad (3.223)$$

Finally, from Equation 3.221, using the fact that $\mathbf{P}(n)\mathbf{r}(n)$ equals the gain vector $\mathbf{k}(n)$, the equation given above can be rewritten as

$$\begin{aligned}\tilde{\mathbf{w}}(n) &= \tilde{\mathbf{w}}(n-1) - \mathbf{k}(n)[x(n) - \mathbf{r}^T(n)\tilde{\mathbf{w}}(n-1)] \\ &= \tilde{\mathbf{w}}(n-1) + \mathbf{k}(n)\alpha(n),\end{aligned}\quad (3.224)$$

where $\tilde{\mathbf{w}}(0) = \mathbf{0}$, and

$$\begin{aligned}\alpha(n) &= x(n) - \mathbf{r}^T(n)\tilde{\mathbf{w}}(n-1) \\ &= x(n) - \tilde{\mathbf{w}}^T(n-1)\mathbf{r}(n).\end{aligned}\quad (3.225)$$

The quantity $\alpha(n)$ is often referred to as the *a priori error*, reflecting the fact that it is the error obtained using the “old” filter (that is, the filter before being updated with the new data at the n^{th} time instant). It is evident that, in the case of ANC applications, $\alpha(n)$ will be the estimated signal of interest $\tilde{v}(n)$ after the filter has converged, that is,

$$\alpha(n) = \tilde{v}(n) = x(n) - \tilde{\mathbf{w}}^T(n-1)\mathbf{r}(n).\quad (3.226)$$

Furthermore, after convergence, the primary noise estimate, that is, the output of the adaptive filter $y(n)$, can be written as

$$y(n) = \tilde{m}(n) = \tilde{\mathbf{w}}^T(n-1)\mathbf{r}(n).\quad (3.227)$$

By substituting Equations 3.189 and 3.227 in Equation 3.226, we get

$$\begin{aligned}\tilde{v}(n) &= v(n) + m(n) - \tilde{m}(n) \\ &= v(n) + m(n) - \tilde{\mathbf{w}}^T(n-1)\mathbf{r}(n) \\ &= x(n) - \tilde{\mathbf{w}}^T(n-1)\mathbf{r}(n).\end{aligned}\quad (3.228)$$

Equation 3.224 gives a recursive relationship to obtain the optimal values of the tap weights, which, in turn, provide the least-squares estimate $\tilde{v}(n)$ of the signal of interest $v(n)$ as in Equation 3.228.

Illustration of application: Figure 3.95 shows plots of the VAG signal of a normal subject [trace (a)] and a simultaneously recorded channel of muscle-contraction interference [labeled as MCI, trace (b)]. The characteristics of the vibration signals in this example are different from those of the signals in Figure 1.58, due to a different recording protocol in terms of speed and range of swinging motion of the leg [42]. The results of adaptive filtering of the VAG signal with the muscle-contraction interference channel as the reference are also shown in Figure 3.95: Trace (c) shows the result of LMS filtering, and trace (d) shows that of RLS filtering. A single-stage LMS

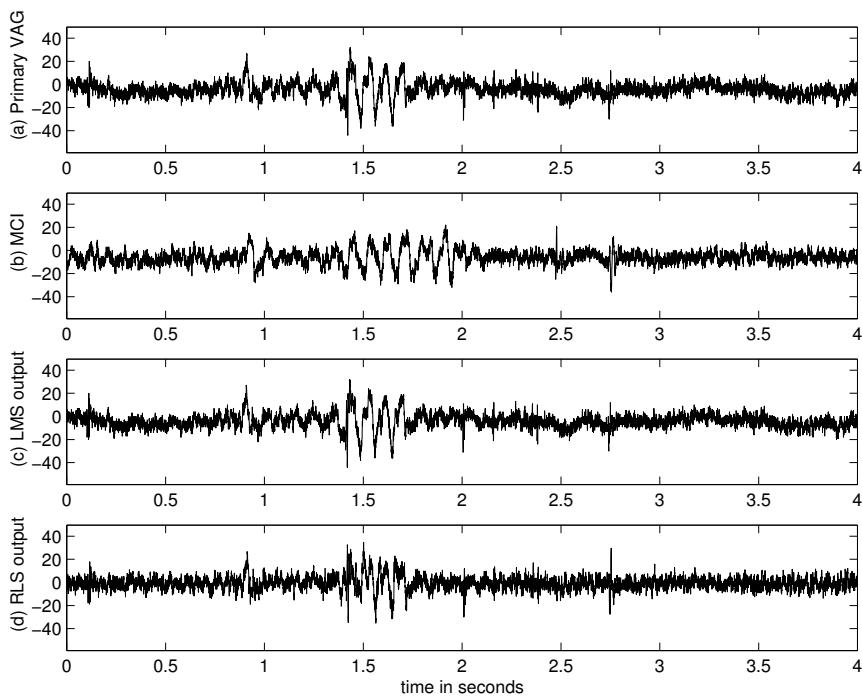


Figure 3.95 (a) VAG signal of a normal subject. (b) Muscle-contraction interference (MCI). (c) Result of LMS filtering. (d) Result of RLS filtering. The recording setup is shown in Figure 1.57.

filter with variable step size $\mu(n)$ as in Equation 3.205 was used; no attempt was made to remove instrumentation noise. The LMS filter used $M = 7$, $\mu = 0.05$, and a forgetting factor $\alpha = 0.98$; other values resulted in poor results. The RLS filter used $M = 7$ and $\lambda = 0.98$.

The relatively low-frequency muscle-contraction interference has been removed better by the RLS filter than by the LMS filter; the latter failed to track the nonstationarities in the interference and has caused additional artifacts in the result. The spectrograms of the primary, reference, and RLS-filtered signals are shown in Figures 3.96, 3.97, and 3.98, respectively. (The logarithmic scale is used to display better the minor differences between the spectrograms.) It is seen that the predominantly low-frequency artifact, indicated by the high energy levels at low frequencies for the entire duration in the spectrograms in Figures 3.96 and 3.97, has been removed by the RLS filter.

3.11 Selecting an Appropriate Filter

In this chapter, we have examined five main approaches to remove noise and interference:

- synchronized or ensemble averaging of multiple realizations of a signal,
- MA filtering,
- frequency-domain filtering,
- Wiener filtering, and
- adaptive filtering.

The first two approaches work directly with the signal in the time domain. Frequency-domain (fixed) filtering is performed on the spectrum of the signal. Note that the impulse response of a filter

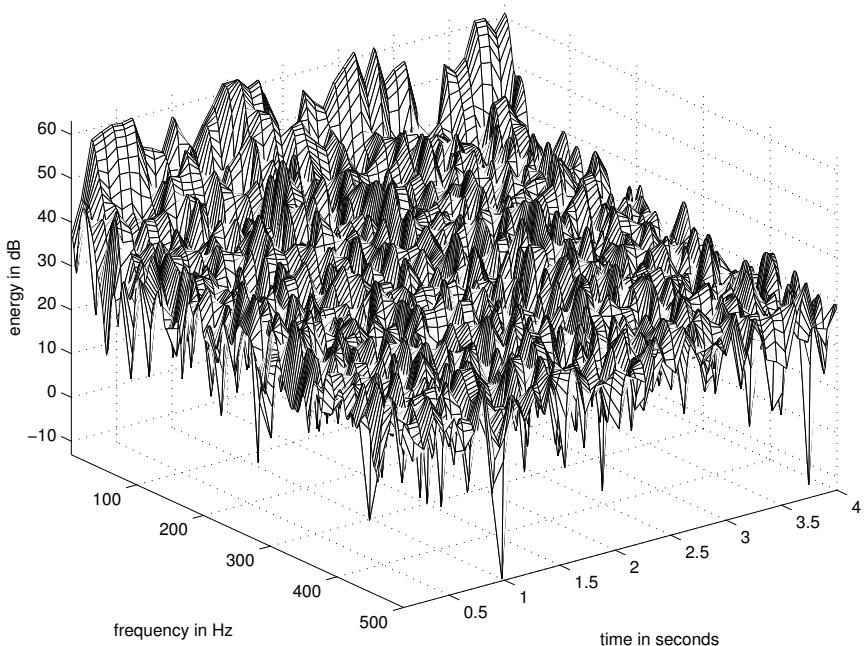


Figure 3.96 Spectrogram of the VAG signal in Figure 3.95 (a). A Hann window of length 256 samples (128 ms) was used; an overlap of 32 samples (16 ms) was allowed between adjacent segments.

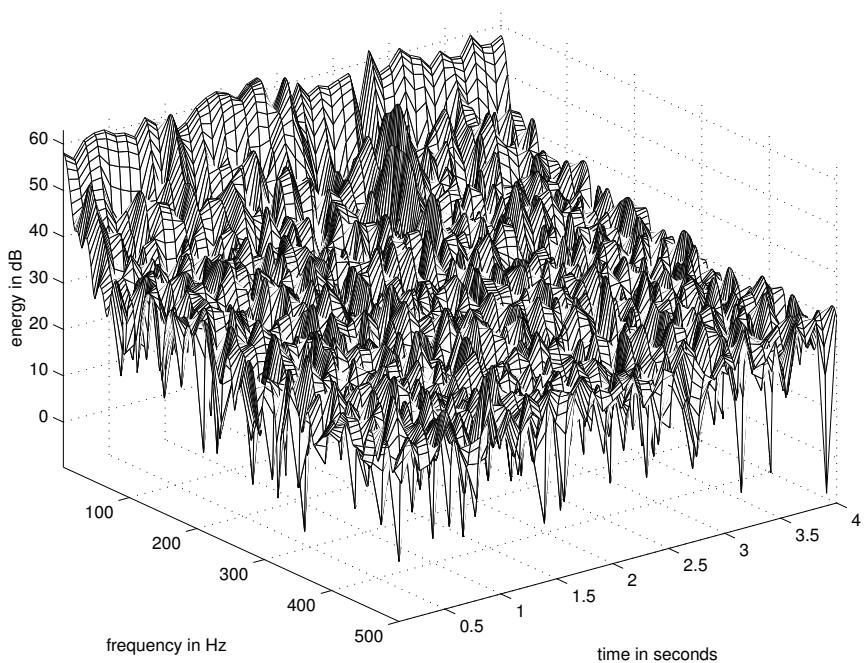


Figure 3.97 Spectrogram of the muscle-contraction interference signal in Figure 3.95 (b). A Hann window of length 256 samples (128 ms) was used; an overlap of 32 samples (16 ms) was allowed between adjacent segments.

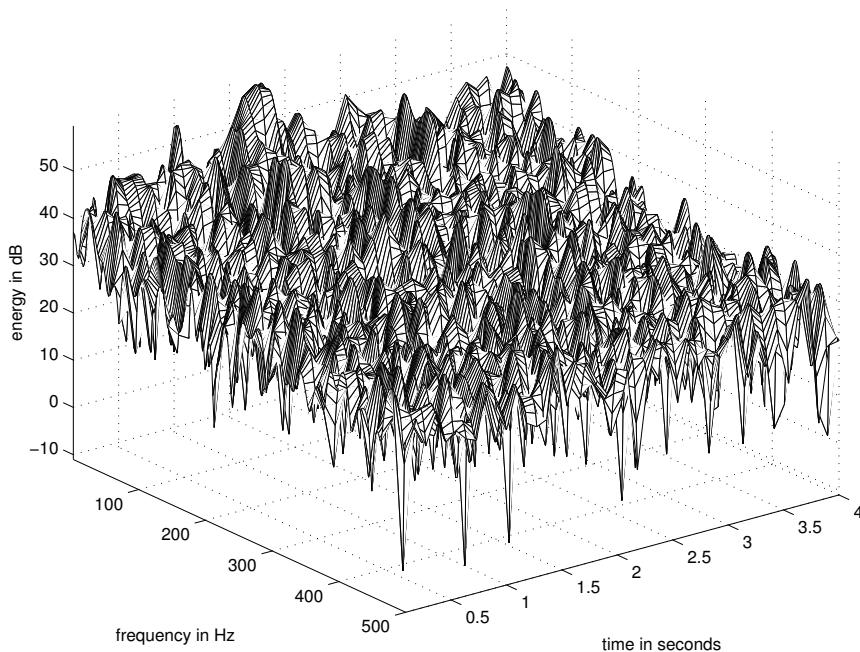


Figure 3.98 Spectrogram of the RLS-filtered VAG signal in Figure 3.95 (d). A Hann window of length 256 samples (128 ms) was used; an overlap of 32 samples (16 ms) was allowed between adjacent segments.

designed in the frequency domain could be used to implement the filter in the time domain as an IIR or FIR filter. Furthermore, time-domain filters may be analyzed in the frequency domain via their transfer function or frequency response to understand better their characteristics and effects on the input signal. The Wiener filter may be implemented either in the time domain as a transversal filter or in the frequency domain. Adaptive filters work directly on the signal in the time domain, but dynamically alter their characteristics in response to changes in the interference; their frequency response varies from one point in time to another.

What are the guiding principles to determine which of these filters is the best for a given application? The following points should assist in making this decision.

Synchronized or ensemble averaging is possible when:

- The signal is statistically stationary, (quasi)periodic, or cyclostationary.
- Multiple realizations or recordings of the signal of interest are available.
- A trigger point or time marker is available, or can be derived, to extract and align the realizations of the signal.
- The noise is a stationary random process that is uncorrelated with the signal and has a zero mean (or a known mean).

Temporal MA filtering is suitable when:

- The signal is statistically stationary at least over the duration of the moving window.
- The noise is a zero-mean random process that is stationary at least over the duration of the moving window and is independent of the signal.

- The signal is a relatively slow (low-frequency) phenomenon.
- Fast, on-line, real-time filtering is desired.

Frequency-domain fixed filtering is applicable when:

- The signal is statistically stationary.
- The noise is a stationary random process that is statistically independent of the signal.
- The signal spectrum is limited in bandwidth compared to that of the noise (or vice versa).
- Loss of information in the spectral band removed by the filter does not seriously affect the signal.
- On-line, real-time filtering is not required (if implemented in the spectral domain via the Fourier transform).

The Wiener filter can be designed if:

- The signal is statistically stationary.
- The noise is a stationary random process that is statistically independent of the signal.
- Specific details (or models) are available regarding the ACFs or the PSDs of the signal and noise.

Adaptive filtering is called for and possible when:

- The noise or interference is nonstationary and not necessarily a random process.
- The noise is uncorrelated with the signal.
- No information is available about the spectral characteristics of the signal and noise, which may also overlap significantly.
- A second source or recording site is available to obtain a reference signal that is strongly correlated with the noise but uncorrelated with the signal.

It is worth noting that an adaptive filter acts as a fixed filter when the signal and noise are stationary. An adaptive filter can also act as a notch or a comb filter when the interference is periodic. It should be noted that all of the filters mentioned above are applicable only when the noise is additive. Techniques such as homomorphic filtering (see Section 4.7) may be used as preprocessing steps if signals combined with operations other than addition need to be separated.

3.12 Application: Removal of Artifacts in ERP Signals

Problem: Propose a method to improve the *SNR* of ERP signals. Suggest methods to measure the improvement in the results obtained.

Solution: Kamath et al. [46] applied synchronized averaging to improve the *SNR* of cortical evoked potentials or ERPs related to electrical and mechanical stimulation of the esophagus. We have seen the details of ensemble averaging or synchronized averaging in Section 3.5; examples of application of synchronized averaging are shown in Figures 3.2 and 3.43. Although improvement in *SNR* was obtained in some experiments conducted by Kamath et al. [46], they also observed that habituation took place as the number of stimuli was increased beyond a certain limit, and that the use of the ERPs obtained after habituation in averaging led to a reduction in the *SNR*.

Let $y_k(n)$ represent one realization or observation of an ERP, with $k = 1, 2, \dots, M$ representing the ensemble index, and $n = 0, 1, 2, \dots, N - 1$ representing the time-sample index. Here, M is the number of realizations of the ERP available, and N is the number of the time samples in each signal. We may express the observed signal as

$$y_k(n) = x_k(n) + \eta_k(n), \quad (3.229)$$

where $x_k(n)$ represents the original uncorrupted signal, and $\eta_k(n)$ represents the noise in the k^{th} realization of the observed signal. The result of ensemble averaging or synchronized averaging is obtained, for each instant of time n , by averaging the M observations of the ERP as

$$\begin{aligned} \bar{y}(n) &= \frac{1}{M} \sum_{k=1}^M y_k(n) \\ &= \frac{1}{M} \sum_{k=1}^M x_k(n) + \frac{1}{M} \sum_{k=1}^M \eta_k(n); \quad n = 0, 1, 2, \dots, N - 1. \end{aligned} \quad (3.230)$$

Figure 3.99 shows superimposed plots of all of the $M = 24$ ERPs available from the experiments conducted by Kamath et al. [46]. While the overall trend of the ERPs is visible, the extent of the noise and inherent variations present in the ERPs can also be observed. Figure 3.100 shows the result of synchronized averaging of all of the 24 ERPs available. The result demonstrates the benefits of averaging in terms of reduced noise and clear depiction of the trends in the ERP signal.

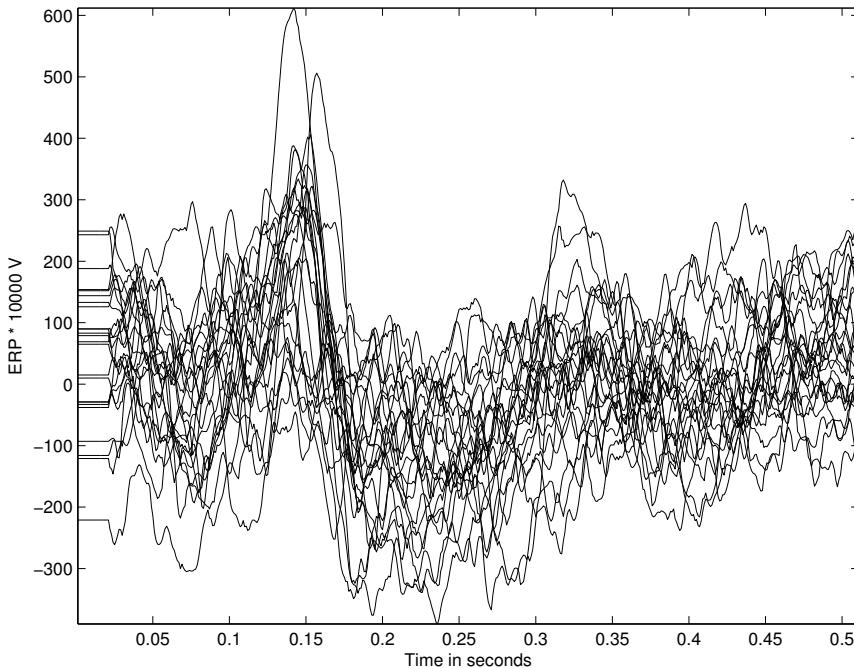


Figure 3.99 Superimposed plots of 24 cortical ERPs related to electrical stimulation of the esophagus. Data courtesy of M.V. Kamath, McMaster University, Hamilton, ON, Canada.

To study the effects of habituation, various sets of the 24 available ERPs were averaged. Figure 3.101 shows three such results, each being the average of eight ERPs. While the average of the first eight ERPs for $k = 1, 2, \dots, 8$ (solid line) appears to be a good result, the result for the next

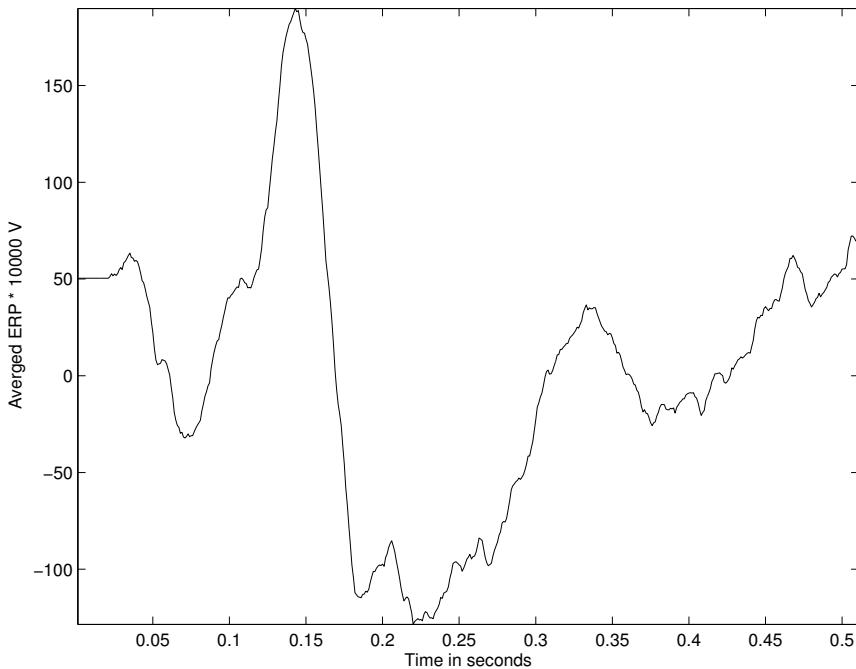


Figure 3.100 Result of synchronized averaging of all of the 24 ERPs shown in Figure 3.99.

set of eight ERPs for $k = 9, 10, \dots, 16$ (dashed line) appears to be noisy. Furthermore, the average of the last set of eight ERPs for $k = 17, 18, \dots, 24$ (dotted line) is not similar to the other results of averaging at all, indicating that habituation (or fatigue) could have resulted in the last eight ERPs being substantially different from the earlier responses.

In Section 3.5, we have seen that synchronized averaging can, theoretically, improve the SNR by the factor of \sqrt{M} . In a practical application, the actual improvement obtained in the SNR may be different. Kamath et al. [46] estimated the SNR by first deriving the power of the noise in the output as

$$\sigma_{\bar{y}}^2 = \frac{1}{NT(M-1)} \sum_{k=1}^M \sum_{n=0}^{N-1} [y_k(n) - \bar{y}(n)]^2. \quad (3.231)$$

Here, $T = 0.001$ s is the sampling interval. The signal power in the output was derived as

$$\sigma_{\bar{y}}^2 = \frac{1}{NT} \left\{ \sum_{n=0}^{N-1} [\bar{y}(n)]^2 \right\} - \frac{\sigma_{\eta}^2}{M}. \quad (3.232)$$

Then, SNR was estimated as

$$SNR = \frac{\sigma_{\bar{y}}^2}{\sigma_{\eta}^2}. \quad (3.233)$$

Kamath et al. [46] also computed the Euclidean distance between the original ERP signals and the averaged signal obtained as

$$D = \frac{1}{M} \sum_{k=1}^M \sqrt{\sum_{n=0}^{N-1} [y_k(n) - \bar{y}(n)]^2}. \quad (3.234)$$

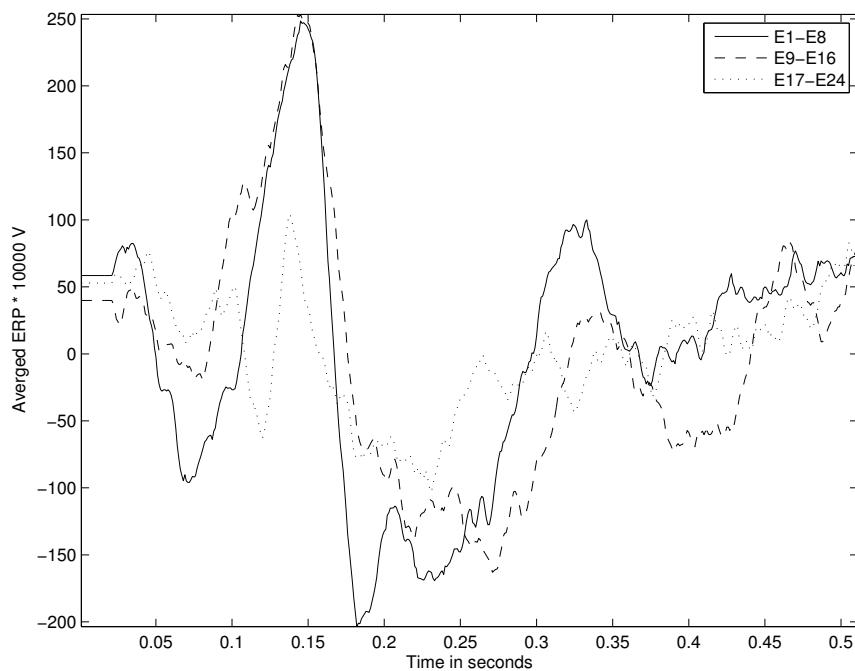


Figure 3.101 Results of synchronized averaging of selected sets of the ERPs shown in Figure 3.99. The three plots shown were obtained by averaging ERPs in groups of eight, for $k = 1, 2, \dots, 8$ (solid line); $k = 9, 10, \dots, 16$ (dashed line); and $k = 17, 18, \dots, 24$ (dotted line).

The SNR for the average of all of the 24 ERPs, shown in Figure 3.100, was found to be 0.37. The $SNRs$ for the three results of averaging eight ERPs, shown in Figure 3.101, were computed to be 0.73, 0.65, and 0.06, respectively, for the first, second, and third sets of eight ERPs. These results confirm, in a quantitative manner, the subjective observations made in the preceding paragraphs: repeated stimulation of a biological system could lead to habituation and fatigue, which could cause the response to be substantially different. Although, based on theoretical considerations, one would expect the averaging of more observations of a certain process to lead to better results with higher SNR , the results of the present study demonstrate otherwise. A judicious design of the experiment is essential, by taking into account various beneficial as well as adverse effects of the methods used.

3.13 Application: Removal of Artifacts in the ECG

Problem: Figure 3.102 (top trace) shows an ECG signal with a combination of baseline drift, high-frequency noise, and power-line interference. Design filters to remove the artifacts.

Solution: The power spectrum of the given signal is shown in the topmost plot in Figure 3.103. Observe the relatively high amount of spectral energy present near DC, from 100 Hz to 500 Hz, and at the power-line frequency and its harmonics located at 60 Hz, 180 Hz, 300 Hz, and 420 Hz. The fundamental component at 60 Hz is lower than the third, fifth, and seventh harmonics, perhaps due to a notch filter included in the signal acquisition system, which has not been effective.

A Butterworth lowpass filter with order $N = 8$ and $f_c = 70$ Hz (see Section 3.7.1 and Equation 3.146), a Butterworth highpass filter of order $N = 8$ and $f_c = 2$ Hz (see Section 3.7.2 and Equation 3.149), and a comb filter with zeros at 60 Hz, 180 Hz, 300 Hz, and 420 Hz (see Section 3.7.3 and Equation 3.152) were applied in series to the signal. The signal spectrum displays

the presence of further harmonics (ninth and eleventh) of the power-line interference at 540 Hz and 660 Hz that have been aliased to the peaks apparent at 460 Hz and 340 Hz, respectively. However, the comb filter in the present example was not designed to remove these components. The lowpass and highpass filters were applied in the frequency domain to the Fourier transform of the signal using the form indicated by Equations 3.146 and 3.149. The comb filter was applied in the time domain using the coefficients in Equation 3.152.

The combined frequency response of the filters is shown in the middle plot in Figure 3.103. The spectrum of the ECG signal after the application of the three filters is shown in the bottom plot in Figure 3.103. The filtered signal spectrum has no appreciable energy beyond about 100 Hz, and displays significant attenuation at 60 Hz.

The outputs after the lowpass filter, the highpass filter, and the comb filter are shown in Figure 3.102. Observe that the baseline drift is present in the output of the lowpass filter, and that the power-line interference is present in the outputs of the lowpass and highpass filters. The final trace is free of all three types of interference. Note, however, that the highpass filter has introduced a noticeable distortion (undershoot) in the P and T waves.

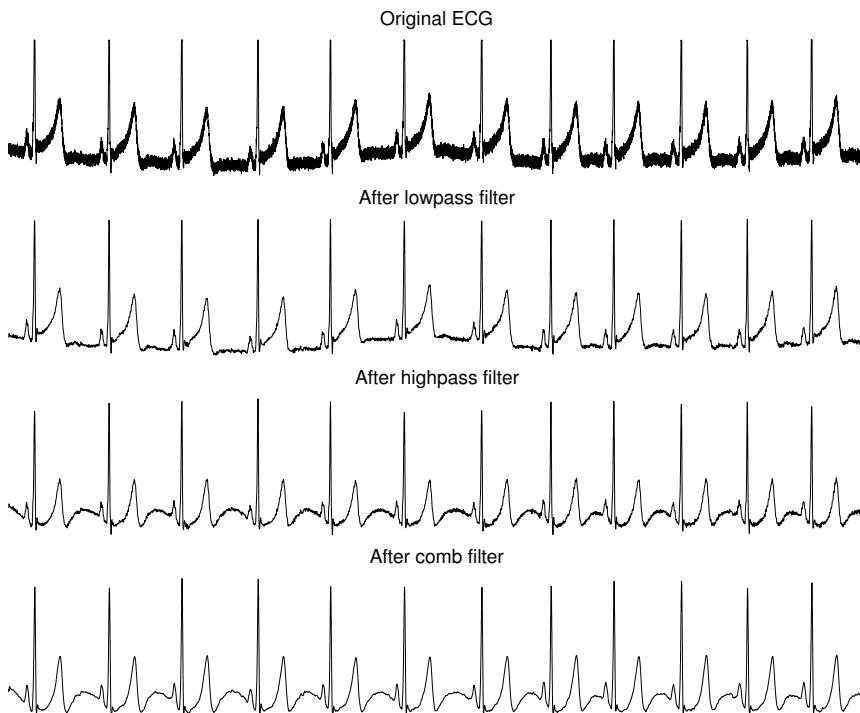


Figure 3.102 ECG signal with a combination of artifacts and its filtered versions. The duration of the signal is 10.7 s.

3.14 Application: Adaptive Cancellation of the Maternal ECG to Obtain the Fetal ECG

Problem: Propose an adaptive noise cancellation filter to remove the maternal ECG signal from the abdominal-lead ECG shown in Figure 3.9 to obtain the fetal ECG. Chest-lead ECG signals of the expectant mother may be used for reference.

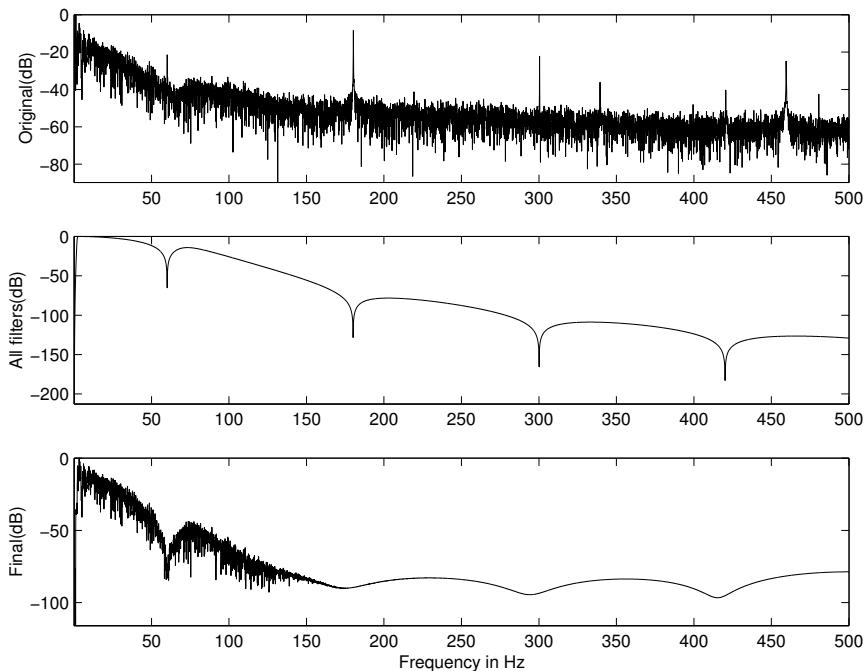


Figure 3.103 Top and bottom plots: Power spectra of the ECG signals in the top and bottom traces of Figure 3.102. Middle plot: Frequency response of the combination of lowpass, highpass, and comb filters. The cutoff frequency of the highpass filter is 2 Hz; the highpass portion of the frequency response is not clearly seen in the plot.

Solution: Widrow et al. [41] described a multiple-reference ANC for removal of the maternal ECG in order to obtain the fetal ECG. The combined ECG was obtained from a single abdominal lead, whereas the maternal ECG was obtained via four chest leads. The model was designed to permit the treatment of not only multiple sources of interference, but also of components of the desired signal present in the reference inputs, and further to consider the presence of uncorrelated noise components in the reference inputs. It should be noted that the maternal cardiac vector is projected on to the axes of different ECG leads in different ways, and hence the appearance and characteristics of the maternal ECG in the abdominal lead would be different from those of the chest-lead ECG signals used as reference inputs.

Each filter channel used by Widrow et al. [41] had 32 taps and a delay of 129 ms. The signals were prefiltered to the bandwidth 3 – 35 Hz, and a sampling rate of 256 Hz was used. The optimal Wiener filter used (see Section 3.9) included transfer functions and cross-spectral vectors between the input source and each reference input. Further extension of the method to more general multiple-source, multiple-reference noise cancelling problems was also discussed by Widrow et al.

The result of cancellation of the maternal ECG from the abdominal lead ECG signal in Figure 3.9 is shown in Figure 3.104. Comparing the two figures, it is seen that the filter output has successfully extracted the fetal ECG and suppressed the maternal ECG. See Widrow et al. [41] for details; see also Ferrara and Widrow [47], Zarzoso and Nandi [48], and Gurve and Krishnan [49] for additional related discussion. See Sections 9.7.2 and 9.11 for other techniques for the same application.

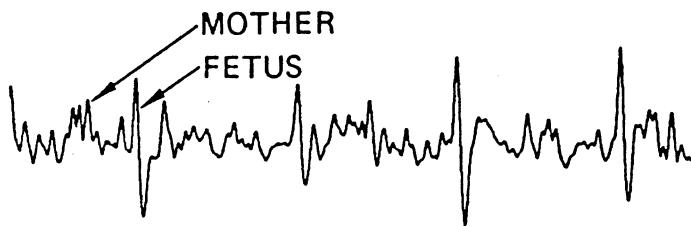


Figure 3.104 Result of adaptive cancellation of the maternal chest ECG from the abdominal ECG in Figure 3.9. The QRS complexes extracted correspond to the fetal ECG. Reproduced with permission from B. Widrow, J.R. Glover, Jr., J.M. McCool, J. Kaunitz, C.S. Williams, R.H. Hearn, J.R. Zeidler, E. Dong, Jr., and R.C. Goodlin, Adaptive noise cancelling: Principles and applications, *Proceedings of the IEEE*, 63(12):1692–1716, 1975. ©IEEE.

3.15 Application: Adaptive Cancellation of Muscle-contraction Interference in VAG Signals

Problem: *Study the applicability of adaptive noise cancellation filters to remove the muscle-contraction interference caused by the rectus femoris in the VAG signal recorded at the patella.*

Solution: Rangayyan et al. [50] conducted a study on the impact of cancellation of muscle-contraction interference on modeling and classification of VAG signals and further classification of the filtered signals as normal or abnormal. Both the LMS (see Section 3.10.2) and the RLS (see Section 3.10.3) methods were investigated, and the RLS method was chosen for its more efficient tracking of nonstationarities in the input and reference signals.

Figure 3.105 shows plots of the VAG signal of a subject with chondromalacia patella of grade II [trace (a)] and a simultaneously recorded channel of muscle-contraction interference [labeled as MCI, trace (b)]. The results of adaptive filtering of the VAG signal with the muscle-contraction interference channel as the reference are also shown in Figure 3.105: Trace (c) shows the result of LMS filtering and trace (d) shows that of RLS filtering. A single-stage LMS filter with variable step size $\mu(n)$ as in Equation 3.205 was used, with $M = 7$, $\mu = 0.05$, and $\alpha = 0.98$. The RLS filter used $M = 7$ and $\lambda = 0.98$.

As in the example in Figure 3.95, it is seen that the muscle-contraction interference has been removed by the RLS filter; however, the LMS filter failed to perform well, due to its limited capabilities in tracking the nonstationarities in the interference. The spectrograms of the primary, reference, and RLS-filtered signals are shown in Figures 3.106, 3.107, and 3.108, respectively. (The logarithmic scale is used to display better the minor differences between the spectrograms.) It is seen that the frequency components of the muscle-contraction interference have been suppressed by the RLS filter.

The primary (original) and filtered VAG signals of 53 subjects were adaptively segmented and modeled in the study of Rangayyan et al. [50] (see Chapter 8). The segment boundaries were observed to be markedly different for the primary and the filtered VAG signals. Parameters extracted from the filtered VAG signals were expected to provide higher discriminant power in pattern classification when compared to the same parameters of the unfiltered or primary VAG signals. However, classification experiments indicated otherwise: The filtered signals gave a lower classification accuracy by almost 10%. It was reasoned that, after removal of the predominantly low-frequency muscle-contraction interference, the transient VAG signals of clinical interest were not modeled well by the prediction-based methods. It was concluded that the adaptive filtering procedure used was not an appropriate preprocessing step before signal modeling for pattern classification. However, it was noted that cancellation of muscle-contraction interference may be a desirable step before auditory or spectral analysis of VAG signals.

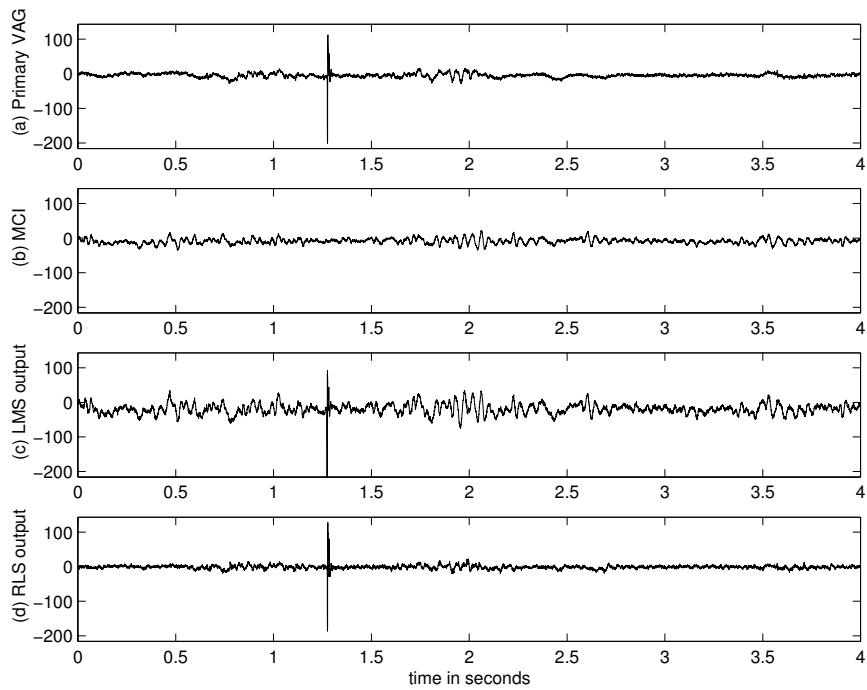


Figure 3.105 Top to bottom: (a) VAG signal of a subject with chondromalacia patella of grade II; (b) muscle-contraction interference (MCI); (c) result of LMS filtering; and (d) result of RLS filtering. The recording setup is shown in Figure 1.57.

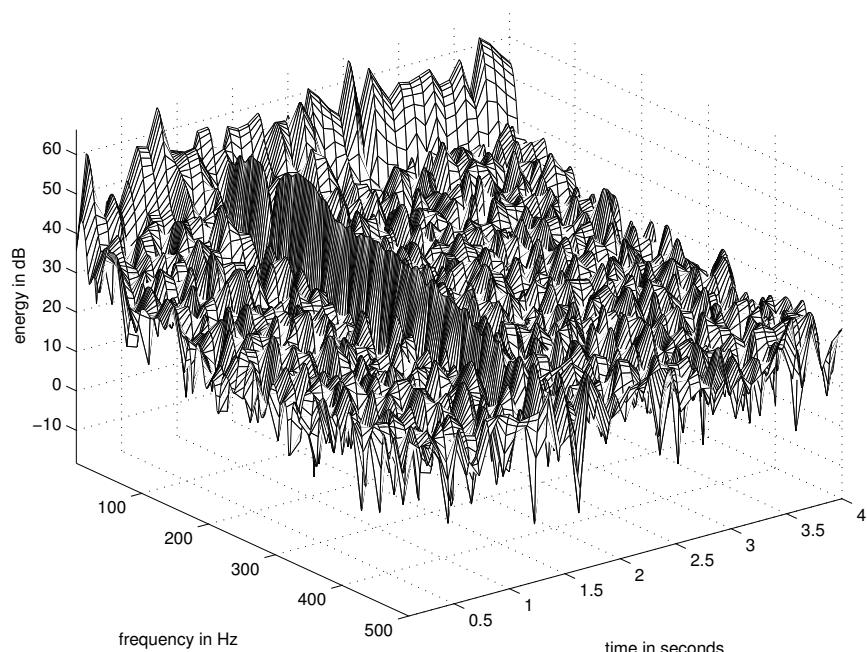


Figure 3.106 Spectrogram of the original VAG signal in Figure 3.105 (a). A Hann window of length 256 samples (128 ms) was used; an overlap of 32 samples (16 ms) was allowed between adjacent segments.

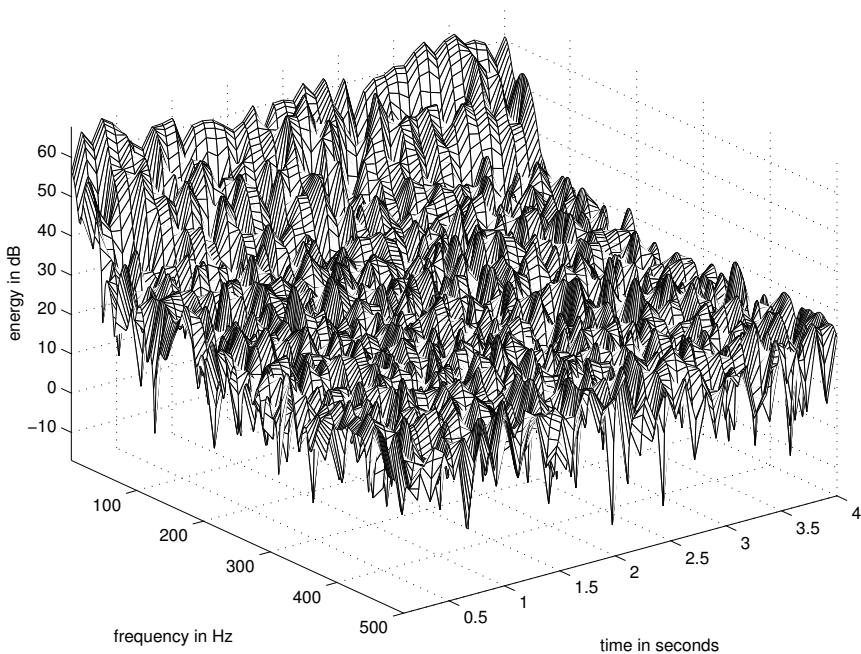


Figure 3.107 Spectrogram of the muscle-contraction interference signal in Figure 3.105 (b). A Hann window of length 256 samples (128 ms) was used; an overlap of 32 samples (16 ms) was allowed between adjacent segments.

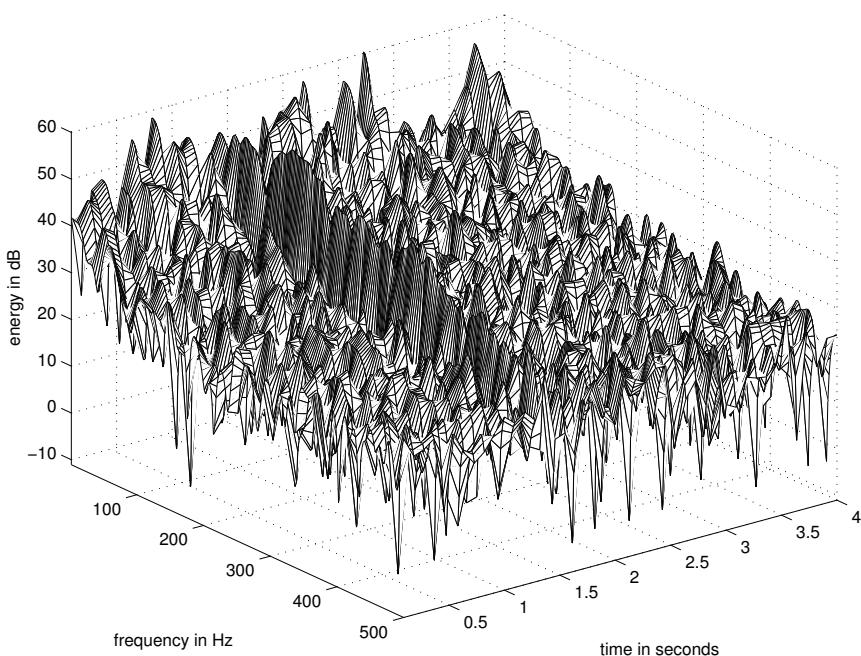


Figure 3.108 Spectrogram of the RLS-filtered VAG signal in Figure 3.105 (d). A Hann window of length 256 samples (128 ms) was used; an overlap of 32 samples (16 ms) was allowed between adjacent segments.

3.16 Remarks

We have investigated problems posed by artifact, noise, and interference of various forms in the acquisition and analysis of several biomedical signals. Random noise, structured interference, and physiological interference were identified and analyzed separately. Attention has been drawn to the different characteristics of various types of noise, such as frequency content and nonstationarity. Fixed, optimal, and adaptive filters were developed in the time and frequency domains for several applications, and guidelines were drawn to assist in choosing the appropriate filter for various types of artifacts. Advanced methods for adaptive denoising based on wavelet and time–frequency decomposition methods have not been discussed in this chapter, but are described by Krishnan and Rangayyan [51] for filtering VAG signals; see Chapter 8 for related material. Another category of filters that has not been considered in this chapter is that of morphological filters [52, 53], which include nonlinear statistics-based operations and could be formulated under certain conditions to include linear filter operations as well.

It is important to observe that each practical problem needs to be studied carefully to determine the type and characteristics of the artifact present; the nature of the signal and its relationship to, or interaction with, the artifact; and the effect of the filter being considered on the desired signal or features computed from the filtered result. Different filters may be suitable for different subsequent steps of signal analysis. It is unlikely that a single filter will address all of the problems and the requirements in a wide variety of practical situations and applications. Regardless of one's expertise in filters, it should be remembered that *prevention is better than cure*: most filters, while removing an artifact, may introduce another. Attempts should be made at the outset to acquire artifact-free signals to the extent possible.

3.17 Study Questions and Problems

Note: Some of the questions deal with the fundamentals of signals and systems, and may require background preparation with other sources such as Lathi [1, 2] or Oppenheim et al. [3]. Such problems are included for the sake of recollection of the related concepts.

1. What are the potential sources of instrumentation and physiological artifacts in recording the PCG signal? Propose nonelectronic methods to prevent or suppress the latter type of artifacts.
2. List four potential sources of instrumentation and physiological artifacts in recording the ECG signal. Describe methods to prevent or remove each artifact. Identify the possible undesired effects of your procedures on the ECG signal.
3. Identify at least three potential sources of physiological artifacts in recording the EEG signal.
4. In recording the EEG in a clinical laboratory, some channels were found to contain the ECG as an artifact. Will simple lowpass or bandpass filtering help in removing the artifact? Why (not)? Propose a scheme to remove the artifact.
5. A biomedical signal is filtered to the range $0 - 150 \text{ Hz}$. Assume the filter to be ideal, and assume any distribution of spectral energy over the bandwidth of the signal. (a) What is the minimum frequency at which the signal should be sampled in order to avoid aliasing errors? (b) A researcher samples the signal at 500 Hz . Draw a schematic representation of the spectrum of the sampled signal. (c) Another researcher samples the signal at 200 Hz . Draw a schematic representation of the spectrum of the sampled signal. Explain the differences between case (b) and case (c).
6. Distinguish between ensemble averages and temporal (time) averages. Identify applications of first-order and second-order averages of both types in EEG analysis.
7. Explain how one may apply ensemble averaging and temporal (time) averaging procedures to process ECG signals. Identify applications of first-order and second-order averages of both types in ECG analysis.
8. Explain how you would apply synchronized averaging to remove noise in (a) ECG signals, (b) event-related (or evoked) potentials, (c) heart sound (PCG) signals, and (d) EMG signals. In each case, explain

- (i) how you will obtain the information required for synchronization of the signals epochs or episodes; (ii) sources of artifacts and how you will deal with them; (iii) limitations and practical difficulties; and (iv) potential for success of the method.
9. Draw a typical ECG waveform over one cardiac cycle indicating the important component waves. How is the waveform affected by passage through the following: (a) A lowpass filter with a cutoff frequency of 40 Hz? (b) A highpass filter with a cutoff frequency of 5 Hz? Draw schematic representations of the expected outputs and explain their characteristics.
10. What is the z -transform of a signal whose samples are given in the series $\{4, 3, 2, 1, 0, -1, 0, 1, 0\}$? (The first sample represents zero time in all the signal sample arrays given in the problems, unless stated otherwise.)
11. A digital filter is used to process a signal at a sampling rate of 2,000 Hz. (a) Draw the unit circle in the complex z -plane and identify the frequencies corresponding to the points $z = (1 + j0)$, $z = (0 + j1)$, $z = (-1 + j0)$, and $z = (0 - j1)$, as well as the point $z = (1 + j0)$ again as approached in the counterclockwise direction. (b) What are the frequencies corresponding to these same points if the sampling rate is 500 Hz?
12. What is the transfer function of an LSI system whose impulse response is given by the series $\{2, 1, 0, 0, -1, 0, 1, 0\}$ for $n = 0, 1, 2, \dots, 7$?
13. The impulse response of a digital filter is $\{1, -2, 1\}$. What will be the response of the filter to the unit step?
14. The impulse response of a filter is $\{3, -2, 2\}$. What will be the response of the filter to the input $\{6, 4, 2, 1\}$?
15. The transfer function of a filter is $H(z) = z^{-1} - 3z^{-2} + 2z^{-4} - z^{-6}$. What is the difference equation relating the output to the input? What is the impulse response of the filter?
16. The impulse response of a filter is given by the series of values $\{3, 2, 1, 0, -1, 0, 0, 1\}$. What is its transfer function?
17. The impulse response of a filter is specified by the series of sample values $\{3, 1, -1, 1\}$. (a) What will be the response of the filter to the input whose sample values are $\{4, 4, 2, 1\}$? (b) Is the filter response obtained by linear convolution or circular convolution of the input with the impulse response? (c) What will be the response with the type of convolution other than the one you indicated as the answer to the previous question? (d) How would you implement convolution of the two signals listed above using the DFT? Which type of convolution will this procedure provide? How would you get the other type of convolution for the signals in this problem via the DFT-based procedure?
18. A biomedical signal is expected to be band-limited to 100 Hz, with significant components of interest up to 80 Hz. However, the signal is contaminated with a periodic artifact with a fundamental frequency of 60 Hz and significant third and fifth harmonics. A researcher samples the signal at 200 Hz without prefiltering the signal. Draw a schematic representation of the spectrum of the signal and indicate the artifactual components. Label the frequency axis clearly in Hz. What kind of a filter would you recommend to remove the artifact?
19. A biomedical signal sampled at 500 Hz was found to have a significant amount of 60 Hz interference. (a) Design a notch filter with two zeros to remove the interference. (b) What is the effect of the filter if a signal sampled at 100 Hz is applied as the input?
20. Two filters with transfer functions $H_1(z) = \frac{1}{3}(1 + z^{-1} + z^{-2})$ and $H_2(z) = 1 - z^{-1}$ are cascaded. (a) What is the transfer function of the complete system? (b) What is its impulse response? (c) What is its gain at DC and at the folding frequency (that is, $f_s/2$)?
21. A filter has the transfer function $H(z) = (1 + 2z^{-1} + z^{-2})/(1 - z^{-2})$. (a) Write the difference equation relating the output to the input. (b) Draw the signal-flow diagram of a realization of the filter. (c) Draw its pole-zero diagram.
22. A digital filter has zeros at $0.5 \pm j0.5$ and poles at $-0.6 \pm j0.3$. (a) Derive the transfer function of the filter. (b) Derive the time-domain difference equation (input-output relationship) of the filter. (c) If the filter is used at the sampling frequency of 1,000 Hz, what are the frequencies at which the gain of the filter is at its maximum and minimum?

23. Two filters with the transfer functions $H_1(z) = \frac{1}{2T}(1 - z^{-2})$ and $H_2(z) = \frac{1}{1-\frac{1}{2}z^{-1}}$ are cascaded.
 (a) What is the transfer function of the complete system? (b) Draw its pole-zero diagram. (c) Write the difference equation relating the output to the input. (d) Draw the signal-flow diagram of a realization of the filter. (e) Compute the first six values of the impulse response of the filter. (f) The filter is used to process a signal sampled at 1,000 Hz. What is its gain at 0, 250, and 500 Hz?
24. A filter is described by the difference equation $y(n) = y(n-1) + \frac{1}{4}x(n) - \frac{1}{4}x(n-4)$. (a) What is its transfer function? (b) Draw the signal-flow diagram of a realization of the filter. (c) Draw its pole-zero diagram.
25. Under what conditions will synchronized averaging fail to reduce noise?
26. A signal sampled at the rate of 100 Hz has the samples $\{0, 10, 0, -5, 0\}$ in mV. The signal is passed through a filter described by the transfer function $H(z) = \frac{1}{T}(1 - z^{-1})$. What will be the output sequence? Plot the output and indicate the amplitude and time scales in detail with appropriate units.
27. A signal sampled at the rate of 100 Hz has the samples $\{0, 10, 0, -5, 0\}$ in mV. It is supposed to be processed by a differentiator with the difference equation $y(n) = \frac{1}{T}[x(n) - x(n-1)]$ and then squared. By mistake the squaring operation is performed before the differentiation. What will be the output sequence? Plot the outputs for both cases and indicate the amplitude and time scales in detail with appropriate units. Explain the differences between the two results.
28. A certain signal analysis technique requires the following operations in order: (a) differentiation, (b) squaring, and (c) lowpass filtering with a filter $H(\omega)$. Considering a generic signal $x(t)$ as the input, write the time-domain and frequency-domain expressions for the output of each stage. Will changing the order of the operations change the final result? Why (not)?
29. A signal sampled at the rate of 100 Hz has the samples $\{0, 10, 0, -5, 0\}$ in mV. The signal is processed by a differentiator with the difference equation $y(n) = \frac{1}{T}[x(n) - x(n-1)]$, and then filtered with a 4-point MA filter. (a) Derive the transfer function and frequency response of each filter and the combined system. (b) Derive the values of the signal samples at each stage. (c) Does it matter which filter is placed first? Why (not)? (d) Plot the output and indicate the amplitude and time scales in detail with appropriate units.
30. Distinguish between ensemble averages and temporal (time) averages. Identify potential applications of first-order and second-order averages of both types in heart sound (PCG) analysis. Explain how you would obtain a trigger for synchronization.
31. Is the heart sound signal (PCG) a stationary signal or not? Provide your answer in the context of one full cardiac cycle and give reasons. If you say that the PCG signal is nonstationary, identify parts (segments) that could possibly be stationary, considering the possibility of murmurs in both systole and diastole.
32. A signal $x(t)$ is transmitted through a channel. The received signal $y(t)$ is a scaled, shifted, and noisy version of $x(t)$ given as $y(t) = \alpha x(t - t_0) + \eta(t)$, where α is a scale factor, t_0 is the time delay, and $\eta(t)$ is noise. Assume that the noise process has zero mean and is statistically independent of the signal process, and that all processes are stationary. Derive expressions for the PSD of $y(t)$ in terms of the PSDs of x and η .
33. A signal $x(n)$ that is observed in an experiment is modeled as a noisy version of a desired signal $d(n)$ as $x(n) = d(n) + \eta(n)$. The noise process η is a zero-mean, unit-variance random process with uncorrelated samples [“white” noise, with $\text{ACF } \phi_\eta(\tau) = \delta(\tau)$] that is statistically independent of the signal process d . The ACF $\phi_d(\tau)$ of d is given by the sequence $\{1.0, 0.6, 0.2\}$, for $\tau = 0, 1, 2$, respectively. Prepare the Wiener-Hopf equation and derive the coefficients of the optimal Wiener filter.
34. Draw the waveform of a typical normal ECG over one cardiac cycle. Label the names of the component waves and give their typical durations and intervals. Draw a version of the ECG waveform including power-line artifact at 60 Hz. Draw another version of the ECG waveform including high-frequency noise. Describe potential causes of the two types of artifact and methods to prevent or remove each artifact.
35. A researcher is designing a Wiener filter and is in need of help in obtaining the autocorrelation matrix, defined as $\Phi = E[\mathbf{x}\mathbf{x}^T]$. Help the researcher by computing the autocorrelation matrix for the sample vector $\mathbf{x} = [1 \ 2 \ 2 \ 1]^T$. Explain the meaning of the operator $E[\cdot]$ and describe how you would implement the operation in practice.

36. A researcher uses a combination of two digital filters in cascade (series). The output of the first filter is the first derivative or difference of the input. The output of the second filter is the average of the current input sample and the preceding input sample. (a) Give the input–output relationship in the time domain (difference equation) for each filter. (b) Derive the transfer function for each filter. (c) Derive the impulse response of the complete system. (d) Derive the transfer function of the complete system. (e) Does it matter which filter is placed first? Explain. (f) Compute the gain of the complete system at 0, $f_s/4$, and $f_s/2$, where f_s is the sampling frequency.
37. A noisy signal $x(n)$ is expressed in vector notation as $\mathbf{x}(n) = \mathbf{d}(n) + \boldsymbol{\eta}(n)$, where $\mathbf{d}(n)$ is the original (ideal) signal, and $\boldsymbol{\eta}(n)$ is random noise that is statistically independent of the signal. All signals are assumed to be second-order stationary. Derive an expression for the ACF matrix of \mathbf{x} in terms of the ACF matrices of \mathbf{d} and $\boldsymbol{\eta}$. Explain each step of your derivation.
38. A random signal $x(t)$ is characterized by its PDF $p_x(x)$. Write the basic definition of the mean-squared value of x based on its PDF. Given an observation of the signal in terms of its samples, expressed as $x(n)$, $n = 0, 1, 2, \dots, N - 1$, give a formula to compute the mean-squared value of x (without the PDF). Give a formula to obtain the RMS value of x . If x represents an electrical signal, what do the mean-squared and RMS values represent?
39. For a discrete-time signal $x(n)$, write the basic definition of the z -transform. The signal $x(n)$ is processed by an LSI system with the impulse response $h(n)$. Write the complete mathematical expression for the output $y(n)$. Showing all steps, derive the relationship between the z -transforms of $x(n)$, $h(n)$, and $y(n)$.
40. A digital filter is specified by the difference equation $y(n) = x(n) - x(n - 2)$, where $x(n)$ is the input and $y(n)$ is the output. Derive the transfer function of the filter. Derive the magnitude and phase parts of the frequency response of the filter. Let the sampling frequency be normalized to unity. Plot the magnitude and phase responses. Indicate the values of the functions at normalized frequency values of 0, 0.25, and 0.5. Explain the nature and effects of the filter.
41. A researcher is designing an experiment to record visual ERPs. Provide advice to the researcher on the following: (a) Identify a potential source of artifact in the form of random noise. Propose a strategy or method to prevent or remove the artifact. (b) Identify a potential source of structured noise. Propose a strategy or method to prevent or remove the artifact. (c) Identify a potential source of physiological interference. Propose a strategy or method to prevent or remove the artifact. No equation is required for your answer to this question.
42. Two LSI filters are specified in terms of their impulse responses as $\delta(n) - \delta(n - 1)$ and $\delta(n) + 2\delta(n - 1) + \delta(n - 2)$. A researcher prepares a new filter by connecting the two filters described above in series. (a) Derive the transfer function of each filter. (b) Derive the transfer function of the combined filter. (c) Derive the impulse response of the combined filter. (d) Does it matter which filter is placed first? Explain and justify your answer. (e) Draw the signal-flow diagram of the combined filter. (f) What is the gain of the combined filter at DC, $f_s/4$, and $f_s/2$, where f_s is the sampling frequency? From these values, give an interpretation of the nature of the combined filter.
43. A signal of interest $x(t)$ is affected by additive noise $\eta(t)$ and is observed as $y(t) = x(t) + \eta(t)$. It is assumed that the processes are random, and that the noise process is statistically independent of the signal process. It is also assumed that the mean of η is zero. Write the basic definition of the expected value (mean) of the random process x based on its PDF. Derive the relationship between the mean of y and the statistics of x and η . Derive the relationship between the variance of y and the statistics of x and η . Interpret and explain your results.
44. A researcher obtains a set of ERP signals $x_k(n)$, $k = 1, 2, \dots, M$, and $n = 0, 1, 2, \dots, N - 1$, where $x_k(n)$ is the k^{th} signal, M is the number of signals recorded, and N is the number of samples in each signal. Help the researcher with the following: (a) Give a step-by-step procedure (algorithm) to obtain the synchronized average of the ERPs. (b) State the conditions and requirements for synchronized averaging to be applicable and yield good results. (c) Give an equation to define the synchronized or ensemble average of the ERPs. (d) Give an equation to define the temporal average of one of the ERPs over a specified window of time. (e) Give an equation to obtain the average Euclidean distance between the result of synchronized averaging and the M ERPs available. Explain the meaning of the Euclidean distance in this context.

45. A researcher is designing an experiment to record electroneurograms. Provide advice to the researcher on the following: (a) Identify a potential source of artifact in the form of random noise. Propose a strategy or method to prevent or remove the artifact. (b) Identify a potential source of structured noise. Propose a strategy or method to prevent or remove the artifact. (c) Identify a potential source of physiological interference. Propose a strategy or method to prevent or remove the artifact. No equation is required for your answer to this question.
46. You are given two signals, $x(n)$ and $y(n)$, each with N samples, $n = 0, 1, 2, \dots, N - 1$. No information is available regarding the PDFs of the related processes. Give an equation to compute the normalized correlation coefficient between the two signals. Explain the meaning and purpose of each part of your equation.
47. A filter is specified with the transfer function $H(z) = \frac{1}{T} \left[\frac{1-z^{-1}}{1-0.95 z^{-1}} \right]$. (a) Derive an expression for the input-output relationship of the filter in the time domain. (b) Draw a signal-flow diagram of the filter using only one delay element. (c) Draw the pole-zero plot of the system. (d) What is the gain of the system at zero frequency and at one-half of the sampling frequency?
48. You are given a set of M signals, $x_k(n)$, $k = 1, 2, \dots, M$, each with N samples, $n = 0, 1, 2, \dots, N - 1$. The index n represents sampled time, and the index k represents the k^{th} signal in the set. (a) Explain the differences between ensemble averages and temporal averages. Give equations to define the following: (b) the ensemble mean at an instant of time $n = n_1$; (c) the temporal mean of the k^{th} signal computed over the period $n = n_1$ to $n = n_2$; and (d) the average signal or template, $\bar{x}(n)$, $n = 0, 1, 2, \dots, N - 1$, computed over the set of signals.
49. You are given two processes that generate two signals, x and y , with the PDFs $p_x(x)$ and $p_y(y)$, respectively. The joint PDF between the two processes is $p_{x,y}(x, y)$. (a) Write an expression to define the mean of the signal x . (b) Write an expression to define the variance of the signal x . (c) Write an expression to define the covariance between the two signals x and y . (d) Write an expression to define the normalized correlation coefficient between the two signals. (e) Give interpretations of the variance and normalized correlation coefficient and indicate practical use or applications of these measures.
50. You are given an ensemble of M signals, $y_k(n)$, $k = 1, 2, \dots, M$, with each signal having N samples, indexed as $n = 0, 1, 2, \dots, N - 1$. Write mathematical equations or expressions for the following: (a) The ensemble or synchronized average, $s(n)$, of the M signals. (b) The total power of the difference, error, or noise between the result of averaging, $s(n)$, and all of the given signals, $y_k(n)$, $k = 1, 2, \dots, M$. (c) The total power of the result, $s(n)$. (d) The SNR of the result.
51. You are given two signals, $x(n)$ and $y(n)$, $n = 0, 1, 2, \dots, N$. Write a mathematical equation or expression to derive the normalized correlation coefficient between the two signals. Suppose that $x(n) = u(n) - u(n - 4)$, where $u(n)$ is the unit step function, and $y(n) = x(n - 8)$. Sketch the signals $x(n)$ and $y(n)$. Compute the correlation coefficient (without normalization) between the two signals and plot the result. Explain the relationships between the signals and the result.
52. A digital filter is specified by the difference equation $y(n) = \frac{1}{4} \sum_{k=0}^{k=3} x(n-k)$, where $x(n)$ is the input, and $y(n)$ is the output. (a) Give an equation for and draw a sketch of the impulse response of the filter. (b) Draw the signal-flow diagram of the filter. (c) Derive the transfer function of the filter. (d) Derive the magnitude and phase parts of the frequency response of the filter. (e) What is the gain of the filter at 0 Hz and $f_s/2 \text{ Hz}$, where f_s is the sampling frequency? (f) Explain the nature and effects of the filter.
53. Considering the acquisition of the ECG signal, identify and describe one potential source or cause of each of the following types of artifact: (a) high-frequency noise, (b) periodic artifact, and (c) a physiological artifact. In each case, explain how the artifact is caused, how the artifact gets combined with the ECG, and how you would remove or prevent the artifact.
54. Write equations to define (a) the convolution of two signals, and (b) the Fourier transform of a signal. Prove that the Fourier transform of the convolution of two signals is equal to the product of the Fourier transforms of the two individual signals. Show and explain all steps. You may use continuous-time or discrete-time notation.
55. Write an equation to define the cross-correlation between two signals. Explain the computational procedures required to obtain the cross-correlation. Derive an expression for the Fourier transform of the

- cross-correlation of two signals in terms of the Fourier transforms of the individual signals. Show all steps. If you use any property of the Fourier transform, give its proof.
56. A biomedical engineer working in a neurophysiology laboratory is frustrated by the appearance of the ECG as an artifact in the EEG of a patient. You are hired to help the engineer. (a) Compare the typical amplitude ranges and frequency bandwidths of the two signals. (b) Would you recommend the use of a fixed lowpass, highpass, or bandpass filter to remove the artifact? If yes, give the essential characteristics of the filter that you recommend. If not, explain your reasons. (c) The engineer has heard about adaptive noise cancelers (ANCs). Draw a schematic (block) diagram of an ANC. In the context of the problem mentioned above, explain what the primary input should be; how and from where on the patient you would obtain an appropriate reference input; and how the various inputs and outputs of the system relate to one another. Explain the basic assumptions made in the design and application of the filter. You do not have to derive any equation in your answer to this problem.
57. A discrete-time signal $x(n)$ is passed through an LSI filter with the impulse response $h(n)$. Write the full expression that gives the output $y(n)$ in terms of $x(n)$ and $h(n)$. Starting with the basic definition of the z -transform, derive the relationship between the z -transforms of $y(n)$, $x(n)$, and $h(n)$.
58. (a) A filter is specified to have a zero in the z -domain at $z = 1$. (i) Derive the transfer function $H_1(z)$ of the filter. (ii) Derive the impulse response $h_1(n)$ of the filter. (iii) Is this a lowpass, highpass, bandpass, or band-reject filter?
 (b) Another filter is specified to have a double zero (or two zeros) at $z = -1$. (i) Derive the transfer function $H_2(z)$ of the filter. (ii) Derive the impulse response $h_2(n)$ of the filter. (iii) Is this a lowpass, highpass, bandpass, or band-reject filter?
 (c) A researcher uses the two filters $H_1(z)$ and $H_2(z)$ as above in cascade (series). (i) Derive the transfer function $H(z)$ of the combined filter. (ii) Derive the impulse response $h(n)$ of the combined filter. (iii) Draw the pole-zero plot for the combined filter. (iv) Draw the signal-flow diagram for the combined filter. (v) Is this a lowpass, highpass, bandpass, or band-reject filter?
59. A biomedical signal is sampled at 400 Hz with no aliasing error. The signal is known to contain power-line artifact at the fundamental frequency of 50 Hz and all of its harmonics up to and including the maximum frequency present in the signal. Draw the unit circle in the z -domain and show the positions of the zeros of a comb filter to remove the artifact. Mark the angle and frequency of each zero. You do not have to derive the transfer function of the filter.
60. Two discrete-time filters are specified in terms of their impulse responses as $h_1(n) = \delta(n) + \delta(n - 1) + \delta(n - 2) + \delta(n - 3)$ and $h_2(n) = \delta(n) - \delta(n - 1)$. The two filters are used in series to filter a signal. Derive and plot the impulse response of the combined filter. Derive the transfer function and frequency response of the combined filter.
61. In the optimization procedure for the derivation of the Wiener filter, the coefficients of the filter are expressed as the vector $\mathbf{w} = [w_0, w_1, w_2, \dots, w_{M-1}]^T$, where M is the order of the filter, and T indicates the transpose. The current input sample $x(n)$ and $M - 1$ previous input samples are expressed in another vector as $\mathbf{x}(n) = [x(n), x(n - 1), \dots, x(n - M + 1)]^T$. (a) Write the full expression of convolution to define the output of the filter in terms of the input signal and the impulse response of the filter. Write the equivalent expression using the vectors as defined above and explain how the two methods lead to the same result. (b) Explain the difference between $\mathbf{x}(n) \mathbf{x}^T(n)$ and $\mathbf{x}^T(n) \mathbf{x}(n)$. (c) What does $E[\mathbf{x}(n) \mathbf{x}^T(n)]$ represent? Write a mathematical expression to give the detailed relationship between an element in the result and the input sample values.
62. A signal $x(n)$ is given in terms of its samples as $\{4, 3, 2, 1, 2, 4, 2, 1\}$, for $n = 0, 1, 2, \dots, 7$. The signal is processed using an LSI filter with the impulse response, $h(n)$, having the sampled values $\{1, 2, 1\}$, for $n = 0, 1, 2$. Give the procedure to compute the output of the filter. Derive the output of the filter showing all steps.
63. Draw a block diagram of the ANC (adaptive noise canceler). State the expected relationships between the inputs to the filter. State the conditions that must be met for optimal functioning of the adaptive filter. Give equations for the output of the filter using both the summation form and the vectorial form for convolution of signals. Explain the composition of the vectors in the vectorial form of the equation.
64. A filter is specified in terms of its pole-zero plot as follows: a zero at $z = 1$ and a zero at $z = -1$.
 (a) Derive the transfer function of the filter. (b) Derive the difference equation and draw a signal-flow

- diagram of the filter. (c) Derive and plot the impulse response of the filter. (d) Derive the magnitude and phase of the frequency response of the filter. (e) Draw a sketch of the magnitude of the frequency response of the filter. Assuming the sampling rate to be 200 Hz, label the frequency axis in Hz. (f) Derive the gain at 0 Hz and 100 Hz and explain the nature of the filter.
65. Draw a schematic sketch of a speech signal including segments of voiced and unvoiced speech. Explain their characteristics. Explain the test for randomness. Indicate the result you would expect if you were to apply the test for randomness to your speech signal example.
 66. A biomedical signal sampled at 240 Hz contains power-line interference at 60 Hz. Design a notch filter to remove the artifact. Draw the unit circle in the complex z -domain. Indicate the values of z and the frequency in Hz at the intersections of the circle with the axes. Mark the locations of the poles and/or zeros in your filter design. Derive the transfer function and the impulse response of the filter. Show all steps.
 67. A researcher is interested in recording vibration signals from the knee joint during swinging movement of the leg. However, muscle vibration signals from the thigh muscle were observed to contaminate the knee-joint vibration signal. Provide recommendations to the researcher on how a filter for ANC may be designed to reduce the muscle artifact. Give a schematic block diagram of the ANC filter. You do not need to derive the mathematical procedures for the filter. Indicate where and which signals need to be provided as input to the ANC filter, and where the filtered output is to be obtained. Which part of the ANC filter has time-varying characteristics? Write an equation to describe the input–output relationship of this part.
 68. Compare the schematic representations of a cell in Figure 1.9 and an LSI system in Figures 3.12 and 3.15. Draw analogies between their inputs, outputs, and characteristics.

3.18 Laboratory Exercises and Projects

Note: Data files related to the exercises are available at the site

<https://github.com/srikrishnan1972/Biomedical-Signal-Analysis>

1. Using any signal of your choice, write a program to compute the DFT of the signal using the FFT algorithm. Pad the signal with zeros to increase the number of samples to (a) $N1 =$ the first power of two greater than the number of samples in the original signal, and (b) $N2 = 2 \times N1$. For each case, compute the power spectrum in dB, with the maximum value normalized to 0 dB. Plot the signal and the two versions of the power spectrum with the frequency axis labeled in Hz. (i) Plot each spectrum over the range $[1, N1]$ or $[1, N2]$ with the frequency axis labeled over the range $[0, f_s]$ Hz, where f_s is the sampling frequency. (ii) Shift the samples in the spectra such that DC is at the center and plot them with the frequency axis labeled over the range $[-f_s/2, f_s/2]$ Hz. (iii) Plot the spectra over the positive frequency axis only, with the frequency axis labeled over the range $[0, f_m = f_s/2]$ Hz. Compare the various plots and explain the similarities and differences between them.
2. A noisy ECG signal is provided in the file `ecg_hfn.dat`. (See also the file `ecg_hfn.m`.) The sampling rate is 1,000 Hz. Write a program to compute (a) the power spectrum of the entire signal, and (b) the power spectrum of a part of the signal containing the ECG over only one cardiac cycle. Plot each signal and its power spectrum with the frequency axis labeled in Hz. Apply a suitable lowpass filter to reduce the noise in the ECG signal and repeat the steps given above. Compare the various spectra and explain the similarities and differences between them.
3. The data file `ecg2x60.dat` contains an ECG signal, sampled at 200 Hz, with a significant amount of 60 Hz power-line artifact. (See also the file `ecg2x60.m`.) (a) Design a notch filter with two zeros to remove the artifact and implement it in MATLAB.[®] (b) Add two poles at the same frequencies as those of the zeros, but with a radius that is less than unity. Study the effect of the poles on the output of the filter as their radius is varied.
4. A noisy ECG signal is provided in the file `ecg_hfn.dat`. (See also the file `ecg_hfn.m`.) The sampling rate of this signal is 1,000 Hz.

Develop a MATLAB® program to perform synchronized averaging as described in Section 3.5. Select a QRS complex from the signal for use as the template and use a suitable threshold on the cross-correlation function in Equation 3.97 for beat detection. Plot the resulting averaged QRS complex. Ensure that the averaged result covers one full cardiac cycle. Plot a sample ECG cycle from the noisy signal for comparison.

Select the QRS complex from a different beat for use as the template and repeat the experiment. Observe the results when the threshold on the cross-correlation function is low (for example, 0.4) or high (for example, 0.95) and comment.

5. Filter the noisy ECG signal in the file `ecg_hfn.dat` (see also the file `ecg_hfn.m`; $f_s = 1,000 \text{ Hz}$) using four different Butterworth lowpass filters (individually) realized with MATLAB® with the following characteristics: (a) Order 2, cutoff frequency 10 Hz . (b) Order 8, cutoff frequency 20 Hz . (c) Order 8, cutoff frequency 40 Hz . (d) Order 8, cutoff frequency 70 Hz . Use “help butter” and “help filter” in MATLAB® to get details about the Butterworth filter.

Compare the results obtained using each of the four Butterworth filters (individually) with those obtained by synchronized averaging (as in the preceding exercise), and comment upon the improvement or distortion in the outputs. Relate your discussions to specific characteristics observed in plots of the signals.

6. The ECG signal in the file `ecg_lfn.dat`, with $f_s = 1,000 \text{ Hz}$, has a wandering baseline (low-frequency artifact). (See also the file `ecg_lfn.m`.) Filter the signal with the derivative-based filters described in Section 3.6.2 and study the results. Study the effect of variation of the position of the pole in the filter in Equation 3.132 on the signal.
7. Filter the signal in the file `ecg_lfn.dat` ($f_s = 1,000 \text{ Hz}$) using Butterworth highpass filters with orders 2 to 8 and cutoff frequencies 0.5 to 5 Hz . (See also the file `ecg_lfn.m`.) Study the efficacy of the filters in removing the baseline artifact and the effect on the ECG waveform itself. Determine the best compromise acceptable.
8. Design a Wiener filter to remove the artifacts in the ECG signal in the file `ecg_hfn.dat`. (See also the file `ecg_hfn.m`.) The equation of the desired filter is given in Equation 3.186. The required model PSDs may be obtained as follows:

Create a piecewise linear model of the desired version of the signal by concatenating linear segments to provide P, QRS, and T waves with amplitudes, durations, and intervals similar to those in the given noisy ECG signal. Compute the PSD of the model signal.

Select a few segments from the given ECG signal that are expected to be isoelectric (for example, the T-P intervals). Compute their PSDs and obtain their average. The selected noise segments should have zero mean or have the mean subtracted out.

Compare the results of the Wiener filter with those obtained by synchronized averaging and lowpass filtering.

References

- [1] Lathi BP. *Signal Processing and Linear Systems*. Berkeley-Cambridge, Carmichael, CA, 1998.
- [2] Lathi BP. *Linear Systems and Signals*. Oxford University Press, New York, NY, 2nd edition, 2005.
- [3] Oppenheim AV, Willsky AS, and Nawab SH. *Signals and Systems*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1997.
- [4] Papoulis A. *Signal Analysis*. McGraw-Hill, New York, NY, 1977.
- [5] Papoulis A. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, NY, 1965.
- [6] Bendat JS and Piersol AG. *Random Data: Analysis and Measurement Procedures*. Wiley, New York, NY, 2nd edition, 1986.
- [7] Auñón JI and Chandrasekar V. *Introduction to Probability and Random Processes*. McGraw-Hill, New York, NY, 1997.

- [8] Ramsey FL and Schafer DW. *The Statistical Sleuth — A Course in Methods of Data Analysis*. Wadsworth Publishing Company, Belmont, CA, 1997.
- [9] Riffenburgh RH. *Statistics in Medicine*. Academic, San Diego, CA, 1993.
- [10] Bailar III JC and Mosteller F, editors. *Medical Uses of Statistics*. NEJM Books, Boston, MA, 2nd edition, 1992.
- [11] Webster JG, editor. *Medical Instrumentation: Application and Design*. Wiley, New York, NY, 3rd edition, 1998.
- [12] Kendall M. *Time-Series*. Charles Griffin, London, UK, 2nd edition, 1976.
- [13] Challis RE and Kitney RI. Biomedical signal processing (in four parts): Part 1. Time-domain methods. *Medical and Biological Engineering and Computing*, 28:509–524, 1990.
- [14] Mintchev MP, Stickel A, and Bowes KL. Dynamics of the level of randomness in gastric electrical activity. *Digestive Diseases and Sciences*, 43(5):953–956, 1998.
- [15] Shannon CE. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.
- [16] Shannon CE. Communication in the presence of noise. *Proceedings of the IRE*, 37:10–21, 1949.
- [17] Rangayyan RM. *Biomedical Image Analysis*. CRC Press, Boca Raton, FL, 2005.
- [18] Oppenheim AV and Schafer RW. *Discrete-time Signal Processing*. Pearson, Englewood Cliffs, NJ, third edition, 2010.
- [19] Oppenheim AV and Schafer RW. *Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1975.
- [20] Tustin A. A method of analysing the behaviour of linear systems in terms of time series. *Journal of the Institution of Electrical Engineers-Part IIA: Automatic Regulators and Servo Mechanisms*, 94(1):130–142, 1947.
- [21] Hall EL. *Computer Image Processing and Recognition*. Academic Press, New York, NY, 1979.
- [22] Cooley JW and Tukey JW. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.
- [23] Ahmed N and Rao KR. *Orthogonal Transforms for Digital Signal Processing*. Springer-Verlag, New York, NY, 1975.
- [24] Ahmed N, Natarajan T, and Rao KR. Discrete cosine transform. *IEEE Transactions on Computer*, C-23:90–93, 1974.
- [25] Tompkins WJ. *Biomedical Digital Signal Processing*. Prentice-Hall, Upper Saddle River, NJ, 1995.
- [26] Jenkins JM, Wu D, and Arzbaecher RC. Computer diagnosis of abnormal cardiac rhythms employing a new P-wave detector for interval measurement. *Computers and Biomedical Research*, 11:17–33, 1978.
- [27] Shanks JL. Recursion filters for digital processing. *Geophysics*, 32(1):33–51, 1967.
- [28] Rabiner LR and Gold B. *Theory and Application of Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1975.
- [29] Hamming RW. *Digital Filters*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1983.
- [30] Antoniou A. *Digital Filters: Analysis, Design, and Applications*. McGraw-Hill, New York, NY, 2nd edition, 1993.
- [31] Williams CS. *Designing Digital Filters*. Prentice-Hall, Englewood Cliffs, NJ, 1986.
- [32] Haykin S. *Modern Filters*. Macmillan, New York, NY, 1989.
- [33] Platt RS, Hajduk EA, Hulliger M, and Easton PA. A modified Bessel filter for amplitude demodulation of respiratory electromyograms. *Journal of Applied Physiology*, 84(1):378–388, 1998.
- [34] Little JN and Shure L. *Signal Processing Toolbox for Use with MATLAB*. The MathWorks, Inc., Natick, MA, 1992.
- [35] Chen YS, Lin PY, and Lin YD. A novel PLI suppression method in ECG by notch filtering with a modulation-based detection and frequency estimation scheme. *Biomedical Signal Processing and Control*, 62:102150, 2020.

- [36] Luo S and Johnston P. A review of electrocardiogram filtering. *Journal of Electrocardiology*, 43(6):486–496, 2010.
- [37] Dougherty ER and Astola J. *An Introduction to Nonlinear Image Processing*. SPIE, Bellingham, WA, 1994.
- [38] Pitas I and Venetsanopoulos AN. Order statistics in digital image processing. *Proceedings of the IEEE*, 80:1893–1923, 1992.
- [39] Wiener NE. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series, with Engineering Applications*. MIT Press, Cambridge, MA, 1949.
- [40] Haykin S. *Adaptive Filter Theory*. Prentice-Hall, Upper Saddle River, NJ, 3rd edition, 1996.
- [41] Widrow B, Glover Jr. JR, McCool JM, Kaunitz J, Williams CS, Hearn RH, Zeidler JR, Dong Jr. E, and Goodlin RC. Adaptive noise cancelling: Principles and applications. *Proceedings of the IEEE*, 63(12):1692–1716, 1975.
- [42] Krishnan S. Adaptive filtering, modeling, and classification of knee joint vibroarthrographic signals. Master’s thesis, Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB, Canada, April 1996.
- [43] Riegler R and Compton Jr. R. An adaptive array for interference rejection. *Proceedings of the IEEE*, 61(6):748–758, 1973.
- [44] Zhang YT, Rangayyan RM, Frank CB, and Bell GD. Adaptive cancellation of muscle contraction interference from knee joint vibration signals. *IEEE Transactions on Biomedical Engineering*, 41(2):181–191, 1994.
- [45] Sesay AB. *ENEL 671: Adaptive Signal Processing*. Unpublished lecture notes, Department of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, Canada, 1995.
- [46] Kamath MV, Tougas G, Hollerbach S, Premji R, Fitzpatrick D, Shine G, and Upton ARM. Estimation of habituation and signal-to-noise ratio of cortical evoked potentials to oesophageal electrical and mechanical stimulation. *Medical and Biological Engineering and Computing*, 35:343–347, 1997.
- [47] Ferrara ER and Widrow B. Fetal electrocardiogram enhancement by time-sequenced adaptive filtering. *IEEE Transactions on Biomedical Engineering*, 29(6):458–460, 1982.
- [48] Zarzoso V and Nandi AK. Noninvasive fetal electrocardiogram extraction: Blind separation versus adaptive noise cancellation. *IEEE Transactions on Biomedical Engineering*, 48(1):12–18, 2001.
- [49] Gurve D and Krishnan S. Separation of fetal-ECG from single-channel abdominal ECG using activation scaled non-negative matrix factorization. *IEEE Journal of Biomedical and Health Informatics*, 24(3):669–680, 2019.
- [50] Rangayyan RM, Krishnan S, Bell GD, Frank CB, and Ladly KO. Impact of muscle contraction interference cancellation on vibroarthrographic screening. In *Proceedings of the International Conference on Biomedical Engineering*, pages 16–19, Kowloon, Hong Kong, June 1996.
- [51] Krishnan S and Rangayyan RM. Automatic de-noising of knee-joint vibration signals using adaptive time-frequency representations. *Medical and Biological Engineering and Computing*, 38(1):2–8, 2000.
- [52] Maragos P and Schafer RW. Morphological filters — Part I: Their set-theoretic analysis and relations to linear shift-invariant filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(8):1153–1169, 1987.
- [53] Maragos P and Schafer RW. Morphological filters — Part II: Their relations to median, order-statistic, and stack filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(8):1170–1184, 1987.

CHAPTER 4

DETECTION OF EVENTS

Biomedical signals carry signatures of physiological events. The part of a signal related to a specific event of interest is often referred to as an *epoch*. Analysis of a signal for the purpose of monitoring or diagnosis requires the identification of epochs and investigation of the corresponding events. Once an event has been identified, the corresponding waveform may be segmented and analyzed in terms of its amplitude, waveshape (morphology), time duration, intervals between events, energy distribution, frequency content, and other characteristics. Detection of events is, therefore, an important step in biomedical signal analysis.

4.1 Problem Statement

A generic problem statement applicable to the theme of this chapter may be formulated as follows:

Given a biomedical signal, identify discrete signal epochs and correlate them with events in the related physiological processes.

In the sections to follow, we will first study a few examples of epochs in different biomedical signals, with the aim of understanding the nature of the related physiological events. Such an understanding will help in the subsequent development of signal processing techniques to emphasize, detect, and analyze epochs.

4.2 Illustration of the Problem with Case Studies

The following sections provide illustrations of several events in biomedical signals. The aim of the illustrations is to develop an appreciation of the nature of signal events. A good understanding of signal events will help in designing appropriate signal processing techniques for their detection.

4.2.1 The P, QRS, and T waves in the ECG

As shown in Section 1.2.5, a cardiac cycle is represented in a period of the repetitive ECG signal as the series of waves labeled P, QRS, and T. If we view the cardiac cycle as a series of events, we have the following epochs in an ECG waveform:

- **The P wave:** Contraction of the atria is triggered by the SA-node impulse. The atria do not possess any specialized conduction system as the ventricles do; as such, contraction of the atrial muscles takes place in a slow squeezing manner, with the excitation stimulus being propagated by the muscle cells themselves. For this reason, the P wave is a slow waveform, with a duration of about 80 ms . The P wave amplitude is much smaller (about $0.1 - 0.2\text{ mV}$) than that of the QRS because the atria are smaller than the ventricles. The P wave is the epoch related to the event of atrial contraction. (Atrial relaxation does not produce any distinct waveform in the ECG as it is overshadowed by the following QRS wave.)
- **The PQ segment:** The AV node provides a delay to facilitate completion of atrial contraction and transfer of blood to the ventricles before ventricular contraction is initiated. The resulting PQ segment, of about 80 ms duration, is thus a “nonevent”; however, it is important in recognizing the baseline as the interval is usually isoelectric.
- **The QRS wave:** The specialized system of Purkinje fibers stimulates contraction of ventricular muscles in a rapid sequence from the apex upward. The almost-simultaneous contraction of the entire ventricular musculature results in a sharp and tall QRS complex of about 1 mV amplitude and $80 - 100\text{ ms}$ duration. The event of ventricular contraction is represented by the QRS wave or epoch.
- **The ST segment:** The normally flat (isoelectric) ST segment is related to the plateau in the action potential of the left ventricular muscle cells (see Figures 1.7 and 1.26). The duration of the plateau in the action potential is about 200 ms ; the ST segment duration is usually about $100 - 120\text{ ms}$. All ventricular myocytes are held in their contracted state in this period, and there is no change in the electrical status of the heart’s muscles; therefore, the ECG in a differential lead is zero. As in the case of the PQ segment, the ST segment may also be called a nonevent. However, myocardial ischemia or infarction could change the action potentials of a portion of the left ventricular musculature and cause the ST segment to be depressed (see Figure 1.50) or elevated. The PQ segment serves as a useful reference when the isoelectric nature of the ST segment needs to be verified.
- **The T wave:** The T wave appears in a normal ECG signal as a discrete wave separated from the QRS by an isoelectric ST segment. However, it relates to the last phase of the action potential of ventricular muscle cells, when the potential returns from the plateau of the depolarized state to the resting potential through the process of repolarization [1]. The T wave is commonly referred to as the wave corresponding to ventricular relaxation. While this is correct, it should be noted that relaxation through repolarization is simply the final phase of contraction: Contraction and relaxation are indicated by the upstroke and downstroke of the same action potential. For this reason, the T wave may be said to relate to a nonspecific event.

The T wave is elusive, being low in amplitude ($0.1 - 0.3\text{ mV}$) and being a slow wave extending over $120 - 160\text{ ms}$. It is almost absent in many ECG recordings. Rather than attempt to detect

the often obscure T wave, one may extract a segment of the ECG 80 – 360 ms from the beginning of the QRS and use it to represent the ST segment and the T wave.

4.2.2 The first and second heart sounds

We observed in Section 1.2.9 that the normal cardiac cycle manifests as a series of the first and second heart sounds — S1 and S2. Murmurs and additional sounds may appear in the presence of cardiovascular diseases or defects. We concentrate on S1, S2, and murmurs only.

- **The first heart sound S1:** S1 reflects a sequence of events related to ventricular contraction — closure of the atrioventricular valves, isovolumic contraction, opening of the semilunar valves, and ejection of the blood from the ventricles [1]. The epoch of S1 is directly related to the event of ventricular contraction.
- **The second heart sound S2:** S2 is related to the end of ventricular contraction, signified by closure of the aortic and pulmonary valves. As we observed in the case of the T wave, the end of ventricular contraction cannot be referred to as a specific event *per se*. However, in the case of S2, we do have the specific events of closure of the aortic and pulmonary valves to relate to, as indicated by the corresponding A2 and P2 components of S2. Unfortunately, separate identification of A2 and P2 is confounded by the fact that they usually overlap in normal signals. If A2 and P2 are separated due to a cardiovascular disorder, simultaneous multisite PCG recordings will be required to identify each component definitively as they may be reversed in order (see Tavel [2] and Rushmer [1]).
- **Murmurs:** Murmurs, if present, could be viewed as specific events. For example, the systolic murmur of aortic stenosis relates to the event of turbulent ejection of blood from the left ventricle through a restricted aortic valve opening. The diastolic murmur in the case of aortic insufficiency corresponds to the event of regurgitation of blood from the aorta back into the left ventricle through a leaky aortic valve.

4.2.3 The dicrotic notch in the carotid pulse

As we saw in Sections 1.2.10 and 1.2.12, closure of the aortic valve causes a sudden drop in aortic pressure that is already on a downward slope at the end of ventricular systole. The dicrotic notch inscribed in the carotid pulse is a delayed, upstream manifestation of the incisura in the aortic pressure wave. The dicrotic notch is a specific signature on the relatively nondescript carotid pulse signal, and may be taken as an epoch related to the event of aortic valve closure (albeit with a time delay); the same event also signifies the end of ventricular systole and ejection as well as the beginning of S2 and diastole.

4.2.4 EEG rhythms, waves, and transients

We have already studied a few basic characteristics of the EEG in Section 1.2.6 and noted the nature of the α , β , δ , and θ waves. We now consider a few events and transients that occur in EEG signals [3–8]. Figure 4.1 shows typical manifestations of the activities described below [3].

- **K-complex:** This is a transient complex waveform with slow waves, sometimes associated with sharp components, and often followed by 14 Hz waves. It occurs spontaneously or in response to a sudden stimulus during sleep, with an amplitude of about 200 μV .
- **Lambda waves:** These are monophasic, positive, sharp waves that occur in the occipital location with an amplitude of less than 50 μV . They are related to eye movement and are associated with visual exploration.

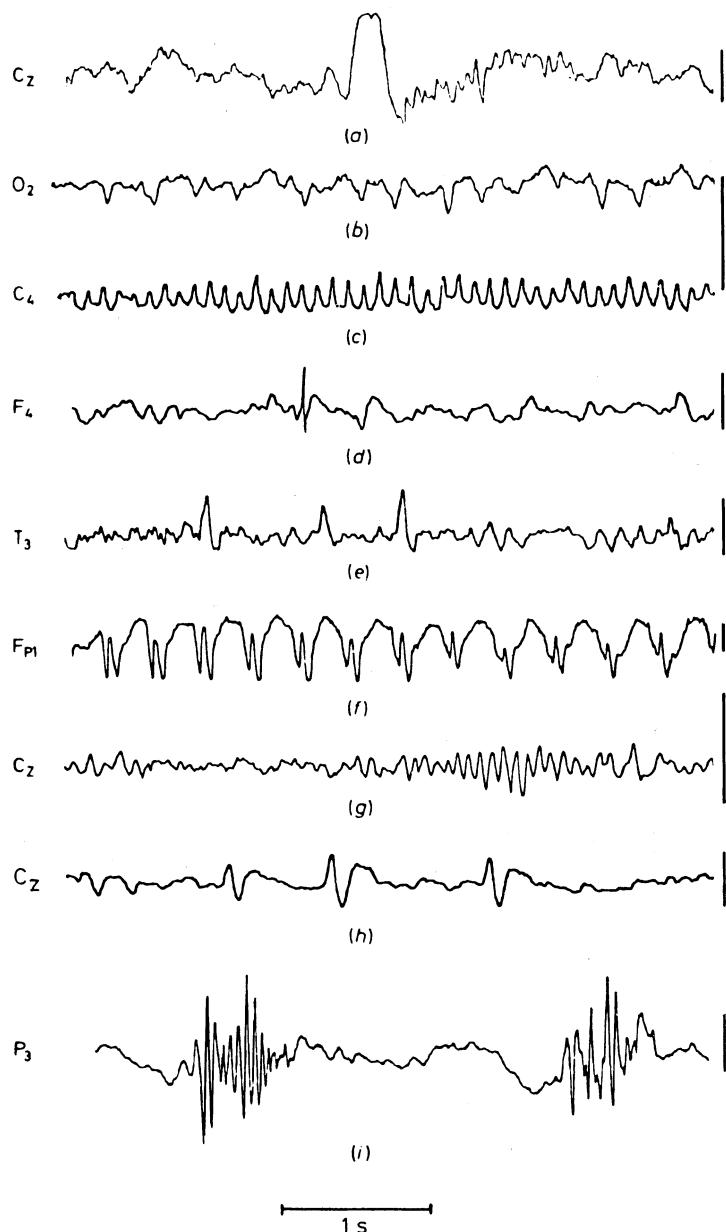


Figure 4.1 Top to bottom: (a) the K-complex; (b) the lambda wave; (c) the mu rhythm; (d) a spike; (e) sharp waves; (f) spike-and-wave complexes; (g) a sleep spindle; (h) vertex sharp waves; and (i) polyspike discharges. The horizontal bar at the bottom indicates a duration of 1 s; the vertical bars at the right indicate 100 μ V. Reproduced with permission from R. Cooper, J.W. Osselton, and J.C. Shaw, *EEG Technology*, 3rd edition, 1980. ©Butterworth Heinemann Publishers, a division of Reed Educational & Professional Publishing Ltd., Oxford, UK.

- **Mu rhythm:** This rhythm appears as a group of waves in the frequency range of $7 - 11 \text{ Hz}$ with an arcade or comb shape in the central location. The mu rhythm usually has an amplitude of less than $50 \mu\text{V}$, and is blocked or attenuated by contralateral movement, thought of movement, readiness to move, or tactile stimulation.
- **Spike:** A spike is defined as a transient with a pointed peak, having a duration in the range of $20 - 30 \text{ ms}$.
- **Sharp wave:** A sharp wave is also a transient with a pointed peak, but with a longer duration than a spike, in the range of $70 - 200 \text{ ms}$.
- **Spike-and-wave rhythm:** A sequence of surface-negative slow waves in the frequency range of $2.5 - 3.5 \text{ Hz}$ and having a spike associated with each wave is referred to as a spike-and-wave rhythm. There could be several spikes of amplitude up to $1,000 \mu\text{V}$ in each complex, in which case the rhythm is called a polyspike-and-wave complex.
- **Sleep spindle:** This is an episodic rhythm at about 14 Hz and $50 \mu\text{V}$, occurring maximally over the frontocentral regions during certain stages of sleep. A spindle is defined, in general, as a short sequence of monomorphic waves having a fusiform appearance [4].
- **Vertex sharp transient or V-wave:** This wave represents a sharp potential that is maximal at the vertex at about $300 \mu\text{V}$ and is negative in relation to the EEG in other areas. It occurs spontaneously during sleep or in response to a sensory stimulus during sleep or wakefulness.

In addition to the above, the term “burst” is used to indicate a phenomenon composed of two or more waves that are different from the principal (background) activity in terms of amplitude, frequency, or waveform. A burst is abrupt and has a relatively short duration [4].

An EEG record is described in terms of [3]

- the most persistent rhythm (for example, α);
- the presence of other rhythmic features, such as δ , θ , or β ;
- discrete features of relatively long duration, such as an episode of spike-and-wave activity;
- discrete features of relatively short duration, such as isolated spikes or sharp waves;
- the activity remaining when all the previous features have been described, referred to as background activity; and
- artifacts, if any, giving rise to ambiguity in interpretation.

Each of the EEG waves or activities is described in chronological sequence in terms of amplitude; frequency, in the case of rhythmic features; waveform, in the case of both rhythmic and transient features; location or spatial distribution on the scalp; incidence or temporal variability; the presence or absence of right-left symmetry in the location of activity; and responsiveness to stimuli, such as eye opening and closure. The EEG record at rest is first described as above; effects of evocative techniques are then specified in the same terms. Behavioral changes, such as the subject becoming drowsy or falling asleep, are also noted [3].

The EEG signals in Figure 1.41 demonstrate the presence of the α rhythm in all the channels. The EEG signals in Figure 1.42 depict spike-and-wave complexes in almost all the channels.

4.3 Detection of Events and Waves

We now see how the knowledge that we have gained so far of several biomedical signal events may be applied to develop signal processing techniques for their detection. Each of the following sections deals with the problem of detection of a specific type of event. The techniques described should find applications in the detection of other events of comparable characteristics.

4.3.1 Derivative-based methods for QRS detection

Problem: Develop signal processing techniques to facilitate detection of the QRS complex, given that it is the sharpest wave in an ECG cycle.

Solution 1: We noted in Section 1.2.5 that the QRS complex has the largest slope (rate of change of voltage) in a cardiac cycle by virtue of the rapid conduction and depolarization characteristics of the ventricles. As the rate of change is given by the derivative operator, the $\frac{d}{dt}$ operation would be the most logical starting point in an attempt to develop an algorithm to detect the QRS complex.

We saw in Section 3.6.2 that the derivative operator enhances the QRS, although the resulting wave does not bear any resemblance to a typical QRS complex. Observe in Figures 3.56 and 3.57 that the slow P and T waves have been suppressed by the derivative operators, while the output is the highest at the QRS. However, given the noisy nature of the results of the derivative-based operators, it is also evident that significant smoothing will be required before further processing can take place.

Balda et al. [9] proposed a derivative-based algorithm for QRS detection, which was further studied and evaluated by Ahlstrom and Tompkins [10], Friesen et al. [11], and Tompkins [12]. The algorithm progresses as follows. In a manner similar to Equation 3.128, the smoothed three-point first derivative $y_0(n)$ of the given signal $x(n)$ is approximated as

$$y_0(n) = |x(n) - x(n - 2)|. \quad (4.1)$$

The second derivative is approximated as

$$y_1(n) = |x(n) - 2x(n - 2) + x(n - 4)|. \quad (4.2)$$

The two results are weighted and combined to obtain

$$y_2(n) = 1.3y_0(n) + 1.1y_1(n). \quad (4.3)$$

The result $y_2(n)$ is scanned with a threshold of 1.0. Whenever the threshold is crossed, the subsequent eight samples are also tested against the same threshold. If at least six of the eight points pass the threshold test, the segment of eight samples is taken to be a part of a QRS complex. The procedure results in a pulse with its width proportional to that of the QRS complex; however, the method is sensitive to noise.

Illustration of application: Figure 4.2 illustrates, in the topmost trace, two cycles of a filtered version of the ECG signal shown in Figure 3.5. The signal was filtered with an eighth-order Butterworth lowpass filter with $f_c = 90 \text{ Hz}$, down-sampled by a factor of 5, and filtered with a notch filter with $f_o = 60 \text{ Hz}$. The effective sampling rate is 200 Hz. The signal was normalized by dividing by its maximum value.

The second and third plots in Figure 4.2 show the derivatives $y_0(n)$ and $y_1(n)$, respectively; the fourth plot illustrates the combined result $y_2(n)$. Observe the relatively high values in the derivative-based results at the QRS locations; the outputs are low or negligible at the P and T wave locations, in spite of the fact that the original signal possesses an unusually sharp and tall T wave. It is also seen that the results have multiple peaks over the duration of the QRS wave, due to the fact that the QRS complex includes three major swings: Q–R, R–S, and S–ST baseline in the present example (an additional PQ baseline–Q swing may also be present in other ECG signals).

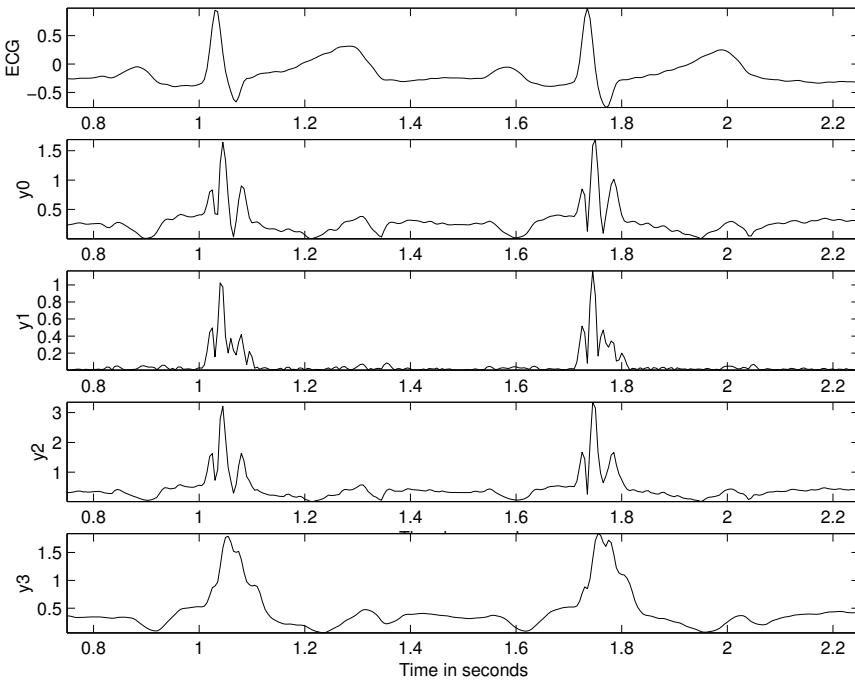


Figure 4.2 From top to bottom: two cycles of a filtered version of the ECG signal shown in Figure 3.5; output $y_0(n)$ of the first-derivative-based operator in Equation 4.1; output $y_1(n)$ of the second-derivative-based operator in Equation 4.2; the combined result $y_2(n)$ from Equation 4.3; and the result $y_3(n)$ of passing $y_2(n)$ through the 8-point MA filter in Equation 3.108.

The last plot in Figure 4.2 shows the smoothed result $y_3(n)$ obtained by passing $y_2(n)$ through the 8-point MA filter in Equation 3.108. We now have a single pulse with amplitude greater than 1.0 over the duration of the corresponding QRS complex. A simple peak-searching algorithm may be used to detect each ECG beat. The net delay introduced by the filters should be subtracted from the detected peak location in order to obtain the corresponding QRS location.

Note that peak searching cannot be performed directly on an ECG signal: The QRS might not always be the highest wave in a cardiac cycle, and artifacts may easily upset the search procedure. Observe also that the ECG signal in the present illustration was filtered to a restricted bandwidth of 90 Hz before the derivatives were computed, and that it is free of baseline drift.

Solution 2: Murthy and Rangaraj [13] proposed a QRS detection algorithm based upon a weighted and squared first-derivative operator and an MA filter. In this method, a filtered-derivative operator was defined as

$$g_1(n) = \sum_{i=1}^N |x(n-i+1) - x(n-i)|^2(N-i+1), \quad (4.4)$$

where $x(n)$ is the ECG signal, and N is the width of a window within which first-order differences are computed, squared, and weighted by the factor $(N-i+1)$. The weighting factor decreases linearly from the current difference to the difference N samples earlier in time and provides a smoothing effect. Further smoothing of the result was performed by an MA filter over M points to obtain

$$g(n) = \frac{1}{M} \sum_{j=0}^{M-1} g_1(n-j). \quad (4.5)$$

With the sampling rate of 100 Hz, the filter window widths were set as $M = N = 8$ samples. The algorithm provides a single peak for each QRS complex and suppresses P and T waves.

Searching for the peak in a processed signal such as $g(n)$ may be accomplished by a simple peak-searching algorithm as follows:

1. Scan a portion of the signal $g(n)$ that may be expected to contain a peak and determine the maximum value g_{\max} , or let g_{\max} be the maximum of $g(n)$ over its entire available duration.
2. Define a threshold as a fraction of the maximum, for example, $Th = 0.5 g_{\max}$.
3. For all $g(n) > Th$, select those samples for which the corresponding $g(n)$ values are greater than a certain predefined number M of preceding and succeeding samples of $g(n)$, that is,

$$\{p\} = \{n \mid \begin{aligned} & [g(n) > Th] \text{ AND} \\ & [g(n) > g(n-i), i = 1, 2, \dots, M] \text{ AND} \\ & [g(n) > g(n+i), i = 1, 2, \dots, M]\}. \end{aligned} \quad (4.6)$$

The set $\{p\}$ defined as above contains the indices of the peaks in $g(n)$.

Additional conditions may be imposed to reject peaks due to artifacts, such as a minimum interval between two adjacent peaks. A more elaborate peak-searching algorithm is described in Section 4.3.2.

Illustration of application: Figure 4.3 illustrates, in the topmost trace, two cycles of a filtered version of the ECG signal shown in Figure 3.5. The signal was filtered with an eighth-order Butterworth lowpass filter with $f_c = 40$ Hz and was down-sampled by a factor of 10. The effective sampling rate is 100 Hz to match the parameters used by Murthy and Rangaraj [13]. The signal was normalized by dividing by its maximum value.

The second and third plots in Figure 4.3 show the outputs of the derivative-based operator and the smoothing filter. Observe that the final output contains a single, smooth peak for each QRS, and that the P and T waves produce no significant output. A simple peak-searching algorithm may be used to detect and segment each beat [13]. The net delay introduced by the filter used should be subtracted from the detected peak location in order to obtain the corresponding QRS location.

4.3.2 The Pan–Tompkins algorithm for QRS detection

Problem: Propose a real-time algorithm to detect QRS complexes in an ongoing ECG signal.

Solution: Pan and Tompkins [14] (see also Tompkins [12]) proposed a real-time QRS detection algorithm based on analysis of the slope, amplitude, and width of QRS complexes. The algorithm includes a series of filters and methods that perform lowpass, highpass, derivative, squaring, integration, adaptive thresholding, and search procedures. Figure 4.4 illustrates the steps of the algorithm in schematic form.

Lowpass filter: The recursive lowpass filter used in the Pan–Tompkins algorithm has integers as its coefficients to reduce computational complexity, with the transfer function defined as

$$H(z) = \frac{1}{32} \frac{(1 - z^{-6})^2}{(1 - z^{-1})^2}. \quad (4.7)$$

(See also Equations 3.120 and 3.121.) The output $y(n)$ is related to the input $x(n)$ as

$$y(n) = 2y(n-1) - y(n-2) + \frac{1}{32} [x(n) - 2x(n-6) + x(n-12)]. \quad (4.8)$$

With the sampling rate being 200 Hz, the filter has a low cutoff frequency of $f_c = 11$ Hz and introduces a delay of 5 samples or 25 ms. The filter provides an attenuation greater than 35 dB at 60 Hz and effectively suppresses power-line interference, if present.

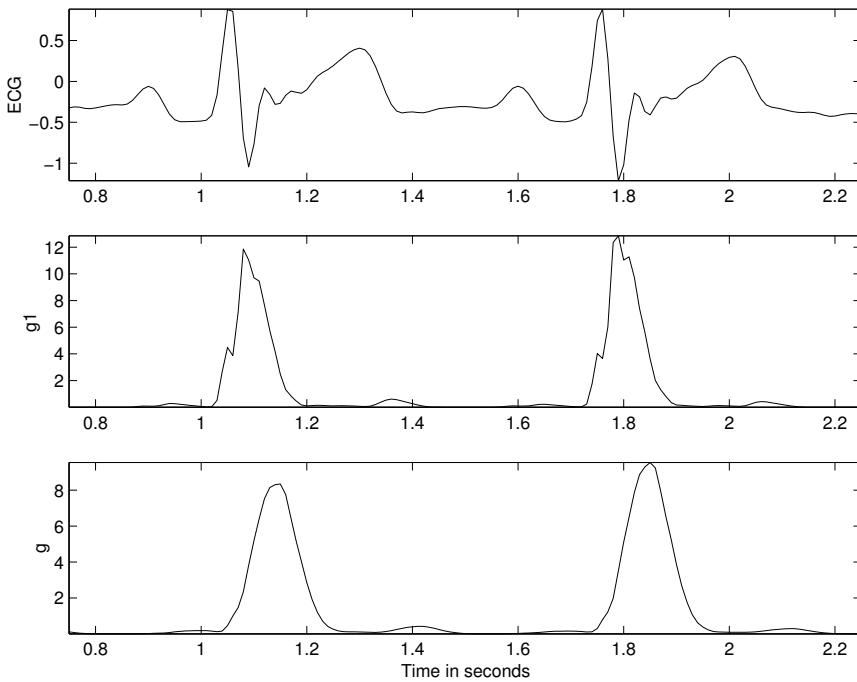


Figure 4.3 From top to bottom: two cycles of a filtered version of the ECG signal shown in Figure 3.5; output $g_1(n)$ of the weighted and squared first-derivative operator in Equation 4.4; output $g(n)$ of the smoothing filter in Equation 4.5.

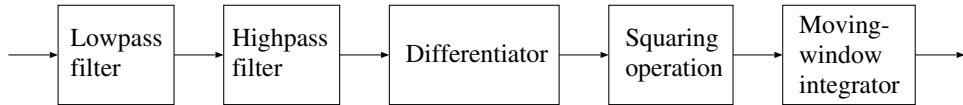


Figure 4.4 Block diagram of the Pan–Tompkins algorithm for QRS detection.

Highpass filter: The highpass filter used in the algorithm is implemented as an allpass filter minus a lowpass filter. The lowpass component has the transfer function

$$H_{lp}(z) = \frac{(1 - z^{-32})}{(1 - z^{-1})}; \quad (4.9)$$

the input–output relationship is

$$y(n) = y(n - 1) + x(n) - x(n - 32). \quad (4.10)$$

The transfer function $H_{hp}(z)$ of the highpass filter is specified as

$$H_{hp}(z) = z^{-16} - \frac{1}{32} H_{lp}(z). \quad (4.11)$$

Equivalently, the output $p(n)$ of the highpass filter is given by the difference equation

$$p(n) = x(n - 16) - \frac{1}{32} [y(n - 1) + x(n) - x(n - 32)], \quad (4.12)$$

with $x(n)$ and $y(n)$ being related as in Equation 4.10. If all of the operations from Equation 4.9 to Equation 4.12 are combined, the input–output relationship of the overall highpass filter is

$$p(n) = p(n - 1) - \frac{1}{32}x(n) + x(n - 16) - x(n - 17) + \frac{1}{32}x(n - 32). \quad (4.13)$$

The highpass filter has a cutoff frequency of 5 Hz and introduces a delay of 80 ms for $f_s = 200$ Hz.

Derivative operator: The derivative operation used by Pan and Tompkins is specified as

$$y(n) = \frac{1}{8} [2x(n) + x(n - 1) - x(n - 3) - 2x(n - 4)] \quad (4.14)$$

and approximates the ideal $\frac{d}{dt}$ operator up to 30 Hz. The derivative procedure suppresses the low-frequency components of the P and T waves, and provides a large gain to the high-frequency components arising from the high slopes of the QRS complex. (See Section 3.6.2 for details on the properties of derivative-based filters.)

Figure 4.5 shows the magnitude of the frequency response for each of the filters in the Pan–Tompkins algorithm for QRS detection described to this point. Observe the response of the highpass filter that was derived from a lowpass filter. The combined procedure has a bandpass nature that is evident from the final response shown in the figure.

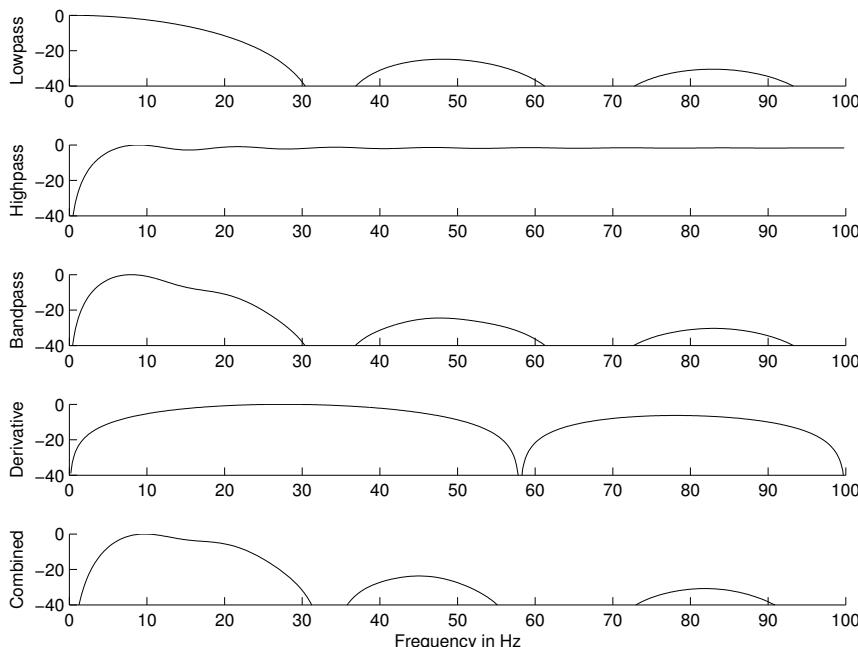


Figure 4.5 Frequency response (magnitude, in dB) of the filters used in the Pan–Tompkins algorithm for QRS detection. Top to bottom: the initial lowpass filter, the highpass filter, the bandpass filter resulting from the combination of the two filters, the derivative operator, and the combination of the bandpass filter and the derivative operator. The bandpass nature of the combined procedure is evident from the final response.

Squaring: The squaring operation makes the result positive and emphasizes large differences resulting from QRS complexes; the small differences arising from P and T waves are suppressed. The high-frequency components in the signal related to the QRS complex are further enhanced.

Integration: As observed in Section 4.3.1, the output of a derivative-based operation will exhibit multiple peaks within the duration of a single QRS complex. The Pan–Tompkins algorithm performs

smoothing of the output of the preceding operations through a moving-window integration filter as

$$y(n) = \frac{1}{N} \{x[n - (N - 1)] + x[n - (N - 2)] + \dots + x(n)\}. \quad (4.15)$$

The choice of the window width N is to be made with the following considerations: too large a value will result in the outputs due to the QRS and T waves being merged, whereas too small a value could yield multiple peaks for a single QRS. A window width of $N = 30$ samples was found to be suitable for $f_s = 200 \text{ Hz}$. Figure 4.6 illustrates the effect of the window width on the output of the integrator and its relationship to the QRS width. (See Section 3.6.1 for details on the properties of MA and integrating filters.)

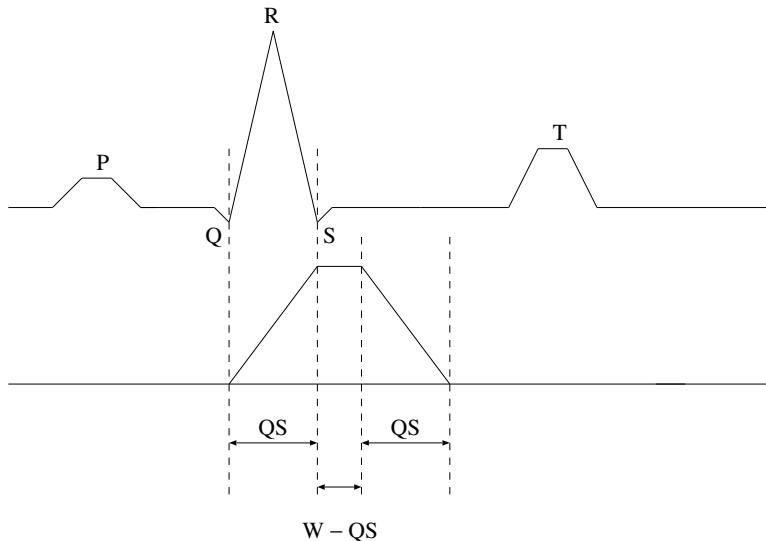


Figure 4.6 The relationship of a QRS complex to the moving-window integrator output. Upper plot: Schematic ECG signal. Lower plot: Schematic representation of the output of the moving-window integrator. QS: QRS complex width. W: width of the integrator window, given as $N/f_s \text{ s}$. It is assumed that the QRS complex will provide a result in the form of a rectangular pulse of width QS before the integrator; then, the output of the integrator is a trapezoid as shown, with a total width of $W+QS$, assuming that $W > QS$. Adapted from Tompkins [12].

Adaptive thresholding: The thresholding procedure in the Pan–Tompkins algorithm adapts to changes in the ECG signal by computing running estimates of signal and noise peaks. A peak is said to be detected whenever the final output $y(n)$ changes direction within a specified time interval. In the following discussion, $SPKI$ represents the peak level that the algorithm has learned to correspond to QRS peaks, and $NPKI$ represents the peak level related to non-QRS events (such as noise, EMG, and various artifacts). $THRESHOLD\ I1$ and $THRESHOLD\ I2$ are two thresholds used to categorize peaks detected as signal (QRS) or noise.

Every new peak detected is categorized as a signal peak or a noise peak. If a peak exceeds $THRESHOLD\ I1$ during the first step of analysis, it is classified as a QRS (signal) peak. If the search-back technique (described in the following paragraph) is used, the peak should be above $THRESHOLD\ I2$ to be called a QRS. The peak levels and thresholds are updated after each peak is detected and classified as

$$\begin{aligned} SPKI &= 0.125 PEAKI + 0.875 SPKI && \text{if } PEAKI \text{ is a signal peak;} \\ NPKI &= 0.125 PEAKI + 0.875 NPKI && \text{if } PEAKI \text{ is a noise peak;} \end{aligned} \quad (4.16)$$

$$\begin{aligned} \text{THRESHOLD } I1 &= NPKI + 0.25(SPKI - NPKI); \\ \text{THRESHOLD } I2 &= 0.5 \text{ THRESHOLD } I1. \end{aligned} \quad (4.17)$$

The updating formula for $SPKI$ is changed to

$$SPKI = 0.25 PEAKI + 0.75 SPKI \quad (4.18)$$

if a QRS is detected in the search-back procedure using $\text{THRESHOLD } I2$.

Search-back procedure: The Pan–Tompkins algorithm maintains two averages of RR intervals: $RR \text{ AVERAGE}1$ is the average of the eight most-recent beats, and $RR \text{ AVERAGE}2$ is the average of the eight most-recent beats having RR intervals within the range specified by $RR \text{ LOW LIMIT} = 0.92 \times RR \text{ AVERAGE}2$ and $RR \text{ HIGH LIMIT} = 1.16 \times RR \text{ AVERAGE}2$. Whenever a QRS is not detected for a certain interval specified as $RR \text{ MISSED LIMIT} = 1.66 \times RR \text{ AVERAGE}2$, the QRS is taken to be the peak between the established thresholds applied in the search-back procedure.

The Pan–Tompkins algorithm performed with a low error rate of 0.68%, or 33 beats per hour on a database of about 116,000 beats obtained from 24-hour records of the ECGs of 48 patients (see Tompkins [12] for details).

If N_B QRS complexes or beats are detected in an ECG signal over a duration of T s, we have

$$HR = 60 \frac{N_B}{T} \quad (4.19)$$

in bpm (on the average). If the average RR interval measured over several beats is RR_a s, we have

$$HR = \frac{60}{RR_a} \quad (4.20)$$

in bpm (on the average). The equation given above may also be used with the beat-to-beat RR interval instead of RR_a to obtain the instantaneous heart rate. In a practical ECG monitoring system, it would be desirable to derive the average heart rate and update its display a few times per minute, such as every 10 s.

Illustration of application: Figure 4.7 illustrates, in the topmost trace, the same ECG signal as in Figure 4.2. The Pan–Tompkins algorithm as above was implemented. The outputs of the various stages of the algorithm are illustrated in sequence in the same figure. The observations to be made are similar to those in the preceding section on the derivative-based method. The derivative operator suppresses the P and T waves and provides a large output at the QRS locations. The squaring operation preferentially enhances large values and boosts high-frequency components. The result still possesses multiple peaks for each QRS and hence needs to be smoothed. The final output of the integrator is a single smooth pulse for each QRS. Observe the shift between the actual QRS location and the pulse output due to the cumulative delays of the various filters. The thresholding and search procedures and their results are not illustrated. More examples of QRS detection are presented in Sections 4.8 and 4.9.

4.3.3 Detection of the P wave in the ECG

Detection of the P wave in the ECG is difficult, because the P wave typically has a low amplitude, has an ill-defined and variable shape, and could be placed in a background of noise and baseline fluctuation of varying size and origin. Quite often, the P wave may not be visible in several ECG leads.

Problem: Propose an algorithm to detect the P wave in the ECG signal, given that it typically occurs before the QRS wave.

Solution: In the method proposed by Hengeveld and van Bemmel [15], VCG signals are processed as follows:

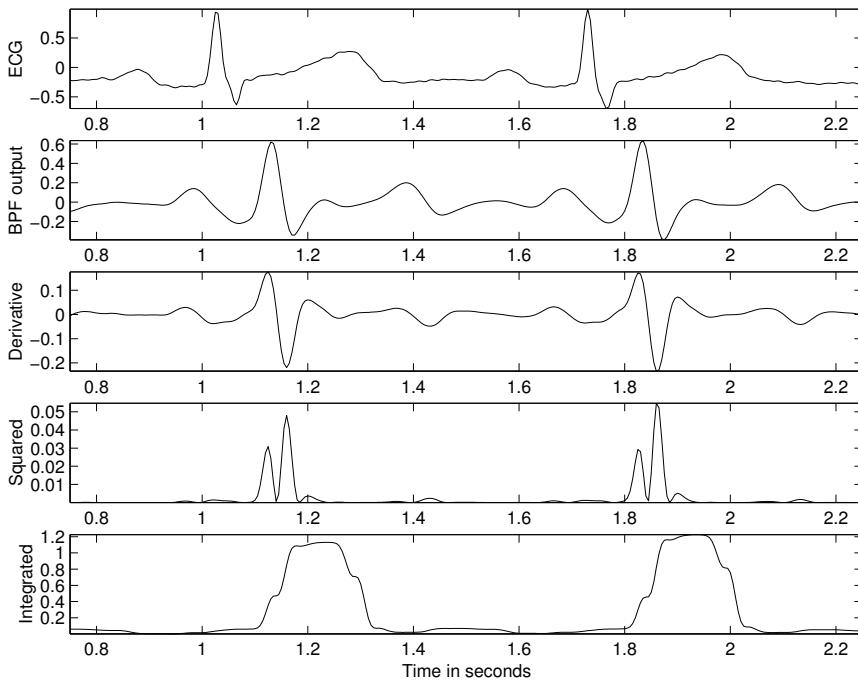


Figure 4.7 Results of the Pan–Tompkins algorithm. From top to bottom: two cycles of a filtered version of the ECG signal shown in Figure 3.5 (the same as that in Figure 4.2); output of the bandpass filter (BPF, a combination of lowpass and highpass filters); output of the derivative-based operator; the result of squaring; and $100 \times$ the result of the final integrator.

1. The QRS is detected, deleted, and replaced with the baseline. The baseline is determined by analyzing a few samples preceding the QRS complex.
2. The resulting signal is bandpass filtered with -3 dB points at 3 Hz and 11 Hz .
3. The end of the preceding T wave is estimated with reference to the current QRS by using $QT_{\max} = \frac{2}{9}RR + 250 \text{ ms}$, where RR is the interval between two successive QRS complexes.
4. The maximum and minimum values are found in all three VCG leads in the search interval from the preceding T wave to the current QRS.
5. The signal is rectified and thresholded at 50% and 75% of the maximum to obtain a ternary (three-level) signal.
6. The cross-correlation of the result is computed with a ternary template derived in a manner similar to the procedure in the previous step from a representative set of P waves.
7. The peak in the cross-correlation corresponds to the P location in the original ECG.

The algorithm overcomes the dominance of the QRS complex by first detecting the QRS and then deleting it. Observe that the cross-correlation is computed not with an original P wave, which we have noted could be obscure and variable, but with a ternary wave derived from the P wave. The ternary wave represents a simplified template of the P wave.

Figure 4.8 illustrates the results of the various stages of the P-finding algorithm of Hengeveld and van Bemmel [15]. Observe that the original ECG signal shown in part (a) of the figure has a P wave

that is hardly discernible. The processed versions of the signal after deleting the QRS, filtering, and rectification are shown in parts (b), (c), and (d). The ternary version in part (e) shows that the P wave has been converted into two pulses corresponding to its upstroke and return parts. The result of cross-correlation with the template in part (f) is shown in part (g). A simple peak-picking algorithm with search limits may be used to detect the peak in the result, and hence determine the P wave position.

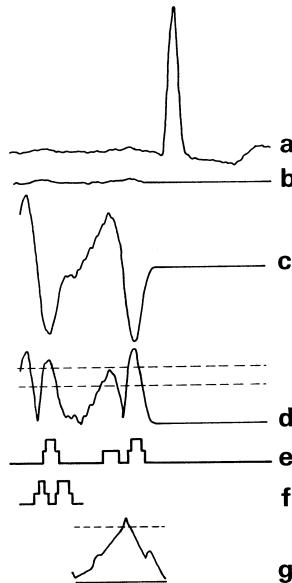


Figure 4.8 Illustration of the results at various stages of the Hengeveld and van Bemmel method for P wave detection. From top to bottom: (a) the original ECG signal; (b) after replacement of the QRS with the baseline; (c) after bandpass filtering; (d) after rectification, with the dashed lines indicating the thresholds; (e) the thresholded ternary signal; (f) the ternary P wave template; and (g) result of cross-correlation between the signals in (e) and (f). Reproduced with permission from S.J. Hengeveld and J.H. van Bemmel, Computer detection of P waves, *Computers and Biomedical Research*, 9:125–132, 1976. ©Academic Press.

Note that the result in part (d) has other waves preceding those related to the P wave. An appropriate search interval should be used so as to disregard the unwanted components.

It should also be noted that the P wave is typically absent in the case of PVCs; dissociated P waves could be present in the case of atrioventricular dissociation; and more P waves than QRS waves would be present in the case of atrial flutter. In such cases, other approaches would be required to detect P waves without regard to QRS waves or to detect atrial activity (P or A waves; see Section 2.2.3) independent of the detection of ventricular activity (QRS waves).

4.3.4 Detection of the T wave in the ECG

Detection of the T wave in the ECG is difficult, because the T wave has a substantially variable amplitude, has a variable shape with possible inversion, and could be placed in a background of noise of varying size and origin. Quite often, the T wave may not be visible in several ECG leads.

Problem: Propose an algorithm to detect the T wave in the ECG signal, given that it occurs after the QRS wave.

Solution: Gritzali et al. [16] proposed a common approach to detect the QRS, T, and P waves in multichannel ECG signals based on a transformation they labeled as the “length” transformation. Given a collection of ECG signals from N simultaneous channels $x_1(t), x_2(t), \dots, x_N(t)$, the

length transformation was defined as

$$L(N, w, t) = \int_t^{t+w} \sqrt{\sum_{j=1}^N \left(\frac{dx_j}{dt} \right)^2} dt, \quad (4.21)$$

where w is the width of the time window over which the integration is performed. In essence, the procedure computes the total squared derivative of the signals across the various channels available, and integrates the summed quantity over a moving time window. The advantage of applying the derivative-based operator across multiple channels of an ECG signal is that P and T waves may be detected even when they are well defined in only one or a few of the channels used.

In the procedure for waveform detection proposed by Gritzali et al., the QRS is first detected by applying a threshold to $L(N, w, t)$, with w set equal to the average QRS width. The onset (start) and offset (end) points of the QRS are represented by a pulse waveform, as indicated in Figure 4.9. The QRS complexes in the signals are then replaced by the isoelectric baseline of the signals, the procedure is repeated with w set equal to the average T duration, and the T waves are detected. The same steps are repeated to detect the P waves. Figure 4.9 illustrates the detection of the QRS, T, and P waves in a three-channel ECG signal. Gritzali et al. also proposed a procedure based on correlation analysis and least-squares modeling to determine the thresholds required.

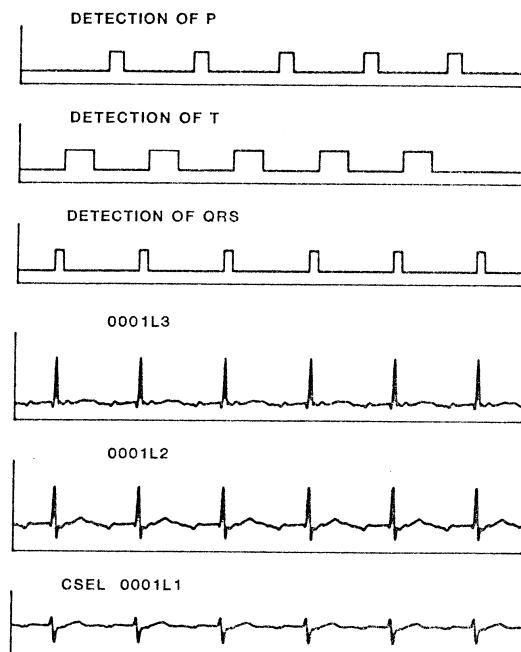


Figure 4.9 Detection of the P, QRS, and T waves in a three-channel ECG signal using the length transformation. The lower three traces show the three ECG channels. The upper three traces indicate the onset and end of the P, QRS, and T waves detected by the procedure in the form of pulse trains. The first P and the last T waves have not been processed. CSEL: CSE library. L1/L2/L3: three ECG leads. Reproduced with permission from F. Gritzali, G. Frangakis, and G. Papakonstantinou, Detection of the P and T waves in an ECG, *Computers and Biomedical Research*, 22:83–91, 1989. ©Academic Press. See Willems et al. [17, 18] for details on the ECG database used by Gritzali et al.

It should be noted that ectopic beats do not typically display a distinct and separate T wave.

4.3.5 Detection of the dicrotic notch

Problem: Propose a method to detect the dicrotic notch in the carotid pulse signal.

Solution: Lehner and Rangayyan [19] proposed a method for detection of the dicrotic notch that used the least-squares estimate of the second derivative $p(n)$ of the carotid pulse signal $y(n)$, defined as

$$p(n) = 2y(n-2) - y(n-1) - 2y(n) - y(n+1) + 2y(n+2). \quad (4.22)$$

Observe that this expression is noncausal; it may be made causal by applying a delay of two samples.

The second derivative was used due to the fact that the dicrotic notch appears as a short wave riding on the downward slope of the carotid pulse signal (see also Starmer et al. [20]). A first-derivative operation would give an almost-constant output for the downward slope. The second-derivative operation removes the effect of the downward slope and enhances the notch itself. The result was squared and smoothed to obtain

$$s(n) = \sum_{k=1}^M p^2(n-k+1)w(k), \quad (4.23)$$

where $w(k) = (M - k + 1)$, $k = 1, 2, \dots, M$, is a linear weighting function, and $M = 16$ samples for $f_s = 256 \text{ Hz}$.

The method yields two peaks for each period of the carotid pulse signal. The first peak in the result represents the onset of the carotid upstroke. The second peak that appears in the result within a cardiac cycle is related to the dicrotic notch. To locate the dicrotic notch, the local minimum in the carotid pulse within a $\pm 20 \text{ ms}$ interval of the second peak was used.

Illustration of application: The upper plot in Figure 4.10 illustrates two cycles of a carotid pulse signal. The signal was lowpass filtered at 100 Hz and sampled at 250 Hz . The result of application of the Lehner and Rangayyan method to the signal is shown in the lower plot. It is evident that the second derivative has successfully accentuated the dicrotic notch. A simple peak-searching algorithm may be used to detect the first and second peaks in the result. The dicrotic notch may then be located by searching for the minimum in the carotid pulse signal within a $\pm 20 \text{ ms}$ interval around the second peak location.

Observe that the result illustrated in Figure 4.10 may benefit from further smoothing by increasing the window width M in Equation 4.23. The window width needs to be chosen in accordance with the characteristics of the signal on hand as well as the lowpass filter and sampling rate used. Further illustration of the detection of the dicrotic notch is provided in Section 4.9.

4.4 Correlation Analysis of EEG Rhythms

EEG signals are usually acquired simultaneously over multiple channels. Event detection in and epoch analysis of EEG signals becomes more complicated than the problems we have seen with the single-channel ECG and carotid pulse signals, due to the need to detect similar or related events across multiple channels. Autocorrelation and cross-correlation techniques in both the time and frequency domains serve such needs.

4.4.1 Detection of EEG rhythms

Problem: Propose a method to detect the presence of the α rhythm in an EEG channel. How would you extend the method to detect the presence of the same rhythm simultaneously in two EEG channels?

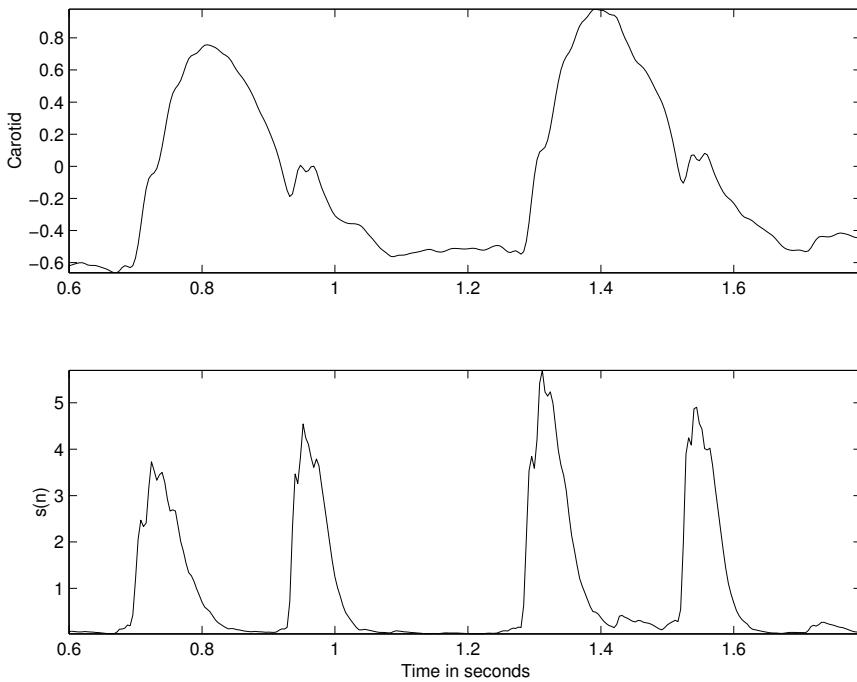


Figure 4.10 Two cycles of a carotid pulse signal and the result of the Lehner and Rangayyan method for detection of the dicrotic notch.

Solution: Two signals may be compared to detect the characteristics present in common between them via their dot product (also known as the inner or scalar product), defined as

$$x \cdot y = \langle x, y \rangle = \sum_{n=0}^{N-1} x(n) y(n), \quad (4.24)$$

where the signals $x(n)$ and $y(n)$ have N samples each. The dot product represents the projection of one signal on to the other, with each signal being viewed as an N -dimensional vector. The dot product may be normalized by the geometric mean of the energies of the two signals to obtain a correlation coefficient as [21]

$$\gamma_{xy} = \frac{\sum_{n=0}^{N-1} x(n) y(n)}{\left[\sum_{n=0}^{N-1} x^2(n) \sum_{n=0}^{N-1} y^2(n) \right]^{1/2}}. \quad (4.25)$$

The means of the signals may be subtracted out, if desired, as in Equation 3.97.

In the case of two continuous-time signals $x(t)$ and $y(t)$, the projection of one signal on to the other is defined as

$$\theta_{xy} = \int_{-\infty}^{\infty} x(t) y(t) dt. \quad (4.26)$$

When a shift or time delay may be present in the occurrence of the epoch of interest in the two signals being compared, it becomes necessary to introduce a time-shift parameter to compute the projection for every possible position of overlap. The shift parameter facilitates searching one signal for the occurrence of an event matching that in the other signal at any time instant within the available duration of the signals. The CCF between two signals for a shift or delay of τ seconds

or k samples may be obtained as

$$\theta_{xy}(\tau) = \int_{-\infty}^{\infty} x(t) y(t + \tau) dt, \quad (4.27)$$

or

$$\theta_{xy}(k) = \sum_n x(n) y(n + k). \quad (4.28)$$

The range of summation in the latter case needs to be limited to the range of the available overlapped data. A scale factor, depending on the number of data samples used, needs to be introduced to obtain the true CCF, but is neglected here (see Section 6.3). An extended version of the correlation coefficient γ_{xy} in Equation 4.25, to include time shift, is provided in Equation 3.97.

When the ACF or the CCF is computed for various shifts, a question arises about the data samples in one of the signal segments beyond the duration of the other. We may append zeros to one of the signals and increase its length by the maximum shift of interest, or we may use the true data samples from the original signal record, if available. The latter method was used wherever possible in the following illustrations. See Section 6.3 for related discussions.

In the case of random signals, we need to take the expectation or sample average of the outer product of the vectors formed by the available samples of the signals. Let $\mathbf{x}(n) = [x(n), x(n-1), \dots, x(n-N+1)]^T$ and $\mathbf{y}(n) = [y(n), y(n-1), \dots, y(n-N+1)]^T$ represent the N -dimensional vectorial form of the two signals $x(n)$ and $y(n)$ with the most-recent N samples being available in each signal at the time instant n . If $\mathbf{x}(n)$ and $\mathbf{y}(n)$ are sample observations of random processes, their CCF is defined as

$$\Theta_{xy} = E[\mathbf{x}(n) \mathbf{y}^T(n)], \quad (4.29)$$

in a manner similar to Equations 3.163 and 3.164. The outer product, which is an $N \times N$ matrix, provides the cross-terms that include all possible delays (shifts) within the duration of the given signals.

All of the equations given above may be modified to obtain the ACF by replacing the second signal y with the first signal x . The signal x is then compared with itself.

The ACF displays peaks at intervals corresponding to the period (and integral multiples thereof) of any periodic or repetitive pattern present in the signal. This property facilitates the detection of rhythms in signals such as the EEG: The presence of the α rhythm would be indicated by a peak in the neighborhood of 0.1 s . The ACF of most signals decays and reaches negligible values after delays of a few milliseconds, except for periodic signals of infinite or indefinite duration for which the ACF will also exhibit periodic peaks. The ACF will also exhibit multiple peaks when the same event repeats itself at regular or irregular intervals. One may need to compute the ACF only up to certain delay limits depending on the expected characteristics of the signal being analyzed.

The CCF displays peaks at the period of any periodic pattern present in *both* of the signals being analyzed. The CCF may, therefore, be used to detect rhythms present in common between two signals, for example, between two channels of the EEG. When one of the functions being used to compute the CCF is a template representing an event, such as an ECG cycle as in the illustration in Section 3.5 or an EEG spike-and-wave complex as in Section 4.4.2, the procedure is known as *template matching*.

Illustration of application: Figure 4.11 shows, in the upper trace, the ACF of a segment of the p4 channel of the EEG in Figure 1.41 over the time interval $4.67 - 5.81\text{ s}$. The ACF displays peaks at time delays of 0.11 s and its integral multiples. The inverse of the delay of the first peak corresponds to 9 Hz , which is within the α rhythm range. (The PSD in the lower trace of Figure 4.11 and the others to follow are described in Section 4.5.) It is, therefore, obvious that the signal segment analyzed contains the α rhythm. A simple peak-searching algorithm may be applied to the ACF to detect the presence of peaks at specific delays of interest or over the entire range of the ACF.

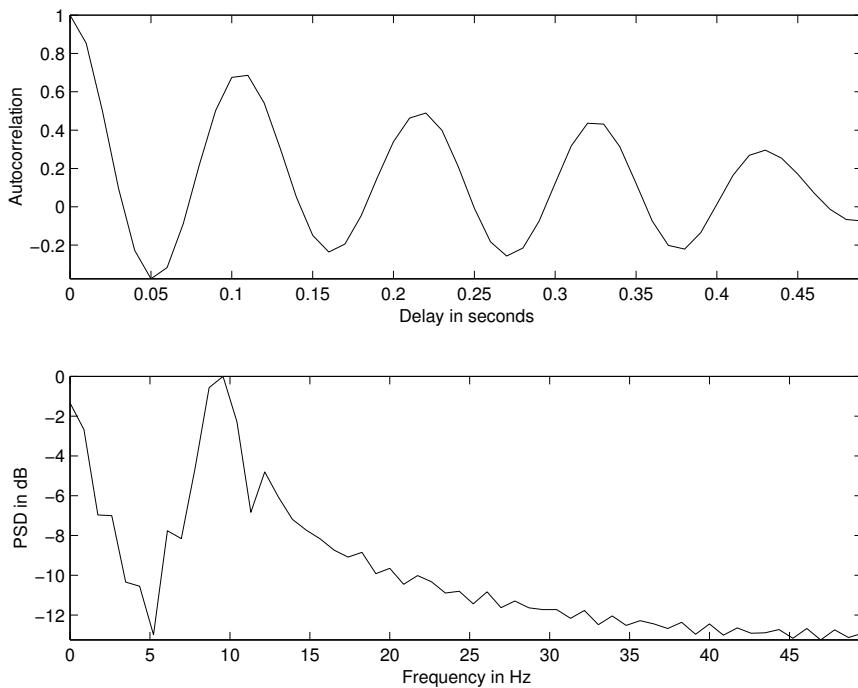


Figure 4.11 Upper trace: ACF of the $4.67 - 5.81$ s portion of the p4 channel of the EEG signal shown in Figure 1.41. Lower trace: The PSD of the signal segment in dB, given by the Fourier transform of the ACF.

To contrast with the preceding example, the upper trace of Figure 4.12 shows the ACF of the $4.2 - 4.96$ s segment of the f3 channel of the EEG in Figure 1.41. The ACF shows no peak in the $0.08 - 1.25$ s region, indicating absence of the α rhythm in the segment analyzed.

Figures 4.13, 4.14, and 4.15 illustrate the CCF results comparing the following portions of the EEG signal shown in Figure 1.41 in order: the p3 and p4 channels over the duration $4.72 - 5.71$ s when both channels exhibit the α rhythm; the o2 and c4 channels over the duration $5.71 - 6.78$ s when the former has the α rhythm but not the latter channel; and the f3 and f4 channels over the duration $4.13 - 4.96$ s when neither channel has α activity. The relative strengths of the peaks in the α range, as described earlier, agree with the joint presence, singular presence, or absence of the α rhythm in the various segments (channels) analyzed.

It should be noted that the segments of EEG signals used in the illustrations in the present section are of short duration. See Section 6.7 for an example of analysis of longer EEG records.

4.4.2 Template matching for EEG spike-and-wave detection

We have seen the use of template matching for the extraction of ECG cycles for use in synchronized averaging in Section 3.5. We now consider another application of template matching.

Problem: Propose a method to detect spike-and-wave complexes in an EEG signal. You may assume that a sample segment of a spike-and-wave complex is available.

Solution: A spike-and-wave complex is a well-defined event in an EEG signal. The complex is composed of a sharp spike followed by a wave with a frequency of about 3 Hz; the wave may contain a half period or a full period of an almost-sinusoidal pattern. One may, therefore, extract an epoch of a spike-and-wave complex from an EEG channel and use it for template matching with the same formula as in Equation 3.97 (see also Barlow [7]). The template may be correlated with the same channel from which it was extracted to detect similar events that appear at a later time, or

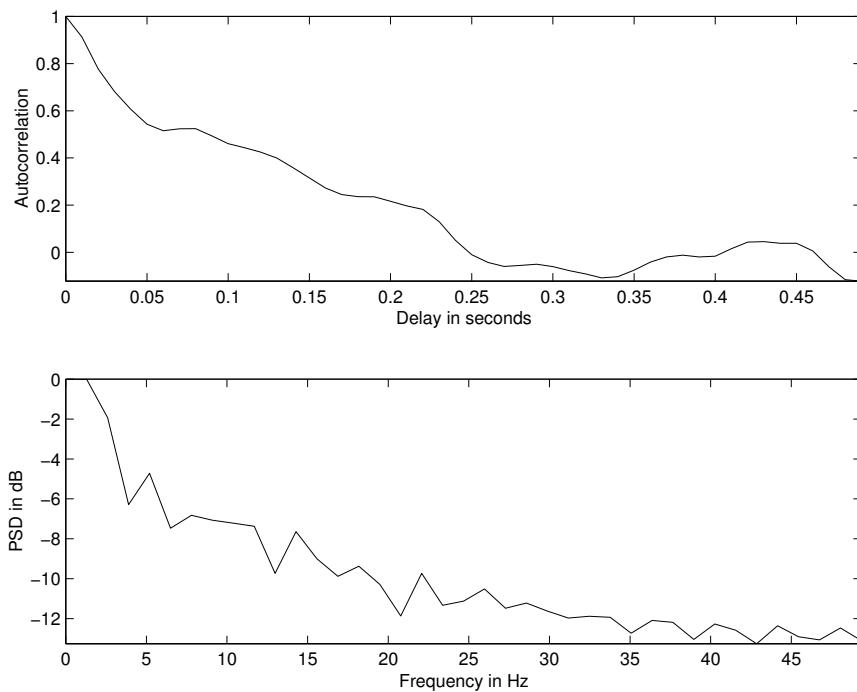


Figure 4.12 Upper trace: ACF of the $4.2 - 4.96\text{ s}$ portion of the f3 channel of the EEG signal shown in Figure 1.41. Lower trace: The PSD of the signal segment in dB.

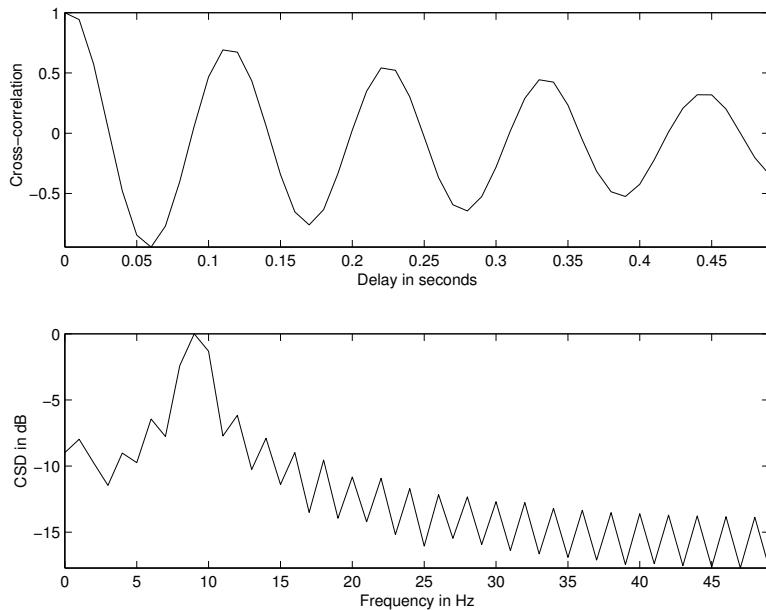


Figure 4.13 Upper trace: CCF between the $4.72 - 5.71\text{ s}$ portions of the p3 and p4 channels of the EEG signal shown in Figure 1.41. Lower trace: The CSD of the signal segments in dB, computed as the Fourier transform of the CCF.

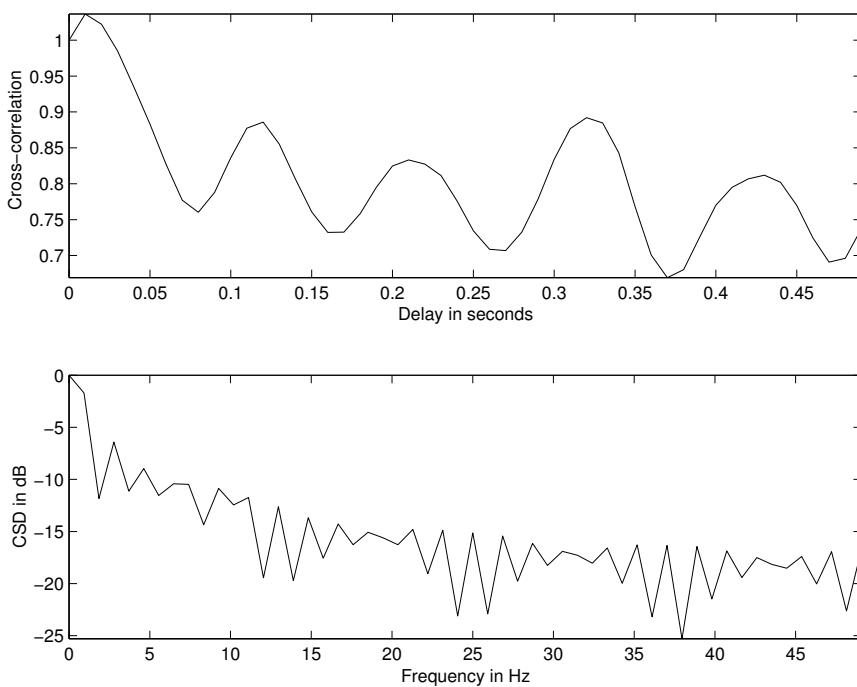


Figure 4.14 Upper trace: CCF between the $5.71 - 6.78\text{ s}$ portions of the o2 and c4 channels of the EEG signal shown in Figure 1.41. Lower trace: The CSD of the signal segments in dB .

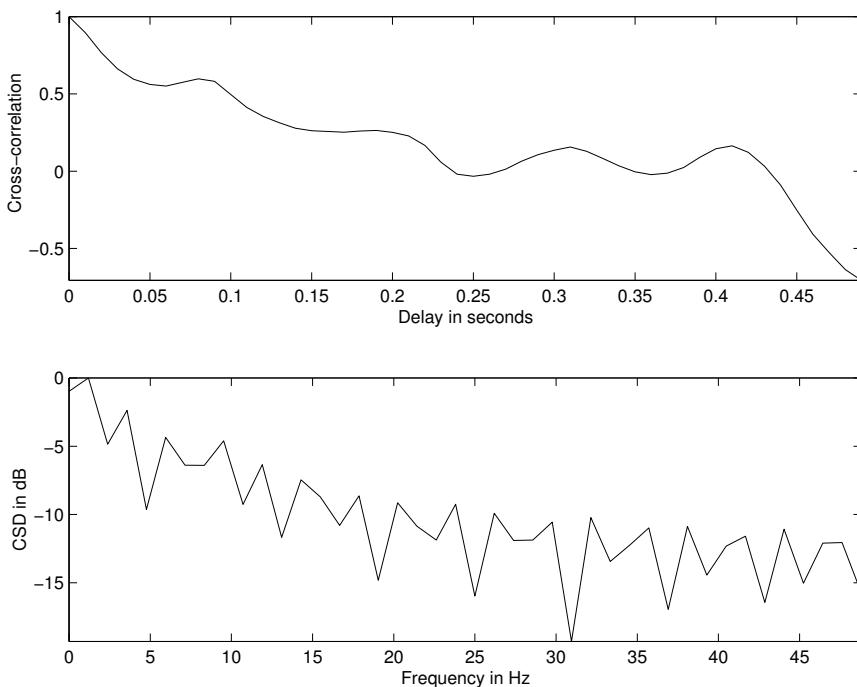


Figure 4.15 Upper trace: CCF between the $4.13 - 4.96\text{ s}$ portions of the f3 and f4 channels of the EEG signal shown in Figure 1.41. Lower trace: The CSD of the signal segments in dB .

with another channel to search for similar events. A simple threshold on the result should yield the time instants where the events appear.

Illustration of application: The c3 channel of the EEG signal in Figure 1.42 is shown in the upper trace of Figure 4.16. The spike-and-wave complex between 0.60 s and 0.82 s in the signal was selected for use as the template, and template matching was performed with the same channel signal using the formula in Equation 3.97. The result in the lower trace of Figure 4.16 demonstrates strong and clear peaks at each occurrence of the spike-and-wave complex in the EEG signal. The peaks in the result occur at the same instants of time as the corresponding spike-and-wave complexes.

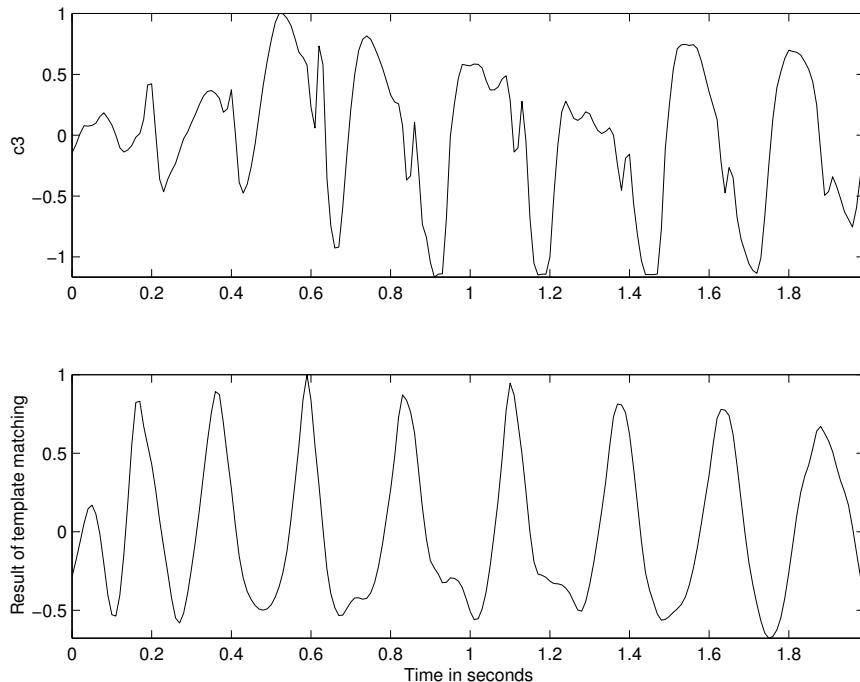


Figure 4.16 Upper trace: the c3 channel of the EEG signal shown in Figure 1.42. Lower trace: result of template matching. The spike-and-wave complex between 0.60 s and 0.82 s in the signal was used as the template.

Figure 4.17 shows the f3 channel of the EEG signal in Figure 1.42, along with the result of template matching, using the same template that was used in the previous example from channel c3. The result shows that the f3 channel also has spike-and-wave complexes that match the template.

The focus in this section has been on detecting rhythms in the EEG. Note that the same approaches could be applied to any signal that has rhythmic repetitions of a basic pattern, such as the ECG and voiced-speech signals.

4.4.3 Detection of EEG rhythms related to seizure

Yadav et al. [22] noted that the recurring nature of seizures of a given patient could be used to design patient-specific templates for the recognition of seizures. They noted that, in most patients, a few types of seizures tend to occur repeatedly, with related similar, though not identical, patterns in the EEG. Figure 4.18 illustrates the temporal evolution of a seizure event on a single-channel EEG, including varying patterns of piecewise stationary rhythms. (Note that the seizure rhythms are substantially different from the α and other rhythms seen in the normal EEG.) Yadav et al. [22] suggested that it is possible to train a patient-specific seizure detector using previously identified

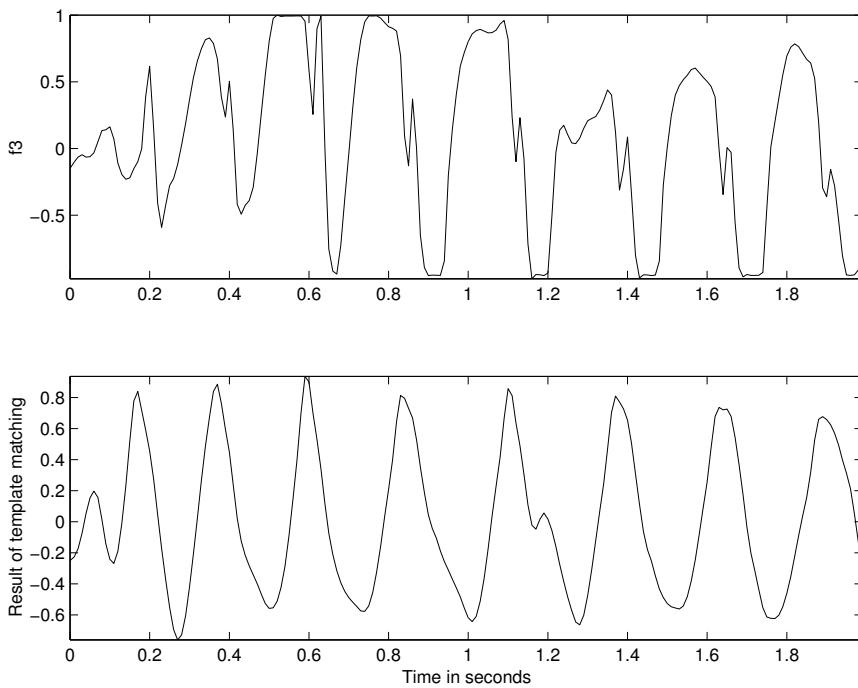


Figure 4.17 Upper trace: the f3 channel of the EEG signal shown in Figure 1.42. Lower trace: result of template matching. The spike-and-wave complex between 0.60 s and 0.82 s in the c3 channel (see Figure 4.16) was used as the template.

templates, and they proposed a model-based patient-specific method using statistically optimal null filters for automatic detection of seizures in intracranial EEG. See Qu and Gotman [23] and Grewal and Gotman [24] for additional details on detection of seizures; see also Sections 8.17 and 9.8.

4.5 Cross-spectral Techniques

The multiple peaks that arise in the ACF or CCF may cause confusion in the detection of rhythms; the analyst may be required to discount peaks that appear at integral multiples of the delay corresponding to a fundamental frequency and other delays. The Fourier-domain equivalents of the ACF or the CCF permit easier and more intuitive analysis in the frequency domain than in the time domain. The notion of rhythms would be easier to associate with frequencies in *cps* or *Hz* than with the corresponding inversely related periods (see also the introductory paragraphs of Chapter 6).

4.5.1 Coherence analysis of EEG channels

Problem: *Describe a frequency-domain approach to study the presence of rhythms in multiple channels of an EEG signal.*

Solution: The Fourier-domain equivalents of the ACF and CCF are the PSD (also known as the autospectrum) and the cross-spectrum (or cross-spectral density — CSD), respectively. The PSD $S_{xx}(f)$ of a signal is related to its ACF via the Fourier transform:

$$S_{xx}(f) = FT[\phi_{xx}(\tau)] = X(f)X^*(f) = |X(f)|^2, \quad (4.30)$$

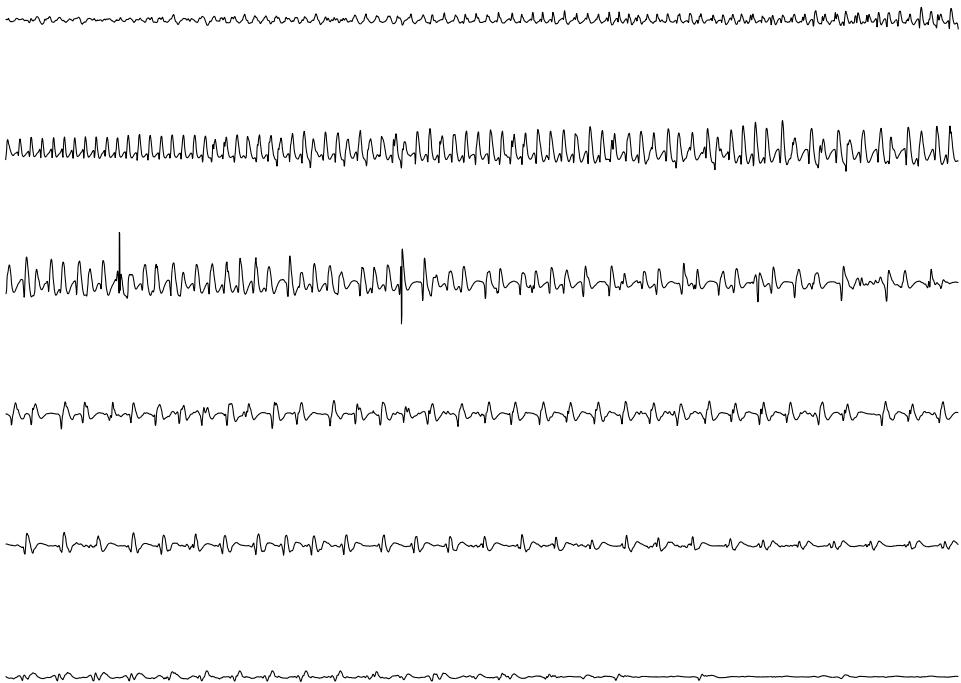


Figure 4.18 The evolution of various rhythms related to a seizure event as seen in an intracranial EEG signal. Each row represents a duration of 15 s and is a continuation of the preceding row. The minimum to maximum range of the signal shown is about 120 μV . Data courtesy of R. Agarwal [22].

where $FT[]$ indicates the Fourier transform. The Fourier transform of the CCF between two signals gives the CSD:

$$S_{xy}(f) = FT[\theta_{xy}(\tau)] = X(f)Y^*(f). \quad (4.31)$$

(For the sake of simplicity, the double-symbol subscripts xx and yy may be replaced by their singular versions, or dropped entirely when not relevant in subsequent discussions.)

The PSD displays peaks at frequencies corresponding to periodic activities in the signal. This property facilitates the detection of rhythms in signals such as the EEG: The presence of the α rhythm would be indicated by a peak or multiple peaks in the neighborhood of 8 – 13 Hz. The PSD may also be studied to locate the presence of activity spread over specific bands of frequencies, such as formants in the speech signal or murmurs in the PCG.

The CSD exhibits peaks at frequencies that are present in both of the signals being compared. The CSD may be used to detect rhythms present in common between two channels of the EEG.

The normalized magnitude of the *coherence spectrum* of two signals is given by [3, 25]

$$\Gamma_{xy}(f) = \left[\frac{|S_{xy}(f)|^2}{S_{xx}(f)S_{yy}(f)} \right]^{1/2}. \quad (4.32)$$

If this expression is computed for two individual signals directly, the magnitude of the result will be equal to unity for all f , which is incorrect. In order to evaluate the magnitude coherence spectrum as defined in Equation 4.32, each spectral density function involved — S_{xy} , S_{xx} , and S_{yy} — must be estimated using an averaging procedure applied to several observations of the processes generating the signals [25–27]. See Section 6.3 for procedures for the estimation of spectral density functions.

The phase of the coherence spectrum is given by $\psi_{xy}(f) = \angle S_{xy}(f)$, which represents the average phase difference (related to the time delay) between frequency components in the two signals.

Rosenberg et al. [28] applied coherence to identify functional coupling between neuronal spike trains. They observed that frequency-domain analysis using coherence as a measure of association can lead to new insight and understanding of interactions between spike trains. Amjad et al. [29] described several methods to estimate the coherence function and its application to the analysis of SMUAP discharges and physiological tremor; see also Farmer et al. [30] and Farmer [31]. Johnston et al. [32] used coherence analysis to study the correlation between neural inputs to motor units innervating different muscles. See Section 6.3 for procedures for the estimation of spectral density functions.

Illustration of application: The coherence between EEG signals recorded from different positions on the scalp depends on the structural connectivity or functional coupling between the corresponding parts of the brain. Investigations into the neurophysiology of seizure discharges and behavior attributable to disorganization of cerebral function may be facilitated by coherence analysis [3]. The symmetry, or lack thereof, between two EEG channels on the left and right sides of the same position (for example, c3 and c4) may be analyzed via the CSD or the coherence function.

The lower traces in Figures 4.11 and 4.12 illustrate the PSDs of EEG segments with and without the α rhythm, respectively. The former shows a strong and clear peak at about 9 Hz, indicating the presence of the α rhythm. Observe that the PSD displays a single peak although the corresponding ACF has multiple peaks at two, three, and four times the delay corresponding to the fundamental period of the α wave in the signal. The PSD in Figure 4.12 exhibits no peak in the α range, indicating the absence of the α rhythm in the signal.

The lower traces in Figures 4.13, 4.14, and 4.15 illustrate the CSDs corresponding to the CCFs in the respective upper traces. Once again, it is easier to deduce the common presence of strong α activity between channels p3 and p4 from the CSD rather than the CCF in Figure 4.13. The single peak at 9 Hz in the CSD is more easily interpretable than the multiple peaks in the corresponding CCF. The CSD in Figure 4.14 lacks a clear peak in the α range, even though the corresponding CCF shows a peak at about 0.1 s, albeit less significant than that in Figure 4.13. The results agree with the fact that one channel has α activity, while the other does not. Finally, the CSD in Figure 4.15 is clearly lacking a peak in the α range; the two signal segments have no α activity. Further methods for the analysis of α activity are presented in Sections 6.3.4, 6.7, and 7.5.2.

4.6 The Matched Filter

When a sample observation or template of a typical version of a signal event is available, it becomes possible to design a filter that is *matched* to the characteristics of the event and maximizes the *SNR* of the output. If a signal that contains repetitions of the event with almost the same characteristics is passed through the *matched filter*, the output should provide peaks at the time instants of occurrence of the event. Matched filters are commonly used for the detection of signals of known characteristics that are buried in noise [33, 34]. They are designed to perform a correlation operation between the input signal and the signal template, and hence are also known as *correlation filters*.

4.6.1 Derivation of the transfer function of the matched filter

In order to derive the transfer function, $H(\omega)$, of the matched filter [33], let the signal $x(t)$ be the input to the matched filter. The Fourier transform of $x(t)$ is

$$X(\omega) = \int_{-\infty}^{\infty} x(t) \exp(-j\omega t) dt. \quad (4.33)$$

The output of the matched filter, $y(t)$, is given by the inverse Fourier transform of $Y(\omega) = X(\omega)H(\omega)$, as follows:

$$\begin{aligned}
y(t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) H(\omega) \exp(+j\omega t) d\omega \\
&= \int_{-\infty}^{\infty} X(f) H(f) \exp(+j 2\pi f t) df.
\end{aligned} \tag{4.34}$$

In the second expression of the equation given above, the frequency variable has been changed from ω in radians per second to f in Hz.

Consider the presence of white noise at the input, with the PSD

$$S_{\eta i}(f) = \frac{P_{\eta i}}{2}, \tag{4.35}$$

where $P_{\eta i}$ is the average noise power at the input. Then, the noise PSD at the output is

$$S_{\eta o}(f) = \frac{P_{\eta i}}{2} |H(f)|^2. \tag{4.36}$$

The average output noise power is

$$P_{\eta o} = \frac{P_{\eta i}}{2} \int_{-\infty}^{\infty} |H(f)|^2 df. \tag{4.37}$$

The RMS value of the noise in the absence of any signal is $\sqrt{P_{\eta o}}$.

Letting $t = t_0$ in Equation 4.34, the magnitude of the instantaneous output signal at $t = t_0$ is

$$M_y = |y(t_0)| = \left| \int_{-\infty}^{\infty} X(f) H(f) \exp(+j 2\pi f t_0) df \right|. \tag{4.38}$$

Thus, the SNR at the output is $\frac{M_y}{\sqrt{P_{\eta o}}}$.

To derive the optimal transfer function of the matched filter, we could maximize the SNR, which is equivalent to maximizing the expression

$$\frac{M_y^2}{P_{\eta o}} = \frac{\text{instantaneous peak power of signal}}{\text{noise mean power}}, \tag{4.39}$$

which represents peak-power SNR [33].

For a given signal $x(t)$, the total energy is a constant, given by

$$E_x = \int_{-\infty}^{\infty} x^2(t) dt = \int_{-\infty}^{\infty} |X(f)|^2 df. \tag{4.40}$$

Let us consider the following ratio:

$$\frac{M_y^2}{E_x P_{\eta o}} = \frac{\left| \int_{-\infty}^{\infty} H(f) X(f) \exp(+j 2\pi f t_0) df \right|^2}{\frac{P_{\eta i}}{2} \int_{-\infty}^{\infty} |H(f)|^2 df \int_{-\infty}^{\infty} |X(f)|^2 df}. \tag{4.41}$$

The quantity E_x is a constant for a given input signal; hence, maximizing the expression in Equation 4.41 is equivalent to maximizing the expression in Equation 4.39.

In order to determine the condition for maximization of the expression in Equation 4.41, recall Schwarz's inequality for two arbitrary complex functions $A(f)$ and $B(f)$:

$$\left| \int_{-\infty}^{\infty} A(f) B(f) df \right|^2 \leq \left[\int_{-\infty}^{\infty} |A(f)|^2 df \right] \left[\int_{-\infty}^{\infty} |B(f)|^2 df \right]. \tag{4.42}$$

For any two real functions $a(t)$ and $b(t)$, the corresponding inequality is

$$\left[\int_{-\infty}^{\infty} a(t) b(t) dt \right]^2 \leq \left[\int_{-\infty}^{\infty} a^2(t) dt \right] \left[\int_{-\infty}^{\infty} b^2(t) dt \right]. \quad (4.43)$$

For any two vectors \mathbf{a} and \mathbf{b} , Schwarz's inequality states that

$$|\mathbf{a} \cdot \mathbf{b}| \leq |\mathbf{a}| |\mathbf{b}|, \quad (4.44)$$

and

$$|\mathbf{a} + \mathbf{b}| \leq |\mathbf{a}| + |\mathbf{b}|. \quad (4.45)$$

In the inequalities stated above, equality is achieved if $\mathbf{a} = K \mathbf{b}$, that is, \mathbf{a} and \mathbf{b} are collinear; if $a(t) = K b(t)$; or if $A(f) = K B^*(f)$, where K is a real constant.

The inequality in Equation 4.42 can be applied to Equation 4.41 by considering $A(f) = H(f)$ and $B(f) = X(f) \exp(+j 2\pi f t_0)$. Then, we have

$$\frac{P_{\eta i} M_y^2}{2 E_x P_{\eta o}} \leq 1, \quad (4.46)$$

because

$$\left| \int_{-\infty}^{\infty} X(f) H(f) \exp(+j 2\pi f t_0) df \right|^2 \leq \int_{-\infty}^{\infty} |X(f)|^2 df \int_{-\infty}^{\infty} |H(f)|^2 df. \quad (4.47)$$

The LHS of Equation 4.47 represents the instantaneous output power, M_y^2 , evaluated in the Fourier domain. Thus, the ratio in Equation 4.41 is maximized when equality is achieved, that is, $A(f) = K B^*(f)$ with the functions $A(f)$ and $B(f)$ as explained above, leading to the condition

$$\begin{aligned} H(f) &= K [X(f) \exp(+j 2\pi f t_0)]^* \\ &= K X^*(f) \exp(-j 2\pi f t_0), \end{aligned} \quad (4.48)$$

which leads to maximal peak output SNR .

Taking the inverse Fourier transform of the last expression given above, we have

$$h(t) = K x[-(t - t_0)]. \quad (4.49)$$

Therefore, the impulse response of the matched filter is a scaled, reversed, and shifted version of the signal of interest.

The output of the matched filter is then obtained by the following steps:

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} X(f) H(f) \exp(+j 2\pi f t) df \\ &= \int_{-\infty}^{\infty} X(f) K X^*(f) \exp(-j 2\pi f t_0) \exp(+j 2\pi f t) df \\ &= K \int_{-\infty}^{\infty} |X(f)|^2 \exp[j 2\pi f(t - t_0)] df \\ &= K \phi_x(t - t_0). \end{aligned} \quad (4.50)$$

The last relationship listed above is based on the property that the Fourier transform of the ACF is equal to the PSD of the signal. [Note: The Fourier transform of $x(-t)$ is $X^*(f)$, and that of $x(t - \tau)$ is $\exp(-j 2\pi f \tau) X(f)$.]

It is seen from the derivations given above that, when the impulse response of the matched filter is related to the desired signal as in Equation 4.49, the output is maximized and is equal to a scaled and delayed version of the ACF of the signal.

Illustration of application: Figure 4.19 shows a signal $x(n)$ composed of three events. The signal may be defined in terms of impulses as

$$\begin{aligned} x(n) = & 3\delta(n-5) + 2\delta(n-6) + \delta(n-7) \\ & + 1.5\delta(n-16) + \delta(n-17) + 0.5\delta(n-18) \\ & + 0.75\delta(n-26) + 0.5\delta(n-27) + 0.25\delta(n-28). \end{aligned} \quad (4.51)$$

From the plot of the signal in Figure 4.19, it is seen that $x(n)$ is composed of three occurrences of a basic signal $g(n)$ that may be defined as

$$g(n) = 3\delta(n) + 2\delta(n-1) + \delta(n-2), \quad (4.52)$$

with the composite signal given by

$$x(n) = g(n-5) + 0.5g(n-16) + 0.25g(n-26). \quad (4.53)$$

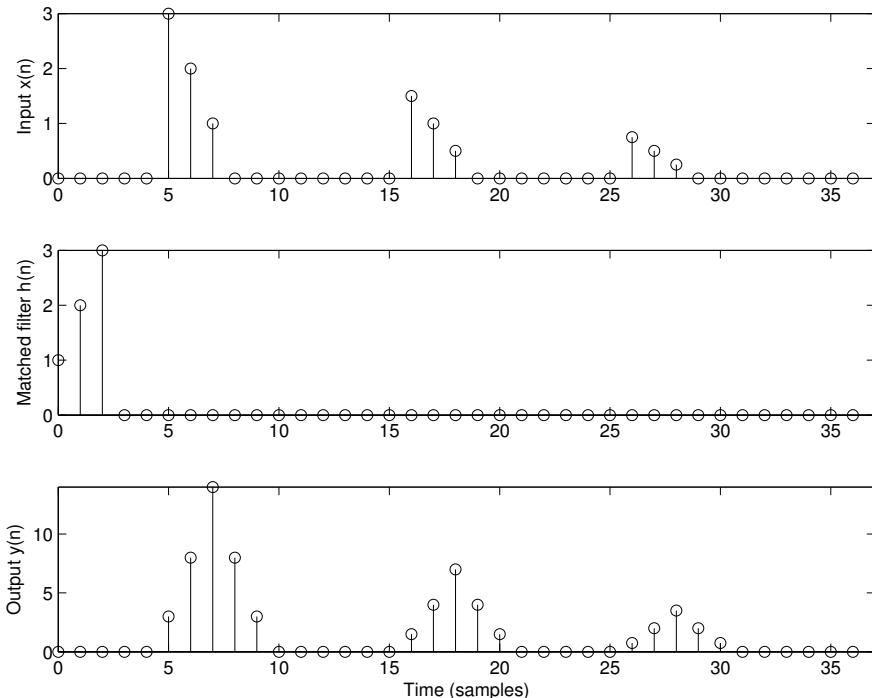


Figure 4.19 Top: A test signal with three similar events or an event with two echoes. Middle: Impulse response of the matched filter. Bottom: Output of the matched filter with peaks at the locations of occurrence of the basic signal pattern.

If we consider $g(n)$ as the basic pattern to be detected by using a matched filter, we need the impulse response of the filter to be $h(n) = K g(-n + n_0)$, where K is a scale factor or gain and n_0

is a delay to obtain a causal $h(n)$. In the present case, we need $n_0 = 2$ samples. Let $K = 1$. Then, we have

$$h(n) = \delta(n) + 2\delta(n-1) + 3\delta(n-2), \quad (4.54)$$

which is plotted in the trace in the middle of Figure 4.19. The trace at the bottom in the same figure shows the output of the matched filter. It is evident that the output possesses peaks at the locations of occurrences of $g(n)$ in the input signal, with a delay of two samples introduced by the filtering process. The values of the peaks are in proportion to the amplitude scale factors of the repeated occurrences of the basic pattern. It can be verified that the output values for the samples 5–9 represent the ACF of $g(n)$.

4.6.2 Detection of EEG spike-and-wave complexes

Problem: Design a matched filter to detect spike-and-wave complexes in an EEG signal. A reference spike-and-wave complex is available.

Solution: Let $x(t)$ be the given reference signal, representing an ideal observation of the event of interest. Let $X(f)$ be the Fourier transform of $x(t)$. Consider passing $x(t)$ through an LTI filter whose impulse response is $h(t)$; the transfer function of the filter is $H(f) = FT[h(t)]$. The output is given by $y(t) = x(t) * h(t)$ or $Y(f) = X(f)H(f)$.

Based on the derivation of the matched filter in Section 4.6.1, it is evident that the output energy is maximized when

$$H(f) = KX^*(f) \exp(-j2\pi ft_0), \quad (4.55)$$

where K is a scale factor, and t_0 is a time instant or delay [33]. This corresponds to the impulse response being

$$h(t) = Kx(t_0 - t). \quad (4.56)$$

Thus, the transfer function of the matched filter is proportional to the complex conjugate of the Fourier transform of the signal event to be detected. In the time domain, the impulse response is simply a *reversed* or *reflected* version of the reference signal that is scaled and delayed. A suitable delay will have to be added to make the filter causal, as determined by the duration of the reference signal.

If the derivation of the impulse response of the matched filter is based upon Equation 4.55 and implemented using the DFT, with the number of samples in the DFT equal to the number of samples N in the reference signal or template x , due to the periodicity of the DFT, the time delay or shift provided inherently will be $(N - 1)$. This is short by one sample; the desired shift is N samples to make the reversed signal causal.

Because the impulse response is a reversed version of $x(t)$, the convolution operation performed by the matched filter is equivalent to correlation: The output is then equal to the cross-correlation between the input and the reference signal. When a portion of an input signal that is different from $x(t)$ matches the reference signal, the output approximates the ACF ϕ_x of the reference signal at the corresponding time delay. The corresponding frequency domain result is

$$Y(f) = X(f)H(f) = X(f)X^*(f) = S_x(f), \quad (4.57)$$

which is the PSD of the reference signal (ignoring the time delay and scale factors). The output is, therefore, maximum at the time instant of occurrence of an approximation to the reference signal. (See also Barlow [7] for related discussion.)

Illustration of application: To facilitate comparison with template matching, the spike-and-wave complex between 0.60 s and 0.82 s in the c3 channel of the EEG in Figure 1.42 was used as the reference signal to derive the matched filter. Figure 4.20 shows the extracted reference signal in the upper trace. The lower trace in the same figure shows the impulse response of the matched filter,

which is simply a time-reversed version of the reference signal. The matched filter was implemented as an FIR filter using the MATLAB® *filter* command.

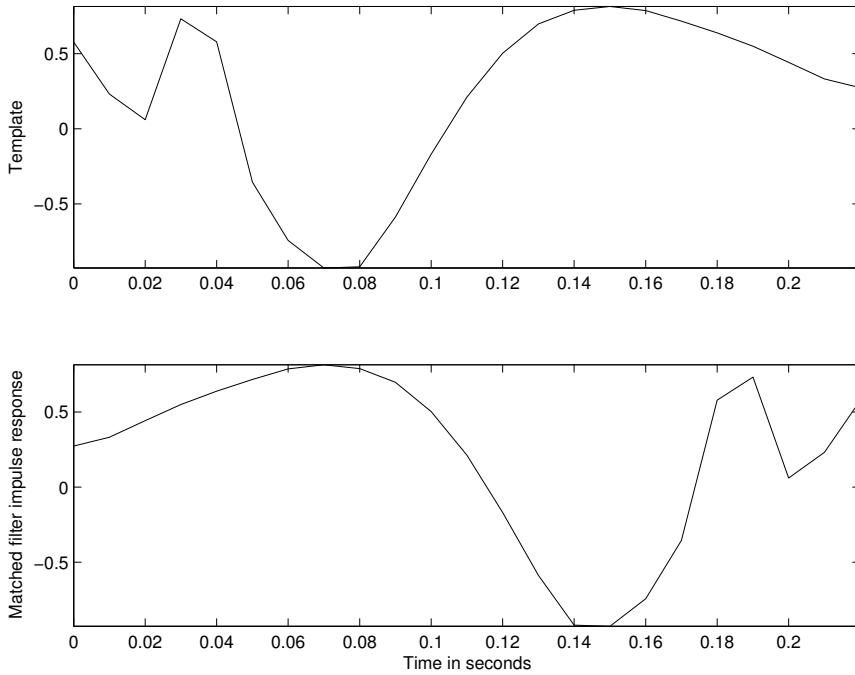


Figure 4.20 Upper trace: The spike-and-wave complex between 0.60 s and 0.82 s in the c3 channel of the EEG signal shown in Figure 1.42. Lower trace: Impulse response of the matched filter derived from the signal segment in the upper trace. Observe that the latter is a time-reversed version of the former.

Figures 4.21 and 4.22 show the outputs of the matched filter applied to the c3 and f3 channels of the EEG in Figure 1.42, respectively. The upper trace in each plot shows the signal, and the lower trace shows the output of the matched filter. It is evident that the matched filter provides a large output for each spike-and-wave complex. Comparing the outputs of the matched filter in Figures 4.21 and 4.22 with those of template matching in Figure 4.16 and 4.17, respectively, we observe that they are similar, with the exception that the results of the matched filter peak with a delay of 0.22 s after the corresponding spike-and-wave complex. The delay corresponds to the duration of the impulse response of the filter. The values differ due to different normalization in Equation 3.97. (Note: MATLAB® provides the command *filtfilt* for zero-phase forward and reverse digital filtering; this method is not considered in the book.)

4.7 Homomorphic Filtering and the Complex Cepstrum

In Chapter 3, we studied linear filters designed to separate signals that were added together. The question asked was the following: Given $y(t) = x(t) + \eta(t)$, how can one extract $x(t)$ only? Given that the Fourier transform is linear, we know that the Fourier transforms of the signals are also combined in an additive manner: $Y(\omega) = X(\omega) + \eta(\omega)$. Therefore, a linear filter will facilitate the separation of $X(\omega)$ and $\eta(\omega)$, with the assumption that they have substantial portions of their energy in different frequency bands.

Suppose now that we are presented with a signal that contains the product of two signals, for example, $y(t) = x(t) p(t)$. From the multiplication or convolution property of the Fourier transform,

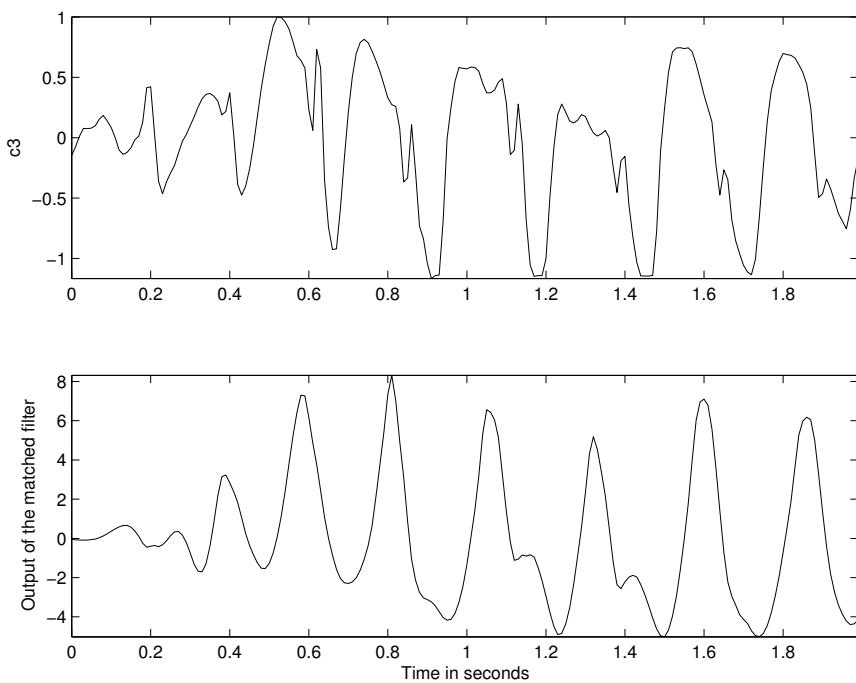


Figure 4.21 Upper trace: The c_3 channel of the EEG signal shown in Figure 1.42, used as input to the matched filter in Figure 4.20. Lower trace: Output of the matched filter. See also Figure 4.16.

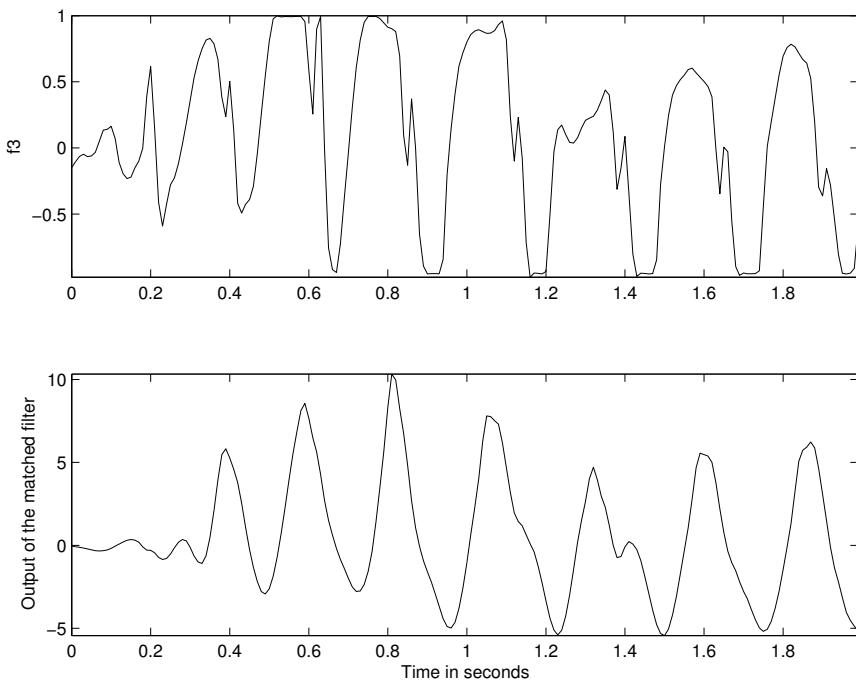


Figure 4.22 Upper trace: The f_3 channel of the EEG signal shown in Figure 1.42, used as input to the matched filter in Figure 4.20. Lower trace: Output of the matched filter. See also Figure 4.17.

we have $Y(\omega) = X(\omega) * P(\omega)$, where $*$ represents convolution in the frequency domain. How would we be able to separate $x(t)$ from $p(t)$?

Furthermore, suppose that we have $y(t) = x(t) * h(t)$, where $*$ stands for convolution as in the case of the passage of the glottal pulse train or random excitation $x(t)$ through the vocal-tract system with the impulse response $h(t)$. The Fourier transforms of the signals are related as $Y(\omega) = X(\omega) H(\omega)$. How would we attempt to separate $x(t)$ and $h(t)$?

4.7.1 Generalized linear filtering

Given that linear filters are well established and understood, it is attractive to consider extending their application to signals that have been combined by operations other than addition, especially by multiplication and convolution as indicated in the preceding paragraphs. An interesting possibility to achieve this is via conversion of the operation combining the signals into addition by one or more transformations. Under the assumption that the transformed signals occupy different portions of the transformed space, linear filters may be applied to separate them. The inverses of the transformations used initially would then take us back to the original space of the signals. This approach was proposed in a series of papers by Bogert et al. [35], Oppenheim et al. [36], and Oppenheim and Schafer [37, 38]. Because the procedure extends the application of linear filters to multiplied and convolved signals, it has been referred to as *generalized linear filtering*. Furthermore, as the operations can be represented by algebraically linear transformations between the input and output vector spaces, they have been called *homomorphic systems*.

As a simple illustration of a homomorphic system for multiplied signals, consider again the signal

$$y(t) = x(t) p(t). \quad (4.58)$$

Given the goal of converting the multiplication operation to addition, it is evident that a simple logarithmic transformation is appropriate:

$$\log[y(t)] = \log[x(t) p(t)] = \log[x(t)] + \log[p(t)]; \quad x(t) \neq 0, p(t) \neq 0 \quad \forall t. \quad (4.59)$$

The logarithms of the two signals are now combined in an additive manner. Taking the Fourier transform, we get

$$Y_l(\omega) = X_l(\omega) + P_l(\omega), \quad (4.60)$$

where the subscript l indicates that the Fourier transform has been applied to a log-transformed version of the signal.

Assuming that the logarithmic transformation has not affected the separability of the Fourier components of the two signals $x(t)$ and $p(t)$, a linear filter (such as a lowpass or a highpass filter) may now be applied to $Y_l(\omega)$ to separate them. An inverse Fourier transform will yield the filtered signal in the time domain. An exponential operation will complete the reversal procedure (if required).

Figure 4.23 illustrates the operations involved in a multiplicative homomorphic system (or filter). The symbol at the input or output of each block indicates the operation that combines the signal components at the corresponding step. A system of this type is useful in image processing, where an image may be treated as the product of an illumination function and a transmittance or reflectance function. The homomorphic filter facilitates separation of the illumination function and correction for nonuniform lighting. The method has been used to achieve simultaneous dynamic range compression and contrast enhancement [36, 39–41].

4.7.2 Homomorphic deconvolution

Problem: Propose a homomorphic filter to separate two signals that have been combined through the convolution operation.

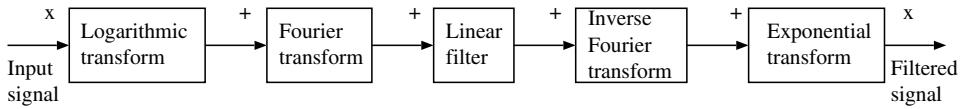


Figure 4.23 Operations involved in a multiplicative homomorphic system or filter. The symbol at the input or output of each block indicates the operation that combines the signal components at the corresponding step.

Solution: Consider the case expressed by the relation

$$y(t) = x(t) * h(t). \quad (4.61)$$

As in the case of the multiplicative homomorphic system, our goal is to convert the convolution operation to addition. From the convolution property of the Fourier transform, we know that

$$Y(\omega) = X(\omega) H(\omega). \quad (4.62)$$

Thus, application of the Fourier transform converts convolution to multiplication. Now, it is readily seen that the multiplicative homomorphic system may be applied to convert the multiplication above to addition. Taking the complex logarithm of $Y(\omega)$, we have

$$\log[Y(\omega)] = \log[X(\omega)] + \log[H(\omega)]; \quad X(\omega) \neq 0, H(\omega) \neq 0 \quad \forall \omega. \quad (4.63)$$

[Note: $\log_e[X(\omega)] = \hat{X}(\omega) = \log_e[|X(\omega)| \angle X(\omega)] = \log_e[|X(\omega)|] + j\angle X(\omega)$, where $|X(\omega)|$ and $\angle X(\omega)$ are the magnitude and phase spectra of $x(t)$.]

A linear filter may now be used to separate the transformed components of x and h , with the assumption as before that they are separable in the transformed space. A series of the inverses of the transformations applied initially will take us back to the original domain.

Whereas the discussion here has been in terms of application of the Fourier transform, the general formulation of the homomorphic filter by Oppenheim and Schafer [39] is in terms of the z -transform. However, the Fourier transform is equivalent to the z -transform evaluated on the unit circle in the z -plane, and the Fourier transform is more commonly used in signal processing than the z -transform.

Figure 4.24 gives a block diagram of the steps involved in a homomorphic filter for convolved signals. The path formed by the first three blocks (in the top row) transforms the convolution operation at the input to addition. The third block with the inverse Fourier transform is used to move back to a pseudo time domain. The last three blocks (in the bottom row) perform the reverse transformation, converting addition to convolution. The filter in between deals with (transformed) signals that are combined by simple addition. Further details of these operations and results are presented in the following section.

4.7.3 Extraction of the vocal-tract response

Problem: Design a homomorphic filter to extract the basic wavelet corresponding to the vocal-tract response from a voiced speech signal.

Solution: We noted in Section 1.2.13 that voiced speech is generated by excitation of the vocal tract, while it is held in a particular form, with a glottal waveform that may be approximated as a series of pulses. The voiced speech signal may, therefore, be expressed in discrete-time terms as $y(n) = x(n) * h(n)$, where $y(n)$ is the speech signal, $x(n)$ is the glottal waveform (excitation sequence), and $h(n)$ is the impulse response of the vocal tract (basic wavelet). The $*$ symbol represents convolution, with the assumption that the vocal-tract filter may be approximated by an LSI filter. We may, therefore, use the homomorphic filter for convolved signals, as introduced in the preceding section, to separate $h(n)$ and $x(n)$.

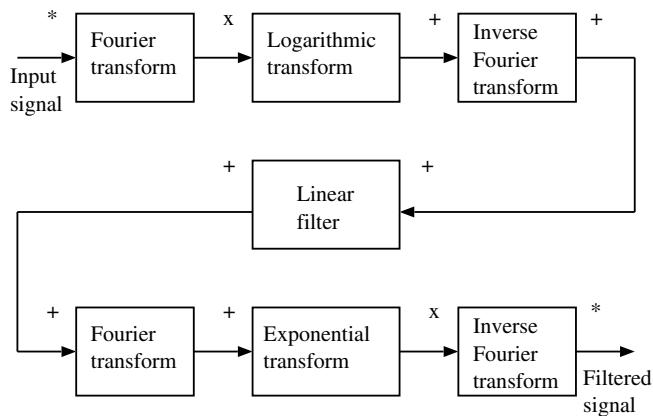


Figure 4.24 Operations involved in a homomorphic filter for convolved signals. The symbol at the input or output of each block indicates the operation that combines the signal components at the corresponding step.

The glottal excitation sequence may be further expressed as $x(n) = p(n) * g(n)$, where $p(n)$ is a train of ideal impulses (Dirac delta functions), and $g(n)$ is a smoothing function, to indicate that the physical vocal-cord system cannot produce ideal impulses but rather pulses of finite duration and slope [39]. This aspect will be neglected in our discussions.

Practical application of the homomorphic filter is not simple. Figure 4.25 gives a detailed block diagram of the procedure [39,42]. Some of the details and practical techniques are explained in the following paragraphs.

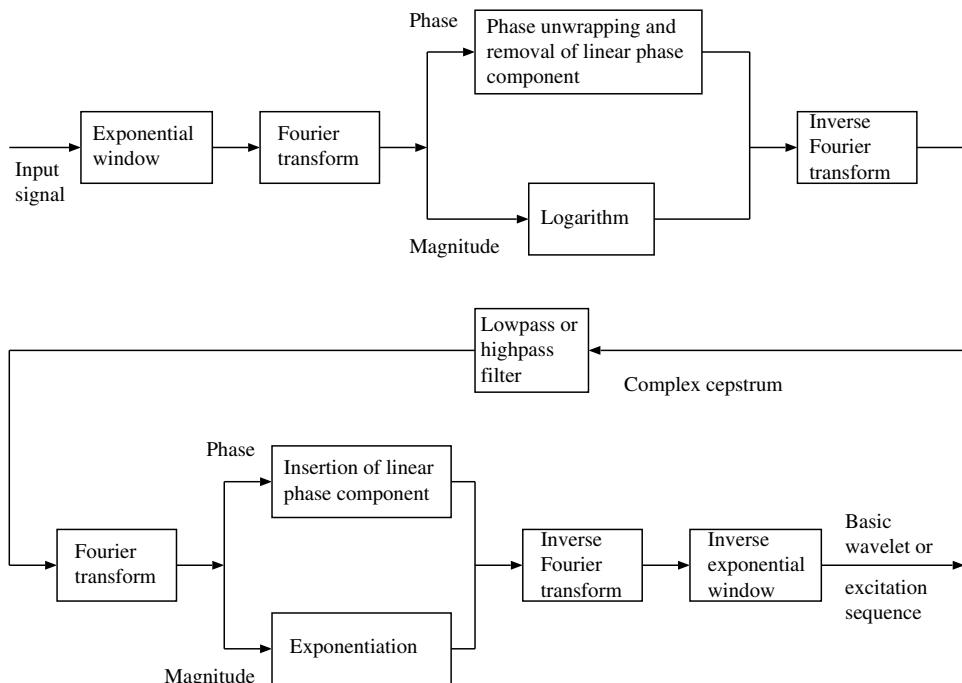


Figure 4.25 Detailed block diagram of the steps involved in deconvolution of signals using the complex cepstrum.

The complex cepstrum: The formal definition of the complex cepstrum states that it is the inverse z -transform of the complex logarithm of the z -transform of the input signal [39, 42]. (The name “cepstrum” was derived by transposing the syllables of the word “spectrum”; other transposed terms [35, 39, 42] are less commonly used.) If $y(n)$ is the input signal, and $Y(z)$ is its z -transform, the complex cepstrum $\hat{y}(n)$ is defined as

$$\hat{y}(n) = \frac{1}{2\pi j} \oint \log[Y(z)] z^{n-1} dz. \quad (4.64)$$

The contour integral performs the inverse z -transform, and should be evaluated within an annular region in the complex z -plane, where $\hat{Y}(z) = \log[Y(z)]$ is single valued and analytic [39, 43]. The unit of n in $\hat{y}(n)$, that is, in the cepstral domain, is referred to as *quefrency*, a term obtained by switching the first two syllables in the term frequency; it is also referred to as pseudo time [35, 38].

Given $y(n) = x(n) * h(n)$, it follows that

$$\hat{Y}(z) = \hat{X}(z) + \hat{H}(z) \text{ or } \hat{Y}(\omega) = \hat{X}(\omega) + \hat{H}(\omega), \quad (4.65)$$

and further that the complex cepstra of the signals are related as

$$\hat{y}(n) = \hat{x}(n) + \hat{h}(n). \quad (4.66)$$

Here, the $\hat{\cdot}$ symbol over a function of z or ω indicates the complex logarithm of the corresponding function of z or ω , whereas the same symbol over a function of time (n) indicates the complex cepstrum of the corresponding signal. It should be noted that, if the original signal $y(n)$ is real, its complex cepstrum $\hat{y}(n)$ is *real*; the prefix *complex* is used to indicate the fact that the preceding z and logarithmic transformations are computed as complex functions. Furthermore, it should be noted that the complex cepstrum is a function of time.

An important consideration in the evaluation of the complex logarithm of $Y(z)$ or $Y(\omega)$ relates to the phase of the signal. The phase spectrum computed as its principal value in the range $[0, 2\pi]$, given by $\arctan \left[\frac{\text{imaginary}\{Y(\omega)\}}{\text{real}\{Y(\omega)\}} \right]$, will almost always have discontinuities that will conflict with the requirements of the inverse z -transformation or inverse Fourier transform to follow later. Thus, $Y(\omega)$ needs to be separated into its magnitude and phase components, the logarithmic operation applied to the magnitude, the phase corrected to be continuous by adding correction factors of $\pm 2\pi$ at discontinuities larger than π , and the two components combined again before the subsequent inverse transformation. Correcting the phase spectrum as above is referred to as *phase unwrapping* [39, 42]. It has been shown that a linear phase term, if present in the spectrum of the input signal, may cause rapidly decaying oscillations in the complex cepstrum [42]. It is advisable to remove the linear phase term, if present, in the phase-unwrapping step. The linear phase term may be added to the filtered result (as a time shift), if necessary.

The complex cepstrum of exponential sequences: Exponential sequences are signals that have a rational z -transform, that is, their z -transforms may be expressed as ratios of polynomials in z [39]. Such signals are effectively represented as weighted sums of exponentials in the time domain. Consider a signal $x(n)$ whose z -transform is expressed as

$$X(z) = A z^r \frac{\prod_{k=1}^{M_I} (1 - a_k z^{-1}) \prod_{k=1}^{M_O} (1 - b_k z)}{\prod_{k=1}^{N_I} (1 - c_k z^{-1}) \prod_{k=1}^{N_O} (1 - d_k z)}, \quad (4.67)$$

with $|a_k|, |b_k|, |c_k|, |d_k| < 1$. The function $X(z)$ possesses M_I zeros that are located inside the unit circle, given by a_k , $k = 1, 2, \dots, M_I$; M_O zeros outside the unit circle, given by $1/b_k$, $k = 1, 2, \dots, M_O$; N_I poles inside the unit circle, given by c_k , $k = 1, 2, \dots, N_I$; and N_O poles outside the unit circle, given by $1/d_k$, $k = 1, 2, \dots, N_O$.

In the process of computing the complex cepstrum, we get [39]

$$\begin{aligned}
\hat{X}(z) &= \log[X(z)] \\
&= \log[A] + \log[z^r] \\
&\quad + \sum_{k=1}^{M_I} \log(1 - a_k z^{-1}) + \sum_{k=1}^{M_O} \log(1 - b_k z) \\
&\quad - \sum_{k=1}^{N_I} \log(1 - c_k z^{-1}) - \sum_{k=1}^{N_O} \log(1 - d_k z).
\end{aligned} \tag{4.68}$$

For real sequences, A is real. If $A > 0$, $\hat{x}(0) = \log[A]$.

The factor z^r indicates a delay or advance by r samples, depending upon the sign of r . Recall the power series

$$\log(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots, \quad \text{if } |x| < 1. \tag{4.69}$$

Thus, we have

$$\log(1 - \alpha z^{-1}) = - \sum_{n=1}^{\infty} \frac{\alpha^n}{n} z^{-n}, \quad \text{if } |z| > |\alpha|, \tag{4.70}$$

and

$$\log(1 - \beta z) = - \sum_{n=1}^{\infty} \frac{\beta^n}{n} z^n, \quad \text{if } |z| < |\beta^{-1}|. \tag{4.71}$$

Expanding all of the log terms in Equation 4.68 into their equivalent power series and taking the inverse z -transform, we get [39]

$$\hat{x}(n) = \begin{cases} \log |A| & \text{for } n = 0; \\ - \sum_{k=1}^{M_I} \frac{a_k^n}{n} + \sum_{k=1}^{N_I} \frac{c_k^n}{n} & \text{for } n > 0; \\ \sum_{k=1}^{M_O} \frac{b_k^{-n}}{n} - \sum_{k=1}^{N_O} \frac{d_k^{-n}}{n} & \text{for } n < 0. \end{cases} \tag{4.72}$$

From the steps shown above, a few properties of the complex cepstrum that are evident and important are as follows [39]:

- The complex cepstrum, $\hat{x}(n)$, decays at least as fast as $1/n$. We have

$$|\hat{x}(n)| < K \left| \frac{\alpha^n}{n} \right|, \quad -\infty < n < \infty, \tag{4.73}$$

where $\alpha = \max(|a_k|, |b_k|, |c_k|, |d_k|)$, and K is a constant.

- $\hat{x}(n)$ will be of infinite duration even if $x(n)$ is of finite duration, and exists for $-\infty < n < \infty$, in general.

- If $x(n)$ is a minimum-phase signal, all of its poles and zeros are inside the unit circle in the z -plane. Then, $\hat{x}(n) = 0$ for $n < 0$; that is, the complex cepstrum is causal.
- If $x(n)$ is a maximum-phase signal, it has no poles or zeros inside the unit circle in the z -plane. Then, $\hat{x}(n) = 0$ for $n > 0$; that is, the complex cepstrum is anticausal.

Limiting ourselves to causal signals of finite energy, we need not consider the presence of poles on or outside the unit circle in the z -plane. However, the z -transform of a finite-energy signal may have zeros outside the unit circle. Such a composite mixed-phase signal may be separated into its minimum-phase component and maximum-phase component by extracting the causal part ($n > 0$) and anticausal part ($n < 0$), respectively, of its complex cepstrum, followed by the inverse procedures. The composite signal is equal to the convolution of its minimum-phase component and maximum-phase component. (See also Section 5.4.2.)

Effect of echoes or repetitions of a wavelet: Let us consider a simplified signal $y(n) = x(n) * h(n)$, where

$$x(n) = \delta(n) + a \delta(n - n_0), \quad (4.74)$$

with a and n_0 being two constants. (The sampling interval T is ignored, or assumed to be normalized to unity in this example.) The signal may also be expressed as

$$y(n) = h(n) + a h(n - n_0). \quad (4.75)$$

The signal thus has two occurrences of the basic wavelet $h(n)$ at $n = 0$ and $n = n_0$. The coefficient a indicates the magnitude of the second appearance of the basic wavelet (called an *echo* in seismic applications), and n_0 indicates its delay (pitch in the case of a voiced speech signal). The topmost plot in Figure 4.26 shows a synthesized signal with a wavelet and an echo at one-half of the amplitude (that is, $a = 0.5$) of the first wavelet arriving at $n_0 = 0.01125$ s.

Taking the z -transform of the signal in Equation 4.75, we have

$$Y(z) = (1 + az^{-n_0})H(z). \quad (4.76)$$

If the z -transform is evaluated on the unit circle, we get the Fourier-transform-based expression

$$Y(\omega) = [1 + a \exp(-j\omega n_0)]H(\omega). \quad (4.77)$$

Taking the logarithm, we have

$$\hat{Y}(\omega) = \hat{H}(\omega) + \log[1 + a \exp(-j\omega n_0)]. \quad (4.78)$$

If $a < 1$, the log term may be expanded in a power series, to get

$$\hat{Y}(\omega) = \hat{H}(\omega) + a \exp(-j\omega n_0) - \frac{a^2}{2} \exp(-2j\omega n_0) + \frac{a^3}{3} \exp(-3j\omega n_0) - \dots. \quad (4.79)$$

Taking the inverse Fourier transform, we get

$$\hat{y}(n) = \hat{h}(n) + a \delta(n - n_0) - \frac{a^2}{2} \delta(n - 2n_0) + \frac{a^3}{3} \delta(n - 3n_0) - \dots. \quad (4.80)$$

The derivation given above shows that the complex cepstrum of a signal with a basic wavelet and an echo is equal to the complex cepstrum of the basic wavelet plus a series of impulses at the echo delay and integral multiples thereof [39, 42]. The amplitudes of the impulses are proportional to the echo amplitude (the factor a) and decay for the higher-order repetitions (if $a < 1$). It may be readily seen that, if the signal has multiple echoes or repetitions of a basic wavelet, the cepstrum will possess multiple impulse trains, with an impulse at the arrival time of each wavelet and integral

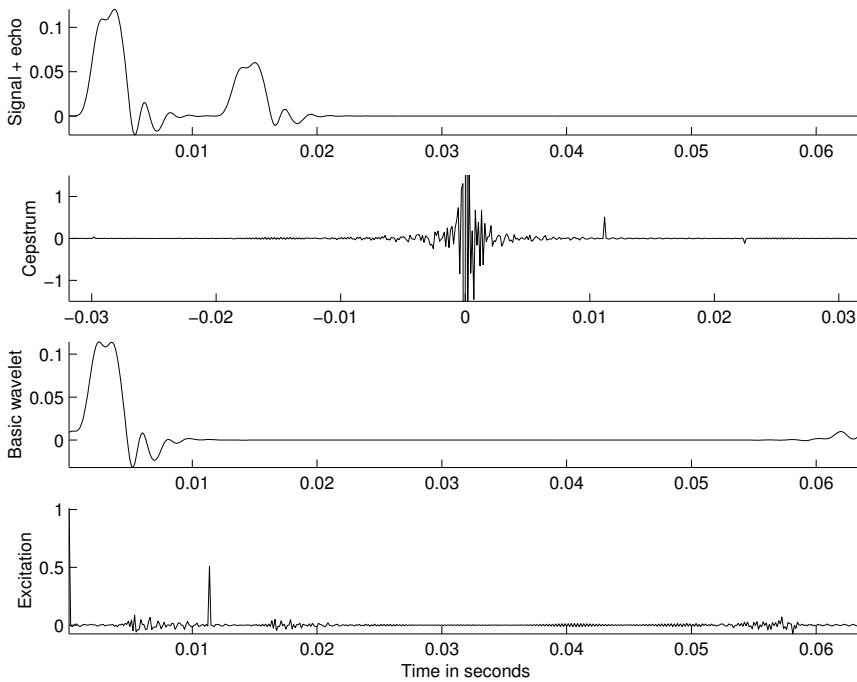


Figure 4.26 From top to bottom: a composite signal with a wavelet and an echo; the complex cepstrum of the signal (the amplitude axis has been stretched to make the peaks at the echo time and its multiples more readily visible; values outside the range ± 1.5 have been clipped); the basic wavelet extracted by shortpass filtering the cepstrum; and the excitation sequence extracted by longpass filtering the cepstrum.

multiples thereof. In the case of a voiced speech signal, the location of the first peak will give the pitch. The second plot in Figure 4.26 shows the complex cepstrum of the signal in the first plot of the same figure. It is seen that the cepstrum has a peak at 0.01125 s , the echo arrival time; a smaller (negative) peak is also seen at twice the echo arrival time.

Under the assumption that the complex cepstrum of the basic wavelet decays to negligible values before the first impulse $a \delta(n - n_0)$ related to the echo, $\hat{h}(n)$ may be extracted from the complex cepstrum $\hat{y}(n)$ of the composite signal by a window that has unit value for $|n| < n_c$, n_c being the cutoff point. (This filter is referred to as a “shortpass” filter as the cepstrum is a function of pseudo time [35, 38]; it may also be called a lowpass filter.) The inverse procedures will yield $h(n)$. The remaining portion of the cepstrum (obtained by “longpass” or highpass filtering) will give $\hat{x}(n)$, which upon application of the inverse procedures will yield $x(n)$. The third and fourth plots in Figure 4.26 show the basic wavelet $h(n)$ and the excitation sequence $x(n)$ extracted by filtering the cepstrum with the cutoff point at $n_c = 0.005\text{ s}$.

In the case where $a \geq 1$, it can be shown that the complex cepstrum will have a train of impulses on its negative time axis, that is, at $(n + kn_0)$, $k = 1, 2, \dots$ [39, 42]. An appropriate exponential weighting sequence may be used to achieve the condition $a < 1$, in which case the impulse train will appear on the positive axis of the cepstrum. If the weighted signal satisfies the minimum-phase condition, the cepstrum will be causal.

The power cepstrum: Several variants of the cepstrum have been proposed in the literature; Childers et al. [42] provide a review of the related techniques. One variant that is commonly used is the *real cepstrum* or the *power cepstrum*, which is defined as the square of the inverse z -transform of the logarithm of the squared magnitude of the z -transform of the given signal. In practice, the z -transform in the definition stated is replaced by the DFT. The power cepstrum has the computational

advantage of not requiring phase unwrapping, but does not facilitate separation of the components of the signal.

By evaluating the inverse z -transform on the unit circle in the z -plane, the power cepstrum $\hat{y}_p(n)$ of a signal $y(n)$ may be defined as

$$\hat{y}_p(n) = \left\{ \frac{1}{2\pi j} \oint \log |Y(z)|^2 z^{n-1} dz \right\}^2. \quad (4.81)$$

If, as before, we consider $y(n) = x(n) * h(n)$, we have $|Y(z)|^2 = |X(z)|^2 |H(z)|^2$, and it follows that $\log |Y(z)|^2 = \log |X(z)|^2 + \log |H(z)|^2$. Applying the inverse z -transform to this relationship, we get

$$\hat{y}_p(n) = \hat{x}_p(n) + \hat{h}_p(n), \quad (4.82)$$

where $\hat{h}_p(n)$ is the power cepstrum of the basic wavelet, and $\hat{x}_p(n)$ is the power cepstrum of the excitation signal. Note that, in the equation given above, the cross-product term was neglected; the cross-term will be zero if the two components of the power cepstra occupy nonoverlapping quefrency ranges. The final squaring operation in Equation 4.81 is omitted in some definitions of the power cepstrum; in such a case, the cross-term does not arise, and Equation 4.82 is valid.

The power cepstrum does not retain the phase information of the original signal. However, it is useful in the identification of the presence of echoes in the signal and estimation of their arrival times. The power cepstrum is related to the complex cepstrum as [42]

$$\hat{y}_p(n) = [\hat{y}(n) + \hat{y}(-n)]^2. \quad (4.83)$$

Let us again consider the situation of a signal with two occurrences of a basic wavelet $h(n)$ at $n = 0$ and $n = n_0$ as in Equations 4.74 and 4.75. Then [42],

$$|Y(z)|^2 = |H(z)|^2 |1 + az^{-n_0}|^2. \quad (4.84)$$

By substituting $z = \exp(j\omega)$ and taking the logarithm of both sides of the equation, we get

$$\begin{aligned} \log |Y(\omega)|^2 &= \log |H(\omega)|^2 + \log[1 + a^2 + 2a \cos(\omega n_0)] \\ &= \log |H(\omega)|^2 + \log(1 + a^2) \\ &\quad + \log \left(1 + \frac{2a}{1 + a^2} \cos(\omega n_0) \right). \end{aligned} \quad (4.85)$$

It is now seen that the logarithm of the PSD of the signal has sinusoidal components (ripples) due to the presence of an echo. The amplitudes and frequencies of the sinusoidal modulation or ripples are related to the amplitude a of the echo and its time delay n_0 . (See Section 7.3 for related discussions and illustrations.)

Illustration of application: A voiced speech signal $y(n)$ is the result of convolution of a slowly varying vocal-tract response $h(n)$ with a relatively fast-varying glottal pulse train $x(n)$, expressed as $y(n) = x(n) * h(n)$. Under these conditions, it may be demonstrated that the contributions of $h(n)$ to the complex cepstrum $\hat{y}(n)$ will be limited to low values of n within the range of the pitch period of the speech signal [39, 44]. The complex cepstrum $\hat{y}(n)$ will possess impulses at the pitch period and integral multiples thereof. Therefore, a filter that selects a portion of the complex cepstrum for low values of n , followed by the inverse transformations, will yield an estimate of $h(n)$.

When the repetitions of the basic wavelet have magnitudes almost equal to (or even greater than) that of the first wavelet in the given signal record, the contributions of the pulse-train component to the complex cepstrum may not decay rapidly and may cause aliasing artifacts when the cepstrum is computed over a finite duration. A similar situation is caused when the delay between the occurrences of the multiple versions of the basic wavelet is a significant portion of the duration of

the given signal. The problem may be ameliorated by applying an exponential weight α^n to the data sequence, with $\alpha < 1$. Childers et al. [42] recommended values of α in the range [0.96, 0.99], depending on the signal characteristics as listed above. Furthermore, they recommend appending or padding zeros to the input signal to facilitate computation of the cepstrum to a longer duration than the signal in order to avoid aliasing errors and ambiguities in time-delay estimates. (See Figure 4.25 for an illustration of the various steps in homomorphic filtering of convolved signals.)

Figure 4.27 illustrates a segment of a voiced speech signal (obtained by filtering the signal shown in Figure 1.55) and the basic wavelet extracted by shortpass filtering of its complex cepstrum with $n_c = 0.003125$ s. The signal was padded with zeros to twice its duration; exponential weighting with $\alpha = 0.99$ was used. It is seen that the basic vocal-tract response wavelet has been successfully extracted. Extraction of the vocal-tract response facilitates spectral analysis without the effect of the quasiperiodic repetitions in the speech signal as indicated by Equation 4.85.

The fourth trace in Figure 4.27 shows the glottal (excitation) waveform extracted by longpass filtering of the cepstrum with the same parameters as in the preceding paragraph. The result shows impulses at the time of arrival of each wavelet in the composite speech signal. The peaks are decreasing in amplitude due to the use of exponential weighting (with $\alpha = 0.99$) prior to computation of the cepstrum. Inverse exponential weighting can restore the pulses to their original levels; however, the artifact at the end of the excitation signal gets amplified to much higher levels than the desired pulses due to progressively higher values of α^{-n} for large n . Hence, the inverse weighting operation was not applied in the present illustration. Regardless, the result indicates that pitch information may also be recovered by homomorphic filtering of voiced speech signals.

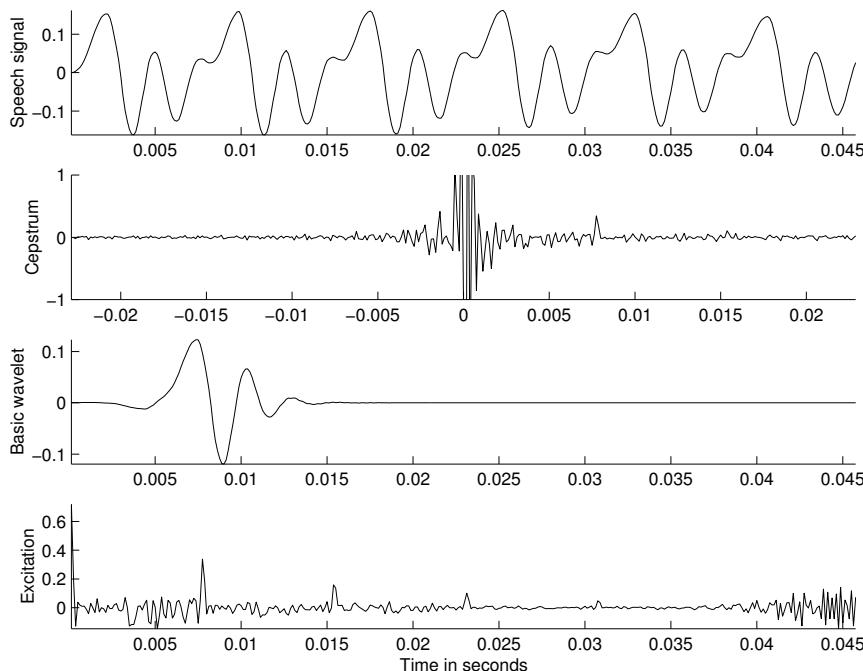


Figure 4.27 From top to bottom: a segment of a voiced speech signal over six pitch periods (extracted from the signal shown in Figure 1.55 and lowpass filtered); the complex cepstrum of the signal (the amplitude axis has been stretched to make the peaks at the echo time and its multiples more readily visible; values outside the range ± 1.0 have been clipped); the (shifted) basic wavelet extracted by shortpass filtering the cepstrum; and the excitation sequence extracted by longpass filtering the cepstrum.

4.8 Application: ECG Rhythm Analysis

Problem: Describe a method to measure the heart rate and average RR interval from an ECG signal.

Solution: Algorithms for QRS detection such as the Pan–Tompkins method described in Section 4.3.2 are useful for ECG rhythm analysis or heart-rate monitoring. The output of the final smoothing filter could be subjected to a peak-searching algorithm to obtain a time marker for each QRS or ECG beat. The search procedure proposed by Pan and Tompkins was explained in Section 4.3.2. The intervals between two such consecutive markers gives the *RR* interval, which could be averaged over a number of beats to obtain a good estimate of the interbeat interval. The heart rate may be computed in *bpm* as 60 divided by the average *RR* interval in seconds. The heart rate may also be obtained by counting the number of beats detected over a certain period, for example, 10 s, and multiplying the result with the required factor (6 in the present case) to get the number of beats over one minute.

The upper plot in Figure 4.28 shows a filtered version of the noisy ECG signal shown in Figure 3.5. The noisy signal was filtered with an eighth-order Butterworth lowpass filter with a cutoff frequency of 90 Hz, and the signal was down-sampled by a factor of five to an effective sampling rate of 200 Hz. The lower plot shows the output of the Pan–Tompkins method. The Pan–Tompkins result was normalized by dividing by its maximum over the data record available (as the present example was computed off-line). A fixed threshold of 0.1 and a blanking interval of 250 ms were used in a simple search procedure, which was successful in detecting all of the beats in the signal. (The blanking interval indicates the period over which threshold checking is suspended once the threshold has been crossed.) The average *RR* interval was computed as 716 ms, leading to an effective heart rate of 84 bpm.

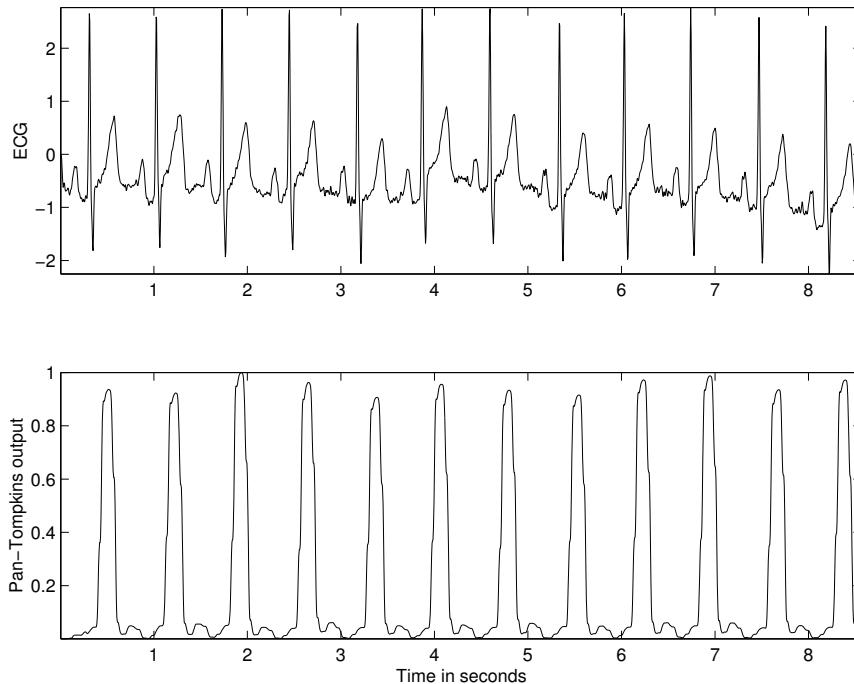


Figure 4.28 Results of the Pan–Tompkins algorithm. Top: lowpass-filtered version of the ECG signal shown in Figure 3.5. Bottom: normalized result of the final integrator.

Results at the various stages of the Pan–Tompkins algorithm for a noisy ECG signal sampled at 200 Hz are shown in Figure 4.29. The bandpass filter has efficiently removed the low-frequency artifact in the signal. The final output has two peaks that are much larger than the others: one at the beginning of the signal due to filtering artifacts, and one at about 7.5 s due to an artifact in the signal. Furthermore, the output has two peaks for the beat with an artifact at 7.5 s. The simple peak-searching procedure as explained in the previous paragraph was applied, which resulted in the detection of 46 beats: one more than the 45 present in the signal due to the artifact at about 7.5 s. The average RR interval was computed to be 446.6 ms, leading to an effective heart rate of 137 bpm.

The illustration demonstrates the need to prefilter a given ECG signal to remove artifacts and the need to apply an adaptive threshold to the output of the Pan–Tompkins algorithm for QRS detection. It is readily seen that direct thresholding of the original ECG signal will not be successful in detecting all of the QRS complexes in the signal.

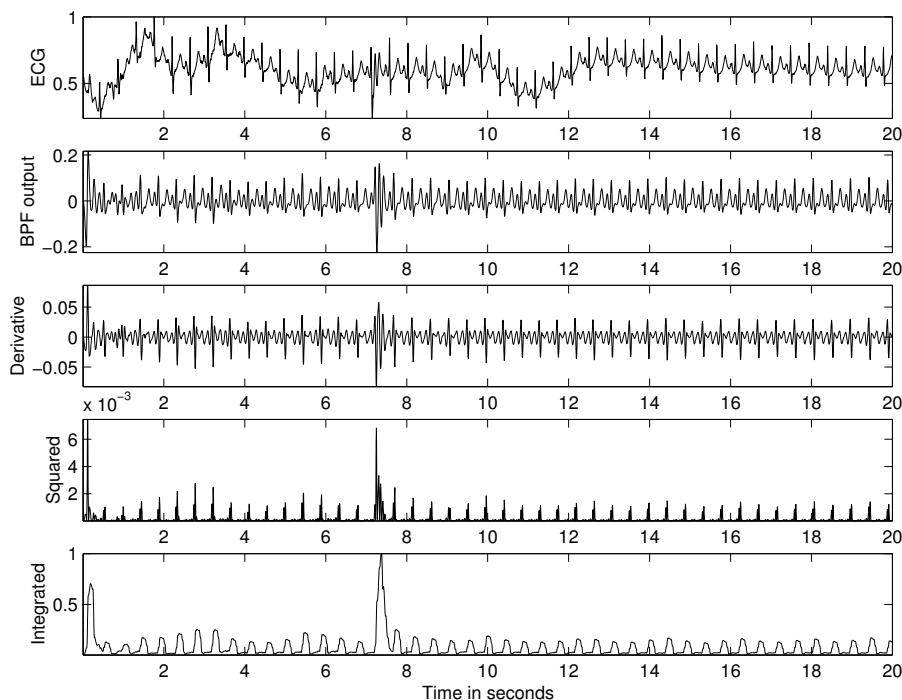


Figure 4.29 Results of the Pan–Tompkins algorithm with a noisy ECG signal. From top to bottom: ECG signal sampled at 200 Hz; output of the bandpass filter (BPF); output of the derivative-based operator; the result of squaring; and normalized result of the final integrator.

4.9 Application: Identification of Heart Sounds

Problem: Outline a signal processing algorithm to identify S1 and S2 in a PCG signal, and further segment the PCG signal into its systolic and diastolic parts. The ECG and carotid pulse signals are available for reference.

Solution: We saw in Section 2.3 how the ECG and carotid pulse signals could be used to demarcate the onset of S1 and S2 in the PCG; the procedure, however, was not based on signal processing but on visual analysis of the signals. We have, in the present chapter, studied signal processing

techniques to detect the QRS complex in the ECG and the dicrotic notch in the carotid pulse signal. We may, therefore, use these methods to transfer the timing information from the ECG and carotid pulse signals to the PCG signal. In order to perform this task, we need to recognize a few timing relationships between the signals [2, 19].

The beginning of S1 may be taken to be the same instant as the beginning of the QRS. The QRS itself may be detected using one of the three methods described in the present chapter, such as the Pan–Tompkins method.

Detection of the beginning of S2 is more involved. Let the heart rate be $HR \text{ bpm}$. The pre-ejection period, PEP , is defined as the interval from the beginning of the QRS to the onset of the corresponding carotid upstroke. The rate-corrected PEP is defined as $PEPC = PEP + 0.4HR$, with the periods in ms . $PEPC$ is in the range of $131 \pm 13 ms$ for normal adults [2].

The ejection time, ET , is the interval from the onset of the carotid upstroke to the dicrotic notch. The rate-corrected ejection time in ms is $ETC = ET + 1.6HR$, and is in the ranges of $395 \pm 13 ms$ for normal adult males and $415 \pm 11 ms$ for normal adult females.

Using $PEPC_{\max} = 144 ms$ and $HR_{\min} = 60 \text{ bpm}$, we get $PEP_{\max} = 120 ms$. With $HR_{\min} = 60 \text{ bpm}$ and $ETC_{\max} = 425 ms$, we get $ET_{\max} = 329 ms$. Based on the parameters defined in the preceding sentences, the maximum interval between a QRS wave and the corresponding dicrotic notch can be approximated to be $380 ms$. The procedure proposed by Lehner and Rangayyan [19] for detection of the dicrotic notch recommends searching the output of the derivative-based method described in Section 4.3.5 in a $500 ms$ interval after the QRS. After the dicrotic notch is detected, we need to subtract the time delay (S2–D) between the beginning of S2 and the dicrotic notch to get the time instant where S2 begins. Lehner and Rangayyan [19] measured the average S2–D delay over the PCG and carotid pulse signals of 60 pediatric patients to be $42.6 \pm 5 ms$.

The following procedure may be used to segment a PCG signal into its systolic and diastolic parts.

1. Use the Pan–Tompkins method described in Section 4.3.2 to locate the QRS complexes in the ECG.
2. Identify one period of the PCG as the interval between two successive QRS locations. Note that the delay introduced by the filters used in the Pan–Tompkins method needs to be subtracted from the detected peak locations to obtain the starting points of the QRS complexes.
3. Use the Lehner and Rangayyan method described in Section 4.3.5 to detect the dicrotic notch in the carotid pulse signal.
4. Let the standardized S2–D delay be its mean $+2 \times SD$ as reported by Lehner and Rangayyan [19], that is, $52.6 ms$. Subtract the standardized S2–D delay from the detected dicrotic notch location to obtain the onset of S2.
5. Use the S1–S2 interval to obtain the systolic part of the PCG cycle.
6. Use the interval between the S2 point and the next detected S1 to obtain the diastolic part of the PCG cycle.

Figures 4.30 and 4.31 illustrate the results of application of the procedure described above to the PCG, ECG, and carotid pulse signals of a normal subject and a patient with a split S2, systolic ejection murmur, and opening snap of the mitral valve. (Clinical diagnosis indicated the possibility of ventricular septal defect, pulmonary stenosis, or pulmonary hypertension for the 14-month-old female patient with murmur.) The peak positions detected in the output of the Pan–Tompkins method (the third trace in each figure) and the output of the Lehner and Rangayyan method (the fifth trace) have been marked with the * symbol. A threshold of 0.75 times the maximum value was used to detect the peaks in the output of the Pan–Tompkins method, with a blanking interval of $250 ms$.

The QRS and dicrotic notch positions have been marked on the ECG and carotid pulse traces with the triangle and diamond symbols, respectively. The S1 and S2 positions are marked on the PCG trace with triangles and diamonds, respectively. The filter delays and timing relationships between the three channels of signals described previously have been accounted for in the process of marking the events. Note how the results of event detection in the ECG and carotid pulse signals have been transferred to locate the corresponding events in the PCG. Lehner and Rangayyan [19] used a similar procedure to break PCG signals into systolic and diastolic segments; the segments were then analyzed separately in the time and frequency domains. (See also Sections 6.3.6 and 7.10.)

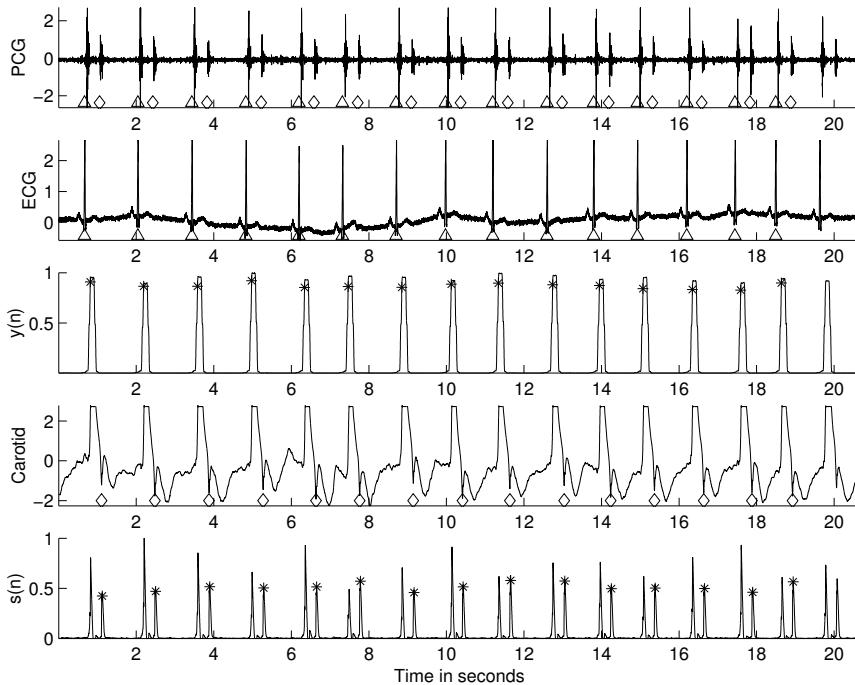


Figure 4.30 Results of segmentation of a PCG signal into systolic and diastolic parts using the ECG and carotid pulse signals for reference. From top to bottom: the PCG signal of a normal subject (male, 23 years); the ECG signal; $y(n)$, the output of the Pan–Tompkins method for detection of the QRS after normalization to the range $[0, 1]$; the carotid pulse signal; and $s(n)$, the output of the Lehner and Rangayyan method for detection of the dicrotic notch, normalized to the range $[0, 1]$. The peaks detected in the outputs of the two methods have been identified with * marks. The QRS and the dicrotic notch positions have been marked with the triangle and diamond symbols, respectively. The S1 and S2 positions are marked on the PCG trace with triangles and diamonds, respectively. The last cardiac cycle was not processed.

4.10 Application: Detection of the Aortic Component of the Second Heart Sound

Heart sounds are preferentially transmitted to different locations on the chest. The aortic and pulmonary components A2 and P2 of S2 are best heard at the aortic area (to the right of the sternum, in the second right intercostal space) and the pulmonary area (left parasternal line in the third left intercostal space), respectively (see Figure 1.33). A2 is caused by the closure of the aortic valve at the end of systole, and is usually louder than P2 at all locations on the chest. Earlier theories on the genesis of heart sounds attributed the sounds to the opening and closing actions of the valve leaflets *per se*. The more commonly accepted theory is that described by Rushmer [1]; see Section 1.2.9.

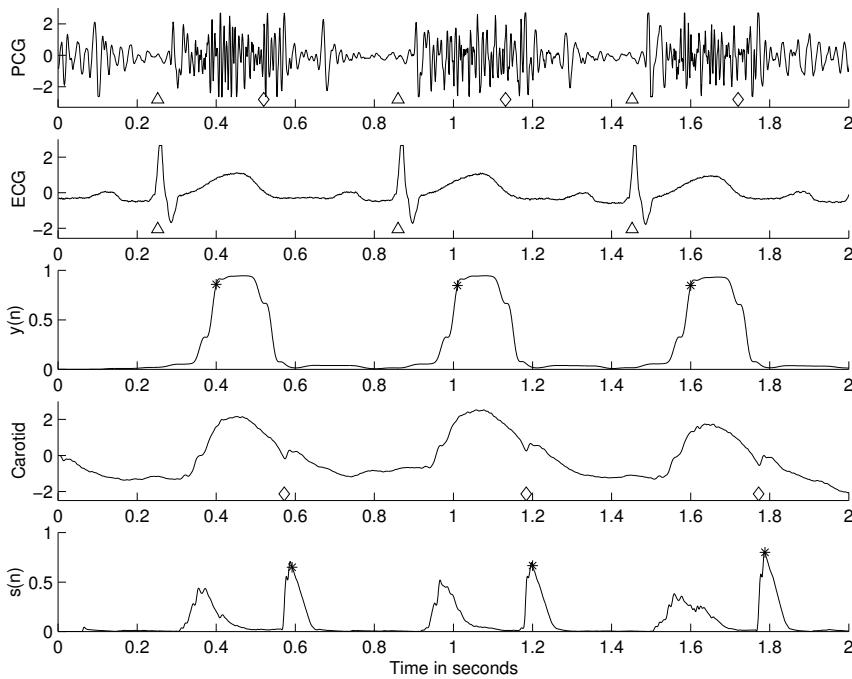


Figure 4.31 Results of segmentation of a PCG signal into systolic and diastolic parts using the ECG and carotid pulse signals for reference. From top to bottom: the PCG signal of a patient (female, 14 months) with a split S2, systolic ejection murmur, and opening snap of the mitral valve; the ECG signal; $y(n)$, the output of the Pan–Tompkins method for detection of the QRS; the carotid pulse signal; $s(n)$, the output of the Lehner and Rangayyan method for detection of the diastolic notch. The peaks detected in the outputs of the two methods have been identified with * marks. The QRS and the diastolic notch positions have been marked with the triangle and diamond symbols, respectively. The S1 and S2 positions are marked on the PCG trace with triangles and diamonds, respectively. The values of $y(n)$ and $s(n)$ have been normalized to the range [0, 1].

The relative timing of A2 and P2 depends on the pressure differences across the corresponding valves in the left and right ventricular circulatory systems. In a normal individual, the timing of P2 with reference to A2 varies with respiration; the timing of A2 itself is independent of respiration. The pulmonary pressure (intrathoracic pressure) is decreased during inspiration, leading to a delayed closure of the pulmonary valve and hence an increased (audible and visible) gap between A2 and P2 [1, 2, 45]. The gap is closed, and A2 and P2 overlap during expiration in normal individuals. A2 and P2 have individual durations of about 50 ms. The normal inspiratory gap between A2 and P2 is of the order of 30 – 40 ms, although splits as long as 100 ms have been recorded [2].

A split in S2 longer than 40 ms during sustained expiration is considered to be abnormal [2]. Complete right bundle-branch block could cause delayed activation of the right ventricle and, therefore, delayed pulmonary valve closure, a delayed P2, and hence a widely split S2. Some of the other conditions that cause a wide split in S2 are atrial septal defect, ventricular septal defect, and pulmonary stenosis. Left bundle-branch block leads to delayed left ventricular contraction and aortic valve closure (with reference to the right ventricle and the pulmonary valve), causing A2 to appear after P2, and *reversed splitting* of the two components. Some of the other conditions that could cause reversed splitting of S2 are aortic insufficiency and abnormally early pulmonary valve closure. It is thus seen that identification of A2 and P2 and their temporal relationships could assist in the diagnosis of several cardiovascular defects and diseases.

MacCanon et al. [46] conducted experiments on a dog for direct detection and timing of aortic valve closure. They developed a catheter with an electrical contacting device that could be placed at the aortic valve to detect the exact moment of closure of the aortic valve. They also measured the aortic pressure and the PCG at the third left intercostal space. It was demonstrated that the aortic valve closes at least 5 – 13 ms before the incisura appears in the aortic pressure wave. (See Figures 1.49 and 1.50 for illustrations of aortic pressure waves recorded from a dog.) The conclusion reached was that S2 is caused not by the collision of the valve leaflets themselves, but due to the rebound of the aortic blood column and walls after valve closure. MacCanon et al. also hypothesized that the relative high-frequency characteristics of the incisura and S2 result from elastic recoil of the aortic wall and valve in reaction to the distention by the rebounding aortic blood column.

Stein et al. [47] and Stein and Sabbah [48] conducted experiments in which intracardiac and intraarterial sounds were recorded and analyzed. Their experiments indicated that S2 begins *after* the aortic valve closes. They argued that the intensity of S2 depends on, among other factors, the distensibility of the aortic and pulmonary valves; hemodynamic factors that cause the valves to distend and vibrate; viscosity of the blood and its ability to inhibit diastolic valve motion; and the configuration of the aorta, the pulmonary artery, and the ventricles. It was demonstrated that the pulmonary valve, due to its larger surface area than that of the aortic valve, is more distensible and hence produces a larger sound than the aortic valve even for the same pressure gradient across the valve. In the case of pulmonary hypertension, it was argued that the pulmonary valve would distend further at a higher speed: the rate of development of the diastolic pressure gradient across the closed pulmonary valve would be higher than that in normal cases.

Problem: *Given that S2 is made up of an aortic component A2 and a pulmonary component P2 with variable temporal relationships, propose a method to detect only A2.*

Solution: We have seen in the preceding section how the dicrotic notch in the carotid pulse signal may be used to detect the beginning of S2. The technique is based on the direct relationship between aortic valve closure and the aortic incisura, and consequently the dicrotic notch, as explained above. Now, if we were to detect and segment S2 over several cardiac cycles and several respiratory cycles, we could perform synchronized averaging of S2. A2 should appear at the same instant in every S2 segment and should be strengthened by the synchronized averaging process. Assuming that the subject is breathing normally, P2, on the other hand, would appear at different times, and should be cancelled out (suppressed) by the averaging process.

Figure 4.32 shows segments of duration of 300 ms containing S2 segmented from nine successive cardiac cycles of the PCG of a patient with atrial septal defect. The PCG signal was segmented using the ECG and carotid pulse signals for reference in a method similar to that illustrated in Figures 4.30 and 4.31. The PCG signal was recorded at the second left intercostal space, which is closer to the pulmonary area than to the aortic area. The nine S2 segments clearly show the fixed timing of A2 and the variable timing of P2. The last plot is the average of S2 segments extracted from 21 successive cardiac cycles. The averaged signal displays A2 very well, while P2 has been suppressed.

The detection of A2 could have been better had the PCG been recorded at the aortic area, where A2 would be stronger than P2. Once A2 is detected, it could be subtracted from each S2 record to obtain individual estimates of P2. Sarkady et al. [49], Baranek et al. [50], and Durand et al. [51] proposed averaging techniques as above with or without envelope detection (but without the use of the carotid pulse); the methods were called aligned ensemble averaging to detect wavelets or coherent detection and averaging.

Nigam and Priemer [52] proposed a method to extract the A2 and P2 components from S2 via blind source separation (BSS) techniques. Xu et al. [53, 54] used time–frequency representations of transient nonlinear chirp signals to model A2 and P2 and derive the A2–P2 splitting interval; they showed that the splitting interval is strongly correlated with the pulmonary arterial pressure measured by catheterization.

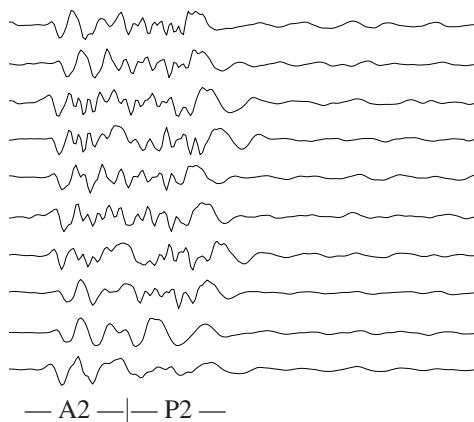


Figure 4.32 Synchronized averaging of S2 to detect A2 and suppress P2. The figure displays nine consecutive segments of S2 over a duration of 300 ms each, including S2 and the subsequent part of the corresponding diastole, of a patient (female, 7 years) with atrial septal defect leading to a variable split in S2. The trace at the bottom is the average of the segments over 21 consecutive beats. The expected time intervals of A2 and P2 are indicated below the last trace.

4.11 Remarks

We have now established links between the characteristics of certain epochs in a number of biomedical signals and the corresponding physiological or pathological events in the biomedical systems of concern. We have seen how derivative-based operators may be applied to detect QRS complexes in the ECG signal as well as the dicrotic notch in the carotid pulse signal. The utility of correlation and spectral density functions in the detection of rhythms and events in EEG signals was also demonstrated. We have studied how signals with repetitions of a certain event or wavelet, such as a voiced speech signal, may be analyzed using the complex cepstrum and homomorphic filtering. Finally, we also saw how events detected in one signal may be used to locate the corresponding or related events in other signals: the task of detecting S1 and S2 in the PCG was made simpler by using the ECG and carotid pulse signals, in which the QRS wave and the dicrotic notch can be detected more readily than the heart sounds themselves.

Event detection is an important step that is required before we may attempt to analyze the corresponding waves or wavelets in more detail. After a specific wave of interest has been detected, isolated, and extracted, methods targeted to the expected characteristics of the event may be applied for specifically focused analysis of the corresponding physiological or pathological event. Analysis of the event is then not hindered or obscured by other events or artifacts in the acquired signal.

4.12 Study Questions and Problems

1. Prove that the ACF $\phi_{xx}(\tau)$ of a stationary function $x(t)$ is maximum at $\tau = 0$. *Hint:* Start with $E[\{x(t + \tau) \pm x(t)\}^2] \geq 0$.
2. For a stationary process x , prove that the ACF is even-symmetric, that is, $\phi_{xx}(\tau) = \phi_{xx}(-\tau)$. You may use the expectation or time-average definition of the ACF.
3. Starting with the continuous time-average definition of the ACF, prove that the Fourier transform of the ACF is the PSD of the signal.
4. What are the Fourier-domain equivalents of the ACF and the CCF? Describe their common features and differences. List their applications in biomedical signal analysis.

5. A signal $x(t)$ is transmitted through a channel. The received signal $y(t)$ is a scaled, shifted, and noisy version of $x(t)$ given as $y(t) = \alpha x(t - t_0) + \eta(t)$ where α is a scale factor, t_0 is the time delay, and $\eta(t)$ is noise. Assume that the noise process has zero mean and is statistically independent of the signal process, and that all processes are stationary. Derive expressions for the mean and the ACF of $y(t)$ in terms of the statistics of x and η .
6. Derive an expression for the ACF of the signal $x(t) = \sin(\omega_0 t)$. Use the time-average definition of the ACF. From the ACF, derive an expression for the PSD of the signal. Show all steps.
7. A rhythmic episode of a theta wave in an EEG signal is approximated by a researcher to be a sine wave of frequency 5 Hz. The signal is sampled at 100 Hz. Draw a schematic representation of the ACF of the episode for delays up to 0.5 s. Label the time axis in samples and in seconds. Draw a schematic representation of the PSD of the episode. Label the frequency axis in Hz.
8. The values of a signal sampled at 100 Hz are given by the series $\{0, 0, 0, 0, 10, 10, 10, 0, 0, 0, 0, 0, 5, 5, 5, 0, 0, 0, 0, 0, -3, -3, -3, 0, 0, 0\}$. An investigator performs template matching with the pattern $\{0, 5, 5, 5, 0\}$. The first sample in each array stands for zero time. Plot the output of the template matching operation and interpret the result. Label the time axis in seconds.
9. You are given a signal with the samples $\{0, 0, 2, 2, 3, -3, 2, 0, 0\}$ and a template with the samples $\{1, -1\}$. Perform the template matching operation and derive the sample values for the output. Provide an interpretation of the result.
10. Discuss the similarities and differences between the problems of (i) detection of spike transients in EEG signals, and (ii) the detection of QRS complexes in ECG signals.
11. You have been hired to develop a heart-rate monitor for use in a coronary-care unit. Design a system to accept a patient's ECG signal, filter it to remove artifacts and noise, sample the signal, measure the heart rate, and set off alarms as appropriate. Provide a block diagram of the system, with details of the signal processing steps to be performed in each block. Specify the important parameters for each processing step.
12. A biphasic signal is given by the series of samples $x(n) = \{0, 1, 2, 1, 0, -1, -2, -1, 0\}$ for $n = 0, 1, 2, \dots, 8$. (a) Draw a plot of $x(n)$. (b) Compose a signal $y(n)$ defined as $y(n) = 3x(n) + 2x(n - 12) - x(n - 24)$. Draw a plot of $y(n)$. (c) Design a matched filter to detect the presence of $x(n)$ in $y(n)$. Explain how the impulse response $h(n)$ and the frequency response $H(\omega)$ of the filter are related to $x(n)$. Plot $h(n)$. (d) Compute the output of the filter and plot it. Interpret the output of the filter.
13. A researcher uses the derivative operator (filter) specified as $w(n) = x(n) - x(n - 1)$, where $x(n)$ is the input and $w(n)$ is the output. The result is then passed through the MA filter $y(n) = \frac{1}{3}[w(n) + w(n - 1) + w(n - 2)]$, where $y(n)$ is the final output desired. (a) Derive the transfer functions (in the z-domain) of the two filters individually as well as that of the combination. (b) Does it matter which of the two filters is placed first? Why (not)? (c) Derive the impulse response of each filter and that of the combination. Plot the three signals. (d) The signal described by the samples $\{0, 0, \dots, 0, 6, 6, 6, 6, 6, 6, 6, 6, 0, 0, \dots\}$ is applied to the system. Derive the values of the final output signal. Explain the effect of the operations on the signal.
14. You have been hired to develop two software packages for the analysis of 10-channel EEG signals for the following purposes: (a) Detection of the presence of the alpha rhythm: (i) in any one single channel and (ii) jointly in a pair of left-right channels. (b) Detection of spike-and-wave complexes of a prespecified shape in any channel.

Design two signal processing packages to address the two problems mentioned above. For each package, provide the following details: (i) A schematic block diagram representing the various algorithms or signal processing steps that you recommend. Include at least three distinct and nontrivial procedures in your design. (ii) Explain each block or part of your package. Describe the reason or logic behind your recommendation. (iii) Write at least two nontrivial equations related to your procedures. (iv) Provide graphical illustrations representing a typical EEG signal as it is processed by each block of your package.

15. The template of a signal is specified by the series of samples $\{2, 3, -1\}$ for $n = 0, 1, 2$. (a) Design a matched filter to detect the signal. Give the impulse response and transfer function of the filter. (b) The signal specified by the series of samples $\{1, -1, -2, 4, 6, -2, 1, -1, 0\}$ is applied to the matched filter. Compute the output and explain its characteristics.

16. Describe the significance of the P wave in the analysis of ECG signals. Describe a method for the detection of P waves in an ECG signal. Explain the purpose and reasoning behind each step of your algorithm. Give at least one nontrivial equation representing a procedure in your algorithm. Draw schematic sketches representing a sample input signal and the corresponding output at each stage of your method.
17. Draw a block diagram representing the various steps in the Pan–Tompkins method to detect QRS complexes in ECG signals. Explain the purpose and nature of each step in the procedure, including detection of the peaks in the output corresponding to the QRS complexes. No equations are required in your answer to this question.
Draw a schematic sketch of a noisy ECG signal including three cardiac cycles with high-frequency and low-frequency noise, and illustrate how it is modified by each step. Explain how the result could be used to estimate the heart rate of a patient.
18. A researcher in signal processing seeks your help to design filters to obtain a smoothed estimate of the first derivative (difference) of a signal digitized at the sampling rate of 200 Hz. Help the researcher with the following: (a) Give the difference equation and impulse response of a filter to compute the first derivative (difference) of the signal. (b) Give the transfer function of the derivative filter. Derive the magnitude and phase parts of the frequency response of the filter. (c) Draw a sketch of the magnitude of the frequency response of the derivative filter. Explain the nature of the filter by deriving the gain at 0 Hz and 100 Hz. (d) Give the difference equation and impulse response of a filter to compute the average of the current input sample and the previous input sample (that is, an MA filter of length 2). (e) Give the transfer function of the MA filter. Derive the magnitude and phase parts of the frequency response of the filter. (f) Draw a sketch of the magnitude of the frequency response of the MA filter. Explain the nature of the filter by deriving the gain at 0 Hz and 100 Hz. (g) Derive the impulse response and the difference equation of a combined filter with the derivative and MA filters in series. (h) Derive the transfer function and plot the pole–zero diagram of the combined filter. (i) Derive the magnitude and phase parts of the frequency response of the combined filter. Draw a sketch of the magnitude of the frequency response of the filter. Explain the nature of the filter by deriving the gain at 0 Hz and 100 Hz. (j) Draw a signal-flow diagram of the combined filter.
19. Describe a procedure for the detection of QRS complexes in an ECG signal. Give a schematic block diagram of the important steps in the procedure. Include at least four steps. Give one suitable equation for each step of your algorithm either in the discrete-time domain as a difference equation or in the z-domain as a transfer function. Sketch an ECG signal with at least two beats and show the corresponding output at the various steps of your procedure. Explain the effect of each step of the procedure on the signal.

4.13 Laboratory Exercises and Projects

Note: Data files related to the exercises are available at the site

<https://github.com/srikrishnan1972/Biomedical-Signal-Analysis>

1. Implement the Pan–Tompkins method for QRS detection. You may employ a simple threshold-based method to detect QRS complexes as the procedure will be run off-line.
Apply the procedure to the signals in the files ECG3.dat, ECG4.dat, ECG5.dat, and ECG6.dat, sampled at a rate of 200 Hz (see the file ECGS.m). Compute the averaged heart rate and QRS width for each record. Verify your results by measuring the parameters visually from plots of the signals.
2. The files eeg1-xx.dat (where xx indicates the channel name) give eight simultaneously recorded channels of EEG signals with the alpha rhythm. (You may read the signals using the MATLAB® program in the file eeg1.m.) The sampling rate is 100 Hz per channel. Cut out a portion of a signal with a clear presence of the alpha rhythm for use as a template or reference signal. Perform cross-correlation of the template with running (short-time) windows of various channels and study the use of the results for the detection of the presence of the alpha rhythm.
3. The files eeg2-xx.dat (where xx indicates the channel name) give 10 simultaneously recorded channels of EEG signals with spike-and-wave complexes. (You may read the signals using the MATLAB® program

in the file eeg2.m.) The sampling rate is 100 Hz per channel. Cut out one spike-and-wave complex from any EEG channel and use it as a template. Perform template matching by cross-correlation or by designing a matched filter. Apply the procedure to the same channel from which the template was selected as well as to other channels. Study the results and explain how they may be used to detect spike-and-wave complexes.

4. The files pec1.dat, pec33.dat, and pec52.dat give three-channel recordings of the PCG, ECG, and carotid pulse signals (sampled at 1,000 Hz ; you may read the signals using the program in the file plotpec.m). The signals in pec1.dat (adult male) and pec52.dat (male subject, 23 years) are normal; the PCG signal in pec33.dat has systolic murmur, and is of a patient suspected to have pulmonary stenosis, ventricular septal defect, and pulmonary hypertension (female, 14 months).
Apply the Pan-Tompkins method for QRS detection to the ECG channel and the Lehner and Rangayyan method to detect the dicrotic notch in the carotid pulse channel. Extrapolate the timing information from the ECG and carotid pulse channels to detect the onset of S1 and S2 in the PCG channel. What are the corrections required to compensate the delays between the corresponding events in the three channels?
5. Write a program to compute the ACF of a given signal. Apply the program to analyze (a) an ECG signal over several cardiac cycles, and (b) a voiced-speech signal. What information are you able to obtain from the peaks in the ACF?
6. Write a program to perform template matching. Apply the program to detect all QRS complexes in an ECG signal with several cardiac cycles, including PVCs. Modify the program to detect each occurrence of the basic wavelet in a voiced-speech signal. Apply the program to ECG and voiced-speech signals with noise and study the performance of the method in the presence of noise.
7. Write a program to design a matched filter to detect all QRS complexes in an ECG signal with several cardiac cycles. Apply the method to an ECG signal including PVCs and analyze the results. Modify the program to detect each occurrence of the basic wavelet in a voiced-speech signal. Apply the program to ECG and speech signals with noise and study the performance of the matched filter in the presence of noise.
8. Compute the PSD of an ECG signal spanning several cardiac cycles. Extract a part of the same signal spanning only one cardiac cycle and compute its PSD with the same number of samples as the previous one. Compare the PSDs and explain the similarities and differences between them.
9. Let the signal spanning one cardiac cycle from the previous exercise be called $x(n)$ with N samples. Prepare a signal $d(n)$ with a train of 20 impulses with a period of N samples. Obtain the signal $y_1(n) = x(n) * d(n)$ by convolving $x(n)$ and $d(n)$ in the time domain. Compute the PSDs of $x(n)$, $d(n)$, and $y_1(n)$ with the same number of samples. Compute $Y_2(f) = X(f)D(f)$ by multiplying the Fourier transforms of the signals $x(n)$ and $d(n)$ obtained with the same number of samples. Obtain $y_2(n)$ as the inverse Fourier transform of $Y_2(f)$ and compute the related PSD. Plot all of the signals and the PSDs. Analyze the signals and PSDs and explain their characteristics.
10. Compute the power cepstrum of each of the signals in the previous exercise. Prepare a signal with two cycles of the ECG and compute its cepstrum. Analyze the signals and their cepstra and explain their characteristics.

References

- [1] Rushmer RF. *Cardiovascular Dynamics*. WB Saunders, Philadelphia, PA, 4th edition, 1976.
- [2] Tavel ME. *Clinical Phonocardiography and External Pulse Recording*. Year Book Medical, Chicago, IL, 3rd edition, 1978.
- [3] Cooper R, Osseston JW, and Shaw JC. *EEG Technology*. Butterworths, London, UK, 3rd edition, 1980.
- [4] Kooi KA, Tucker RP, and Marshall RE. *Fundamentals of Electroencephalography*. Harper & Row, Hagerstown, MD, 2nd edition, 1978.
- [5] Hughes JR. *EEG in Clinical Practice*. Butterworth, Woburn, MA, 1982.
- [6] Dumermuth G, Huber PJ, Kleiner B, and Gasser T. Numerical analysis of electroencephalographic data. *IEEE Transactions on Audio and Electroacoustics*, 18(4):404–411, 1970.

- [7] Barlow JS. Computerized clinical electroencephalography in perspective. *IEEE Transactions on Biomedical Engineering*, 26(7):377–391, 1979.
- [8] Bodenstein G and Praetorius HM. Feature extraction from the electroencephalogram by adaptive segmentation. *Proceedings of the IEEE*, 65(5):642–652, 1977.
- [9] Balda RA, Diller G, Deardorff E, Doue J, and Hsieh P. The HP ECG analysis program. In van Beimmel JH and Willems JL, editors, *Trends in Computer-processed Electrocardiograms*, pages 197–205. North Holland, Amsterdam, The Netherlands, 1977.
- [10] Ahlstrom ML and Tompkins WJ. Digital filters for real-time ECG signal processing using microprocessors. *IEEE Transactions on Biomedical Engineering*, 32:708–713, 1985.
- [11] Friesen GM, Jannett TC, Jadallah MA, Yates SL, Quint SR, and Nagle HT. A comparison of the noise sensitivity of nine QRS detection algorithms. *IEEE Transactions on Biomedical Engineering*, 37(1):85–97, 1990.
- [12] Tompkins WJ. *Biomedical Digital Signal Processing*. Prentice-Hall, Upper Saddle River, NJ, 1995.
- [13] Murthy ISN and Rangaraj MR. New concepts for PVC detection. *IEEE Transactions on Biomedical Engineering*, 26(7):409–416, 1979.
- [14] Pan J and Tompkins WJ. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, 32:230–236, 1985.
- [15] Hengeveld SJ and van Beimmel JH. Computer detection of P waves. *Computers and Biomedical Research*, 9:125–132, 1976.
- [16] Gritzali F, Frangakis G, and Papakonstantinou G. Detection of the P and T waves in an ECG. *Computers and Biomedical Research*, 22:83–91, 1989.
- [17] Willems JL, Arnaud P, van Beimmel JH, Bourdillon PJ, Brohet C, Volta SD, Andersen JD, Degani R, Denis B, Demeester M, Dudeck J, Harms FMA, Macfarlane PW, Mazzocca G, Meyer J, Michaelis J, Pardaens J, Pöppel SJ, Reardon BC, van Eck HJR, de Medina EOR, Rubel P, Talmon JL, and Zywietz C. Assessment of the performance of electrocardiographic computer programs with the use of a reference data base. *Circulation*, 71(3):523–534, 1985.
- [18] Willems JL, Arnaud P, van Beimmel JH, Bourdillon PJ, Degani R, Denis B, Harms FMA, Macfarlane PW, Mazzocca G, Meyer J, van Eck HJR, de Medina EOR, and Zywietz C. Establishment of a reference library for evaluating computer ECG measurement programs. *Computers and Biomedical Research*, 18:439–457, 1985.
- [19] Lehner RJ and Rangayyan RM. A three-channel microcomputer system for segmentation and characterization of the phonocardiogram. *IEEE Transactions on Biomedical Engineering*, 34:485–489, 1987.
- [20] Starmer CF, McHale PA, and Greenfield Jr. JC. Processing of arterial pressure waves with a digital computer. *Computers and Biomedical Research*, 6:90–96, 1973.
- [21] Jenkins JM, Wu D, and Arzbaecher RC. Computer diagnosis of abnormal cardiac rhythms employing a new P-wave detector for interval measurement. *Computers and Biomedical Research*, 11:17–33, 1978.
- [22] Yadav R, Swamy MNS, and Agarwal R. Model-based seizure detection for intracranial EEG recordings. *IEEE Transactions on Biomedical Engineering*, 59(5):1419–1428, 2012.
- [23] Qu H and Gotman J. A patient-specific algorithm for the detection of seizure onset in long-term EEG monitoring: Possible use as a warning device. *IEEE Transactions on Biomedical Engineering*, 44(2):115–122, 1997.
- [24] Grewal S and Gotman J. An automatic warning system for epileptic seizures recorded on intracerebral EEGs. *Clinical Neurophysiology*, 116:2460–2472, 2005.
- [25] Bendat JS and Piersol AG. *Random Data: Analysis and Measurement Procedures*. Wiley, New York, NY, 2nd edition, 1986.
- [26] Cadzow JA and Solomon, Jr. OM. Linear modeling and the coherence function. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(1):19–48, 1987.
- [27] Dobie RA and Wilson MJ. Analysis of auditory evoked potentials by magnitude-squared coherence. *Ear and Hearing*, 10(1):2–13, 1989.

- [28] Rosenberg JR, Amjad AM, Breeze P, Brillinger DR, and Halliday DM. The Fourier approach to the identification of functional coupling between neuronal spike trains. *Progress in Biophysics and Molecular Biology*, 53:1–31, 1989.
- [29] Amjad AM, Halliday DM, Rosenberg JR, and Conway BA. An extended difference of coherence test for comparing and combining several independent coherence estimates: Theory and application to the study of motor units and physiological tremor. *Journal of Neuroscience Methods*, 73:69–79, 1997.
- [30] Farmer SF, Bremner FD, Halliday DM, Rosenberg JR, and Stephens JA. The frequency content of common synaptic inputs to motoneurones studied during voluntary isometric contraction in man. *Journal of Physiology*, 470:127–155, 1993.
- [31] Farmer SF. Rhythmicity, synchronization and binding in human and primate motor systems. *Journal of Physiology*, 509.1:3–14, 1998.
- [32] Johnston JA, Formicone G, Hamm TM, and Santello M. Assessment of across-muscle coherence using multi-unit vs. single-unit recordings. *Experimental Brain Research*, 207:269–282, 2010.
- [33] Schwartz M. *Information Transmission, Modulation, and Noise*. McGraw-Hill, New York, NY, 3rd edition, 1980.
- [34] Wade JG. *Signal Coding and Processing: An Introduction Based on Video Systems*. Ellis Horwood, Chichester, England, 1987.
- [35] Bogert BP, Healy MJR, and Tukey JW. The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. In Rosenblatt M, editor, *Proceedings of the Symposium on Time Series Analysis*, pages 209–243. Wiley, New York, NY, 1963.
- [36] Oppenheim AV, Schafer RW, and Stockham Jr. TG. Nonlinear filtering of multiplied and convolved signals. *Proceedings of the IEEE*, 56(8):1264–1291, 1968.
- [37] Oppenheim AV and Schafer RW. Homomorphic analysis of speech. *IEEE Transactions on Audio and Electroacoustics*, AU-16(2):221–226, 1968.
- [38] Oppenheim AV and Schafer RW. From frequency to quefrency: A history of the cepstrum. *IEEE Signal Processing Magazine*, 21:95–99, 106, September 2004.
- [39] Oppenheim AV and Schafer RW. *Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1975.
- [40] Gonzalez RC and Woods RE. *Digital Image Processing*. Prentice-Hall, Upper Saddle River, NJ, 2nd edition, 2002.
- [41] Rangayyan RM. *Biomedical Image Analysis*. CRC Press, Boca Raton, FL, 2005.
- [42] Childers DG, Skinner DP, and Kemereit RC. The cepstrum: A guide to processing. *Proceedings of the IEEE*, 65(10):1428–1443, 1977.
- [43] Oppenheim AV, Willsky AS, and Nawab SH. *Signals and Systems*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1997.
- [44] Rabiner LR and Schafer RW. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, 1978.
- [45] Luisada AA and Portaluppi F. *The Heart Sounds — New Facts and Their Clinical Implications*. Praeger, New York, NY, 1982.
- [46] MacCanon DM, Arevalo F, and Meyer EC. Direct detection and timing of aortic valve closure. *Circulation Research*, 14:387–391, 1964.
- [47] Stein PD, Sabbah HN, Anbe DT, and Khaja F. Hemodynamic and anatomic determinants of relative differences in amplitude of the aortic and pulmonary components of the second heart sound. *American Journal of Cardiology*, 42:539–544, 1978.
- [48] Stein PD and Sabbah H. Intensity of the second heart sound: Relation of physical, physiological and anatomic factors to auscultatory evaluation. *Henry Ford Hospital Medical Journal*, 28(4):205–209, 1980.
- [49] Sarkady AA, Clark RR, and Williams R. Computer analysis techniques for phonocardiogram diagnosis. *Computers and Biomedical Research*, 9:349–363, 1976.

- [50] Baranek HL, Lee HC, Cloutier G, and Durand LG. Automatic detection of sounds and murmurs in patients with Ionescu-Shiley aortic bioprostheses. *Medical and Biological Engineering and Computing*, 27:449–455, 1989.
- [51] Durand LG, de Guise J, Cloutier G, Guardo R, and Brais M. Evaluation of FFT-based and modern parametric methods for the spectral analysis of bioprosthetic valve sounds. *IEEE Transactions on Biomedical Engineering*, 33(6):572–578, 1986.
- [52] Nigam V and Priemer R. A dynamic method to estimate the time split between the A2 and P2 components of the S2 heart sound. *Physiological Measurement*, 27:553–567, 2006.
- [53] Xu JP, Durand LG, and Pibarot P. Extraction of the aortic and pulmonary components of the second heart sound using a nonlinear transient chirp signal model. *IEEE Transactions on Biomedical Engineering*, 48(3):277–283, 2001.
- [54] Xu JP, Durand LG, and Pibarot P. A new, simple, and accurate method for non-invasive estimation of pulmonary arterial pressure. *Heart*, 88:76–80, 2002.

CHAPTER 5

ANALYSIS OF WAVESHAPE AND WAVEFORM COMPLEXITY

Certain biomedical signals, such as the ECG and the carotid pulse, have simple waveshapes (although the QRS wave is often referred to as a *complex*!). The readily identifiable signatures of the ECG and the carotid pulse are modified by abnormal events and pathological processes. Hence, analysis of waveshapes could be useful in the diagnosis of various diseases.

Signals such as the EMG and the PCG do not have waveshapes that may be identified easily. EMG signals are complex interference patterns of innumerable SMUAPs. PCG signals represent vibration waves that do not possess specific waveshapes. Regardless, even the complexity of the waveforms in some signals, such as the EMG and the PCG, does vary in relation to physiological and pathological phenomena. Analyzing the waveform complexity of such signals may assist in gaining an understanding of the processes that they reflect.

5.1 Problem Statement

Explain how waveshapes and waveform complexity in biomedical signals relate to the characteristics of the underlying physiological and pathological phenomena. Propose techniques to parameterize and analyze the signal features you identify.

As in the preceding chapters, the problem statement given above is generic and represents the theme of the present chapter. The following section presents illustrations of the problem with case studies that provide more specific definitions of the problem with a few signals of interest. The remaining sections of the chapter describe techniques to address the stated problems. It should be noted that, although signal analysis techniques are proposed in the context of specific signals and applications, they may find applications in other fields where signals with comparable characteristics and behavior are encountered.

5.2 Illustration of the Problem with Case Studies

5.2.1 The QRS complex in the case of bundle-branch block

We saw in Section 1.2.5 that the His bundle and its branches conduct the cardiac excitation pulse from the AV node to the ventricles. A block in one of the bundle branches causes asynchrony between the contraction of the left and the right ventricles. This, in turn, causes a staggered summation of the action potentials of the myocytes of the left and the right ventricles over a longer-than-normal duration. The result is a longer and possibly jagged QRS complex, as illustrated by the ECG of a patient with right bundle-branch block in Figure 1.30.

5.2.2 The effect of myocardial ischemia on QRS waveshape

The occlusion of a coronary artery, or a branch thereof, due to the deposition of fat, calcium, and other substances, results in reduced blood supply to a portion of the cardiac musculature. The part of the myocardium served by the affected artery then suffers from ischemia, that is, lack of blood supply. Complete blockage of an artery leads to myocardial infarction when the affected tissues die. The deceased myocytes cannot contract any more and no longer produce action potentials.

The action potential of an under-nourished ventricular myocyte reflects altered repolarization characteristics: The action potential is of smaller amplitude and shorter duration [1, 2] than the normal case. The result of the summation of the action potentials of all of the active ventricular myocytes will then be different from the normal QRS complex. The primary change reflected in the ECG is a modified ST segment that is either elevated or depressed, depending on the lead used and the position of the affected region; the T wave may also be inverted. Chronic myocardial infarction causes a return to a normal ST segment and a pronounced Q wave [3].

5.2.3 Ectopic beats

Ectopic beats are generated by cardiac tissues that possess abnormal pacing capabilities. Ectopic beats originating from focal points on the atria could cause altered P waveshapes due to different paths of propagation of the excitation pulse and hence different activation sequences of atrial muscle fibers. However, the QRS complex of atrial ectopic beats will appear normal because the conduction of the excitation past the AV node would be normal.

Ectopic beats originating on the ventricles (that are necessarily premature beats, that is, PVCs) typically possess bizarre waveshapes due to widely differing paths of conduction and excitation of the ventricular muscle fibers. Figure 1.29 illustrates an ECG signal with PVCs. PVCs typically lack a preceding P wave; however, an ectopic beat triggered during the normal AV node delay will demonstrate a normal preceding P wave. PVCs triggered by ectopic foci close to the AV node may possess near-normal QRS shape as the path of conduction may be almost the same as that in the case of a normal impulse from the AV node. On the other hand, beats triggered by ectopic foci near the apex of the heart could take widely different paths of propagation, resulting in substantially different QRS waveshapes. In addition to waveshape, the preceding and succeeding *RR* intervals play important roles in determining the nature of ectopic beats.

5.2.4 Complexity of the EMG interference pattern

We saw in Section 1.2.4 that motor units are recruited by two mechanisms — spatial and temporal recruitment — in order to produce increasing levels of contraction and muscular force output. As more and more motor units are brought into action and their individual firing rates increase (within certain limits), the SMUAPs of the active motor units overlap and produce a complex interference pattern. Figures 1.20 and 1.21 illustrate an EMG signal obtained from the crural diaphragm of a

dog during one normal breath cycle. The increasing complexity of the waveform with deeper levels of inspiration is clearly seen in the expanded plot in Figure 1.21.

Although a surface EMG interference pattern is typically too complex for visual analysis, the general increase in the level of activity (“busy-ness”) may be readily seen. It used to be common practice in EMG laboratories to feed EMG signals to an amplified speaker: low levels of activity when the SMUAPs are not overlapping (that is, separated in time) result in discrete “firing” type of sounds; increasing levels of contraction result in increased “chatter” in the sound produced. EMG signals may be analyzed to derive parameters of waveform complexity that increase with increasing muscular contraction, and hence to obtain correlates to mechanical activity that are derived from its electrical manifestation.

5.2.5 PCG intensity patterns

Although the vibration waves in a PCG signal may not be amenable to direct visual analysis, the general intensity pattern of the signal over a cardiac cycle may be readily appreciated either by auscultation or visual inspection. Certain cardiovascular diseases and defects alter the relative intensity patterns of S1 and S2, cause additional sounds or murmurs, and/or split S2 into two distinct components, as described in Section 1.2.9. While many cardiovascular diseases and defects may cause systolic murmurs, the intensity pattern or envelope of the murmur could assist in arriving at a specific diagnosis. It should also be noted that definitive diagnosis based on the PCG would usually require comparative analysis of PCG signals from a few positions on the chest. Figures 1.44, 1.46, 4.30, and 4.31 illustrate PCG signals of a normal subject and patients with systolic murmur, split S2, and opening snap of the mitral valve. The differences in the overall intensity patterns of the signals are obvious. However, signal processing techniques are desirable to convert the signals into positive-valued envelopes that could be treated as distributions of signal energy over time. Such a transformation permits the treatment of signal intensity patterns as distribution or density functions, which lends to the computation of various statistical measures and moments.

5.3 Analysis of ERPs

The most important parameter extracted from a visual ERP is the timing or latency of the first major positivity; since the average of this latency is about 120 *ms* for normal adults, it is referred to as P120 (see Figure 3.43). The latencies of the troughs before and after P120, called N80 and N145, respectively, are also of interest. The amplitudes of the ERP at the corresponding instants are of lesser importance. Delays in the latencies that are well beyond the normal range could indicate abnormalities in the visual system. Asymmetries in the latencies of the left and right parts of the visual system could also be indicative of disorders.

The lowest trace in Figure 3.43 is an averaged flash visual ERP recorded from a normal adult male subject. The signal has been labeled to indicate the N80, P120, and N145 points; the corresponding actual latencies for the subject are 85, 100.7, and 117 *ms*, respectively.

Auditory ERPs are weaker and more complex than visual ERPs, requiring averaging over several hundred or a few thousand stimuli. Auditory ERPs are analyzed for the latencies and amplitudes of several peaks and troughs. Clinical ERP analysis is usually performed manually, since there is no pressing need for signal processing techniques beyond synchronized averaging. Advanced analysis of ERPs for motor imagery tasks is discussed in Section 9.12.

5.4 Morphological Analysis of ECG Waves

The waveshape of an ECG cycle could be changed by many different abnormalities, including myocardial ischemia or infarction, bundle-branch block, and ectopic beats. It is not possible to

propose a single analysis technique that can assist in categorizing all possible abnormal causes of change in waveshape. The following sections address a few illustrative cases.

5.4.1 Correlation coefficient

Problem: Propose a general index to indicate altered QRS waveshape. You are given a normal QRS template.

Solution: Jenkins et al. [4] applied the correlation coefficient γ_{xy} as defined in Equation 4.25 to classify ECG cycles as normal beats or beats with abnormal morphology. A normal beat was used as a template to compute γ_{xy} for each detected beat. They found that most normal beats possessed γ_{xy} values above 0.9, and that PVCs and beats with abnormal shape had considerably lower values. A threshold of 0.9 was used to assign a code to each beat as 0 for abnormal or 1 for normal in terms of waveshape. Figure 2.2 shows an ECG signal with five abnormal beats that have the first symbol in the 4-symbol code shown for each beat as 0, indicating an abnormal shape due to generation by an ectopic focus or due to aberrant conduction of a pulse generated by the SA node. The normal beats have the first symbol of the code as 1, indicating high correlation with the normal template. See Section 5.8 for related discussion.

5.4.2 The minimum-phase correspondent and signal length

The normal ECG signal contains epochs of activity where the signal's energy is concentrated. Discounting the usually low-amplitude P and T waves, most of the energy of a normal ECG signal is concentrated within the interval of about 80 ms that is spanned by the QRS complex. The normally isoelectric PQ, ST, and TP segments contain no energy as the signal amplitude is typically zero over the corresponding intervals. We have observed that certain abnormal conditions cause the QRS to widen or the ST segment to possess a nonzero value. In such a case, it could be said that the energy of the signal is being spread over a longer duration. Let us now consider how we may capture this information, and investigate if it may be used for waveshape analysis.

Problem: Investigate the effect of the distribution of energy over the time axis on a signal's characteristics. Propose measures to parameterize the effects and study their use in the classification of ECG beats.

Solution: A signal $x(t)$ may be seen as a distribution of the amplitude of a certain variable over the time axis. The square of the signal, that is, $x^2(t)$, may be interpreted as the instantaneous energy of the signal-generating process. The function $x^2(t)$, $0 \leq t \leq T$, may then be viewed as an energy distribution or density function, with the observation that the total energy of the signal is given by

$$E_x = \int_0^T x^2(t) dt. \quad (5.1)$$

Such a representation facilitates the definition of moments of the energy distribution, leading to a centroidal time

$$t_{\bar{x}} = \frac{1}{E_x} \int_0^T t x^2(t) dt, \quad (5.2)$$

and dispersion of energy about the centroidal time

$$\sigma_{t_{\bar{x}}}^2 = \frac{1}{E_x} \int_0^T (t - t_{\bar{x}})^2 x^2(t) dt. \quad (5.3)$$

Observe the similarity between the equations given above and Equations 3.1 and 3.3: The normalized function

$$p_x(t) = \frac{1}{E_x} x^2(t) \quad (5.4)$$

is now treated as a PDF. Other moments may also be defined to characterize and study the distribution of $x^2(t)$ over the time axis. The preceding equations have been stated in continuous time for the sake of generality; they are valid for discrete-time signals, with a simple change of t to n and $\int dt$ to \sum_n .

Minimum-phase signals: The distribution of the energy of a signal over its duration is related to its amplitude spectrum and, more importantly, to its phase spectrum. The notion of minimum phase is useful in analyzing the related signal characteristics. The minimum-phase property of signals may be explained in both the time and frequency domains [5–11].

In the time domain, a signal $x(n)$ is a minimum-phase signal if both the signal and its inverse $x_i(n)$ are one-sided (that is, completely causal or anticausal) signals with finite energy, that is, $\sum_{n=0}^{\infty} x^2(n) < \infty$ and $\sum_{n=0}^{\infty} x_i^2(n) < \infty$. [Note: The inverse of a signal is defined such that $x(n) * x_i(n) = \delta(n)$; equivalently, we have $X_i(z) = \frac{1}{X(z)}$.]

Some of the important properties of a minimum-phase signal are [5–11]:

- For a given amplitude spectrum there exists one and only one minimum-phase signal.
- Of all finite-energy, one-sided signals with identical amplitude spectra, the energy of the minimum-phase signal is optimally concentrated toward the origin, and the signal has the smallest phase lag and phase-lag derivative at each frequency.
- The z -transform of a minimum-phase signal has all of its poles and zeros inside the unit circle in the z -plane.
- The complex cepstrum of a minimum-phase signal is causal (see also Section 4.7.3).

The extreme example of a minimum-phase signal is the delta function $\delta(t)$, which has all of its energy concentrated at $t = 0$. The magnitude spectrum of the delta function is real and equal to unity for all frequencies; the phase lag at every frequency is zero.

Minimum-phase and maximum-phase components: A signal $x(n)$ that does not satisfy the minimum-phase condition, referred to as a composite signal or a mixed-phase signal, may be split into its minimum-phase component and maximum-phase component by filtering its complex cepstrum $\hat{x}(n)$ [5, 11–13]. To obtain the minimum-phase component, the causal part of the complex cepstrum (see Section 4.7.3) is chosen as follows:

$$\hat{x}_{\min}(n) = \begin{cases} 0, & n < 0, \\ 0.5 \hat{x}(n), & n = 0, \\ \hat{x}(n), & n > 0. \end{cases} \quad (5.5)$$

Application of the inverse procedures yields the minimum-phase component $x_{\min}(n)$. Similarly, the maximum-phase component is obtained by application of the inverse procedures to the anticausal part of the cepstrum, selected as

$$\hat{x}_{\max}(n) = \begin{cases} \hat{x}(n), & n < 0, \\ 0.5 \hat{x}(n), & n = 0, \\ 0, & n > 0. \end{cases} \quad (5.6)$$

The minimum-phase and maximum-phase components of a signal satisfy the following relationships:

$$\hat{x}(n) = \hat{x}_{\min}(n) + \hat{x}_{\max}(n) \quad (5.7)$$

and

$$x(n) = x_{\min}(n) * x_{\max}(n). \quad (5.8)$$

The minimum-phase correspondent: A mixed-phase signal may be converted to a minimum-phase signal that has the same spectral magnitude as the original signal by filtering the complex cepstrum of the original signal as

$$\hat{x}_{\text{MPC}}(n) = \begin{cases} 0, & n < 0, \\ \hat{x}(n), & n = 0, \\ \hat{x}(n) + \hat{x}(-n), & n > 0, \end{cases} \quad (5.9)$$

and applying the inverse procedures [5, 11–13]. The result is known as the *minimum-phase correspondent* (MPC) of the original signal. The MPC possesses optimal concentration of energy near the origin under the constraint imposed by the specified magnitude spectrum (of the original mixed-phase signal).

Observe that $\hat{x}_{\text{MPC}}(n)$ is equal to twice the even part of $\hat{x}(n)$ for $n > 0$. This leads to a simpler procedure to compute the MPC, as follows. Let us assume $\hat{X}(z) = \log X(z)$ to be analytic over the unit circle in the z -plane. We can write $\hat{X}(\omega) = \hat{X}_R(\omega) + j\hat{X}_I(\omega)$, where the subscripts R and I indicate the real and imaginary parts, respectively. $\hat{X}_R(\omega)$ and $\hat{X}_I(\omega)$ are the log-magnitude and phase spectra of $x(n)$, respectively. Now, the inverse Fourier transform of $\hat{X}_R(\omega)$ is equal to the even part of $\hat{x}(n)$, defined as $\hat{x}_e(n) = [\hat{x}(n) + \hat{x}(-n)]/2$. Thus, we have

$$\hat{x}_{\text{MPC}}(n) = \begin{cases} 0, & n < 0, \\ \hat{x}_e(n), & n = 0, \\ 2\hat{x}_e(n), & n > 0. \end{cases} \quad (5.10)$$

This result means that we do not need to compute the complex cepstrum, which requires the unwrapped phase spectrum of the signal, but need only to compute a *real cepstrum* using the log-magnitude spectrum. Furthermore, given that the PSD is the Fourier transform of the ACF, we have $\log\{\text{FT}[\phi_{xx}(n)]\} = 2\hat{X}_R(\omega)$. It follows that, in the cepstral domain, $\hat{\phi}_{xx}(n) = 2\hat{x}_e(n)$, and therefore [13]

$$\hat{x}_{\text{MPC}}(n) = \begin{cases} 0, & n < 0, \\ 0.5\hat{\phi}_{xx}(n), & n = 0, \\ \hat{\phi}_{xx}(n), & n > 0, \end{cases} \quad (5.11)$$

where $\hat{\phi}_{xx}(n)$ is the cepstrum of the ACF $\phi_{xx}(n)$ of $x(n)$.

Signal length: The notion of signal length (SL), as introduced by Berkout [7], is different from signal duration. The duration of a signal is the extent of time over which the signal exists, that is, the signal has nonzero values (neglecting periods within the total duration of the signal where the signal's amplitude could be zero). SL relates to how the energy of a signal is distributed over its duration. SL depends on both the magnitude and phase spectra of the signal. For one-sided signals, minimum SL implies minimum phase; the converse is also true [7].

The general definition of the SL of a signal $x(n)$ is given as [7]

$$SL = \frac{\sum_{n=0}^{N-1} w(n) x^2(n)}{\sum_{n=0}^{N-1} x^2(n)}, \quad (5.12)$$

where $w(n)$ is a nondecreasing, positive weighting function with $w(0) = 0$. The definition of $w(n)$ depends on the application and the desired characteristics of SL . It is readily seen that samples of the signal away from the origin $n = 0$ receive progressively heavier weights given by $w(n)$. The definition of SL as above may be viewed as a normalized moment of $x^2(n)$. If $w(n) = n$, we get the centroidal time instant of $x^2(n)$ as in Equation 5.2.

For a given amplitude spectrum and hence total energy, the minimum-phase signal has its energy optimally concentrated near the origin. Therefore, the minimum-phase signal will have the lowest SL of all signals with the specified amplitude spectrum. Signals with increasing phase lag have their energy spread over a longer time duration and will have larger SL due to the increased weighting by $w(n)$.

Illustration of application: The QRS-T wave is the result of the spatiotemporal summation of the action potentials of ventricular myocytes. The duration of normal QRS-T waves is in the range of $350 - 400\text{ ms}$, with the QRS itself limited to about 80 ms due to rapid and coordinated depolarization of the ventricular muscle fibers via the Purkinje fibers. However, PVCs, in general, have QRS-T complexes that are wider than normal, that is, they have their energy distributed over longer time spans within their total duration. This is due to different and possibly slower and disorganized excitation sequences triggering the ventricular muscle fibers: Ectopic triggers may not get conducted through the Purkinje system, and may be conducted through the ventricular muscle fibers themselves. Furthermore, PVCs do not, in general, display separate QRS and T waves; that is, they lack an isoelectric ST segment.

Regardless of the distinctions described above, normal ECG beats and PVCs have similar amplitude spectra, indicating that the difference between the signals may lie in their phase. SL depends on both the amplitude spectrum and the phase spectrum of the given signal, and it parameterizes the distribution of energy over the duration of the signal. Based on the arguments given above, Murthy and Rangaraj [11] proposed the application of SL to classify ECG beats as normal or ectopic (or PVC, along with the use of the RR interval to indicate prematurity). Furthermore, to overcome ambiguities in the determination of the onset of each beat, they computed the SL of the MPC of the ECG signals (segmented so as to include the P, QRS, and T waves of each cycle). Use of the MPC resulted in a “rearrangement” of the signal such that the dominant QRS wave appeared at the origin in the MPC.

Figure 5.1 illustrates a normal ECG signal and three PVCs of a patient with multiple ectopic foci generating PVCs of widely differing shapes [11]. The figure also illustrates the corresponding MPCs and lists the SL values of all of the signals. The SL values of the MPCs of the abnormal waves are higher than the SL of the MPC of the normal signal (see the right-hand column of signals in Figure 5.1). The SL values of the original PVCs do not exhibit such a separation from the SL of the normal signal (see the left-hand column of signals in Figure 5.1). Ambiguities due to the presence of baseline segments of variable lengths at the beginning of the signals have been overcome by the use of the MPCs. In each case, the MPCs have the most-dominant wave at the origin, reflecting a rearrangement of energy in order to meet the minimum-phase criteria.

Figure 5.2 shows plots of the RR intervals and SL values computed using the original ECG signals and their MPCs for several beats of the same patient whose representative ECG waveforms are illustrated in Figure 5.1 [11]. The SL values of the normal signals and the ectopic beats exhibit a significant overlap in the range $28 - 35$ [plot (a) in Figure 5.2]. However, the SL values of the MPCs of the PVCs are higher than those of the normal beats, which facilitates their classification [plot (b) in Figure 5.2].

Murthy and Rangaraj [11] applied their QRS detection method (described in Section 4.3.1) to ECG signals of two patients with ectopic beats, and used the SL of MPC to classify the beats with a linear discriminant function (described in Section 10.4.1). They analyzed 208 beats of the first patient (whose signals are illustrated in Figures 5.1 and 5.2): 132 out of 155 normals and 48 out of 53 PVCs were correctly classified; one beat was missed by the QRS detection algorithm. Misclassification of normal beats as PVCs was attributed to wider-than-normal QRS complexes and depressed ST segments in some of the normal beats of the patient (see Figure 5.2). The signal of the second patient included 89 normals and 18 PVCs, all of which were detected and classified correctly. It was observed that computation of the MPC was not required in the case of the second patient: The SL values of the original signals provided adequate separation between the normal and ectopic beats. The segments of normal ECG cycles used by Murthy and Rangaraj included the P

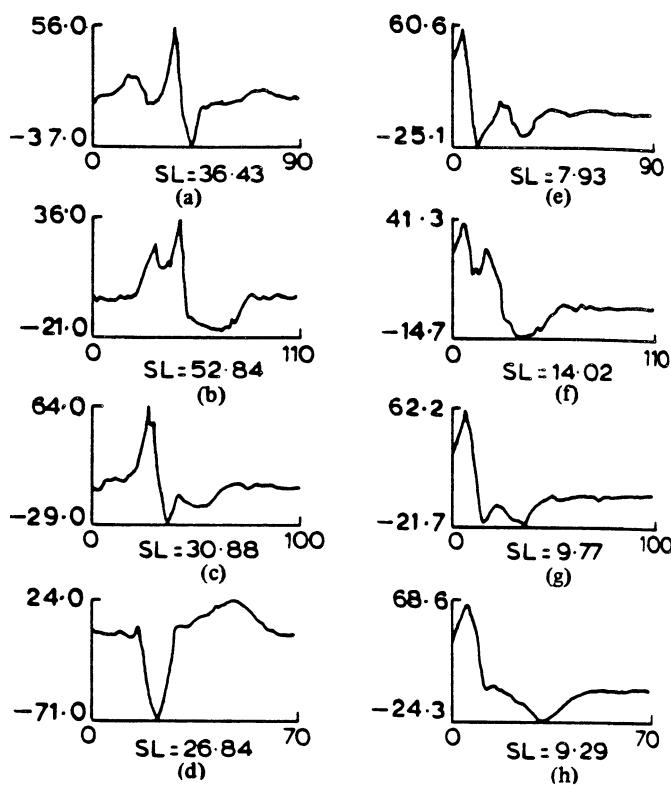


Figure 5.1 (a) A normal ECG beat and (b)–(d) three ectopic beats (PVCs) of a patient with multiple ectopic foci. (e)–(h) MPCs of the signals in (a)–(d). The SL values of the signals are also indicated [11]. Note that the abscissa is labeled in samples, with a sampling interval of 10 ms. The ordinate is not calibrated. The signals have different durations and amplitudes although plotted to the same size. Reproduced with permission from I.S.N. Murthy and M.R. Rangaraj, New concepts for PVC detection, *IEEE Transactions on Biomedical Engineering*, 26(7):409–416, 1979. ©IEEE.

wave; better results could be obtained using only the QRS and T waves since most PVCs do not include a distinct P wave and essentially correspond to the QRS and T waves in a normal ECG signal.

It should be noted that the QRS width may be increased by other abnormal conditions such as bundle-branch block; the definition of SL as above would lead to higher SL for wider-than-normal QRS complexes. Furthermore, ST segment elevation or depression would be interpreted as the presence of energy in the corresponding time interval in the computation of SL . Abnormally large T waves could also lead to SL values that are larger than those for normal signals. More sophisticated logic and other parameters in addition to SL could be used to rule out these possibilities and affirm the classification of a beat as an ectopic beat.

5.4.3 ECG waveform analysis

Measures such as the correlation coefficient and SL described in the preceding sections provide general parameters that could assist in comparing waveforms. The representation, however, is in terms of gross features, and many different waveforms could possess the same or similar feature values. Detailed analysis of ECG waveforms requires the use of several features or measurements for accurate categorization of various QRS complex shapes and correlation with cardiovascular

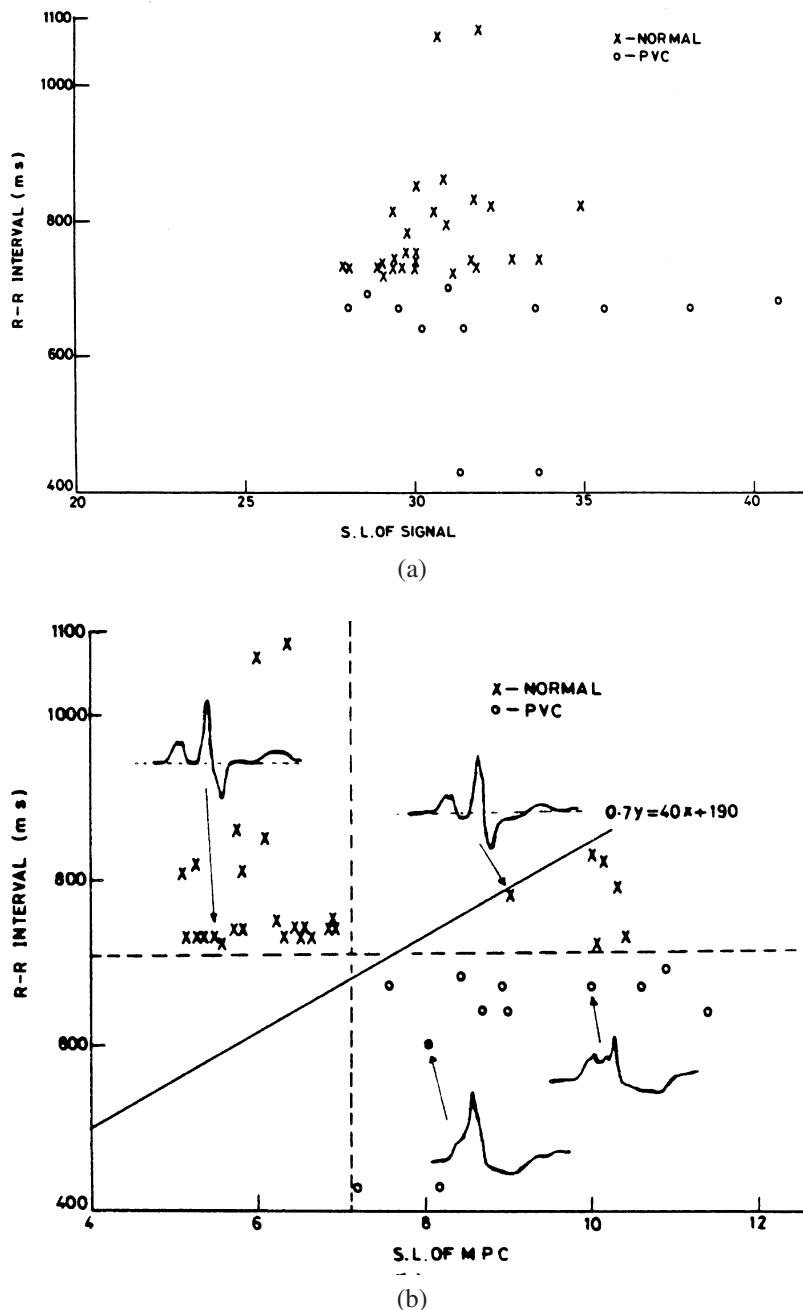


Figure 5.2 (a) Plot of RR and SL values of several beats of a patient with multiple ectopic foci (as in Figure 5.1). (b) Same as (a) but with the SL of the MPCs of the signals. A few representative ECG cycles are illustrated. The linear discriminant (decision) function used to classify the beats is also shown. Reproduced with permission from I.S.N. Murthy and M.R. Rangaraj, New concepts for PVC detection, *IEEE Transactions on Biomedical Engineering*, 26(7):409–416, 1979. ©IEEE.

diseases. Since the ECG waveform depends on the lead system used, sets of features may have to be derived for multiple-lead ECGs, including as many as 12 leads that are commonly used in clinical practice.

The steps required for ECG waveform analysis may be expressed as [14]:

1. Detection of ECG waves, primarily the QRS complex, and possibly the P and T waves.
2. Delimitation of wave boundaries, including the P, QRS, and T waves.
3. Measurement of interwave intervals, such as RR, PQ, QT, ST, QQ, and PP intervals.
4. Characterization of the morphology (shape) of the waves.

The last step given above may be achieved using parameters such as the correlation coefficient and SL as described earlier, or via detailed measurements of the peaks of the P, Q, R, S, and T waves (some of which could be negative); the durations of the P, Q, R, S, QRS, and T waves; and the interwave intervals defined above [14]. The nature of the PQ and ST segments, in terms of their being isoelectric or not (in case of the latter, as being positive, negative, elevated, or depressed), should also be documented. However, a large number of such features would make the development of further pattern classification rules difficult.

Cox et al. [14] and Nolle [15] proposed four measures to characterize QRS complexes, as illustrated in Figure 5.3 and defined as follows:

1. *Duration* — the duration or width of the QRS complex.
2. *Height* — the maximum amplitude minus the minimum amplitude of the QRS complex.
3. *Offset* — the positive or negative vertical distance from the midpoint of the baseline to the center of the QRS complex. The baseline is defined as the straight line connecting the temporal boundary points of the QRS complex. The center is defined as the midpoint between the highest and lowest bounds in amplitude of the QRS complex.
4. *Area* — the area under the QRS waveform rectified with respect to a straight line through the midpoint of the baseline.

Since the measures are independent of time, they are not sensitive to the preceding procedures for the detection of fiducial markers.

The measures of Cox et al. [14] and Nolle [15] were used to develop a system for arrhythmia monitoring, known as “Argus” for Arrhythmia Guard System, for use in coronary-care units. Figure 5.4 shows the grouping of more than 200 QRS complexes of a patient with multifocal PVCs into 16 dynamic families by Argus using the four features defined above [14]. The families labeled 00, 01, 02, 04, 06, and 10 were classified as normal beats by Argus (163 beats which were all classified as normals by a cardiologist; 91% of the normals were correctly labeled by Argus). PVCs of different shapes from more than two ectopic foci form the remaining families, with some of them having shapes close to those of the patient’s normal beats. Of the 52 beats in the remaining families, 96% were labeled as PVCs by the cardiologist; Argus labeled 85% of them as PVCs, 13% as not PVCs, and 2% as borderline beats [15]. Some noteworthy points of one of the clinical tests of Argus with over 50,000 beats are as follows: 85% of 45,364 normal beats detected and classified correctly, with 0.04% beats missed; 78% of 4,010 PVCs detected and classified correctly, with 5.3% beats missed; and 38 normals (less than 0.1% of the beats) falsely labeled as PVCs.

For further discussions on the analysis and classification of ECG signals, see Ye et al. [16] and Ince et al. [17].

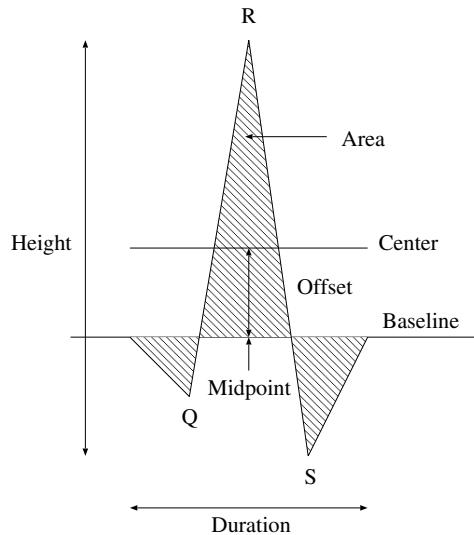


Figure 5.3 ECG waveform features used by Cox et al. [14] and Nolle [15].

5.5 Envelope Extraction and Analysis

Signals with complex patterns such as the EMG and PCG may not permit direct analysis of their waveshape. In such cases, the intricate high-frequency variations may not be of interest; rather, the general trends in the level of the overall activity might convey useful information. Considering, for example, the EMG in Figure 1.20, observe that the general signal level increases with the level of activity (breathing). As another example, the PCG in the case of aortic stenosis, as illustrated in Figure 1.46, demonstrates a diamond-shaped systolic murmur: The *envelope* of the overall signal carries important information. Let us, therefore, consider the problem of extraction of the envelope of a seemingly complex signal.

Problem: Formulate algorithms to extract the envelope of an EMG or PCG signal to facilitate analysis of trends in the level of activity or energy in the signal.

Solution: The first step required in order to derive the envelope of a signal with positive and negative deflections is to obtain the absolute value of the signal at each time instant, that is, perform full-wave rectification. This procedure will create abrupt discontinuities at time instants when the original signal values change sign, that is, at zero-crossings. The discontinuities create high-frequency components of significant magnitude. This calls for the application of a lowpass filter with a relatively low bandwidth in the range of 0 – 10 or 0 – 50 Hz to obtain smooth envelopes of EMG and PCG signals. An MA filter may be used to perform lowpass filtering, leading to the basic definition of a time-averaged envelope as

$$y(t) = \frac{1}{T_a} \int_{t-T_a}^t |x(t)| dt, \quad (5.13)$$

where T_a is the duration of the MA filtering window.

In a procedure similar in principle to that described above, Lehner and Rangayyan [18] applied a weighted MA filter to the squared PCG signal to obtain a smoothed energy distribution curve $E(n)$ as

$$E(n) = \sum_{k=1}^M x^2(n-k+1) w(k), \quad (5.14)$$

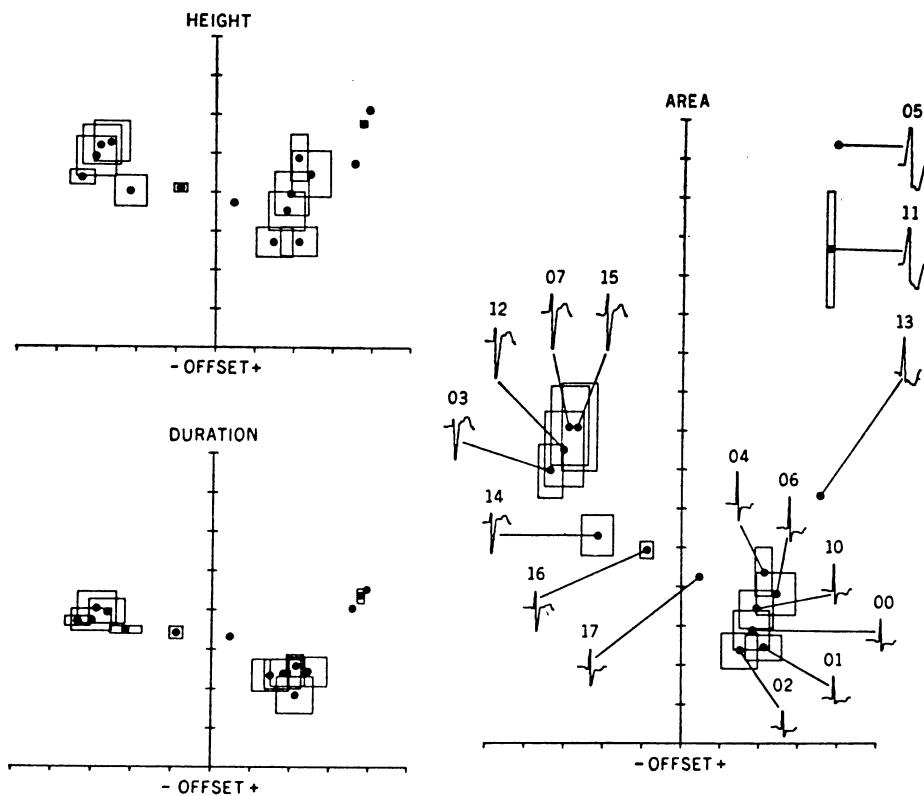


Figure 5.4 Use of four features to catalog QRS complexes into one of 16 dynamic families of similar complexes enclosed by four-dimensional boxes. The waveforms of typical members of each family are shown in the area-versus-offset feature plane. The family numbers displayed are in the octal (base eight) system. The families labeled 00, 01, 02, 04, 06, and 10 were classified as normal beats, with the others being PVCs or borderline beats. Reproduced with permission from J.R. Cox, Jr., F.M. Nolle, and R.M. Arthur, Digital analysis of the electroencephalogram, the blood pressure wave, and the electrocardiogram, *Proceedings of the IEEE*, 60(10):1137–1164, 1972. ©IEEE.

where $x(n)$ is the PCG signal, $w(k) = M - k + 1$, and $M = 32$ with the signal sampled at 1,024 Hz. Observe that the difference between energy and power is simply a division by the time interval being considered, which may be treated as a scale factor or ignored.

The envelope represents the total averaged activity (such as electrical or acoustic) within the averaging window. An improved filter such as a Bessel filter [19] may be required if a smooth envelope is desired. The filter should strike a balance between the need to smooth discontinuities in the rectified or squared signal and the requirement to maintain good sensitivity to represent relevant changes in signal level or power. This procedure is known as envelope detection or amplitude demodulation. A few related procedures and techniques are described in the following sections.

5.5.1 Amplitude demodulation

Amplitude modulation (AM) of signals for radio transmission involves multiplication of the signal $x(t)$ to be transmitted by the carrier $\cos(\omega_c t)$, where ω_c is the carrier frequency [20–22]. The AM signal is given as $y(t) = x(t) \cos(\omega_c t)$. {Recollect that the Fourier transform of $x(t) \exp(j\omega_c t)$ is $X(\omega - \omega_c)$, and that of $x(t) \cos(\omega_c t)$ is $[X(\omega - \omega_c) + X(\omega + \omega_c)]/2$.} If the exact carrier wave used

at the transmitting end were available at the receiving end as well (including the phase), *synchronous demodulation* becomes possible by multiplying the received signal $y(t)$ with the carrier. We then have the demodulated signal as

$$x_d(t) = y(t) \cos(\omega_c t) = x(t) \cos^2(\omega_c t) = \frac{1}{2}x(t) + \frac{1}{2}x(t) \cos(2\omega_c t). \quad (5.15)$$

The AM component at $2\omega_c$ may be removed by a lowpass filter, which will leave us with the desired signal $x(t)$.

If $x(t)$ is always positive, or a DC bias is added to meet this requirement, it becomes readily apparent that the envelope of the AM signal is equal to $x(t)$. A simple *asynchronous demodulation* procedure that does not require the carrier then becomes feasible: We only need to follow the envelope of $y(t)$. Given that the carrier frequency ω_c is far greater than the maximum frequency present in $x(t)$, the positive envelope of $y(t)$ may be extracted by performing half-wave rectification. A lowpass filter with an appropriate time constant to “fill the gaps” between the peaks of the carrier wave will provide a good estimate of $x(t)$. The difference between the use of a full-wave rectifier or a half-wave rectifier (that is, the larger gaps between the peaks of the carrier wave available after the latter type of rectification) can be smoothed over by increasing the time constant of the filter. The main differences between various envelope detectors lie in the way the rectification operation is performed and in the lowpass filter used [20, 21].

In a related procedure known as complex demodulation, a given arbitrary signal is demodulated to derive the time-varying amplitude and phase characteristics of the signal for each frequency (band) of interest [23–25]. In this approach, an arbitrary signal $x(t)$ is expressed as

$$x(t) = a(t) \cos[\omega_o t + \psi(t)] + x_r(t), \quad (5.16)$$

where ω_o is the frequency of interest; $a(t)$ and $\psi(t)$ are the time-varying amplitude and phase of the component at ω_o , respectively; and $x_r(t)$ is the remainder of the signal $x(t)$ after the component at ω_o has been removed. It is assumed that $a(t)$ and $\psi(t)$ vary slowly in relation to the frequencies of interest. The signal $x(t)$ may be equivalently expressed in terms of complex exponentials as

$$x(t) = \frac{1}{2} a(t) [\exp\{j[\omega_o t + \psi(t)]\} + \exp\{-j[\omega_o t + \psi(t)]\}] + x_r(t). \quad (5.17)$$

In the procedure of complex demodulation, the signal is shifted in frequency by $-\omega_o$ via multiplication with $2 \exp(-j\omega_o t)$, to obtain the result $y(t)$ as

$$\begin{aligned} y(t) &= 2x(t) \exp(-j\omega_o t) \\ &= a(t) \exp[j\psi(t)] + a(t) \exp\{-j[2\omega_o t + \psi(t)]\} + 2x_r(t) \exp(-j\omega_o t). \end{aligned} \quad (5.18)$$

The second term in the expression above is centered at $2\omega_o$, whereas the third term is centered at ω_o ; only the first term is placed at DC. Therefore, a lowpass filter may be used to extract the first term, to obtain the final result $y_o(t)$ as

$$y_o(t) \approx a(t) \exp[j\psi(t)]. \quad (5.19)$$

The desired entities may then be extracted as $a(t) \approx |y_o(t)|$ and $\psi(t) \approx \angle y_o(t)$.

The frequency resolution of the method depends on the bandwidth of the lowpass filter used. The procedure may be repeated at every frequency of interest. The result may be interpreted as the envelope of the signal for the specified frequency. The method was applied for the analysis of HRV by Shin et al. [23] and for the analysis of heart rate and arterial pressure variability by Hayano et al. [24].

In applying envelope detection to biomedical signals such as the PCG and the EMG, it should be noted that there is no underlying RF carrier wave in the signal: The envelope rides on relatively

high-frequency acoustic or electrical activity that has a composite spectrum. The difference in frequency content between the envelope and the “carrier activity” will not be comparable to that in AM. Regardless, we could expect at least a 10-fold difference in frequency content: The envelope of an EMG or PCG signal may have an extremely limited bandwidth of $0 - 20\text{ Hz}$, whereas the underlying signal has components up to at least 200 Hz , if not to $1,000\text{ Hz}$. Application of envelope detection to the analysis of EMG related to respiration is illustrated in Section 5.10; an application to PCG analysis is described in the following section.

5.5.2 Synchronized averaging of PCG envelopes

The ECG and PCG form a good signal pair for synchronized averaging: The latter could be averaged over several cardiac cycles using the former as the trigger. However, the PCG is not amenable to direct synchronized averaging as the vibration waves may interfere in a destructive manner and cancel themselves out. Karpman et al. [26] proposed to first rectify the PCG signal, smooth the result using a lowpass filter, and then perform synchronized averaging of the nonnegative envelopes so obtained using the ECG as the trigger. The PCG envelopes were averaged over up to 128 cardiac cycles to get repeatable averaged envelopes. It should be noted that while synchronized averaging can reduce the effects of noise, breathing, coughing, and other types of artifacts, it can also smudge the time boundaries of cardiac events if the heart rate is not constant during the averaging procedure.

Figure 5.5 illustrates the envelopes obtained for a normal case and seven cases of systolic murmur due to various cardiovascular diseases and defects. The typical diamond-shaped envelope in the case of aortic stenosis results in an envelope shaped like an isosceles triangle due to rectification. Mitral regurgitation results in a rectangular holosystolic (spanning the entire systolic period) murmur envelope.

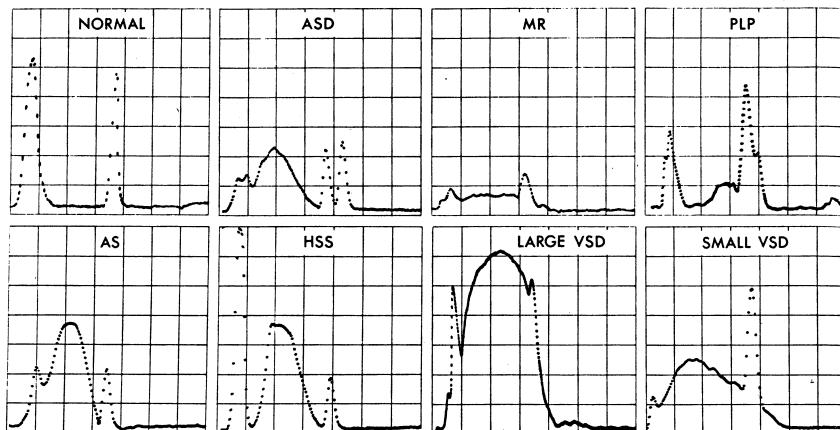


Figure 5.5 Averaged envelopes of the PCG signals of a normal subject and patients with systolic murmur due to aortic stenosis (AS), atrial septal defect (ASD), hypertrophic subaortic stenosis (HSS), rheumatic mitral regurgitation (MR), ventricular septal defect (VSD), and mitral regurgitation with posterior leaflet prolapse (PLP). Reproduced with permission from L. Karpman, J. Cage, C. Hill, A.D. Forbes, V. Karpman, and K. Cohn, Sound envelope averaging and the differential diagnosis of systolic murmurs, *American Heart Journal*, 90(5):600–606, 1975. ©American Heart Association.

Karpman et al. analyzed 400 cases of systolic murmurs due to six types of diseases and defects, and obtained an accuracy of 89% via envelope analysis. They proposed a decision tree to classify systolic murmurs based on the shape of a given murmur’s envelope as well as its relation to S1 and S2, which is illustrated in Figure 5.6.

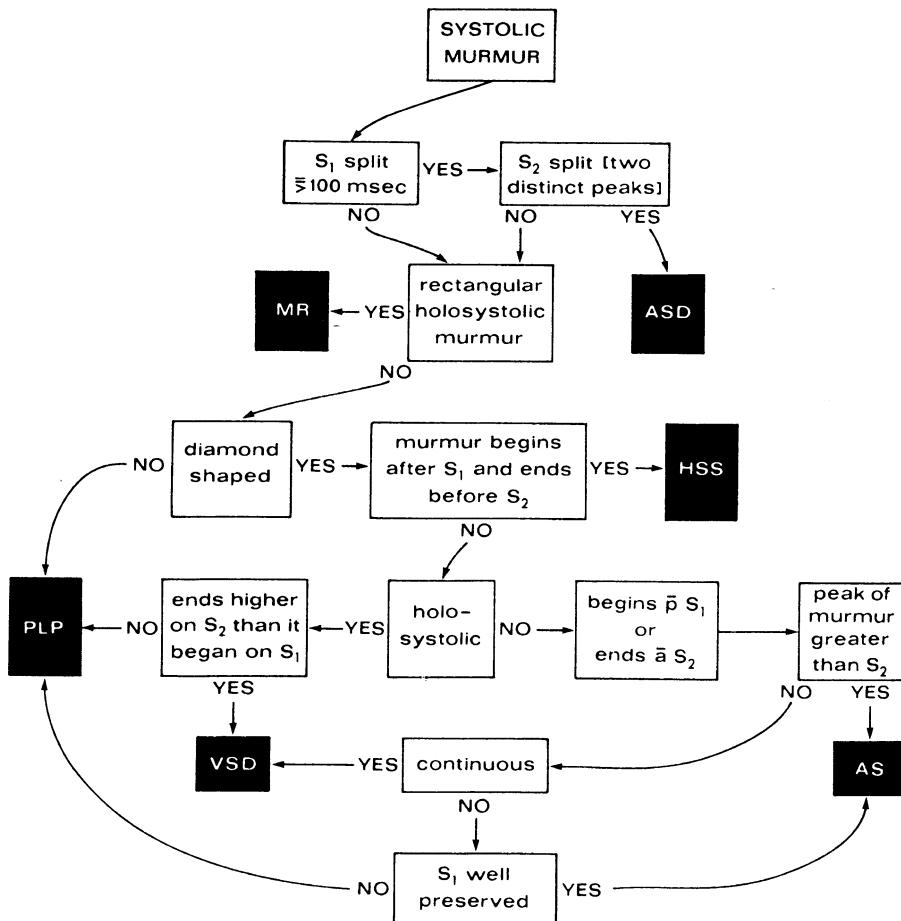


Figure 5.6 Decision tree to classify systolic murmurs based on envelope analysis. For details on the abbreviations used, refer to the caption of Figure 5.5. $\bar{p}S_1$, after S_1 ; $\bar{a}S_2$, before S_2 . Reproduced with permission from L. Karpman, J. Cage, C. Hill, A.D. Forbes, V. Karpman, and K. Cohn, Sound envelope averaging and the differential diagnosis of systolic murmurs, *American Heart Journal*, 90(5):600–606, 1975. ©American Heart Association.

5.5.3 The envelopogram

Sarkady et al. [27] proposed a Fourier-domain algorithm to obtain envelopes of PCG signals. They defined the *envelopogram estimate* as the magnitude of the analytic signal $y(t)$ formed using the PCG, $x(t)$, and its Hilbert transform, $x_H(t)$, as

$$y(t) = x(t) + jx_H(t). \quad (5.20)$$

(Note: An analytic function is a complex function of time having a Fourier transform that vanishes for negative frequencies [5, 28].) The Hilbert transform of a signal is defined as the convolution of the signal with $\frac{1}{\pi t}$, that is,

$$x_H(t) = \int_{-\infty}^{\infty} \frac{x(\tau)}{\pi(t - \tau)} d\tau. \quad (5.21)$$

The Fourier transform of $\frac{1}{\pi t}$ is $-j \operatorname{sgn}(\omega)$, where

$$\operatorname{sgn}(\omega) = \begin{cases} -1, & \omega < 0 \\ 0, & \omega = 0 \\ 1, & \omega > 0 \end{cases}, \quad (5.22)$$

is the signum function. Then, we have $Y(\omega) = X(\omega)[1 + \operatorname{sgn}(\omega)]$. $Y(\omega)$ is a one-sided or single-sideband function of ω containing only positive-frequency terms.

Based on the definitions and properties described above, Sarkady et al. [27] proposed the following algorithm to obtain the envelopogram estimate:

1. Compute the DFT of the PCG signal.
2. Set the negative-frequency terms to zero; that is, $X(k) = 0$ for $\frac{N}{2} + 2 \leq k \leq N$, with the DFT indexed $1 \leq k \leq N$.
3. Multiply the positive-frequency terms, that is, $X(k)$ for $2 \leq k \leq \frac{N}{2} + 1$, by 2; the DC term $X(1)$ remains unchanged.
4. Compute the inverse DFT of the result.
5. Obtain an estimate of the envelopogram as the magnitude of the result.

The procedure described above, labeled also as complex demodulation by Sarkady et al., yields a high-resolution envelope of the input signal. Envelopograms and PSDs computed from PCG signals over single cardiac cycles tend to be noisy and are affected by respiration and muscle noise. Sarkady et al. recommended synchronized averaging of both envelopograms and PSDs of PCGs over several cycles. A similar method was used by Baranek et al. [29] to obtain the envelopes of PCG signals for the detection of the aortic component A2 of S2.

See Guerrero et al. [30] for illustrations of the use of the Hilbert transform to obtain envelopes of BCG signals and their application to detect sleep-related breathing disorders.

Illustration of application: The topmost plots in Figures 5.7 and 5.8 show one cycle each of the PCG signals of a normal subject and of a patient with systolic murmur, split S2, and opening snap of the mitral valve. The PCG signals were segmented using the Pan-Tompkins method to detect the QRS complexes in the ECG signal, as illustrated in Figures 4.30 and 4.31 for the same signals. The envelopograms of the PCG cycles illustrated and the averaged envelopograms (over 16 beats for the normal case and 26 beats for the case with murmur) obtained using the method of Sarkady et al. [27] are shown in the second and third plots of Figures 5.7 and 5.8, respectively. Observe that while a split S2 is visible in the individual signal and envelopogram illustrated in Figure 5.7, the split is not clearly seen in the averaged envelopogram and envelope, due to variations related to breathing over the duration of the signal and averaging.

Furthermore, based upon the method of Karpman et al. [26], the averaged envelopes were computed by taking the absolute value of the signal over each cardiac cycle, smoothing with a Butterworth lowpass filter with $N = 8$ and $f_c = 50 \text{ Hz}$, and synchronized averaging. The last plots in Figures 5.7 and 5.8 show the averaged envelopes. (The Butterworth filter has introduced a small delay in the envelope.) The averaged envelopograms and averaged envelopes for the normal case display the envelopes of S1 and S2; the individual components of S1 and S2 have been smoothed over and merged in the averaged results. The averaged envelopograms and averaged envelopes for the case with murmur clearly demonstrate the envelopes of S1, the systolic murmur, the split S2, and the opening snap of the mitral valve.

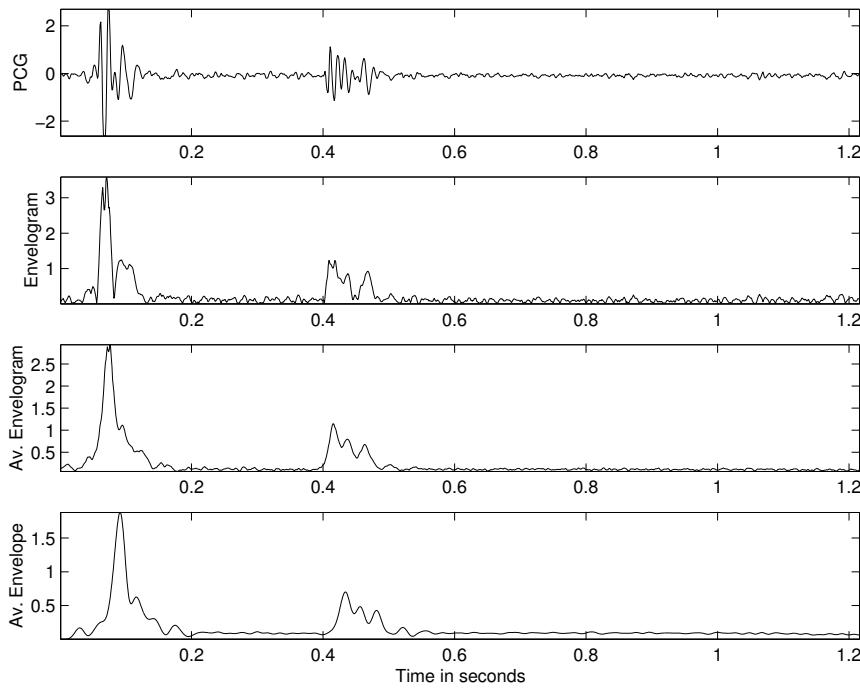


Figure 5.7 Top to bottom: PCG signal of a normal subject (male, 23 years); envelopgram estimate of the signal shown; averaged envelopgram over 16 cardiac cycles; averaged envelope over 16 cardiac cycles. The PCG signal starts with S1. See Figure 4.30 for an illustration of segmentation of the same signal. Av. stands for average.

5.6 Analysis of Activity

Problem: Propose measures of waveform complexity or activity that may be used to analyze the extent of variability in signals such as the PCG and EMG.

Solution: The samples of a given EMG or PCG signal may, for the sake of generality, be treated as a random variable x . Then, the variance $\sigma_x^2 = E[(x - \mu_x)^2]$ represents an averaged measure of the variability or *activity* of the signal about its mean. If the signal has zero mean, or is preprocessed to meet the same condition, we have $\sigma_x^2 = E[x^2]$; that is, the variance is equal to the average power of the signal. Taking the square root, we get the *SD* of the signal as equal to its *RMS* value. Thus, the *RMS* value could be used as an indicator of the level of activity about the mean of the signal. A much simpler indicator of activity is the number of zero-crossings within a specified interval; the zero-crossing rate (*ZCR*) increases as the high-frequency content of the signal increases. A few measures related to the concepts introduced above are described in the following sections, with illustrations of application.

5.6.1 The *RMS* value

The *RMS* value of a signal $x(n)$ over its total duration of N samples is given by

$$RMS = \left[\frac{1}{N} \sum_{n=0}^{N-1} x^2(n) \right]^{\frac{1}{2}}. \quad (5.23)$$

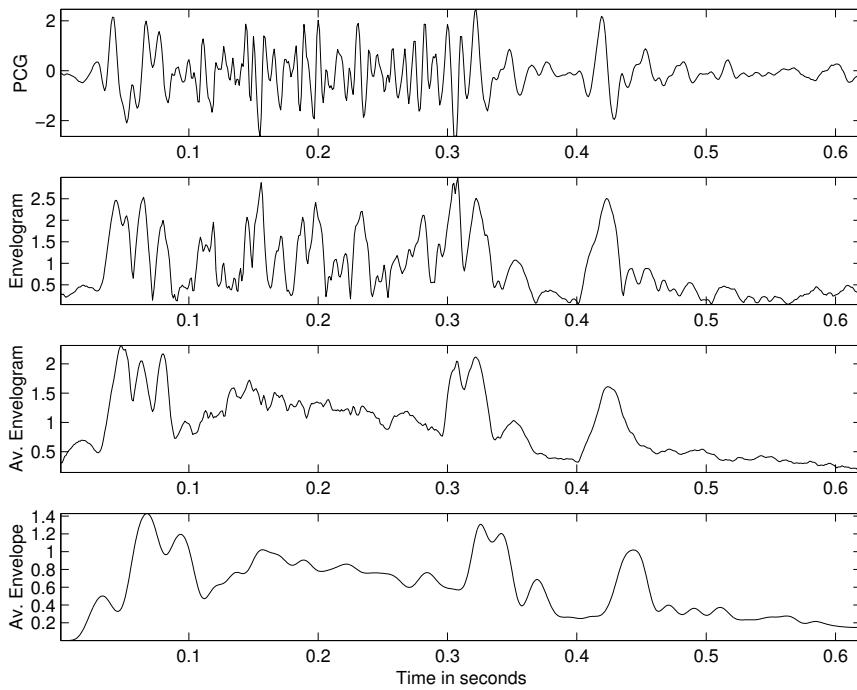


Figure 5.8 Top to bottom: PCG signal of a patient (female, 14 months) with systolic murmur (approximately $0.1 - 0.28$ s), split S2 ($0.28 - 0.38$ s), and opening snap of the mitral valve ($0.4 - 0.43$ s); envelopgram estimate of the signal shown; averaged envelopgram over 26 cardiac cycles; averaged envelope over 26 cardiac cycles. The PCG signal starts with S1. Av. stands for average. See Figure 4.31 for an illustration of segmentation of the same signal.

This global measure of signal level (related to power), however, is not useful for the analysis of trends in nonstationary signals. A running estimate of the *RMS* value of the signal computed over a causal window of M samples, defined as

$$RMS(n) = \left[\frac{1}{M} \sum_{k=0}^{M-1} x^2(n-k) \right]^{\frac{1}{2}}, \quad (5.24)$$

could serve as a useful indicator of the average power of the signal as a function of time. The duration of the window M needs to be chosen in accordance with the bandwidth of the signal, with $M \ll N$. Such an approach for computing running parameters of signals falls under the general scheme of *short-time analysis* of nonstationary signals [31].

Illustration of application: Gerbarg et al. [32, 33] derived power-versus-time curves of PCG signals by computing the average power in contiguous segments of duration 10 ms, and used the curves to identify systolic and diastolic segments of the signals. They noted that, within an interval of 10 s of a PCG signal, at least one diastolic segment would be longer than the corresponding systolic segment, and that all systolic segments in the interval would have approximately the same duration. Innocent (physiological) systolic murmurs in children were observed to be limited to the first and middle thirds of the systolic interval between S1 and S2, whereas pathological systolic murmurs due to mitral regurgitation were noted to be holosystolic. Based on these observations, Gerbarg et al. computed ratios of the mean power of the last third of systole to the mean power of systole and also to a certain “standard” noise level. A ratio was also computed of the mean energy of systole to the mean energy of the PCG over the complete cardiac cycle. Agreement in the range

of 78 – 91% was obtained between computer classification based on the three ratios defined above and clinical diagnosis of mitral regurgitation in different groups of subjects.

The use of the RMS value for the analysis of EMG and VMG signals, and thereby analysis of muscular activity, is illustrated in Section 5.11.

5.6.2 Zero-crossing rate

An intuitive indication of the “busy-ness” of a signal is provided by the number of times it crosses the zero-activity line or some other reference level. ZCR is defined as the number of times the signal crosses the reference within a specified interval. However, ZCR could be easily affected by DC bias, baseline wander, and low-frequency artifacts. For these reasons, it would be advisable to measure the ZCR of the derivative of the signal, which would be similar to the definition of turning points in the test for randomness described in Section 3.2.1. Saltzberg and Burch [34] discuss the relationship between ZCR and moments of PSDs and their application to EEG analysis.

Illustrations of application: In spite of its simplicity, ZCR has been used in practical applications such as speech signal analysis to perform speech-versus-silence decision and to discriminate between voiced and unvoiced sounds [31] (see also Figure 3.1) and PCG analysis for the detection of murmurs. Jacobs et al. [35] used ZCR to perform normal-versus-abnormal classification of PCG signals using the ECG as a trigger, and obtained correct classification rates of 95% for normals (58/61) and 94% for abnormalities (77/82). They indicated a decision limit of 20 zero-crossings in a cardiac cycle. Yokoi et al. [36] proposed a screening system based on measurements of the maximum amplitude and ZCR in 8 – ms segments of PCG signals (sampled at 2 $k\text{Hz}$). They obtained correct classification rates of 98% with 4,809 normal subjects and 76% with 1,217 patients with murmurs.

5.6.3 Turns count

Willison [37] proposed to analyze the level of activity in EMG signals by determining the number of spikes occurring in the interference pattern (see also Goodgold and Eberstein [38], Fuglsang-Frederiksen and Måansson [39], and Dowling et al. [40]). Instead of counting zero-crossings, Willison’s method investigates the significance of every change in phase (direction or slope) of the EMG signal called a *turn*. Turns greater than 100 μV are counted, with the threshold selected so as to avoid counting insignificant fluctuations due to noise. The method is similar to counting turning points as in the test for randomness described in Section 3.2.1, but is expected to be robust in the presence of noise due to the threshold imposed. See Figure 5.9 for an illustration of the differences between zero-crossings, turning points, and turns count. The method is not directly sensitive to SMUAPs, but significant phase changes caused by superimposed SMUAPs are counted. Willison [37] found that EMG signals of subjects with myopathy possessed higher turns counts than those of normal subjects at comparable levels of volitional effort.

Illustration of application: The topmost plot in Figure 5.10 illustrates the EMG signal over two breath cycles from the crural diaphragm of a dog recorded via implanted fine-wire electrodes [19]. The subsequent plots illustrate, in top-to-bottom order, the short-time RMS values, the turns count by Willison’s procedure, and the smoothed envelope of the signal. The RMS and turns count values were computed using a causal moving window of duration 70 ms (210 samples). The window duration needs to be chosen to strike a balance between the extent of smoothing desired in the turns count series and the accuracy in reflecting the nonstationary nature of the signal; in the present example, nonstationarity is related to the increasing level of EMG activity with inspiration. The envelope was obtained by taking the absolute value of the signal (equivalent to full-wave rectification) followed by a Butterworth lowpass filter of order $N = 8$ and cutoff frequency $f_c = 8 \text{ Hz}$. It is seen that all three of the derived features demonstrate the expected increasing trend with the level

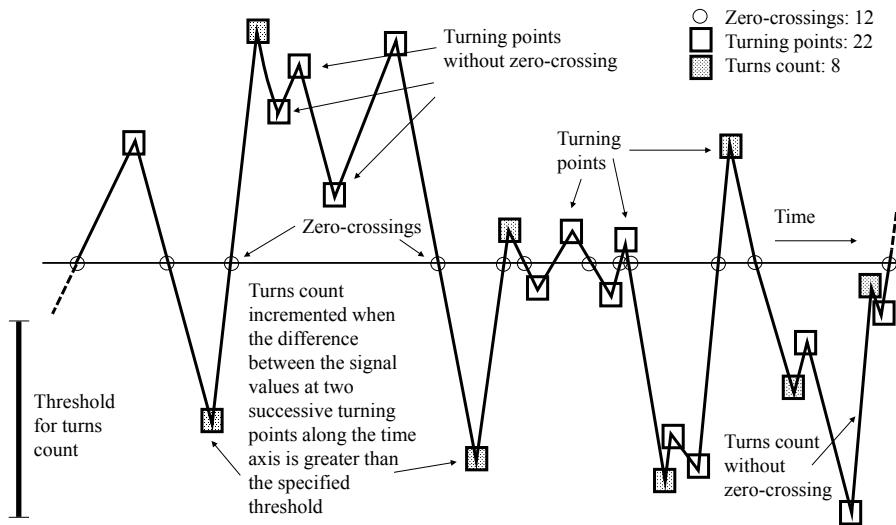


Figure 5.9 Differences between zero-crossings, turning points, and turns count.

of contraction (inspiration), and can serve as correlates or indicators of muscle contraction and the concomitant EMG complexity. The results may be further smoothed (lowpass filtered) if desired.

Figure 5.11 illustrates one $70 - ms$ segment of the EMG signal in Figure 5.10 with the boundary points of the significant turns as detected by Willison's procedure marked by the “*” symbol. The procedure was implemented by first computing the derivative of the EMG signal and detecting points of change in its sign. A turn was marked wherever the EMG signal differed by at least $100 \mu V$ between successive points of sign change in the derivative. Observe from Figure 5.11 that the EMG signal need not cross the zero line to cause a turns count, and that zero-crossings with voltage swings of less than $100 \mu V$ are not counted as turns.

5.6.4 Form factor

Based on the notion of variance as a measure of signal activity, Hjorth [41–43] (see also Cooper et al. [44]) proposed a method for the analysis of EEG waves. In this method, short-time segments of duration 1 s or longer are analyzed, and three parameters are computed. The first parameter is called *activity* and is simply the variance σ_x^2 of the signal segment $x(n)$. The second parameter, called *mobility* M_x , is computed as the square root of the ratio of the activity of the first derivative of the signal to the activity of the original signal:

$$M_x = \left[\frac{\sigma_{x'}^2}{\sigma_x^2} \right]^{\frac{1}{2}} = \frac{\sigma_{x'}}{\sigma_x}, \quad (5.25)$$

where x' stands for the first derivative of x . The third parameter, called *complexity* or the *form factor* (FF), is defined as the ratio of the mobility of the first derivative of the signal to the mobility of the original signal itself:

$$FF = \frac{M_{x'}}{M_x} = \frac{\sigma_{x''}/\sigma_{x'}}{\sigma_{x'}/\sigma_x}, \quad (5.26)$$

where x'' stands for the second derivative of the signal. The complexity of a sinusoidal wave is unity; other waveforms have complexity values increasing with the extent of variations present in them.

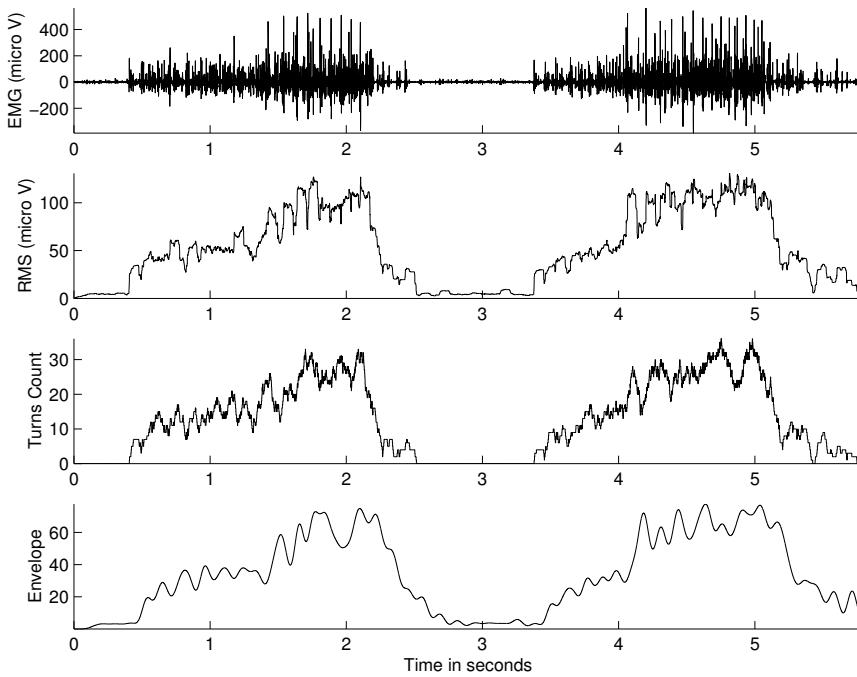


Figure 5.10 Top to bottom: EMG signal over two breath cycles from the crural diaphragm of a dog recorded via implanted fine-wire electrodes; short-time *RMS* values; turns count using Willison’s procedure; and smoothed envelope of the signal. The *RMS* and turns count values were computed using a causal moving window of $70 - ms$ duration. The threshold used to detect turns is $100 \mu V$. EMG signal courtesy of R.S. Platt and P.A. Easton, Department of Clinical Neurosciences, University of Calgary.

Hjorth [42,43] described the mathematical relationships among the activity, mobility, complexity, and PSD of a signal, and applied them to model EEG signal generation. Binnie et al. [45, 46] described the application of *FF* and spectrum analysis to EEG analysis for the detection of epilepsy. However, because the computation of *FF* is based on the first and second derivatives of the signal and their variance, the measure is sensitive to noise. A complex and relatively wideband signal such as the EMG is not amenable to analysis via *FF*. Application of *FF* to discriminate between normal and ectopic ECG beats is illustrated in Section 5.7.

We have explored a few measures to characterize waveform complexity in this section. Many authors have proposed several other diverse measures and interpretations of waveform or system complexity in the literature, examples of which include features based on nonlinear dynamics and the correlation dimension [47], and the embedding dimension of time-varying dynamic systems [48].

5.7 Application: Parameterization of Normal and Ectopic ECG Beats

Problem: *Develop a parameter to discriminate between normal ECG waveforms and ectopic beats (PVCs).*

Solution: We have observed several times that ectopic beats, due to the abnormal propagation paths of the associated excitation pulses, typically possess waveforms that are significantly different from those of the normal QRS waveforms of the same subject. More often than not, ectopic beats have bizarre and complex waveshapes. The form factor *FF* described in Section 5.6.4 parameterizes the notion of waveform complexity, providing a value that increases with complexity.

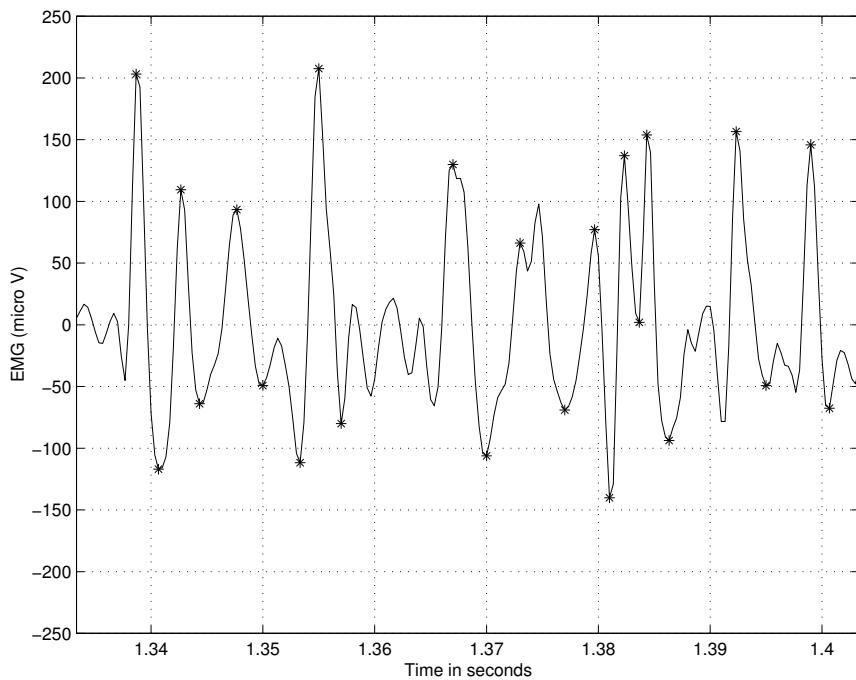


Figure 5.11 Illustration of the detection of turns in a $70 - ms$ window of the EMG signal in Figure 5.10. The segments of the signal between pairs of “**” marks include significant turns. The threshold used to detect turns is $100 \mu V$.

Therefore, FF appears to be a suitable measure to discriminate between normal and ectopic beats. Note that the RR interval by itself cannot indicate ectopic beats, as the RR interval could vary due to sinus arrhythmia and conduction problems, as well as due to HRV.

Figure 5.12 displays a segment of the ECG of a patient with ectopic beats; the segment illustrates the initiation of an episode of *ventricular bigeminy*, where every normal beat is followed by an ectopic beat [3]. The ECG of the patient was processed using the Pan–Tompkins algorithm for QRS detection (see Section 4.3.2). QRS marker points were detected using a simple threshold applied to the output of the Pan–Tompkins algorithm. Each beat was segmented at points $160\ ms$ before and $240\ ms$ after the detected marker point; the diamond and circle symbols on the ECG in Figure 5.12 indicate the starting and ending points of the corresponding beats. The FF value as given by Equation 5.26 was computed for each segmented beat. The RR interval (in ms) and the FF value are shown for each beat in Figure 5.12. It can be readily seen that the FF values for the PVCs are higher than those for the normal beats.

Note from Figure 5.12 that the RR intervals for the PVCs are lower than those for the normal beats, and that the normal beats that follow the PVCs have higher-than-normal RR intervals due to the compensatory pause. Pattern classification of the ECG beats in this example as normal or PVCs using RR , FF , and other features is described in Section 10.11.

5.8 Application: Analysis of Exercise ECG

Problem: Develop an algorithm to analyze changes in the ST segment of the ECG during exercise.

Solution: Hsia et al. [49] developed a method to analyze changes in the ST segment of the ECG signal while the subject performed physical exercises. The analysis was performed as part of

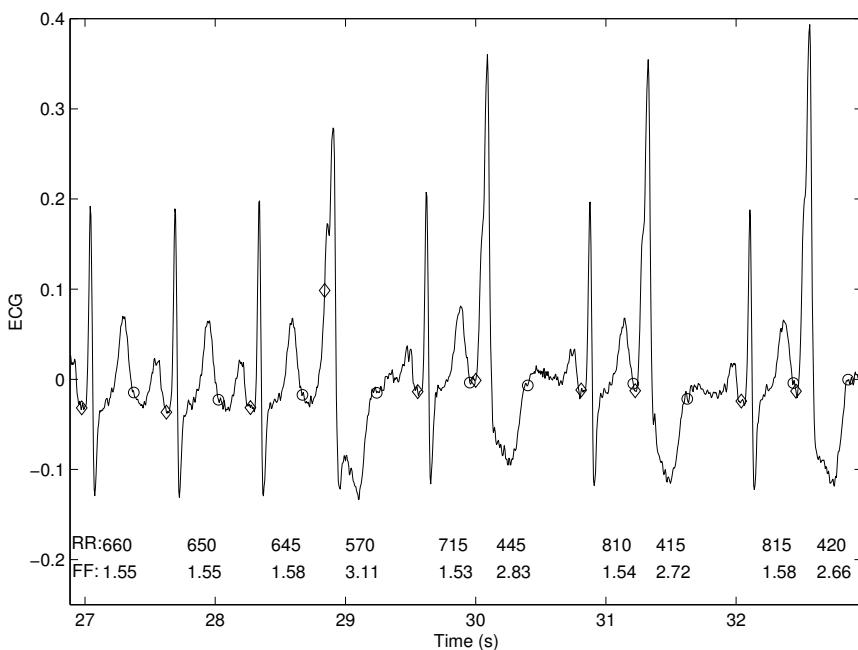


Figure 5.12 Segment of the ECG of a patient (male, 65 years) with ectopic beats. The diamond and circle symbols indicate the starting and ending points, respectively, of each QRS-T wave obtained using the Pan-Tompkins algorithm. The *RR* interval (in ms) and the *FF* value are shown for each beat.

a radionuclide ventriculography (gated blood-pool imaging) procedure. In this procedure, nuclear medicine images are obtained of the left ventricle before and after the patient performs exercises on a treadmill or bicycle ergometer. Images are obtained at different phases of the cardiac cycle by gating the gamma ray emission data with reference to the ECG; image data for each phase are averaged over several cardiac cycles [50]. Analysis of exercise ECG is complicated due to baseline artifacts caused by the effects of respiration, skin resistance changes due to perspiration, and soft tissue movement affecting electrode contact. Detection of changes in the ST segment in the presence of such artifacts poses a major challenge.

One of the main parameters used by Hsia et al. is related to the correlation coefficient as defined in Equation 3.97. The measure, however, is affected by baseline variations. To address this, a modified correlation coefficient was defined as

$$\gamma_{xy} = \frac{\sum_{n=0}^{N-1} [x(n)][y(n) - \Delta]}{\sqrt{\sum_{n=0}^{N-1} [x(n)]^2 \sum_{n=0}^{N-1} [y(n) - \Delta]^2}}. \quad (5.27)$$

Here, $x(n)$ is the template, $y(n)$ is a segment of the ECG signal being analyzed, Δ is a baseline correction factor defined as the difference between the baseline of $y(n)$ and the baseline of $x(n)$, and N is the duration (number of samples) of the template and the signal segment being analyzed. The template was generated by averaging up to 20 QRS complexes that met a specified *RR* interval constraint.

Hsia et al. proposed a method to establish the baseline of each ECG beat by searching for the PQ segment by backtracking from the detected R point (trigger for the purpose of gating the image data). The region of three consecutive samples with the minimum change (maximum flatness) preceding the QRS was taken to represent the baseline of the beat. (*Note:* The PQ segment is typically isoelectric (after high-and low-frequency noise in the ECG has been removed), whereas the

ST segment is variable in the presence of certain cardiac abnormalities.) The search procedure also established the width of the QRS complex to be used in template matching (N in Equation 5.27). Beats with $\gamma_{xy} < 0.85$ were considered to be abnormal. The baseline correction factor in Equation 5.27 provided the robustness required.

Groups of 16 successive normal beats were aligned and averaged to obtain a representative waveform. The ST segment level was computed as the difference between a reference ST point and the isoelectric level of the current averaged beat. The averaging procedure included a condition to reject beats with abnormal morphology, such as PVCs. The ST reference point was defined as $R + 64 \text{ ms} + \max(4, \frac{200-HR}{16}) \times 4 \text{ ms}$ or $S + 44 \text{ ms} + \max(4, \frac{200-HR}{16}) \times 4 \text{ ms}$, where R or S indicates the position of the R or S wave of the present beat in ms , and HR is in bpm . Significant ST level differences were reported by the algorithm, with elevation or depression of the ST segment by more than 0.1 mV with reference to the baseline being considered to be significant. Furthermore, the slope of the ST segment was computed using two samples before and two samples after the detected ST point as described above (a duration of 16 ms , with the sampling rate of 250 Hz).

In addition to the analysis of the ST segment, the method of Hsia et al. performed rhythm analysis, identified PVCs and other abnormal beats, and assisted in the rejection of gamma ray emission data related to abnormal beats from the imaging procedure. The combined use of nuclear medicine imaging and ECG analysis can improve the accuracy of the diagnosis of myocardial ischemia.

5.9 Application: Quantitative Analysis of the EMG in Relation to Force Exerted

Problem: Propose methods for parametric analysis of the variations in the EMG signal with respect to the force exerted by a muscle.

Solution: We have seen several examples demonstrating increasing levels of power and complexity of the EMG signal with increasing muscular activity or force exerted; see, for example, Figures 1.20, 1.21, 1.22, 1.23, and 5.10. We have also studied how such variations in a signal may be represented quantitatively using measures such as the *RMS* value, *ZCR*, and turns count. Let us now explore the relationships between these entities in a formal and mathematical manner.

Figure 5.13 shows the EMG and force signals as in Figure 1.22, but with automatic delineation of the portions of the force signal corresponding to the five intervals of muscular contraction. To perform this task, starting from the first sample, the point where the force signal increased beyond 10% MVC was identified. Then, the next point where the signal dropped below 10% MVC was identified. This process was repeated until the end of the signal. To refine the definition of each interval of contraction, a threshold was defined as 0.7 times the maximum level of contraction within the interval. Then, the smaller extent of each interval previously identified, within which the force remained above the threshold, was detected. The final intervals detected by this procedure are labeled in Figure 5.13 with “**” marks.

Within each interval of the force signal identified as above as well as the corresponding interval in the EMG signal, the average force exerted, the *RMS* value, *ZCR*, and the turns count divided by the time duration of the interval (referred to as the turns count rate or *TCR*) were computed. The threshold to detect significant turns was set at $100 \mu\text{V}$.

Figure 5.14 demonstrates the important difference between zero-crossings and turns. The cumulative number of zero-crossings detected keeps increasing with time even during the intervals when no muscular force is exerted, due to noisy variations in the EMG signal during the periods of rest. The threshold to detect significant turns, which was set at $100 \mu\text{V}$ in this experiment, disregards the small turns associated with the background variations in the EMG signal, as well as small variations during periods of contraction. Thus, the cumulative number of significant turns detected is seen to increase only during the periods of contraction and not during the periods of rest. Therefore, we may expect *TCR* to be a better indicator of muscular force exerted than *ZCR*.

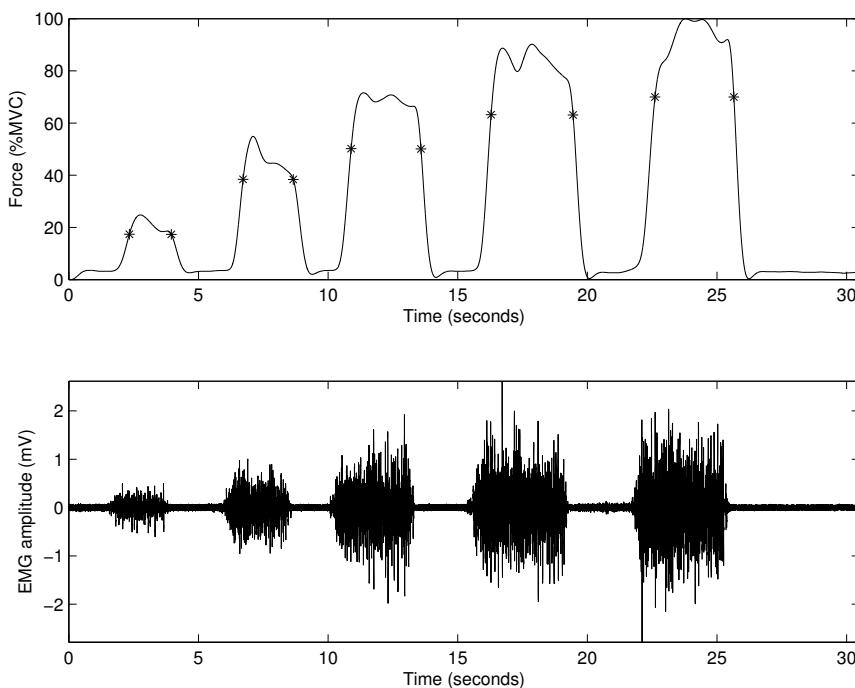


Figure 5.13 Force and EMG signals recorded from the forearm muscle of a subject using surface electrodes. The portions of muscular contraction analyzed have been automatically identified in the force signal and delimited by the “*” marks.

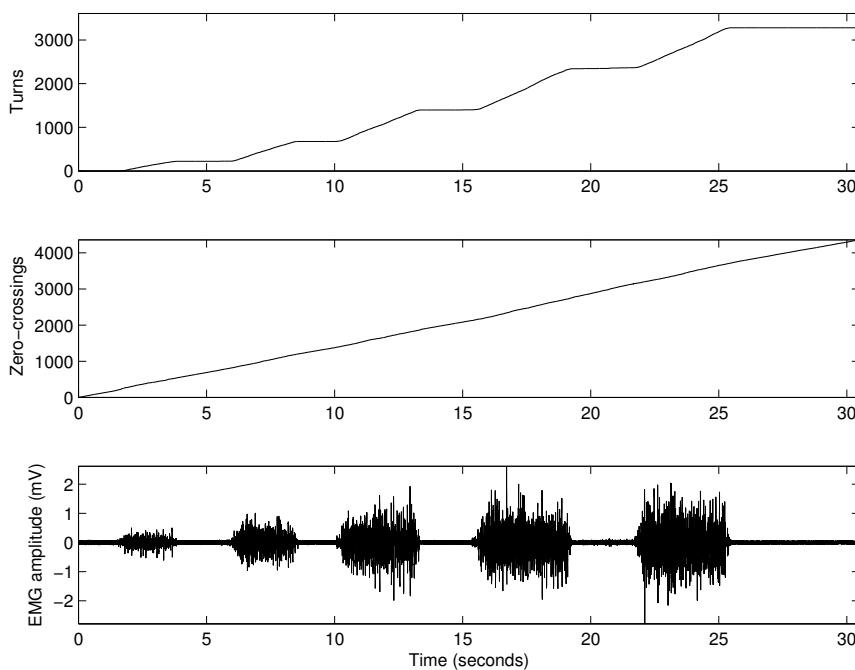


Figure 5.14 Bottom to top: EMG signal, cumulative count of zero-crossings, and cumulative number of significant turns. See also Figure 5.13.

In order to analyze the variation of the EMG signal with force in a quantitative manner, a linear fit was computed to represent the variation of each EMG parameter versus force. The linear models (straight-line fits) obtained are shown in the plots in Figures 5.15, 5.16, and 5.17 for the *RMS*, *ZCR*, and *TCR* values, respectively, as compared with the force in %MVC. To obtain a measure of fit between each parameter and force, the correlation coefficient was computed as [51]

$$r^2 = \frac{\left[\sum_{n=1}^{n=N} x(n)y(n) - N\bar{x}\bar{y} \right]^2}{\left[\sum_{n=1}^{n=N} x^2(n) - N\bar{x}^2 \right] \left[\sum_{n=1}^{n=N} y^2(n) - N\bar{y}^2 \right]}, \quad (5.28)$$

where N is the number of samples of x or y representing the variables *RMS*, *ZCR*, *TCR*, or force ($N = 5$ in the present study); \bar{x} is the mean of x . The values of r^2 for *RMS*, *ZCR*, and *TCR*, as compared with the force in %MVC, are given in the captions of Figures 5.15, 5.16, and 5.17, respectively. The value of *ZCR* shows limited variation and low correlation with force. It is evident that *RMS* and *TCR* have a high degree of correlation and linear behavior with respect to force.

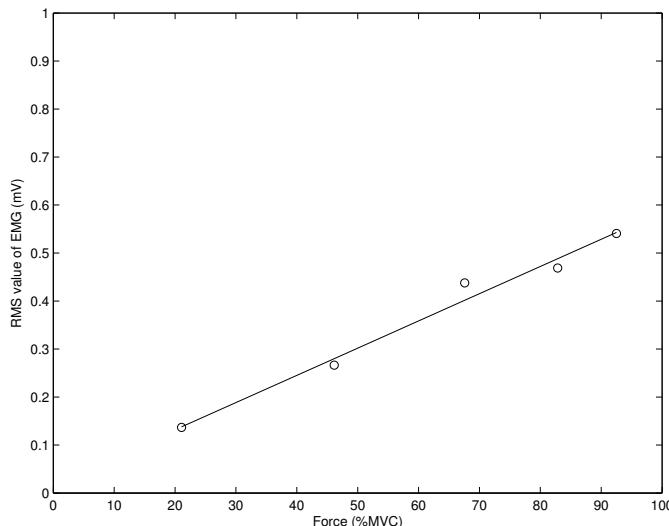


Figure 5.15 Variation of the *RMS* value of the EMG signal in Figure 5.13 with the average muscular force exerted in each of the five periods of contraction. The “o” marks indicate the measured values, and the straight line indicates the computed linear model. $r^2 = 0.98$.

The study described above is limited due to the use of a single EMG signal and a simple method to delimit the intervals of muscular contraction. It is seen in Figure 5.13 that the muscular force exerted is not constant within each interval. For robust analysis, it is necessary to repeat the experiment several times with many subjects and analyze a large number of samples of the parameters. See Section 5.11 for further related discussion and illustration of the relationships between EMG and force.

5.10 Application: Analysis of Respiration

Problem: Propose a method to relate EMG activity to airflow during respiration.

Solution: Platt et al. [19] recorded EMG signals from the parasternal intercostal and crural diaphragm muscles of dogs. One EMG signal was obtained from a pair of electrodes mounted at a fixed distance of 2 mm placed between fibers in the third left parasternal intercostal muscle

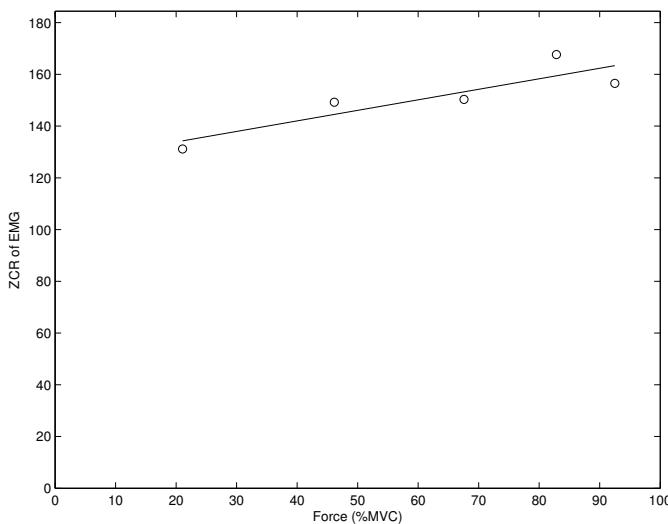


Figure 5.16 Variation of the *ZCR* of the EMG signal in Figure 5.13 with the average muscular force exerted in each of the five periods of contraction. The “o” marks indicate the measured values, and the straight line indicates the computed linear model. $r^2 = 0.78$.

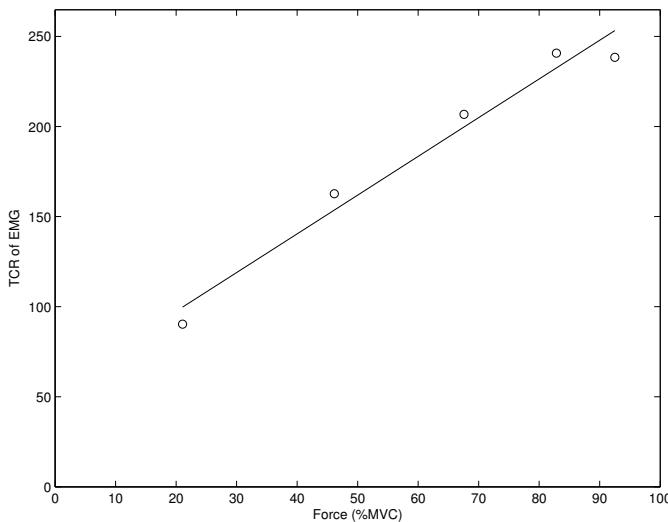


Figure 5.17 Variation of the *TCR* of the EMG signal in Figure 5.13 with the average muscular force exerted in each of the five periods of contraction. The “o” marks indicate the measured values and the straight line indicates the computed linear model. $r^2 = 0.97$.

about 2 cm from the edge of the sternum. The crural diaphragm EMG was obtained via fine-wire electrodes sewn in-line with the muscle fibers and placed 10 mm apart. During the signal acquisition experiment, the dog breathed through a snout mask, and a pneumotachograph was used to measure airflow. Figures 1.20, 1.21, and 5.10 show samples of the crural EMG signal.

Although the EMG signal is commonly used in many physiological studies including analysis of respiration, the intricate variations in the signal are often not of interest. A measure of the total or integrated electrical activity, ideally reflecting the global activity in the pool of active motor units of the muscle, would serve the purposes of most analyses [19]. As the EMG signal is nonstationary,

short-time measures are called for. The smoothed envelope of the EMG signal is commonly used under these circumstances.

Platt et al. observed that the filters commonly used for smoothing rectified EMG signals had poor high-frequency attenuation, resulting in noisy envelopes. They proposed a modified Bessel filter for application to the EMG signal after full-wave rectification; the filter severely attenuated frequencies beyond 20 Hz with gain $< -70 \text{ dB}$, and yielded EMG envelopes that were much smoother than those given by other filters.

The EMG envelopes derived by Platt et al. agreed well with the inspiratory airflow pattern. Figure 5.18 shows plots of the parasternal intercostal EMG signal over two breath cycles, the corresponding filtered envelope, and the airflow pattern. Figure 5.19 shows the correlation between the filtered EMG envelope amplitude and the airflow in liters per second. It is evident that the envelope extracted by this method is an excellent correlate of inspiratory airflow.

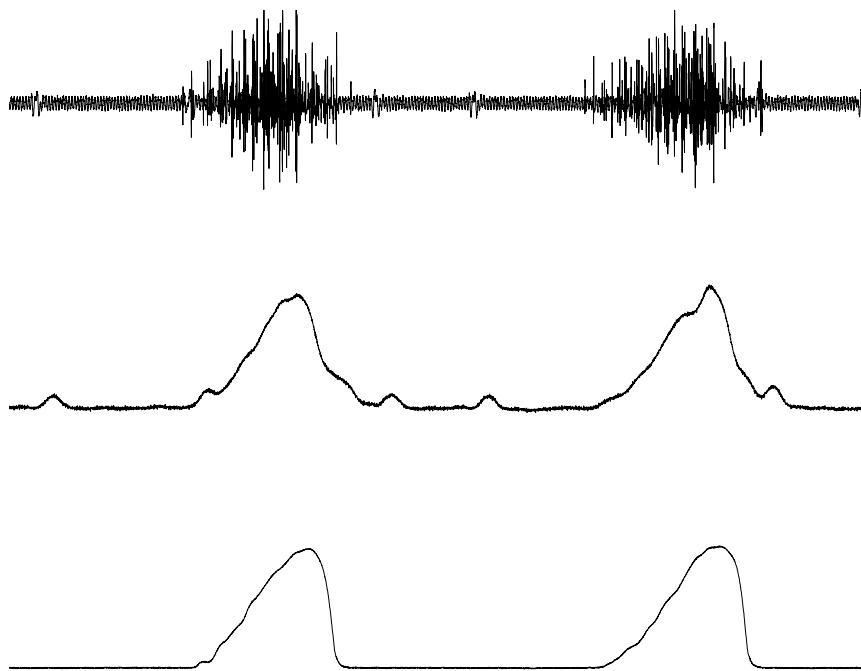


Figure 5.18 Top to bottom: EMG signal over two breath cycles from the parasternal intercostal muscle of a dog recorded via implanted electrodes; EMG envelope obtained with the modified Bessel filter with a time constant of 100 ms; and inspiratory airflow. The duration of the signals plotted is 5 s. The several minor peaks appearing in the envelope are related to the ECG that appears as an artifact in the EMG signal. Data courtesy of R.S. Platt and P.A. Easton, Department of Clinical Neurosciences, University of Calgary [19].

5.11 Application: Electrical and Mechanical Correlates of Muscular Contraction

Problem: Derive parameters from the electrical and mechanical manifestations of muscular activity that correlate with the level of contraction or force produced.

Solution: Zhang et al. [52,53] studied the usefulness of simultaneously recorded EMG and VMG signals in the analysis of muscular force produced by contraction. In their experimental procedure, the subjects performed isometric contraction (that is, with no movement of the associated leg) of the rectus femoris (thigh) muscle to different levels of torque with a Cybex II dynamometer. Four

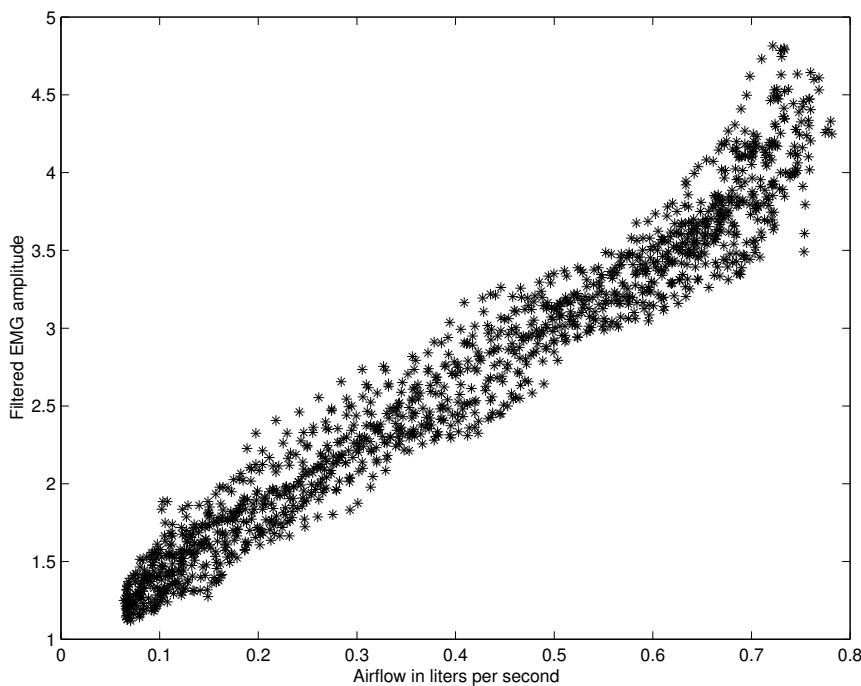


Figure 5.19 Correlation between EMG amplitude obtained from the Bessel-filtered envelope versus inspiratory airflow. The EMG envelope was filtered using a modified Bessel filter with a time constant of 100 ms. Data courtesy of R.S. Platt and P.A. Easton, Department of Clinical Neurosciences, University of Calgary [19].

levels of contraction were performed from 20% to 80% of the MVC level of the individual subject. The experiments were performed at three knee-joint angles of 30°, 60°, and 90°. Each contraction was held for a duration of about 6 s, and the subjects rested in between experiments to prevent the development of muscle fatigue. The VMG signal was recorded using a Dytran 3115a accelerometer, and surface EMG signals were recorded using disposable $Ag - AgCl$ electrodes. The VMG signals were filtered to the bandwidth 3 – 100 Hz, and the EMG signals were filtered to 10 – 300 Hz. The VMG and EMG signals were sampled at 250 Hz and 1,000 Hz, respectively. Figure 2.4 illustrates sample recordings of the VMG and EMG signals at two levels of contraction.

RMS values were computed for each contraction level over a duration of 5 s. Figure 5.20 shows the variation of the *RMS* values of the EMG and VMG signals acquired at a knee-joint angle of 60° and averaged over four subjects. The almost-linear trends of the *RMS* values of both the signals with muscular contraction indicate the usefulness of the derived parameter in the analysis of muscular activity. It should, however, be noted that the relationship between *RMS* values and contraction may not follow the same (linear) pattern for different muscles. Figure 5.21 shows the *RMS*-versus-%MVC relationships for three muscles: The relationship is linear for the FDI, whereas it is nonlinear for the biceps and deltoid muscles [54].

5.12 Application: Statistical Analysis of VAG Signals

Problem: Knee-joint VAG signals possess different levels of waveform complexity depending upon the state of the cartilage covering the joint surfaces. Explore the use of statistical parameters that characterize variability for parametric representation and classification of VAG signals.

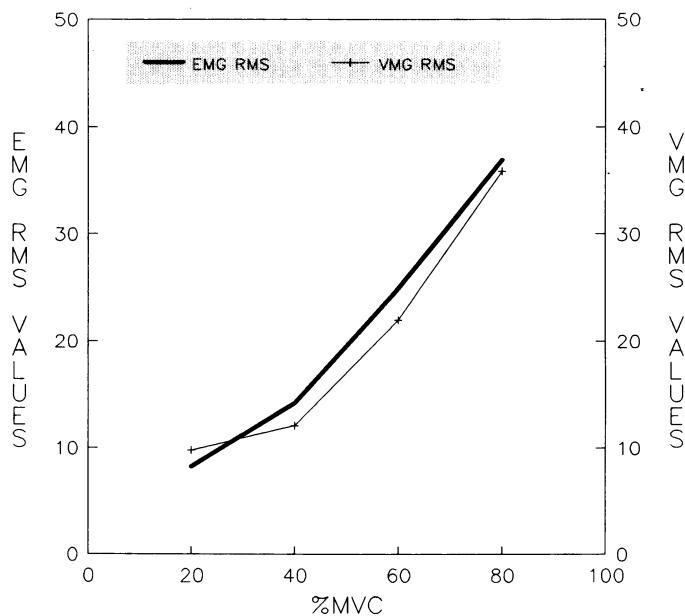


Figure 5.20 *RMS* values of the VMG and EMG signals for four levels of contraction of the rectus femoris muscle at 60° knee-joint angle averaged over four subjects. Reproduced with permission from Y.T. Zhang, C.B. Frank, R.M. Rangayyan, and G.D. Bell, Relationships of the vibromyogram to the surface electromyogram of the human rectus femoris muscle during voluntary isometric contraction, *Journal of Rehabilitation Research and Development*, 33(4): 395–403, 1996. ©Department of Veterans Affairs.

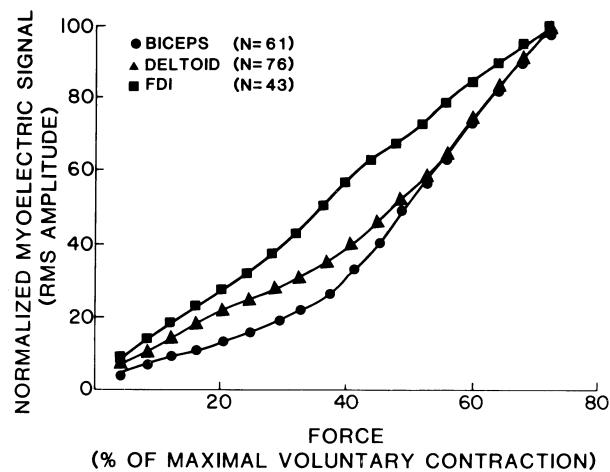


Figure 5.21 EMG *RMS* value versus level of muscle contraction expressed as a percentage of MVC for each subject. The relationship is displayed for three muscles. FDI: first dorsal interosseus. N: number of muscles (subjects) in the study. Reproduced with permission from J.H. Lawrence and C.J. de Luca, Myoelectric signal versus force relationship in different human muscles, *Journal of Applied Physiology*, 54(6):1653–1659, 1983. ©American Physiological Society.

Solution: As described in Section 1.2.14, a normal knee joint has smooth cartilage surfaces and produces almost no sound or vibration; see also Section 8.2.3. Regardless, when a VAG signal is recorded from a normal knee joint, a signal with some variations is obtained. Under pathological conditions, the cartilage surfaces are eroded and could generate additional sounds; see Figure 8.2. We could expect the statistics of the variability of normal and abnormal VAG signals to demonstrate differences due to the different nature of the underlying causes. Rangayyan and Wu [55–57] explored the use of several statistical parameters for the purpose of screening VAG signals. The related background and methods are described in the following paragraphs.

5.12.1 Acquisition of knee-joint VAG signals

The dataset used in the studies of Rangayyan and Wu [55–57] consists of 89 signals, with 51 from normal volunteers (22 male and 29 female, age 28 ± 9.5 years) and 38 from subjects with knee-joint pathology (20 male and 18 female, age 35 ± 13.8 years). The normal condition of the volunteers was established by clinical examination and history. The abnormal signals were collected from symptomatic patients scheduled to undergo arthroscopy independent of the VAG studies. Informed consent was obtained from each subject. The experimental protocol [58] was approved by the Conjoint Health Research Ethics Board of the University of Calgary.

The abnormal cases in the dataset include chondromalacia of different grades at the patella, meniscal tear, tibial chondromalacia, and anterior cruciate ligament injuries, as confirmed during arthroscopic examination. Due to the inadequacy of the data available for classification of the signals into various types or stages of pathology, studies using this dataset have been limited to screening, that is, classification of the signals and the corresponding knee joints as normal or abnormal.

Figure 5.22 shows examples of normal and abnormal VAG signals. Each signal was normalized to the amplitude range $[0, 1]$. The abnormal signal exhibits a higher degree of overall variation, activity, or complexity than the normal signal.

5.12.2 Estimation of the PDFs of VAG signals

In the work of Rangayyan and Wu [57], models of the PDFs of VAG signals were derived using the Parzen-window approach [59, 60] to represent their basic statistical characteristics. A histogram was computed by combining all of the normal signals into one group. The histogram, denoted by $h_n(x_l)$, with x_l , $l = 0, 1, 2, \dots, L - 1$, represents the L bins used to quantize the range of the values of the signal x . Rangayyan and Wu used $L = 100$ bins to represent the normalized range of $[0, 1]$ for the VAG signal values. A large value for L could result in several bins with negligible or zero counts, whereas a small value could cause diminished differences between the histograms for normal and abnormal signals. The value of $L = 100$ was selected based on an analysis of the results with several values of L . Similarly, a histogram $h_a(x_l)$ was obtained by pooling together all of the abnormal signals. Each histogram was normalized by dividing by the total number of samples in the population to have unit area under the normalized histogram. A Gaussian fit was then obtained for each of the two histograms, denoted as $g_n(x_l)$ and $g_a(x_l)$.

The Parzen-window approach to obtain a nonparametric estimate of the PDF from a collection of samples was applied as follows [59, 60]. Consider the situation where we have a set of independent samples, $Z = \{z_1, z_2, \dots, z_K\}$, with an unknown underlying PDF $p(z)$. A nonparametric estimate of $p(z)$ from Z is provided by the function [59, 60]

$$\hat{p}(z) = \frac{1}{K} \sum_{k=1}^K \kappa(z - z_k), \quad (5.29)$$

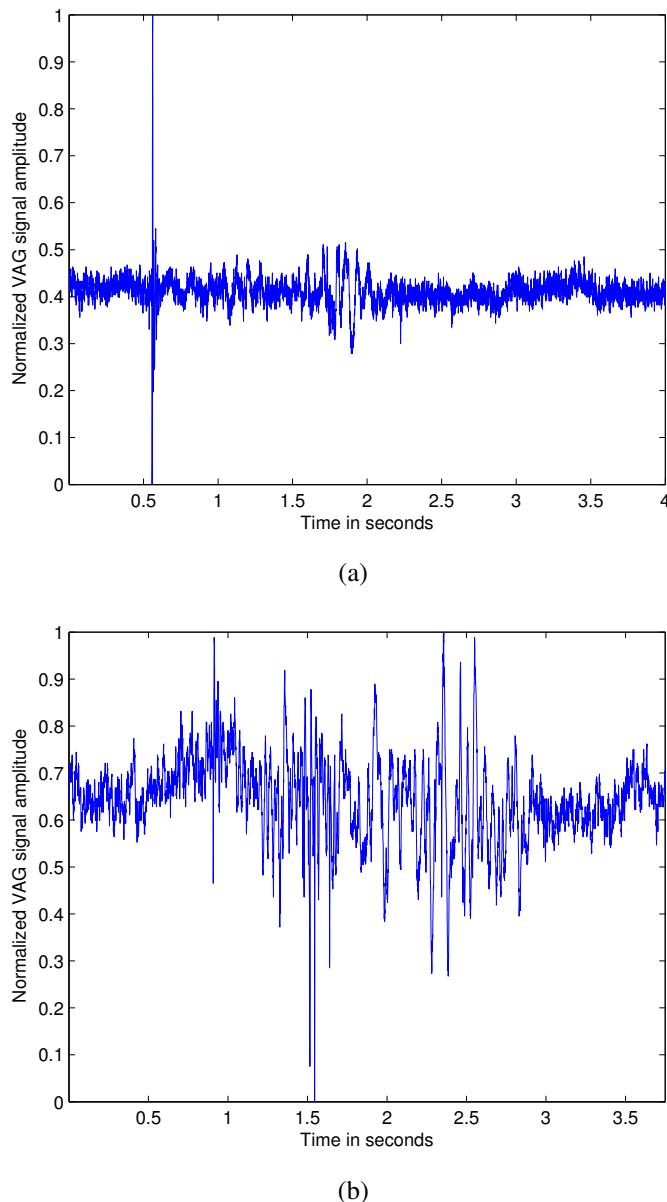


Figure 5.22 VAG signal of (a) a normal subject and (b) a subject with knee-joint pathology. Reproduced from R.M. Rangayyan and Y.F. Wu, Screening of knee-joint vibroarthrographic signals using probability density functions estimated with Parzen windows, *Biomedical Signal Processing & Control*, January 2010, 5(1):53–58, with permission from Elsevier. ©Elsevier.

where κ is a window or kernel function that integrates to unity. Rangayyan and Wu [57] used

$$\kappa(z - z_k) = \frac{1}{\sigma_P \sqrt{2\pi}} \exp \left[-\frac{(z - z_k)^2}{2\sigma_P^2} \right]. \quad (5.30)$$

The Parzen-window PDF was estimated using the quantized values of the signals, x_l , $l = 0, 1, 2, \dots, L - 1$, with $L = 100$ levels, in the normalized range $[0, 1]$. Experiments were conducted with the value of the parameter σ_P in Equation 5.30 varying over the range $[0.01, 0.1]$ in steps of 0.01; the final value was set equal to 0.04.

Figure 5.23 shows the Parzen-window estimates of the PDFs of the normal and abnormal VAG signals shown in Figure 5.22. Figure 5.24 shows the Parzen-window estimates of the PDFs of the 51 normal and 38 abnormal VAG signals in the dataset used. The amplitude of the VAG signals has been normalized to the range $[0, 1]$. The figures also show the normalized histogram and the Gaussian fit for each case.

The following observations were made from the PDFs of the normal and abnormal VAG signals:

- In general, the Parzen-window PDFs are closer to the normalized histograms of the VAG signals than the corresponding Gaussian models. The closeness of the fit depends upon the value of σ_P used: The smaller the value of σ_P , the better the fit. However, a larger value of σ_P is desirable to obtain smooth model PDFs that do not include irrelevant details of the histograms of the signals. A balance needs to be achieved between these two requirements.
- The PDF models for the abnormal signals indicate that the abnormal signals have higher probabilities of occurrence of higher values within the normalized range $[0, 1]$.
- Based on the differences between the PDF models for the normal and abnormal signals, it should be possible to classify VAG signals using parameters derived from their PDFs.
- The apparent differences between the PDF models suggest the existence of different underlying signal-generation processes or statistical models for normal and abnormal VAG signals.

The following section provides details on the derivation of parameters from PDFs for classification of VAG signals.

5.12.3 Screening of VAG signals using statistical parameters

VAG signals related to various types of knee-joint pathology have been observed to possess a larger extent of variability over the duration of a swing cycle of the leg than normal VAG signals [55–57, 61–66]. To characterize this nature of VAG signals, Moussavi et al. [62] used the variance of the means of the segments of a given VAG signal; the signals were segmented adaptively using the RLS algorithm. Rangayyan and Wu [57] explored the use of several statistical parameters [67], including FF (see Section 5.6.4), skewness, kurtosis, and entropy [55] as well as an adaptive turns count and the variance of the MS value [56] (see Section 3.2.1), for screening of VAG signals. Other methods that have been proposed for the analysis of VAG signals include AR modeling [62, 64, 65, 68], cepstral coefficients [64], time–frequency distributions (TFDs) [58], and wavelet packet decomposition [63]; see Wu et al. [66] for a review.

As described in Section 3.2.1, various statistical measures may be computed based on the moments of the PDF of a given signal, $p_x(x_l)$, with x_l , $l = 0, 1, 2, \dots, L - 1$, representing the L bins used to span the range of the values of the signal x . The k^{th} central moment of the PDF $p_x(x_l)$ is defined as

$$m_k = \sum_{l=0}^{L-1} (x_l - \mu)^k p_x(x_l), \quad (5.31)$$

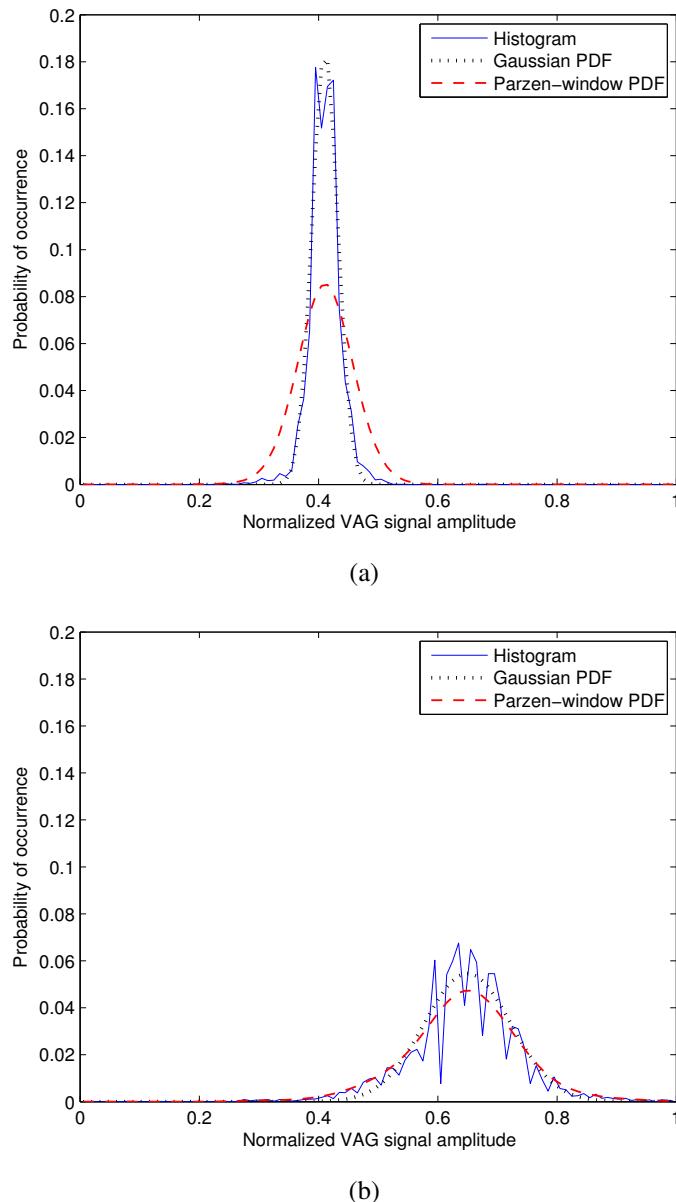


Figure 5.23 Nonparametric Parzen-window estimates of the PDFs of the VAG signals in Figure 5.22 of (a) a normal subject and (b) a patient with knee-joint pathology. The amplitude has been normalized to the range [0, 1]. The figure also shows the normalized histogram and the Gaussian fit for each case. The parameters of the Gaussian fit are (a) mean = 0.4107, $SD = 0.0216$; and (b) mean = 0.6499, $SD = 0.0730$. Reproduced from R.M. Rangayyan and Y.F. Wu, Screening of knee-joint vibroarthrographic signals using probability density functions estimated with Parzen windows, *Biomedical Signal Processing & Control*, January 2010, 5(1):53–58, with permission from Elsevier. ©Elsevier.

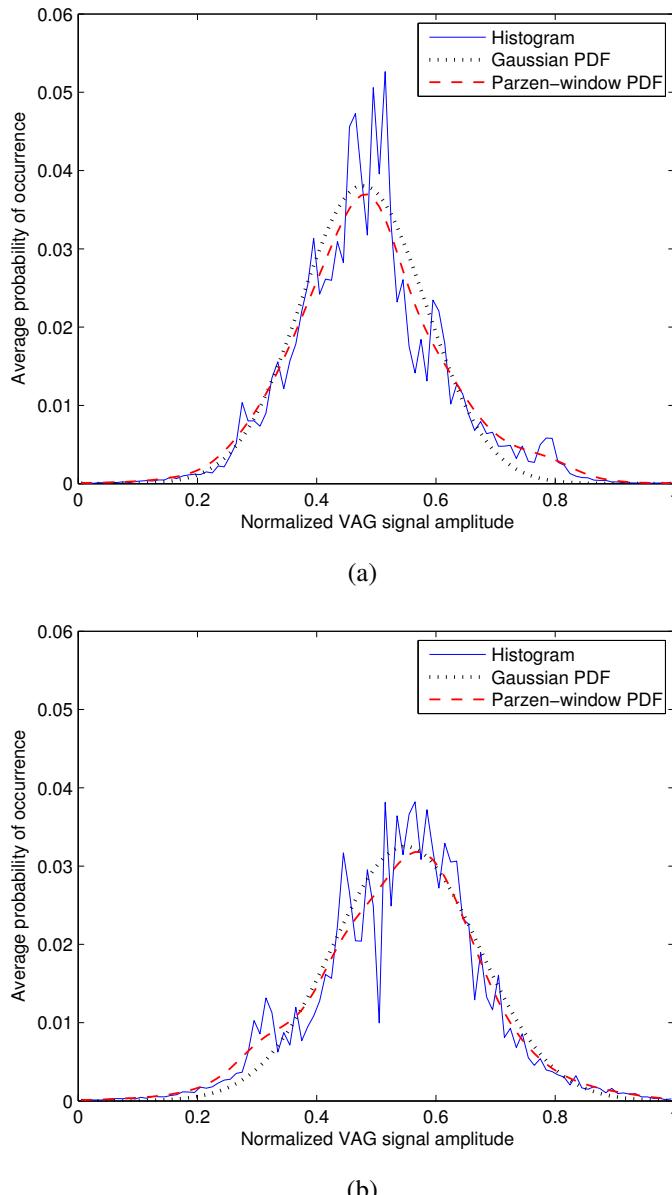


Figure 5.24 Nonparametric Parzen-window estimates of the PDFs of VAG signals derived from (a) VAG signals of 51 normal volunteers and (b) VAG signals of 38 subjects with knee-joint pathology. The amplitude has been normalized to the range [0, 1]. The figure also shows the normalized histogram and the Gaussian fit for each case. The parameters of the Gaussian fit are (a) mean = 0.4778, $SD = 0.1047$; and (b) mean = 0.5495, $SD = 0.1227$. Reproduced from R.M. Rangayyan and Y.F. Wu, Screening of knee-joint vibroarthrographic signals using probability density functions estimated with Parzen windows, *Biomedical Signal Processing & Control*, January 2010, 5(1):53–58, with permission from Elsevier. ©Elsevier.

where μ is the mean value, given by

$$\mu = \sum_{l=0}^{L-1} x_l p_x(x_l). \quad (5.32)$$

See Section 3.2.1 for the definition of several related statistical parameters.

The Kullback–Leibler distance or divergence (KLD) between two PDFs $p_1(x_l)$ and $p_2(x_l)$ is defined as [59]

$$KLD(p_1, p_2) = \sum_{l=0}^{L-1} p_2(x_l) \ln \left[\frac{p_2(x_l)}{p_1(x_l)} \right]. \quad (5.33)$$

Rangayyan and Wu [57] computed the KLD between the PDF estimated for the signal to be classified and the PDF models for the normal and abnormal VAG signals. In obtaining the PDF models for the normal and abnormal signals with the limited dataset available, the leave-one-out (LOO) method was used: the signal to be classified was left out of the procedure to obtain the Parzen-window model for the corresponding class. In this manner, the signal being tested does not contribute to the training process.

Using only the KLD feature obtained with the Parzen-window models, a normal-versus-abnormal classification accuracy of 73% was obtained with the dataset of 89 VAG signals used. The sensitivity and specificity of classification (screening) were calculated to be 68.42% and 76.47%, respectively. The use of KLD with the Gaussian models resulted in poorer performance, with overall accuracy of 69.7%. Note that the Gaussian model does not facilitate the characterization of asymmetric or skewed PDFs.

Pattern classification using the set of parameters (KLD, K, H, μ, σ) led to an overall accuracy of 77.5%, sensitivity of 71.1%, and specificity of 82.4% using radial basis functions (RBF) with LOO. Other combinations of the parameters gave poorer results. Mu et al. [69] obtained better classification performance using some of the features described above but with advanced methods including a genetic algorithm for feature selection and the strict two-surface proximal classifier.

Given the nonstationary nature of VAG signals, it would be advantageous to compute some of the parameters described above using segments of VAG signals instead of their full duration [55, 56]. See Sections 6.6, 9.9, 10.4.2, 10.8.1, and 10.12 for descriptions of other methods for the analysis of VAG signals.

5.13 Application: Fractal Analysis of the EMG in Relation to Force

Problem: *Fractals exhibit patterns of various levels of complexity. Can fractal analysis be applied to characterize the variations in an EMG signal with respect to the force exerted by a muscle?*

Solution: Several biomedical systems and signals exhibit varying levels of complexity in their characteristics and patterns that vary across different states of health and disease. The fractal dimension (FD) can be used to represent such complexity in a quantitative manner. Various methods are available to derive estimates of FD from a given signal, as described in the following sections.

5.13.1 Fractals in nature

The term “fractal” was coined by Mandelbrot [70] to represent patterns that possess self-similarity at several scales or levels of magnification [50, 71–78]. Fractal geometry has received considerable attention as a mathematical tool to model randomness in nature. Self-similarity refers to the property of an object that has a substructure that resembles its superstructure: Under several levels of magnification, a fractal appears to have the same form. A common example of a self-similar fractal in nature is a fern that has the same pattern repeated at multiple scales in its leaves; broccoli and

cauliflower also exhibit fractal patterns. On the other hand, a Euclidean form, such as a square or a triangle, does not demonstrate its original form when magnified. The Cantor bar, the Koch curve, the Sierpinski triangle, and the Hilbert curve are well-known fractals [71, 77].

A fractal is typically formed using a recursive procedure, whereas a Euclidean shape is defined by an algebraic formula. Fractals possess irregular and complex forms that cannot be described adequately using simple Euclidean geometry or shapes. It is readily understood that the dimension of a straight line is unity, that of a square is two, and that of a cube is three. However, a VMG or VAG signal that is a 1D function of time, but occupies the 2D space in a more complicated manner than a straight line, may be considered to possess a fractional dimension that is between unity and two. Fractal analysis could be used to model and analyze apparently complex or complicated patterns. However, natural objects may not possess self-similarity at all scales. A real-life pattern does not possess infinite levels of detail or granularity. Furthermore, a natural pattern, when magnified, may appear similar but not exactly identical to its original unmagnified form. This characteristic is referred to as statistical self-similarity. A natural fractal-like object may be viewed as an approximate fractal with statistical self-similarity over a finite range of scales. Fractal analysis is a part of a relatively new set of advanced techniques for nonlinear analysis of biomedical signals [79–81].

5.13.2 Fractal dimension

Consider a self-similar pattern that exhibits a number of self-similar parts represented by the variable a , at the reduction factor of $1/s$. The reduction factor is related to the measurement scale. The self-similarity dimension (D) is defined in a power-law relationship as [71]

$$a = \frac{1}{s^D}. \quad (5.34)$$

Applying the log and solving for D , we get

$$D = \frac{\log(a)}{\log(1/s)}. \quad (5.35)$$

The slope of a plot of the log of the number of self-similar parts, $\log(a)$, versus the log of the reduction factor, $\log(1/s)$, can provide an estimate of D . A straight-line approximation could be fitted to estimate the slope. Two popular methods to estimate FD are the ruler method and the box-counting method [71], which are described in the following paragraphs.

The ruler method: Using rulers of different length, the total length of a pattern or signal can be estimated to different levels of accuracy. If a large ruler is used, the small details in the given pattern are ignored. As smaller and smaller rulers are used, finer details get measured. The measured or estimated length increases and improves in accuracy as the size of the ruler decreases. An estimate of FD is obtained from the slope of a straight-line fit to a plot of the log of the measured length versus the log of the measuring unit or ruler size.

Let u be the length measured with a ruler of size s . The precision of measurement is represented by $1/s$. A fractal is expected to satisfy the power law [71]

$$u = c \frac{1}{s^d}, \quad (5.36)$$

where c is a constant of proportionality, and $FD = 1 + d$. Applying the log, we have

$$\log(u) = \log(c) + d \log(1/s). \quad (5.37)$$

The slope of a plot of $\log(u)$ versus $\log(1/s)$ provides an estimate of FD as $FD = 1 + d$.

If we let $u = ns$, where n is the number of times the ruler is used to measure the length u with the ruler of size s , then

$$\log(n) = \log(c) + (1+d) \log(1/s). \quad (5.38)$$

The slope of a plot of $\log(n)$ versus $\log(1/s)$ directly provides an estimate of FD .

The box-counting method: The box-counting method [71, 77, 82–85] involves partitioning the pattern or signal space into square boxes of equal size, and counting the number of boxes that contain a part of the signal. The process is repeated by partitioning the signal space into smaller and smaller squares. The log of the number of boxes counted is plotted against the log of the magnification index for each stage of partitioning. The slope of the straight line fitted to the plot gives an estimate of the FD .

Higuchi's method: Higuchi [86] proposed an algorithm to estimate the FD of a signal by computing measures of the length of the signal using multiple versions of the signal reconstructed for varying measurement intervals. Given a signal $x(n)$, $n = 1, 2, \dots, N$, the method creates new signals as [86, 87]

$$x_k(m) = x(m), x(m+k), x(m+2k), \dots, x\left(m + \lfloor \frac{N-m}{k} \rfloor k\right), \quad (5.39)$$

for various values of k , where k and m are integers with $m = 1, 2, \dots, k$. Here, m represents the initial point, and k is the interval between the points of $x_k(m)$. The length of each derived signal is computed as

$$L(m, k) = \frac{1}{k} \frac{N-1}{k \lfloor \frac{N-m}{k} \rfloor} \sum_{i=1}^{\lfloor \frac{N-m}{k} \rfloor} |x(m+ik) - x[m+(i-1)k]| \quad (5.40)$$

and

$$L(k) = \frac{1}{k} \sum_{m=1}^k L(m, k). \quad (5.41)$$

The slope of a straight-line fit to a log–log plot of $L(k)$ against $1/k$ gives the FD of the original signal.

When applying the ruler or box-counting methods to a signal or pattern, it is important to select an appropriate range of the size of the ruler or the box [71, 77, 84]. The range of size needs to be related to the range of the values of the signal for both the independent variable (such as time) and the dependent variable (such as voltage). It could be convenient to normalize both ranges of a given signal to $[0, 1]$ and to define the ruler or box size in relation to the normalized variables.

For discussions on other methods of fractal analysis, see Section 6.6, Cabral and Rangayyan [77], and Banik et al. [88].

5.13.3 Fractal analysis of physiological signals

Fractal properties have been observed in several physiological structures and processes. Many anatomical structures have fractal-like appearance, such as the coronary arteries, venous branching patterns, bronchial trees, certain muscle fiber bundles, and the His–Purkinje network in the ventricles [50, 89]. The branching pattern of the His–Purkinje network of conduction pathways provides an efficient way to distribute the depolarization stimulus to the ventricles. The electogenesis of the QRS complex in the ECG has been modeled using a fractal-like conduction system. The normal QRS complex has been shown to follow an inverse-power-law distribution of frequency content in the log–log scale [90]; this property has been noted as being consistent with depolarization of the myocardium by a self-similar branching network [89]. Studies of such branching networks used to depolarize a network of cells via computer modeling have shown that, after 10 generations of

branching, it is possible to simulate realistic QRS complexes [91]. The frequency content of QRS complexes has been shown to be affected by changes in the geometry of the branching network.

Goldberger et al. [90] and Goldberger and West [92] showed that the rhythm of a healthy heart is not highly regular, but that it is a temporal fractal with a high degree of variability in the heart rate. The PSD of a time-series representation of the heart rate follows the inverse power law (see Sections 6.6, 7.9, and 8.12 for related discussions). It has been shown that, in the case of time series of heart rate, the loss of physiological complexity can lead to greater regularity. The phenomena associated with fractals, self-similar scaling, $1/f$ noise, and inverse-power-law distributions offer interesting models as well as useful methods to characterize physiological processes and biomedical signals [92].

Li et al. [93] applied a method of fractal and wavelet-based spectral analysis to analyze EEG signals of rats. It was hypothesized that variations in FD -related parameters, such as the Hurst and spectral exponents, can be used to describe the dynamic characteristics of the brain in different states, in particular before seizures. They found that the method revealed characteristic signs of an approaching seizure, including the emergence of long-range correlation and decrease in the FD value. Liang et al. [94] applied the Hurst exponent, extracted from EEG recordings, as a measure of the effects of anesthesia on brain activity. The maximal overlap discrete wavelet transform (DWT) was used to suppress the effects of artifacts in the EEG, and the scaling properties of the data in designated frequency bands were calculated prior to estimation of the Hurst exponent. It was observed that the Hurst exponent decreased (especially in low-frequency bands) when anesthesia deepened, and that it is a useful measure for estimation of the depth of anesthesia.

Shah et al. [95] studied acceleration signals obtained from finger joints of patients with calcium pyrophosphate deposition disease, rheumatoid arthritis, or spondyloarthropathy of the finger joint. The ruler method was used to estimate FD . The results showed that there were significant differences between the FD values of acceleration signals from patients in the three categories. Specifically, the FD values of acceleration signals from the patients with calcium pyrophosphate deposition disease (1.709 ± 0.097) were higher than those of patients with rheumatoid arthritis (1.6 ± 0.069) or spondyloarthropathy (1.569 ± 0.081).

There is increasing interest in nonlinear dynamical analysis of biomedical signals related to fractals, chaos, and nonlinear modeling [77, 79–81]; see Stam [96] for a detailed review of nonlinear dynamical analysis of the EEG and other signals. The recently developed methods of nonlinear dynamical analysis are expected to make it possible to study self-organization, pattern formation, and attractors of trajectories in the state space of nonlinear and complex systems that may not be captured by traditional linear methods.

For discussions on the application of fractal analysis to mammographic images, see Rangayyan and Nguyen [97], Cabral and Rangayyan [77], and Banik et al. [88].

5.13.4 Fractal analysis of EMG signals

We have seen several examples of the increasing complexity of EMG signals with increasing force and studied many methods to capture such complexity in quantitative terms; see, for example, Figure 1.22 and Sections 5.9, 5.10, and 5.11. Figure 5.25 shows the variation of FD , obtained using Higuchi's method, for the EMG signal shown in Figure 1.22. The EMG signal was first filtered with a Butterworth lowpass filter of 6th order and cutoff frequency of 300 Hz. Segments of duration 1 s were cut for each level of contraction to estimate FD . It is evident that FD increases with the level of contraction (except for the last level of contraction) with high correlation.

Anmuth et al. [98] applied fractal analysis to surface EMG signals obtained from the FDI at various levels of isometric contraction. A correlation coefficient of 0.991 was obtained between pooled FD and %MVC data for 10 subjects.

Gitter and Czerniecki [99] studied the use of FD , estimated using the box-counting method, in the analysis of EMG signals. They obtained needle EMG signals at various levels of contraction of

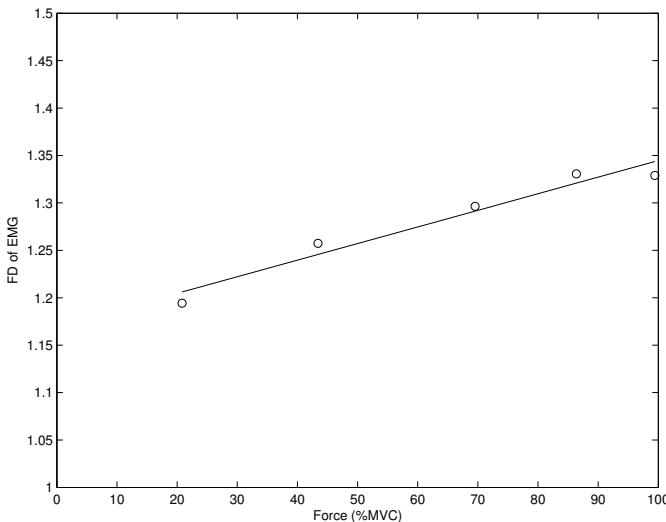


Figure 5.25 Variation of FD with the level of muscle contraction for the EMG signal shown in Figure 1.22; $r^2 = 0.95$. Figure courtesy of Faraz Oloumi.

the biceps brachii muscle. FD values, averaged over three repetitions of each level of contraction, were obtained in the range [1.1, 1.4] and exhibited positive correlation with force. However, while the correlation was high with $r = 0.94$ in the range [10, 70] %MVC, it was low with $r = 0.35$ in the range [70, 90] %MVC.

Gupta et al. [100] studied the variation of the FD of surface EMG signals of the biceps brachii during flexion and extension of the arm at various speeds and with different loads placed on the palm. Using linear regression, they found FD to be correlated well to both the load ($r = 0.99$) and the rate of flexion-extension ($r = 0.98$).

Muscle fatigue is known to cause reductions in the mean and median frequency of the related EMG signals. Ravier et al. [101] applied the $1/f$ model (see Section 6.6.1) to analyze EMG signals related to muscle force and fatigue. They found that the right slope of the EMG spectrum (beyond the peak frequency in the range of 60 – 90 Hz) decreased with increasing levels of force, but there was no substantial effect of fatigue on the slope. However, Talebinejad et al. [102] applied multi-fractal detrended fluctuation analysis methods to obtain optimal Hurst exponents, which showed a high degree of correlation with the progress of fatigue; see also Talebinejad et al. [103, 104].

5.14 Remarks

We have now reached the stage in our study where we can derive parameters from segments of biomedical signals. In the present chapter, we focused our attention on characteristics that could be observed or derived in the time domain. The parameters considered were designed with the aim of discriminating between different types of waveshapes, or of representing change in waveform complexity through the course of a physiological or pathological process. We have seen how the various parameters explored in the present chapter can help in distinguishing between normal and ectopic ECG beats, and how certain measures can serve as quantities that are correlated with levels of physiological activity, such as respiration and muscular contraction.

It should be borne in mind that, in most practical applications, a single parameter or a couple of measures may not adequately serve the purposes of signal analysis or diagnostic decision-making. A single parameter such as FF or SL may assist in distinguishing some types of PVCs from normal

ECG beats; however, several cardiovascular diseases and defects may cause changes in the ECG signal that may lead to similar variations in the *FF* or *SL* values. A practical application would need to maintain a broad scope of analysis and use several parameters to detect various possible abnormalities. As always, an investigator should consider the possibility that a parameter observed to be useful in, for example, ECG analysis in the time domain, may serve the needs in the analysis of some other signal, such as the PCG or EMG, in a different domain.

5.15 Study Questions and Problems

1. Prove that the value of *FF* of a sinusoidal wave is equal to unity.
2. The following discrete-time signals are defined over the interval 0 to 10 s with the sampling frequency being 1 Hz: (a) $x_1(n) = u(n) - u(n - 5)$. (b) $x_2(n) = 2u(n - 3) - 2u(n - 8)$. (c) $x_3(n) = u(n - 2) - u(n - 9)$. (d) $x_4(n) = u(n - 2) - u(n - 10)$. $u(n)$ is the discrete-time unit step function. The *SL* of a signal $x(n)$ is defined as

$$SL = \frac{\sum_{n=0}^{N-1} w(n) x^2(n)}{\sum_{n=0}^{N-1} x^2(n)},$$

where $w(n)$ is a nondecreasing weighting function, and N is the number of samples in the signal. Let $w(n) = n$, $n = 0, 1, 2, \dots, N - 1$. Draw sketches of each signal with the weighting function $w(n)$ superimposed. Compute the *SL* values for the four signals given. Interpret your results and compare the characteristics of the four signals in terms of their *SL* values.

3. A needle EMG signal under low levels of muscle contraction was observed to contain a mixture of three trains of MUAPs. One of the trains contains quasiperiodic occurrences of a monophasic MUAP, the second contains occurrences of a biphasic MUAP, and the third contains occurrences of a triphasic MUAP. It was also observed that the MUAPs do not overlap in the EMG signal. Propose a signal analysis procedure to: (a) detect the occurrence (location in time) of each MUAP of each type individually and (b) determine the firing rate of each motor unit. Note that each MUAP needs to be detected and labeled as being one of monophasic, biphasic, or triphasic type.

Your solution should include: (i) plots of the EMG signal (make up one according to the description above) with labels for the components; (ii) plots of the signal at various stages of your procedure; (iii) equations for important steps of your signal analysis procedure; and (iv) statements describing the reason or logic behind each step you propose.

4. A researcher is attempting to develop a digital signal processing system for the acquisition and analysis of heart sound signals (PCG signals). Assist the researcher in addressing the following concerns and problems: (a) What are the typical bandwidths of normal PCG signals and those with murmurs? What is the recommended sampling frequency? (b) What are the sources of artifacts that one has to consider in recording PCG signals? Name one physiological source and one other source, and recommend techniques to limit or eliminate both. (c) How can one identify the locations of S1 and S2? Which other biomedical signals would you recommend for assistance in this problem? Draw schematic diagrams of the signals and identify the corresponding cardiac events and timing relationships. (d) Propose a technique to obtain the envelope of the PCG signal. List all steps of the method you propose and provide the required parameters. (e) Draw schematic PCG signals and their envelopes over one cardiac cycle for a normal case, a case with systolic murmur, and a case with diastolic murmur. Identify each event in each case.
5. You are given a database of SMUAPs containing several types of normal and abnormal patterns. Each signal record has one SMUAP. The patterns and features of interest are: (i) monophasic SMUAPs, (ii) biphasic SMUAPs, (iii) triphasic SMUAPs, (iv) polyphasic SMUAPs with more than three phases.
 (a) Propose two parameters (features) to help in separating the four classes of SMUAPs. Give the required equations or procedures and explain their relationship to the signal characteristics described above. Describe conditions or preprocessing steps that are required in order for your methods to work well. (b) Draw a schematic plot of the feature-vector space and demarcate regions where you expect features of the four SMUAP types to lie. (c) State decision rules to classify the four SMUAP types using the two measures you propose.

6. Why is the ST segment of the ECG relevant in diagnosis? Recommend signal analysis techniques for the analysis of ST segment variations in clinical applications.
7. A researcher new to the field of biomedical signal analysis is assigned a project on the analysis of heart sounds and murmurs. The researcher is provided with a database of PCG signals of four types: (a) normal, (b) systolic murmur, (c) diastolic murmur, and (d) systolic and diastolic murmur.

The researcher surfs the internet and obtains two programs to compute the ZCR using a moving window of 20 ms and a filtered envelope of a given signal. The sampling rate used is 1,000 Hz. Although the programs generated graphs of the given signal, the ZCR , and the smoothed envelope, the researcher encountered difficulties in interpreting the results.

Assist the researcher by providing the following information: (i) For each type of PCG signal listed above, draw plots of a typical PCG signal, its ZCR as a function of time, and the averaged envelope. Explain the important expected characteristics of the results. (ii) Estimate the ranges of ZCR values (using a moving window of 20 ms) for typical normal heart sounds and murmurs. Explain your procedure and assumptions.

8. A researcher has obtained a system to record the EMG signal from a muscle that is directly involved in breathing (respiration). Help the researcher in understanding the nature of the signal and developing a signal processing system to analyze the signal as follows: (a) Draw a schematic representation of the EMG over two cycles of respiration, showing the ranges of inspiration and expiration. (b) Propose an algorithm to obtain an envelope of the EMG signal. Describe the nature and purpose of each step in your algorithm. Give an equation for each step, and plot the result of the operation on the EMG signal. (c) Propose a method to obtain a measure of activity (to indicate how "busy" the signal is) as a function of time, in order to characterize the variation in the signal with breathing. Give an equation and describe the nature of the method. Plot the result of the operation on the EMG signal. (d) The peak values in the envelope and activity functions over a given cycle of respiration are expected to be correlated with the peak air flow to the lungs. Propose algorithms to detect the peak envelope and peak activity values. (e) For each procedure that you propose in items (b), (c), and (d) above, discuss potential artifacts that could mislead your procedure and indicate how you would prevent them.
9. A researcher is developing methods to extract features from ECG signals to distinguish between normal beats and PVCs. Draw typical waveforms of the two types of beats and explain the major differences between them in qualitative terms. Propose two quantitative measures or features to characterize the differences between normal beats and PVCs. Give an equation or a step-by-step algorithm to compute each feature.
10. A student new to the field of biomedical engineering approaches you seeking assistance to develop methods to analyze the relationship between the force produced by contracting a muscle and the corresponding EMG signal. Help the student by providing detailed responses to the following questions and requests:
 - (a) How can one acquire EMG signals corresponding to different levels of force? (i) How many electrodes are required? (ii) Where and how should the electrodes be placed? (iii) How can one get another signal or variable that is directly proportional to the force? (iv) What should be the settings for the bandwidth of the filters for the EMG signal? (v) What should be the settings for the bandwidth of the filters for the force signal? (vi) What should be the sampling frequency? (vii) Give a step-by-step experimental procedure to acquire the required signals. (viii) Draw a schematic diagram to illustrate the expected nature of the EMG and force signals.
 - (b) (i) What are the precautions to be taken in recording the signals? (ii) List one potential source each for random noise, structured noise, and physiological artifact that could corrupt the EMG signal. (iii) For each source of noise or artifact that you identify, recommend a procedure to prevent the same. (iv) For each source of noise or artifact that you identify, recommend a postacquisition procedure to remove the same.
 - (c) (i) What kind of measures or parameters may one derive from an EMG signal so that the measures vary in proportion to muscular force? (ii) Give equations to define three parameters that are suitable for this purpose. For each parameter, give a step-by-step procedure to compute the parameter from the EMG signal. (iii) Explain the expected relationships between the three parameters that you recommend and muscular force. Draw a plot with the force on the abscissa (x -axis) and the proposed parameters on the ordinate (y -axis).

11. Given a sampled signal, $x(n)$, $n = 0, 1, 2, \dots, N - 1$, write an equation to compute the *variance*, σ_x^2 , of the signal. Give the definitions of the parameters *mobility* and *form factor*. Explain how the three parameters mentioned above may be computed and used to analyze the relationship between the force generated by muscular contraction and the related EMG signal.
12. Explain the concept of an envelope of a signal. Give a step-by-step algorithm to compute an envelope of a signal. Include at least two equations for parts of the procedure. Draw a normal PCG signal over one cardiac signal and a corresponding envelope. Explain the relationships between the two and suggest a use of the envelope in a practical application.

5.16 Laboratory Exercises and Projects

Note: Data files related to the exercises are available at the site

<https://github.com/srikrishnan1972/Biomedical-Signal-Analysis>

1. The signal in the file emg_dog2.dat was recorded from the crural diaphragm of a dog using fine-wire electrodes sewn in-line with the muscle fibers and placed 10 mm apart. The signal represents two cycles of breathing, and has been sampled at 10 kHz. (See also the file emg_dog2.m.)

Write a MATLAB® program to perform full-wave rectification (absolute value) or half-wave rectification (threshold at zero, with the mean value of the signal being zero). Apply a lowpass Butterworth filter of order four and cutoff frequency in the range 5 to 20 Hz to the result. Analyze and evaluate the results with the two methods of rectification and at least two different lowpass cutoff frequencies. Compare the results with the envelope provided in the file emg_dog2_env.dat.

2. The *RMS* value of a signal within a specific duration is related to the average power level of the signal. Write a MATLAB® program to compute the *RMS* value at each instant for the EMG signal in the file emg_dog2.dat by using a causal short-time analysis window of duration in the range 50 – 150 ms. Use at least two different window durations and analyze the results. (See also the file emg_dog2.m.)

3. Develop a program to compute the turns count in causal moving windows of duration in the range 50 – 150 ms. Apply the method to the EMG signal in the file emg_dog2.dat. (See also the file emg_dog2.m.) Study the results for different thresholds in the range 0 – 200 μV . Compare the envelope, *RMS*, and turns count curves in terms of their usefulness as representatives of inspiratory airflow (data provided in the file emg_dog2_flo.dat).

4. The file safety.wav contains the speech signal for the word “safety” uttered by a male speaker, sampled at 8 kHz. (See also the file safety.m.) The signal has a significant amount of background noise (as it was recorded in a computer laboratory). Develop procedures to derive short-time *RMS*, turns count, and ZCR in moving windows of duration in the range 10 – 100 ms. Study the variations in the parameters in relation to the voiced, unvoiced, and silence (background noise) portions of the signal.

What do you expect the results to be if the procedures are applied to the first derivative of the signal? Confirm your assertions or expectations by performing the study.

5. Develop a program to derive the envelopogram. Apply the procedure to the PCG signals in the files pec1.dat, pec33.dat, and pec52.dat. (See the file plotpec.m.) For each of the same signals, derive the *RMS* values in moving windows of duration 10, 20, and 30 ms, and plot the results as approximate envelopes of the signals.

Extend the procedure to average the envelopograms and envelopes over several cardiac cycles using the ECG as the trigger. How will you handle the variations in the duration (number of samples) of the signals from one beat to another?

Compare and study the various results for the purpose of detection of S1 and S2 in PCG signals.

6. The ECG signal in the file ecgpvc.dat contains a large number of PVCs, including episodes of bigeminy. (See the file ecgpvc.m.) Apply the Pan–Tompkins procedure to detect and segment each beat. Label each beat as normal or PVC by visual inspection. Record the number of beats missed, if any, by your detection procedure.

Compute the *RR* interval and the form factor *FF* for each beat. Use a duration of 80 samples (400 ms) spanning the QRS-T portion of each beat to compute *FF*. The P wave need not be considered in the present exercise.

Compute the mean and *SD* of the *FF* and *RR* values for the normal beats and the PVCs. Evaluate the variation of the two parameters within and between the two categories of beats.

7. Obtain EMG signals related to various levels of muscular activity or force. You may use the data files EMGforce.txt and EMGforce2.txt as well as the program EMGforce.m provided; the sampling rate is 2,000 Hz per channel. Normalize the force signal such that the minimum value is zero and the maximum value (corresponding to MVC) is 100. Filter the force signal to remove noise and artifacts. Plot the EMG signal (in mV) and normalized force (in %MVC) against the time axis.

Develop a program for automatic identification of portions (segments) corresponding to each level of contraction within which the force remains close to the corresponding peak values. For each segment of the EMG signal identified as above, compute four suitable parameters. Ensure that your list of EMG features includes the following: *RMS* value, form factor, and turns count. Compute also the average force (in %MVC) for each segment.

Plot the values of the various parameters versus force in %MVC. Label the axes with the appropriate units. Analyze the results in terms of statistical variation of the parameters in relation to force. Using the *polyfit* function in MATLAB®, obtain a straight-line (linear) fit to represent the variation of each EMG parameter versus force. Use *polyval* to evaluate the values of the dependent variable given by the model for the available values of the independent variable. Superimpose the linear models (straight-line fits) obtained on the plots of the parameters in the preceding step. Analyze the results.

Compute the correlation coefficient, *r*, with *r*² given by the formula

$$r^2 = \frac{\left[\sum_{n=1}^{N} x(n)y(n) - N\bar{x}\bar{y} \right]^2}{\left[\sum_{n=1}^{N} x^2(n) - N\bar{x}^2 \right] \left[\sum_{n=1}^{N} y^2(n) - N\bar{y}^2 \right]},$$

where *N* is the number of samples of *x* or *y* representing each of the EMG parameters or force, and \bar{x} is the mean of *x*. Using *r*, analyze the goodness of fit for each parameter and discuss the appropriateness of the linear model. Tabulate the parameters of the linear model and *r* for each of the EMG parameters. Analyze the results and describe your findings. Perform all of the above steps with the two EMG signals mentioned above. Analyze and compare the results.

References

- [1] Webster JG, editor. *Medical Instrumentation: Application and Design*. Wiley, New York, NY, 3rd edition, 1998.
- [2] Wallace AG. Electrophysiology of the myocardium. In *Clinical Cardiopulmonary Physiology*. Grune & Stratton, New York, NY, 3rd edition, 1969.
- [3] Rushmer RF. *Cardiovascular Dynamics*. WB Saunders, Philadelphia, PA, 4th edition, 1976.
- [4] Jenkins JM, Wu D, and Arzbaecher RC. Computer diagnosis of abnormal cardiac rhythms employing a new P-wave detector for interval measurement. *Computers and Biomedical Research*, 11:17–33, 1978.
- [5] Oppenheim AV and Schafer RW. *Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1975.
- [6] Berkhou AJ. On the minimum phase criterion of sampled signals. *IEEE Transactions on Geoscience Electronics*, 11:186–198, 1973.
- [7] Berkhou AJ. On the minimum-length property of one-sided signals. *Geophysics*, 38:657–672, 1978.
- [8] Amazeen RL, Moruzzi RL, and Feldman CL. Phase detection of R-waves in noisy electrocardiograms. *IEEE Transactions on Biomedical Engineering*, 19(1):63–66, 1972.
- [9] Ulrych TJ and Lasserre M. Minimum-phase. *Canadian Journal of Exploration Geophysicists*, 2:22–32, 1966.

- [10] Treitel S and Robinson EA. The stability of digital filters. *IEEE Transactions on Geoscience Electronics*, 2:6–18, 1964.
- [11] Murthy ISN and Rangaraj MR. New concepts for PVC detection. *IEEE Transactions on Biomedical Engineering*, 26(7):409–416, 1979.
- [12] Childers DG, Skinner DP, and Kemerait RC. The cepstrum: A guide to processing. *Proceedings of the IEEE*, 65(10):1428–1443, 1977.
- [13] Oppenheim AV, Kopec GE, and Tribble JM. Signal analysis by homomorphic prediction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):327–332, 1976.
- [14] Cox Jr. JR, Nolle FM, and Arthur RM. Digital analysis of the electroencephalogram, the blood pressure wave, and the electrocardiogram. *Proceedings of the IEEE*, 60(10):1137–1164, 1972.
- [15] Nolle F. *Argus, A Clinical Computer System for Monitoring Electrocardiographic Rhythms*. PhD thesis, Washington University School of Medicine, Saint Louis, MO, December 1972.
- [16] Ye C, Vijaya Kumar BVK, and Coimbra MT. Heartbeat classification using morphological and dynamic features of ECG signals. *IEEE Transactions on Biomedical Engineering*, 59(10):2930–2941, 2012.
- [17] Ince T, Kiranyaz S, and Gabbouj M. A generic and robust system for automated patient-specific classification of ECG signals. *IEEE Transactions on Biomedical Engineering*, 56(5):1415–1426, 2009.
- [18] Lehner RJ and Rangayyan RM. A three-channel microcomputer system for segmentation and characterization of the phonocardiogram. *IEEE Transactions on Biomedical Engineering*, 34:485–489, 1987.
- [19] Platt RS, Hajduk EA, Hulliger M, and Easton PA. A modified Bessel filter for amplitude demodulation of respiratory electromyograms. *Journal of Applied Physiology*, 84(1):378–388, 1998.
- [20] Lathi BP. *Signal Processing and Linear Systems*. Berkeley-Cambridge, Carmichael, CA, 1998.
- [21] Oppenheim AV, Willsky AS, and Nawab SH. *Signals and Systems*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1997.
- [22] Lathi BP. *Linear Systems and Signals*. Oxford University Press, New York, NY, 2nd edition, 2005.
- [23] Shin SJ, Tapp WN, Reisman SS, and Natelson BH. Assessment of autonomic regulation of heart rate variability by the method of complex demodulation. *IEEE Transactions on Biomedical Engineering*, 36(2):274–283, 1989.
- [24] Hayano J, Taylor JA, Yamada A, Mukai S, Hori R, Asakawa T, Yokoyama K, Watanabe Y, Takata K, and Fujinami T. Continuous assessment of hemodynamic control by complex demodulation of cardiovascular variability. *American Journal of Physiology*, 264:H1229–H1238, 1993.
- [25] Bloomfield P. *Fourier Analysis of Time Series: An Introduction*. Wiley, New York, NY, 1976.
- [26] Karpman L, Cage J, Hill C, Forbes AD, Karpman V, and Cohn K. Sound envelope averaging and the differential diagnosis of systolic murmurs. *American Heart Journal*, 90(5):600–606, 1975.
- [27] Sarkady AA, Clark RR, and Williams R. Computer analysis techniques for phonocardiogram diagnosis. *Computers and Biomedical Research*, 9:349–363, 1976.
- [28] Bendat JS and Piersol AG. *Random Data: Analysis and Measurement Procedures*. Wiley, New York, NY, 2nd edition, 1986.
- [29] Baranek HL, Lee HC, Cloutier G, and Durand LG. Automatic detection of sounds and murmurs in patients with Ionescu-Shiley aortic bioprostheses. *Medical and Biological Engineering and Computing*, 27:449–455, 1989.
- [30] Guerrero G, Kortelainen JM, Palacios E, Bianchi AM, Tachino G, Tenhunen M, Méndez MO, and van Gils M. Detection of sleep-disordered breathing with pressure bed sensor. In *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1342–1345, Osaka, Japan, July 2013.
- [31] Rabiner LR and Schafer RW. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, 1978.
- [32] Gerbarg DS, Holcomb Jr. FW, Hofler JJ, Bading CE, Schultz GL, and Sears RE. Analysis of phonocardiogram by a digital computer. *Circulation Research*, 11:569–576, 1962.

- [33] Gerborg DS, Taranta A, Spagnuolo M, and Hofler JJ. Computer analysis of phonocardiograms. *Progress in Cardiovascular Diseases*, 5(4):393–405, 1963.
- [34] Saltzberg B and Burch NR. Period analytic estimates of moments of the power spectrum: A simplified EEG time domain procedure. *Electroencephalography and Clinical Neurophysiology*, 30:568–570, 1971.
- [35] Jacobs JE, Horikoshi K, and Petrovick MA. Feasibility of automated analysis of phonocardiograms. *Journal of the Audio Engineering Society*, 17(1):49–54, 1969.
- [36] Yokoi M, Uozumi Z, Okamoto N, Mizuno Y, Iwatsuka T, Takahashi H, Watanabe Y, and Yasui S. Clinical evaluation on 5 years' experience of automated phonocardiographic analysis. *Japanese Heart Journal*, 18(4):482–490, 1977.
- [37] Willison RG. Analysis of electrical activity in health and dystrophic muscle in man. *Journal of Neurology, Neurosurgery, and Psychiatry*, 27:386–394, 1964.
- [38] Goodgold J and Eberstein A. *Electrodiagnosis of Neuromuscular Diseases*. Williams and Wilkins, Baltimore, MD, 3rd edition, 1983.
- [39] Fuglsang-Frederiksen A and Månssohn A. Analysis of electrical activity of normal muscle in man at different degrees of voluntary effort. *Journal of Neurology, Neurosurgery, and Psychiatry*, 38:683–694, 1975.
- [40] Dowling MH, Fitch P, and Willison RG. A special purpose digital computer (BIOMAC 500) used in the analysis of the human electromyogram. *Electroencephalography and Clinical Neurophysiology*, 25:570–573, 1968.
- [41] Hjorth B. EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology*, 29:306–310, 1970.
- [42] Hjorth B. The physical significance of time domain descriptors in EEG analysis. *Electroencephalography and Clinical Neurophysiology*, 34:321–325, 1973.
- [43] Hjorth B. Time domain descriptors and their relation to a particular model for generation of EEG activity. In Dolce G and Kunkel H, editors, *CEAN: Computerised EEG Analysis*, pages 3–8. Gustav Fischer, Stuttgart, Germany, 1975.
- [44] Cooper R, Osselton JW, and Shaw JC. *EEG Technology*. Butterworths, London, UK, 3rd edition, 1980.
- [45] Binnie CD, Batchelor BG, Bowring PA, Darby CE, Herbert L, Lloyd DSL, Smith DM, Smith GF, and Smith M. Computer-assisted interpretation of clinical EEGs. *Electroencephalography and Clinical Neurophysiology*, 44:575–585, 1978.
- [46] Binnie CD, Batchelor BG, Gainsborough AJ, Lloyd DSL, Smith DM, and Smith GF. Visual and computer-assisted assessment of the EEG in epilepsy of late onset. *Electroencephalography and Clinical Neurophysiology*, 47:102–107, 1979.
- [47] Hornero R, Espino P, Alonso A, and López M. Estimating complexity from EEG background activity of epileptic patients. *IEEE Engineering in Medicine and Biology Magazine*, 18(6):73–79, November/December 1999.
- [48] Celka P, Mesbah M, Keir M, Boashash B, and Colditz P. Time-varying dimension analysis of EEG using adaptive principal component analysis and model selection. In *World Congress on Medical Physics and Biomedical Engineering*, pages 1404–1407. IFMBE/IEEE, Chicago, IL, 2000.
- [49] Hsia PW, Jenkins JM, Shimoni Y, Gage KP, Santinga JT, and Pitt B. An automated system for ST segment and arrhythmia analysis in exercise radionuclide ventriculography. *IEEE Transactions on Biomedical Engineering*, 33(6):585–593, 1986.
- [50] Rangayyan RM. *Biomedical Image Analysis*. CRC Press, Boca Raton, FL, 2005.
- [51] Weisstein EW. *Correlation Coefficient*. From MathWorld—A Wolfram Web Resource, <https://mathworld.wolfram.com/CorrelationCoefficient.html>.
- [52] Zhang YT, Frank CB, Rangayyan RM, and Bell GD. A comparative study of vibromyography and electromyography obtained simultaneously from active human quadriceps. *IEEE Transactions on Biomedical Engineering*, 39(10):1045–1052, 1992.

- [53] Zhang YT, Frank CB, Rangayyan RM, and Bell GD. Relationships of the vibromyogram to the surface electromyogram of the human rectus femoris muscle during voluntary isometric contraction. *Journal of Rehabilitation Research and Development*, 33(4):395–403, 1996.
- [54] Lawrence JH and de Luca CJ. Myoelectric signal versus force relationship in different human muscles. *Journal of Applied Physiology*, 54(6):1653–1659, 1983.
- [55] Rangayyan RM and Wu YF. Screening of knee-joint vibroarthrographic signals using statistical parameters and radial basis functions. *Medical and Biological Engineering and Computing*, 46(3):223–232, 2008.
- [56] Rangayyan RM and Wu YF. Analysis of vibroarthrographic signals with features related to signal variability and radial basis functions. *Annals of Biomedical Engineering*, 37(1):156–163, 2009.
- [57] Rangayyan RM and Wu YF. Screening of knee-joint vibroarthrographic signals using probability density functions estimated with Parzen windows. *Biomedical Signal Processing and Control*, 5(1):53–58, 2010.
- [58] Krishnan S, Rangayyan RM, Bell GD, and Frank CB. Adaptive time-frequency analysis of knee joint vibroarthrographic signals for noninvasive screening of articular cartilage pathology. *IEEE Transactions on Biomedical Engineering*, 47(6):773–783, June 2000.
- [59] Duda RO, Hart PE, and Stork DG. *Pattern Classification*. Wiley, New York, NY, 2nd edition, 2001.
- [60] Fukunaga K. *Introduction to Statistical Pattern Recognition*. Academic, San Diego, CA, 2nd edition, 1990.
- [61] Ladly KO, Frank CB, Bell GD, Zhang YT, and Rangayyan RM. The effect of external loads and cyclic loading on normal patellofemoral joint signals. *Special Issue on Biomedical Engineering, Defence Science Journal (India)*, 43:201–210, July 1993.
- [62] Moussavi ZMK, Rangayyan RM, Bell GD, Frank CB, Ladly KO, and Zhang YT. Screening of vibroarthrographic signals via adaptive segmentation and linear prediction modeling. *IEEE Transactions on Biomedical Engineering*, 43(1):15–23, 1996.
- [63] Umapathy K and Krishnan S. Modified local discriminant bases algorithm and its application in analysis of human knee joint vibration signals. *IEEE Transactions on Biomedical Engineering*, 53(3):517–523, March 2006.
- [64] Rangayyan RM, Krishnan S, Bell GD, Frank CB, and Ladly KO. Parametric representation and screening of knee joint vibroarthrographic signals. *IEEE Transactions on Biomedical Engineering*, 44(11):1068–1074, 1997.
- [65] Krishnan S, Rangayyan RM, Bell GD, Frank CB, and Ladly KO. Adaptive filtering, modelling, and classification of knee joint vibroarthrographic signals for non-invasive diagnosis of articular cartilage pathology. *Medical and Biological Engineering and Computing*, 35(6):677–684, 1997.
- [66] Wu YF, Krishnan S, and Rangayyan RM. Computer-aided diagnosis of knee-joint disorders via vibroarthrographic signal analysis: A review. *Critical Reviews in Biomedical Engineering*, 38(2):201–224, 2010.
- [67] Marques de Sá JP. *Applied Statistics using SPSS, STATISTICA, and MATLAB*. Springer, Berlin, Germany, 2003.
- [68] Tavathia S, Rangayyan RM, Frank CB, Bell GD, Ladly KO, and Zhang YT. Analysis of knee vibration signals using linear prediction. *IEEE Transactions on Biomedical Engineering*, 39(9):959–970, 1992.
- [69] Mu T, Nandi AK, and Rangayyan RM. Screening of knee-joint vibroarthrographic signals using the strict 2-surface proximal classifier and genetic algorithm. *Computers in Biology and Medicine*, 38(10):1103–1111, 2008.
- [70] Mandelbrot BB. *Fractal Geometry of Nature*. WH Freeman, San Francisco, CA, 1983.
- [71] Peitgen HO, Jürgens H, and Saupe D. *Chaos and Fractals: New Frontiers of Science*. Springer, New York, NY, 2004.
- [72] Liu SH. Formation and anomalous properties of fractals. *IEEE Engineering in Medicine and Biology Magazine*, 11(2):28–39, June 1992.

- [73] Deering W and West BJ. Fractal physiology. *IEEE Engineering in Medicine and Biology Magazine*, 11(2):40–46, June 1992.
- [74] Schepers HE, van Beek JHGM, and Bassingthwaite JB. Four methods to estimate the fractal dimension from self-affine signals. *IEEE Engineering in Medicine and Biology Magazine*, 11(2):57–64, June 1992.
- [75] Fortin C, Kumaresan R, Ohley W, and Hoefer S. Fractal dimension in the analysis of medical images. *IEEE Engineering in Medicine and Biology Magazine*, 11(2):65–71, June 1992.
- [76] Goldberger AL, Rigney DR, and West BJ. Chaos and fractals in human physiology. *Scientific American*, 262:42–49, February 1990.
- [77] Cabral TM and Rangayyan RM. *Fractal Analysis of Breast Masses in Mammograms*. Morgan & Claypool and Springer, 2012.
- [78] Voss RF. Fractals in nature: From characterization to simulation. In Peitgen HO and Saupe D, editors, *The Science of Fractal Images*. Springer-Verlag, New York, NY, 1988.
- [79] Akay M, editor. *Nonlinear Biomedical Signal Processing: Volume I, Fuzzy Logic, Neural Networks, and New Algorithms*. IEEE, New York, NY, 2000.
- [80] Akay M, editor. *Nonlinear Biomedical Signal Processing: Volume II, Dynamic Analysis and Modeling*. IEEE, New York, NY, 2001.
- [81] Cerutti S and Marchesi C, editors. *Advanced Methods of Biomedical Signal Processing*. IEEE and Wiley, New York, NY, 2011.
- [82] Dubuc B, Roques-Carmes C, Tricot C, and Zucker SW. The variation method: A technique to estimate the fractal dimension of surfaces. In *Proceedings of SPIE, Volume 845: Visual Communication and Image Processing II*, volume 845, pages 241–248, 1987.
- [83] Weinstein RS and Majumdar S. Fractal geometry and vertebral compression fractures. *Journal of Bone and Mineral Research*, 9(11):1797–1802, 1994.
- [84] Coelho RC, Cesar Junior RM, and Costa LF. Assessing the fractal dimension and the normalized multiscale bending energy for applications in neuromorphometry. In *Proceedings Simpósio Brasileiro de Computação Gráfica e Processamento de Imagens (SIBGRAPI-96)*, pages 353–554, Caxambu, Brazil, November 1996.
- [85] Sedivy R, Windischberger Ch., Svozil K, Moser E, and Breitenecker G. Fractal analysis: An objective method for identifying atypical nuclei in dysplastic lesions of the cervix uteri. *Gynecologic Oncology*, 75:78–83, 1999.
- [86] Higuchi T. Approach to an irregular time series on the basis of the fractal theory. *Physica D*, 31:277–283, 1988.
- [87] Gómez C, Mediavilla A, Hornero R, Abásolo D, and Fernández A. Use of the Higuchi's fractal dimension for the analysis of MEG recordings from Alzheimer's disease patients. *Medical Engineering and Physics*, 31:306–313, 2009.
- [88] Banik S, Rangayyan RM, and Desautels JEL. *Computer-aided Detection of Architectural Distortion in Prior Mammograms of Interval Cancer*. Morgan & Claypool and Springer, 2013.
- [89] Goldberger AL. Fractal mechanisms in the electrophysiology of the heart. *IEEE Engineering in Medicine and Biology Magazine*, 11:47–52, 1992.
- [90] Goldberger AL, Bhargava V, West BJ, and Mandell AJ. On a mechanism of cardiac electrical stability: The fractal hypothesis. *Biophysical Journal*, 48:525–528, 1985.
- [91] Abboud S, Berenfeld O, and Sadeh D. Simulation of high-resolution QRS complex using a ventricular model with a fractal conduction system. *Circulation Research*, 68:1751–1760, 1991.
- [92] Goldberger AL and West BJ. Fractals in physiology and medicine. *The Yale Journal of Biology and Medicine*, 60:421–435, 1987.
- [93] Li X, Polygiannakis J, Kapiris P, Peratzakis A, Eftaxias K, and Yao X. Fractal spectral analysis of pre-epileptic seizures in terms of criticality. *Journal of Neural Engineering*, 2:11–16, 2005.

- [94] Liang Z, Li D, Ouyang G, Wang Y, Voss LJ, Sleigh JW, and Li X. Multiscale rescaled range analysis of EEG recordings in sevoflurane anesthesia. *Clinical Neurophysiology*, 123:681–688, 2012.
- [95] Shah EN, Reddy NP, and Rothschild BM. Fractal analysis of acceleration signals from patients with CPPD, rheumatoid arthritis, and spondyloarthropathy of the finger joint. *Computer Methods and Programs in Biomedicine*, 77(3):233–239, 2005.
- [96] Stam CJ. Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field. *Clinical Neurophysiology*, 116:2266–2301, 2005.
- [97] Rangayyan RM and Nguyen TM. Fractal analysis of contours of breast masses in mammograms. *Journal of Digital Imaging*, 20(3):223–237, September 2007.
- [98] Anmuth CJ, Goldberg G, and Mayer NH. Fractal dimension of electromyographic signals recorded with surface electrodes during isometric contractions is linearly correlated with muscle activation. *Muscle & Nerve*, 17(8):953–954, 1994.
- [99] Gitter JA and Czerniecki MJ. Fractal analysis of the electromyographic interference pattern. *Journal of Neuroscience Methods*, 58:103–108, 1995.
- [100] Gupta V, Suryanarayanan S, and Reddy NP. Fractal analysis of surface EMG signals from the biceps. *International Journal of Medical Informatics*, 45:185–192, 1997.
- [101] Ravier P, Buttelli O, Jennane R, and Couratier P. An EMG fractal indicator having different sensitivities to changes in force and muscle fatigue during voluntary static muscle contractions. *Journal of Electromyography and Kinesiology*, 15:210–221, 2005.
- [102] Talebnejad M, Chan ADC, and Miri A. Fatigue estimation using a novel multi-fractal detrended fluctuation analysis-based approach. *Journal of Electromyography and Kinesiology*, 20:433–439, 2010.
- [103] Talebnejad M, Chan ADC, Miri A, and Dansereau RM. Fractal analysis of surface electromyography signals: A novel power spectrum-based method. *Journal of Electromyography and Kinesiology*, 19:840–850, 2009.
- [104] Talebnejad M, Chan ADC, and Miri A. Multiplicative multi-fractal modeling of electromyography signals for discerning neuropathic conditions. *Journal of Electromyography and Kinesiology*, 20:1244–1248, 2010.

CHAPTER 6

FREQUENCY-DOMAIN CHARACTERIZATION OF SIGNALS AND SYSTEMS

Many biomedical systems exhibit innate rhythms and periodicity that are more readily expressed and appreciated in terms of frequency than temporal measures. As a basic example, consider cardiac function: We express cardiac rhythm more conveniently in terms of *beats per minute* — a measure of the frequency of occurrence or the rate of repetition — than in terms of the duration of a beat or the interval between beats in seconds (the *RR* interval). A cardiac rhythm expressed as 72 *bpm* is more easily understood than a statement of the corresponding *RR* interval as 0.833 *s*. By the same token, the notion of an EEG rhythm is more easily conveyed by description in *cycles per second* in lay terms, or in *Hertz (Hz)* in technical terms. Even engineers would find a frequency-domain expression easier to appreciate than a time-domain description, such as an alpha rhythm having a frequency of 11.5 *Hz* versus the equivalent period of 0.087 *s* or 87 *ms*.

When the signal being studied is made up of discrete (that is, separate and distinct) events in time, such as the ECG, PPG, or a train of SMUAPs, the basic rhythm or rate of activity present in the signal can be assessed directly in the time domain. On the other hand, signals such as the PCG display complex or complicated patterns in the time domain that do not facilitate ready appreciation of their frequency-domain characteristics; furthermore, the time-domain waveforms may differ from one occurrence of the signal (one heart beat) to another.

The PCG provides an interesting example of a signal with multiple frequency-domain features: In addition to the beat-to-beat periodicity or rhythm, the heart sounds within a cardiac cycle exhibit *resonance*. Due to the multicompartmental nature of the cardiac system, we should expect heart sounds to possess multiple resonance frequencies; this leads to the need to describe the PCG not only in terms of a rhythm (the heart rate) or a single resonance frequency, but also a composite *spectrum* of several dominant or resonance frequencies. Furthermore, constrained flow of blood through an orifice such as a septal defect or across a stenosed valve acting as a baffle could lead to turbulence, resulting in *wideband noise*. In the case of noise-like murmurs, we would be able to

identify neither rhythms nor resonance frequencies: The need arises to consider the distribution of the signal's energy or power over a wide band of frequencies, leading to the notion of the power *spectral density* function.

We have seen in Chapter 3 that it is often more convenient and meaningful to describe filters in terms of their transfer function or frequency response — $H(s)$, $H(z)$, $H(\omega)$, or $H(f)$ — than in terms of their impulse response $h(t)$ or the time-domain input–output relationship (difference equation). Furthermore, we have seen in Section 4.4 that it is easier to interpret the PSDs of EEG waves than it is to interpret their theoretically equivalent ACFs. The Fourier and other transforms provide an invertible or reversible transformation from the time domain to the frequency domain (and vice versa). Therefore, it may be argued that no new information is created by transforming a given signal from the time domain to the frequency domain. However, the distribution of the energy or power of the signal in the frequency domain that is provided by the Fourier transform — the spectrum or PSD of the signal — facilitates better analysis and description of the frequency-domain characteristics of the signal. The PSD of a signal is not only useful in analyzing the signal, but also in designing amplifiers, filters, data-acquisition and transmission systems, and signal processing systems to treat the signal appropriately. We have seen in Section 3.9 that we require not only the signal PSD but also the noise PSD in order to be able to implement the Wiener filter.

The treatment of biomedical signals as stochastic processes provides flexibility and a sense of generality in analysis, but imposes conditions and requirements in the estimation of their statistics including the ACF and PSD. In the present chapter, we investigate methods to estimate the PSD and frequency-domain parameters of biomedical signals and systems. We also study methods to derive spectral parameters that can characterize the given signal as well as the system that generated the signal. The motivation for the study, as always, will be to distinguish between normal and abnormal signals or systems, and the potential use of the methods in diagnosis.

6.1 Problem Statement

Investigate the potential use of the Fourier spectrum and parameters derived thereof in the analysis of biomedical signals. Identify physiological and pathological processes that could modify the frequency content of the corresponding signals. Outline the signal processing tasks needed to perform spectral analysis of biomedical signals and systems.

As in the preceding chapters, the problem statement given above is generic and represents the theme of the present chapter. The various signal analysis techniques described and the examples used for illustration in the following sections will address the points raised in the problem statement, with attention to specific problems and techniques.

6.2 Illustration of the Problem with Case Studies

6.2.1 The effect of myocardial elasticity on heart sound spectra

The first and second heart sounds — S1 and S2 — are typically composed of low-frequency components; this is to be expected due to the fluid-filled and elastic nature of the cardiohemic system. Sakai et al. [1] processed recorded heart sound signals by using tunable bandpass filters (with a bandwidth of 20 Hz, tuned over the range 20 – 40 Hz to 400 – 420 Hz), and estimated the frequency distributions of S1 and S2. They found the heart sound spectra to be maximum in the 20 – 40 Hz band, that S1 had a tendency to demonstrate peaks at lower frequencies than those of S2, and that S2 exhibited a “gentle peaking” between 60 and 220 Hz.

Gerbarg et al. [2,3] developed a computer program to simulate a filter bank and obtained averaged power spectra of S1 and S2 of 1,000 adult males, 32 high-school children, and 75 patients in a hospital. The averaged PSDs of S1 and S2 obtained by them indicated peak power in the range

60 – 70 Hz, and relative power levels lower than –10 dB beyond 150 Hz. The PSD of S2 displayed slightly more high-frequency energy than that of S1.

Frome and Frederickson [4] applied the FFT to the analysis of first and second heart sounds. They described how segmented S1 and S2 data may be combined into a single complex signal and how a single FFT may be used to obtain the FFTs of the two signals. Computer data processing techniques were described to obtain smoothed, averaged periodograms (described in Section 6.3.2) of S1 and S2 separately.

Yoganathan et al. [5] applied the FFT for the analysis of S1 of 29 normal subjects. The FFT spectra of 250 ms windows containing S1 were averaged over 15 beats for each subject. It was found that the frequency spectrum of S1 contains peaks in a low-frequency range (10 – 50 Hz) and a medium-frequency range (50 – 140 Hz) [5]. In a related study [6], the spectrum of S2 was observed to contain peaks in low-frequency (10 – 80 Hz), medium-frequency (80 – 220 Hz), and high-frequency ranges (220 – 400 Hz). It has been suggested that the resonance peaks in the spectra may be related to the elastic properties of the heart muscles and the dynamic events causing the various components of S1 and S2 (see Section 1.2.9).

Adolph et al. [7] used a dynamic spectrum analyzer to study the frequency content of S1 during the isovolumic contraction period. The center frequency of a filter with 20 Hz bandwidth was initially set to 30 Hz and then varied in 10 Hz increments up to 70 Hz. The outputs of the filters were averaged over the same (prerecorded) 10 consecutive beats. Finally, the ratios of the average peak voltage of the filtered outputs to that of the total S1 signal during the isovolumic contraction period were computed.

Adolph et al. hypothesized that the frequency content of S1 during the isovolumic contraction period should depend on the relative contributions of the mass and elasticity of the left ventricle. The mass of the left ventricle with its blood content remains constant during isovolumic contraction. Therefore, it was reasoned that the frequency content of S1 should decrease (that is, shift toward lower frequencies) in the case of diseases that reduce ventricular elasticity, such as myocardial infarction.

Figure 6.1 shows averaged S1 spectra for normal subjects and patients with acute or healed myocardial infarction; it is seen that the reduced elasticity due to myocardial infarction has reduced the relative content of power near 40 Hz. However, Adolph et al. also noted that an increase in ventricular mass as in the case of trained athletes, or a reduction in elasticity combined with an increase in the mass as in the case of myocardopathy, could also cause a similar shift in the frequency content of S1. Regardless, they found that frequency analysis of S1 was of value in differentiating acute pulmonary embolism from acute myocardial infarction. Clarke et al. [8] also found reduction in the spectral energy of S1 to be a common accompaniment of myocardial ischemia.

6.2.2 Frequency analysis of murmurs to diagnose valvular defects

As we noted in Section 1.2.9, cardiovascular valvular defects and diseases cause high-frequency, noise-like sounds known as murmurs. Murmurs are often the only indicators of the early stages of certain cardiovascular diseases; prompt diagnosis could prevent further deterioration of the condition and possible complications.

We noted in Section 5.6.2 that zero-crossing analysis in the time domain was applied to assist in the detection of murmurs by Jacobs et al. [9] and Yokoi et al. [10]. Although ZCR increases with the presence of higher-frequency components, it does not yield a direct measure of the frequency content or the spectrum of the signal.

Application of electronic signal filtering techniques to analyze the frequency content of heart sounds and murmurs was initiated as early as the 1950s. Geckeler et al. [11] and McKusick et al. [12, 13] studied the applicability of the sound spectrograph for the analysis of heart sounds and murmurs. The sound spectrograph was developed in the late 1940s by Bell Telephone Laboratories as a tool to produce what was labeled as *visible speech*. The spectrograph used a bandpass filter

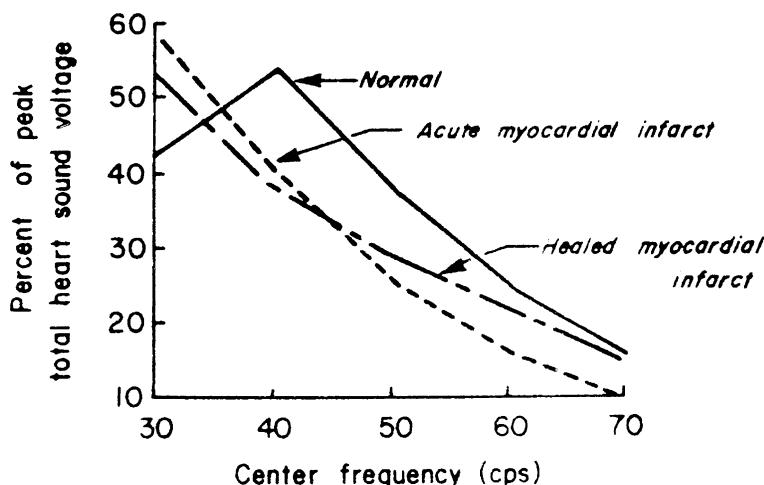


Figure 6.1 First heart sound spectra for normal, acute myocardial infarct, and healed myocardial infarct cases. The latter two cases exhibit an increased percentage of low-frequency components. Reproduced with permission from R.J. Adolph, J.F. Stephens, and K. Tanaka, The clinical value of frequency analysis of the first heart sound in myocardial infarction, *Circulation*, 41:1003–1014, 1970. ©American Heart Association.

(or a bank of bandpass filters) to determine the power of the given signal in each frequency band of interest. The signal was usually recorded and played back repeatedly as the center frequency of the bandpass filter was varied. The output was recorded on heat-sensitive or light-sensitive paper to produce a 2D distribution of the frequency content of the signal at every instant of time as a gray-level image (essentially a TFD; see Section 8.4.1).

Winer et al. [14] proposed iso-intensity contour plotting of the spectrogram instead of using variations in intensity (gray scale); they reported that, whereas normal heart sounds indicated the presence of regularity in the contours of equal intensity, abnormal sounds and murmurs produced irregular contour line structures with extensive “convolutions” and roughness (see Figures 7.36 and 7.37). It was suggested that the cardiospectrograms (or spectral phonocardiography) could provide physiologic and pathologic information beyond that provided by auscultation, without suffering from the psychoacoustic impediments that affect human observers [15].

Yoshimura [16] used a tunable bandpass filter with low and high cutoff frequencies in the range 18 – 1,425 Hz to process recorded PCG signals. It was determined that the diastolic *rumble* of mitral stenosis occupied the range 20 – 200 Hz, whereas the diastolic *blow* of aortic regurgitation spanned a much higher frequency range of 200 – 1,600 Hz (although the characteristic range was 400 – 800 Hz).

Gerbarg et al. [2, 3] developed a computer program to simulate a filter bank and obtain power spectra of heart sounds and murmurs, with the aim of developing a system for the purpose of screening to detect cardiovascular diseases. They argued that innocent (physiological) systolic murmur in children is limited to the first and middle thirds of the systolic interval between S1 and S2, whereas pathological systolic murmur due to mitral regurgitation is holosystolic. Therefore, they computed ratios of the mean power of the *last third of systole* to the mean power of systole and also to a certain “standard” noise level. A ratio was also computed of the mean energy of systole to the mean energy of the PCG over the complete cardiac cycle. Gerbarg et al. obtained 78 – 91% agreement of their computer classification based upon the three ratios defined above with clinical diagnosis of mitral regurgitation in different groups of subjects. Although they did not claim that a fully automated program for the diagnosis of mitral regurgitation had been developed, they indicated that the feasibility

of computer-based diagnosis had been established, and that simulation of human auscultation had been partially achieved.

The specific problem of detection of the murmur due to aortic insufficiency in the presence of the murmur due to mitral stenosis was considered by van Vollenhoven et al. [17]. Aortic insufficiency causes an early diastolic murmur (with a blowing or hissing quality) that is best heard in the aortic area (second right-intercostal space, just right of the sternum), whereas the mid-diastolic rumbling murmur of mitral stenosis is best heard at the apex. A tunable bandpass filter with 50 Hz bandwidth and center frequency tunable in steps of 50 Hz was used by van Vollenhoven et al. to study the frequency content in a 100 ms window during the diastolic phase of recorded PCG signals. They found that the murmur of mitral stenosis was limited in frequency content to less than 400 Hz, whereas the murmur in the case of aortic insufficiency combined with mitral stenosis had more high-frequency energy in the range 300 – 1,000 Hz.

Sarkady et al. [18] suggested synchronized averaging of the PSDs of PCG signals, computed using the FFT algorithm, over several cardiac cycles. Johnson et al. [19, 20] studied FFT-based magnitude spectra of the systolic murmur due to aortic stenosis. They computed the magnitude spectra of systolic windows of duration 86, 170, and 341 ms, and averaged the results over 10 cardiac cycles. Johnson et al. hypothesized that higher murmur frequencies are generated as the severity of aortic stenosis increases. In their study of patients who underwent catheterization and cardiac fluoroscopy, the transvalvular systolic pressure gradient was measured during pull-back of the catheter from the left ventricle through the aortic valve and was found to be in the range 10–140 mm of Hg. Spectral power ratios (described in Section 6.4.2) were computed considering the band 25 – 75 Hz as the constant area (*CA*) related to normal sounds (for reference) and the band 75 – 150 Hz as the predictive area (*PA*) related to murmurs.

Figure 6.2 illustrates the magnitude spectra of four patients with aortic stenosis of different levels of severity. The spectra in the figure are segmented into the *CA* and *PA* parts as described above; the transvalvular systolic pressure gradient (in mm of Hg) and the *PA/CA* spectral power ratio are also shown for each case. Johnson et al. found that the spectral power ratio increased linearly with the transvalvular systolic pressure gradient, and hence correlated well with the severity of aortic stenosis. The importance of recording the PCG in the aortic area, prefiltering the PCG to 25 – 1,500 Hz, and the selection of an appropriate systolic murmur window were discussed by Johnson et al. Although there were confounding factors, it was indicated that the noninvasive PCG-based technique could be useful in identifying the need for catheterization as well as follow-up of patients with aortic stenosis.

6.3 Estimation of the PSD

6.3.1 Considerations in the computation of the ACF

We encountered the ACF and CCF in Equations 3.16, 3.17, 3.20, and 4.28: The first equation cited here provides a general definition of the ACF as a statistical expectation or an integral over an indefinite duration; the last one treats the CCF as the projection of one signal on to another and neglects a scale factor that may be of no consequence in a given application. We now investigate more closely the procedures required to compute the ACF, and hence the PSD, from finite-length signal records.

Let us consider a signal record of N samples: $x(n)$, $n = 0, 1, 2, \dots, N - 1$. In order to compute the time-averaged ACF $\phi_{xx}(m)$ for a delay of m samples, we need to form the product $x(n)x(n \pm m)$ and sum over the available range of data samples. The true ACF is given as $\phi_{xx}(m) = E[x(n)x(n + m)]$. Note that one of the copies of the signal entering the computation of the ACF should be conjugated if the signal is a complex quantity.

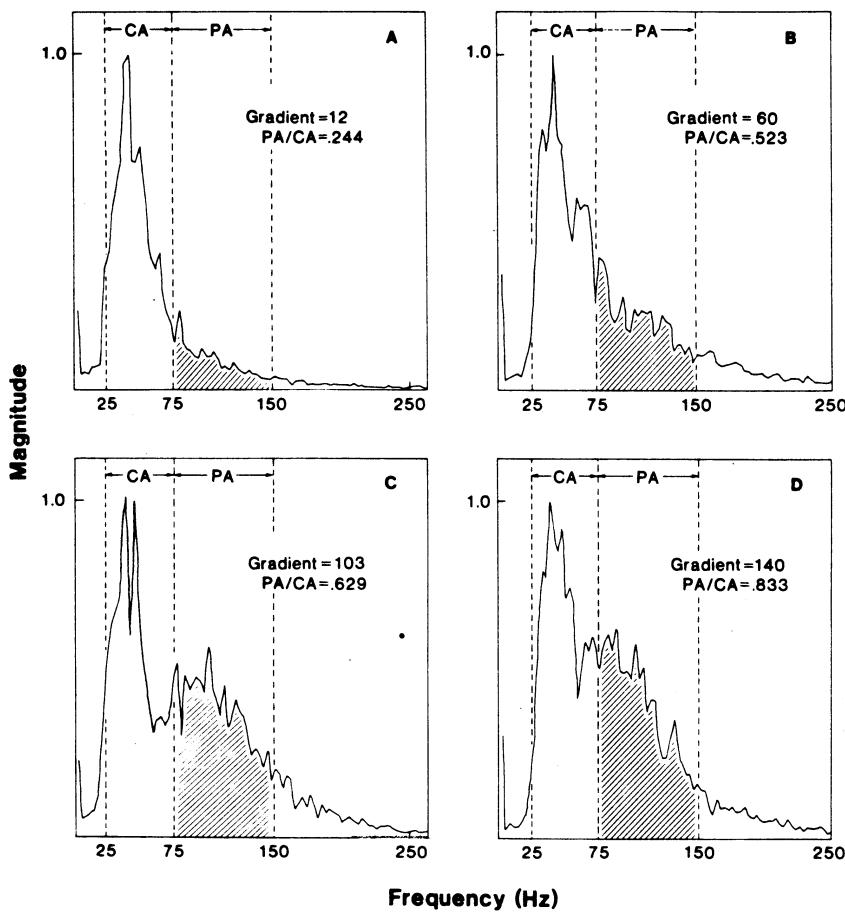


Figure 6.2 Averaged and normalized magnitude spectra of four patients with aortic stenosis of different levels of severity. Each spectrum is segmented into two parts: a constant area CA for reference and a predictive area PA . The transvalvular systolic pressure gradient (measured via catheterization, in $mm\text{ of }Hg$) and the PA/CA spectral power ratio are shown for each case. Reproduced with permission from the American College of Cardiology, from G.R. Johnson, R.J. Adolph, and D.J. Campbell, Estimation of the severity of aortic valve stenosis by frequency analysis of the murmur, *Journal of the American College of Cardiology*, 1(5):1315–1323, 1983. ©Elsevier Science.

It is readily seen that we may sum from $n = 0$ to $n = N - 1$ when computing $\phi_{xx}(0)$ with $x(n)x(n) = x^2(n)$. However, when computing $\phi_{xx}(1)$ with $x(n)x(n+1)$, we can only sum from $n = 0$ to $n = N - 2$. As we apply a linear shift of m samples to one copy of the signal to compute $\phi_{xx}(\pm m)$, m samples of one of the copies of the signal drop out of the window of analysis indicated by the overlap between the two versions of the signal; see Figure 6.3. Therefore, only $N - |m|$ pairs of data samples are available to estimate the ACF for the delay of $\pm m$ samples. We then have a sample-mean estimate of the ACF given by

$$\phi_1(m) = \frac{1}{N - |m|} \sum_{n=0}^{N-|m|-1} x(n)x(n+m). \quad (6.1)$$

The subscript xx has been omitted in the equation given above; the subscript 1 indicates the use of one type of averaging scale factor in estimating the ACF. Oppenheim and Schafer [21] have shown

that $\phi_1(m)$ is a consistent estimate of $\phi_{xx}(m)$: It has zero bias and has a variance that tends to zero as $N \rightarrow \infty$. However, the variance of the estimate becomes exceptionally large as m approaches N : Very few nonzero pairs of overlapping samples are then available to compute the ACF, and the estimate is useless.

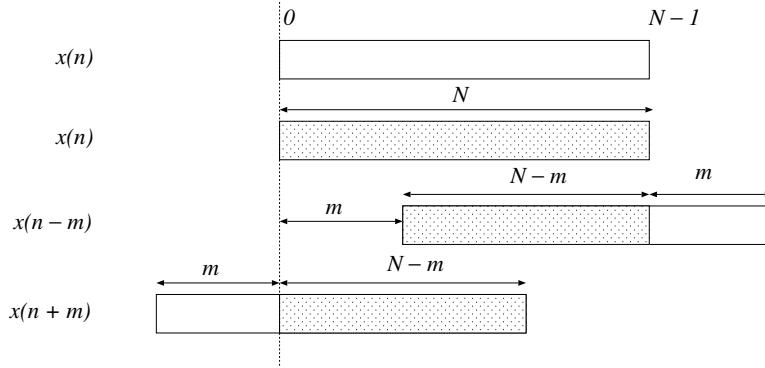


Figure 6.3 The effect of shifting in the computation of the ACF or CCF: as the shift increases, the number of available pairs of overlapping samples is reduced.

An alternative definition of the ACF ignores the lack of $|m|$ nonzero pairs of samples, and applies the same scale factor for all delays, leading to

$$\phi_2(m) = \frac{1}{N} \sum_{n=0}^{N-|m|-1} x(n) x(n+m). \quad (6.2)$$

Note that the upper limit of summation in the above expression could be stated as $N - 1$ with no effect on the result; the first or the last $|m|$ samples of $x(n)$ will not overlap with $x(n+m)$, and result in zero product terms. Oppenheim and Schafer [21] have shown that $\phi_2(m)$ has a bias equal to $\frac{m}{N} \phi_{xx}(m)$: The bias tends to the actual value being estimated as m approaches N , although the variance is almost independent of m and tends to zero as $N \rightarrow \infty$. Regardless, both the ACF estimates are asymptotically unbiased [the bias of $\phi_2(m)$ tends to zero as $N \rightarrow \infty$], and they yield good estimates of the ACF as long as the number of samples N is large and $m \ll N$.

Note that the two ACF estimates $\phi_1(m)$ and $\phi_2(m)$ are interrelated as

$$\phi_2(m) = \frac{N - |m|}{N} \phi_1(m). \quad (6.3)$$

Thus, $\phi_2(m)$ is a scaled version of $\phi_1(m)$. However, since the scaling factor is a function of m , it is commonly referred to as a *window*; more discussion on this interpretation is presented in Section 6.3.2. It should also be observed that the distinction between $\phi_1(m)$ and $\phi_2(m)$ is comparable to that between the unbiased and biased sample variance measures, where the division is by $N - 1$ or N , respectively, with N being the number of samples available.

6.3.2 The periodogram

Since the PSD and the ACF are a Fourier transform pair, we may compute an estimate of the PSD as

$$S_2(\omega) = \sum_{m=-(N-1)}^{N-1} \phi_2(m) \exp(-j\omega m), \quad (6.4)$$

assuming that, indeed, the ACF is computed or available for $|m|$ up to $N - 1$. The Fourier transform of the signal $x(n)$, $n = 0, 1, 2, \dots, N - 1$, is given as

$$X(\omega) = \sum_{n=0}^{N-1} x(n) \exp(-j\omega n). \quad (6.5)$$

It can be shown that

$$S_2(\omega) = \frac{1}{N} |X(\omega)|^2. \quad (6.6)$$

The PSD estimate $S_2(\omega)$ is known as the *periodogram* of the signal $x(n)$ [21]. Oppenheim and Schafer [21] have shown that $S_2(\omega)$ is a biased estimate of the PSD, with

$$E[S_2(\omega)] = \sum_{m=-(N-1)}^{N-1} \frac{N-|m|}{N} \phi_{xx}(m) \exp(-j\omega m). \quad (6.7)$$

If we consider the Fourier transform of $\phi_1(m)$, we get a different estimate of the PSD as

$$S_1(\omega) = \sum_{m=-(N-1)}^{N-1} \phi_1(m) \exp(-j\omega m), \quad (6.8)$$

with the expected value [21]

$$E[S_1(\omega)] = \sum_{m=-(N-1)}^{N-1} \phi_{xx}(m) \exp(-j\omega m). \quad (6.9)$$

Because of the finite limits of the summation, $S_1(\omega)$ is a biased estimate of the PSD.

The two estimates $S_2(\omega)$ and $S_1(\omega)$ may be seen as the Fourier transforms of windowed ACFs, with the window functions being a triangular function, known as the Bartlett window, $w_B(m)$, in the first case, and a rectangular window, $w_R(m)$, in the second case:

$$w_B(m) = \begin{cases} \frac{N-|m|}{N}, & |m| < N, \\ 0, & \text{otherwise,} \end{cases} \quad (6.10)$$

$$w_R(m) = \begin{cases} 1, & |m| < N, \\ 0, & \text{otherwise.} \end{cases} \quad (6.11)$$

Note that the windows defined above have a (nonzero) duration of $(2N - 1)$ samples.

Since the ACF is multiplied with the window function, the PSD is convolved with the Fourier transform of the window function, leading to spectral leakage and loss of resolution (more details on windows follow in Section 6.3.4). The Fourier transforms of the Bartlett and rectangular windows are, respectively [21],

$$W_B(\omega) = \frac{1}{N} \left[\frac{\sin(\omega N/2)}{\sin(\omega/2)} \right]^2 \quad (6.12)$$

and

$$W_R(\omega) = \frac{\sin[\omega(2N - 1)/2]}{\sin(\omega/2)}. \quad (6.13)$$

Oppenheim and Schafer [21] have shown that the periodogram has a variance that does not approach zero as $N \rightarrow \infty$; instead, the variance of the periodogram is of the order of σ_x^4 regardless of N . Thus, the periodogram is not a consistent estimate of the PSD.

6.3.3 The need for averaging PSDs

A common approach to reduce the variance of an estimate is to average over a number of independent estimates or observations. We have seen in Section 3.5 how the variance of the noise in noisy signals may be reduced by synchronized averaging over a number of observations of the corrupted signal. In a similar vein, a number of periodograms may be computed over multiple observations of a signal and averaged to obtain a better estimate of the PSD. It is necessary for the process to be stationary, at least during the period over which the periodograms are computed and averaged.

Problem: *How can we obtain an averaged periodogram when we are given only one signal record of finite duration?*

Solution: Oppenheim and Schafer [21] have described the following procedure, attributed to Bartlett, to average periodograms of segments of the given signal record:

1. Divide the given data sequence $x(n)$, $n = 0, 1, 2, \dots, N - 1$, into K segments of M samples each. We then have the segments given by

$$x_i(n) = x[n + (i - 1)M], \quad 0 \leq n \leq M - 1, \quad 1 \leq i \leq K. \quad (6.14)$$

2. Compute the periodogram of each segment as

$$S_i(\omega) = \frac{1}{M} \left| \sum_{n=0}^{M-1} x_i(n) \exp(-j\omega n) \right|^2, \quad 1 \leq i \leq K. \quad (6.15)$$

The Fourier transform in the equation given above is evaluated as a DFT (using the FFT algorithm) in practice.

3. If the ACF $\phi_{xx}(m)$ is negligible for $|m| > M$, the periodograms of the K segments of duration M samples each may be assumed to be mutually independent. Then, the Bartlett estimate $S_B(\omega)$ of the PSD is obtained as the sample mean of the K independent observations of the periodogram:

$$S_B(\omega) = \frac{1}{K} \sum_{i=1}^K S_i(\omega). \quad (6.16)$$

Oppenheim and Schafer [21] have shown that the expected value of the Bartlett estimate $S_B(\omega)$ is the convolution of the true PSD $S_{xx}(\omega)$ with the Fourier transform of the Bartlett window given in Equation 6.12 (with N replaced by M for the present procedure). The convolution relationship indicates the bias in the estimate, and has the effect of spectral smearing and leakage; the bias may, therefore, be interpreted as a loss in resolution. Although $S_B(\omega)$ is a biased estimate, its variance tends to zero as the number of segments K increases. Therefore, it is a consistent estimate.

When we have a (stationary) signal of fixed duration of N samples, we face limitations on the number of segments K that we may obtain. While the variance of the estimate decreases as K is increased, it should be recognized that there is a concomitant decrease in the number of samples M per segment. As M decreases, the main lobe of the Fourier transform of the Bartlett window (see Equation 6.12) widens; frequency resolution is lost because the estimate is the convolution of the true PSD with the window's frequency response. An illustration of application of the Bartlett procedure is provided at the end of Section 6.3.4.

Cyclostationary signals such as the PCG offer a unique and interesting approach to synchronized averaging of periodograms over a number of cycles, without the trade-off between the reduction of variance and the loss of resolution imposed by segmentation as described above. This is presented as an illustration of application in Section 6.3.6.

6.3.4 The use of windows: spectral resolution and leakage

The Bartlett procedure may be viewed as an ensemble averaging approach to reduce the variance (which may be interpreted as an effect of noise) of the periodogram. Another approach to obtain a smooth spectrum is to convolve the periodogram $S(\omega)$ with a filter or smoothing function $W(\omega)$ in the frequency domain (similar to the use of an MA filter in the time domain). The smoothed estimate $S_s(\omega)$ is given by

$$S_s(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\nu) W(\omega - \nu) d\nu, \quad (6.17)$$

where ν is a temporary variable for integration.

Because the PSD is a nonnegative function, the smoothing function $W(\omega)$ should satisfy $W(\omega) \geq 0$, $-\pi \leq \omega \leq \pi$. The Fourier transform of the Bartlett window $W_B(\omega)$ meets this requirement. Oppenheim and Schafer [21] have shown that the variance of the smoothed periodogram is reduced approximately by the factor

$$\frac{1}{N} \sum_{m=-(M-1)}^{M-1} w^2(m) = \frac{1}{2\pi N} \int_{-\pi}^{\pi} W^2(\omega) d\omega, \quad (6.18)$$

with reference to the variance of the original periodogram; here, N is the total number of samples in the signal, and $(2M - 1)$ is the number of samples in the smoothing window function. A rectangular window offers a variance-reduction factor of approximately $\frac{2M}{N}$, whereas the factor for the Bartlett window is $\frac{2M}{3N}$ [21]. It should be noted that smoothing of the spectrum (reduction of variance) is achieved at the price of loss of frequency resolution.

Since the periodogram is the Fourier transform of the ACF estimate $\phi(m)$, the convolution operation in the frequency domain in Equation 6.17 is equivalent to multiplying $\phi(m)$ with $w(m)$, the inverse Fourier transform of $W(\omega)$. This result suggests that the same PSD estimate as $S_s(\omega)$ may be obtained by applying a window to the ACF estimate and then taking the Fourier transform of the result. As the ACF is an even function, the window should also be even.

Based on the arguments outlined above, Welch [22] (see also Oppenheim and Schafer [21]) proposed a method to average modified periodograms. In Welch's procedure, the given signal is segmented as in the Bartlett procedure, but a window is applied directly to the original signal segments before Fourier transformation. The periodograms of the windowed segments are defined as

$$S_{Wi}(\omega) = \frac{1}{ME_w} \left| \sum_{n=0}^{M-1} x_i(n)w(n) \exp(-j\omega n) \right|^2, \quad i = 1, 2, \dots, K, \quad (6.19)$$

where E_w is the average power of the window given by

$$E_w = \frac{1}{M} \sum_{n=0}^{M-1} w^2(n). \quad (6.20)$$

Note that the duration of the window defined above is M samples. The Welch PSD estimate $S_W(\omega)$ is obtained by averaging the modified periodograms as

$$S_W(\omega) = \frac{1}{K} \sum_{i=1}^K S_{Wi}(\omega). \quad (6.21)$$

Welch [22] showed that, if the segments are not overlapping, the variance of the averaged modified periodogram is inversely proportional to K , the number of segments used. Welch also suggested

that the segments may be allowed to overlap, in which case the modified periodograms are not mutually independent. The spectral window that is effectively convolved with the PSD in the frequency domain is proportional to the squared magnitude of the Fourier transform of the time-domain data window applied. Therefore, no matter which type of a data window is used, the spectral smoothing function is nonnegative, thereby guaranteeing that the PSD estimate will be nonnegative as well.

Some of the commonly used data windows are defined below [21,23]; the windows are of length N samples and causal, defined for $0 \leq n \leq N - 1$.

Rectangular:

$$w(n) = 1. \quad (6.22)$$

Bartlett (triangular):

$$w(n) = \begin{cases} \frac{2n}{N-1}, & 0 \leq n \leq \frac{N-1}{2}, \\ 2 - \frac{2n}{N-1}, & \frac{N-1}{2} \leq n \leq N-1. \end{cases} \quad (6.23)$$

Hamming:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right). \quad (6.24)$$

von Hann (also known as Hann or Hanning):

$$w(n) = \frac{1}{2} \left[1 - \cos\left(\frac{2\pi n}{N-1}\right) \right]. \quad (6.25)$$

Blackman:

$$w(n) = 0.42 - 0.5 \cos\left(\frac{2\pi n}{N-1}\right) + 0.08 \cos\left(\frac{4\pi n}{N-1}\right). \quad (6.26)$$

Figure 6.4 illustrates the rectangular, Bartlett, Hann, and Hamming windows with $N = 256$ samples. A Hann window with $N = 128$ samples is also illustrated (centered with reference to the longer-duration windows).

Use of the tapered windows (all of the above, except the rectangular window) provides the advantage that the ends of the given signal are reduced to zero (with the further exception of the Hamming window, for which the end values are not zero but 0.08). This feature means that there are no discontinuities in the periodic version of the signal encountered in DFT-based procedures. All of the window functions listed above are symmetric (even) functions and, therefore, have a linear phase (or a real spectrum with zero phase if the window is centered at the origin).

Figure 6.5 illustrates the log-magnitude frequency responses of the five window functions shown in Figure 6.4. The frequency responses were computed after padding the windows to a total duration of $L = 2,048$ samples for FFT computation and taking $20 \log_{10}$ of the magnitude of the result. The rectangular window has the narrowest main lobe of width $\frac{4\pi}{N}$; the main lobe is wider at $\frac{8\pi}{N}$ for the Bartlett, Hann, and Hamming windows; it is the widest at $\frac{12\pi}{N}$ for the Blackman window [21]. A reduction in window width will lead to an increase in the main-lobe width, as illustrated by the frequency responses of the two Hann windows. Note that the wider the main lobe, the greater is the spectral smoothing, and hence the loss of spectral resolution is more severe.

The rectangular window has the highest peak side-lobe levels of the five windows shown at -13 dB, with the Bartlett, Hann, Hamming, and Blackman windows having their peak side-lobe levels at -25 dB, -31 dB, -41 dB, and -57 dB, respectively [21]. Higher levels of the side lobes will cause increased spectral leakage (weighted summation of spectral components with significant weights over a wide range of frequencies due to convolution in the frequency domain), resulting

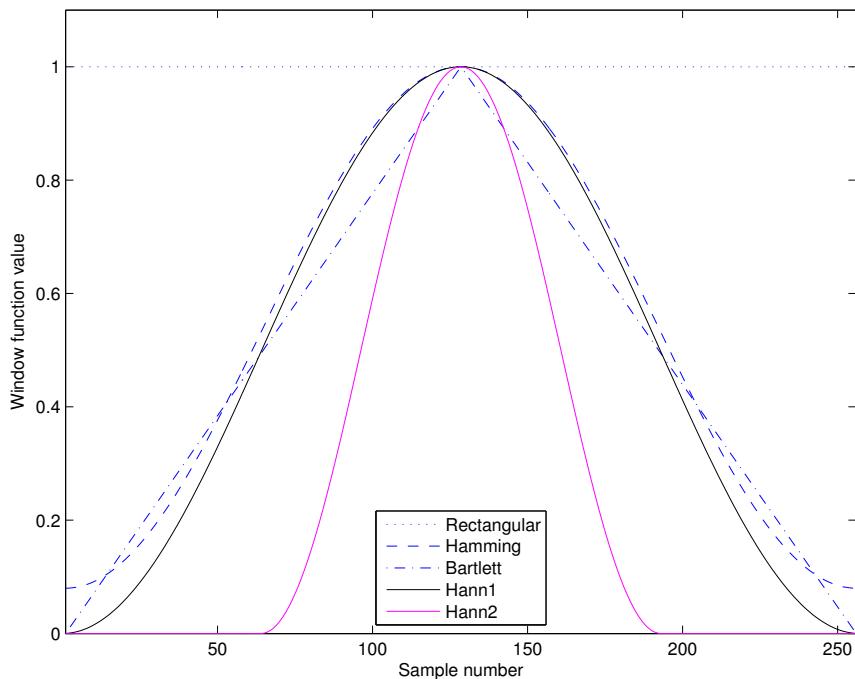


Figure 6.4 Commonly used window functions: rectangular, Bartlett, Hamming, and Hann windows with $N = 256$ (Hann1), and Hann window with $N = 128$ samples (Hann2). All windows are centered at the 128th sample in the figure.

in a more distorted spectrum. Note that reduction in leakage through the use of the tapered windows comes at the price of increased main-lobe width and, therefore, more severe loss of spectral resolution (more smoothing).

See Bingham et al. [24], Welch [22], Kay and Marple [25], Thomson [26], Robinson [27], and Childers et al. [28] for discussions on several issues related to estimation of PSDs.

Illustrations of application: The Welch method of windowing signal segments and averaging their PSDs was applied to the o2 channel of the EEG signal illustrated in Figure 1.41. The number of samples in the signal is $N = 750$, with the sampling frequency being $f_s = 100 \text{ Hz}$. Note that the specific EEG signal record may be assumed to be stationary over its relatively short duration of 7.5 s . The dominant activity in the signal is the alpha rhythm, which appears throughout the duration of the signal record.

The PSD of the entire signal was first computed using no window (that is, the rectangular window was applied implicitly); the FFT array was computed with $L = 1,024$ samples. The top trace in Figure 6.6 illustrates the PSD of the complete signal.

For the first averaged periodogram procedure, the EEG signal was segmented with $M = 64$ samples each, with implicit use of the rectangular window (equivalent to the Bartlett method). A total of $K = 11$ segments were obtained. Each segment was padded with zeros to a length of $L = 1,024$ for the sake of FFT computation. The PSDs of the segments were then averaged, followed by normalization and logarithmic transformation. The second and third plots in Figure 6.6 illustrate the PSD of a sample segment (the 11th segment) and the averaged PSD (the Bartlett estimate), respectively. It is seen that the averaged PSD (third trace) provides a smooth spectral estimate with a clearly dominant peak at approximately 10 Hz , representing the alpha rhythm present in the signal. The PSD of the individual segment (middle trace) displays many peaks and

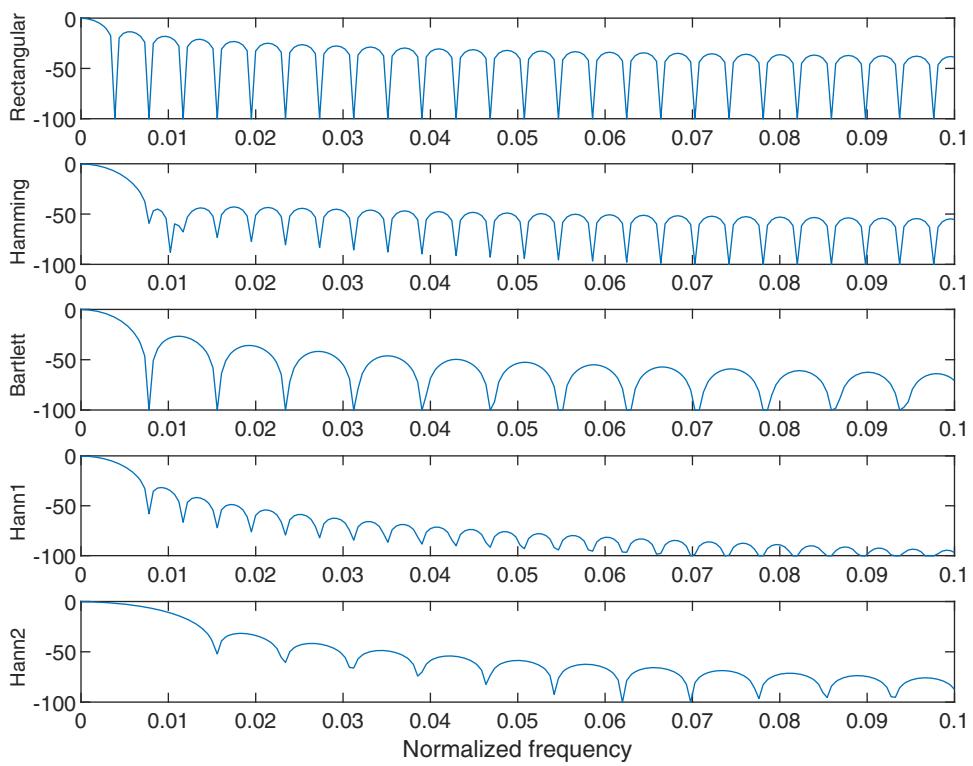


Figure 6.5 Log-magnitude frequency responses of the five windows illustrated in Figure 6.4. The vertical axis is in dB . The plots are on an expanded scale over the limited normalized-frequency range of $[0, 0.1]$ for clear illustration of the main lobes and the side lobes. Discontinuities arise due to the log of the zeros of the responses being $-\infty$. Values less than -100 dB have been set to -100 dB for improved visualization of the responses for the rectangular and Bartlett windows.

valleys that are possibly spurious and not significant, and have been suppressed or smoothed by the averaging process, as evident in the third plot in the figure. The single PSD computed from the entire signal (top trace) exhibits numerous variations that may not be relevant and could confound visual or automated analysis. (Note: Direct comparison of the PSDs is possible since they have the same number of samples, that is, the same frequency sampling rate.)

Figure 6.7 illustrates a second set of PSDs similar to that in Figure 6.6, but with the use of the Hann window in the Welch procedure. The effect of the Hann window is not significant in the case of the PSD of the entire signal (top trace), as the window length is reasonably large ($N = 750$). However, the Hann window has clearly smoothed the multiple (possibly spurious) peaks and valleys in the PSD of the segment illustrated in the middle trace. The wider main lobe of the Hann window's frequency response has caused a more severe loss of frequency resolution (smoothing) than the rectangular window in the case of the corresponding PSD in Figure 6.6. The averaged PSD in the lowest trace of Figure 6.7 clearly illustrates the benefit of the Hann window in the substantially reduced power levels beyond 30 Hz . The lower side-lobe levels of the Hann window have resulted in less spectral leakage than that in the case of the rectangular window as illustrated by the corresponding PSD in Figure 6.6. The price paid, however, is evidenced by the wider peak in the averaged PSD with the Hann window, which spans the range $5 - 15 \text{ Hz}$ at the -10 dB level. The two distinct peaks at about 10 Hz and 12 Hz that are evident in the top traces of Figures 6.6 and 6.7 as well as in the smoothed PSD in the bottom trace of Figure 6.6 are no longer seen separately in

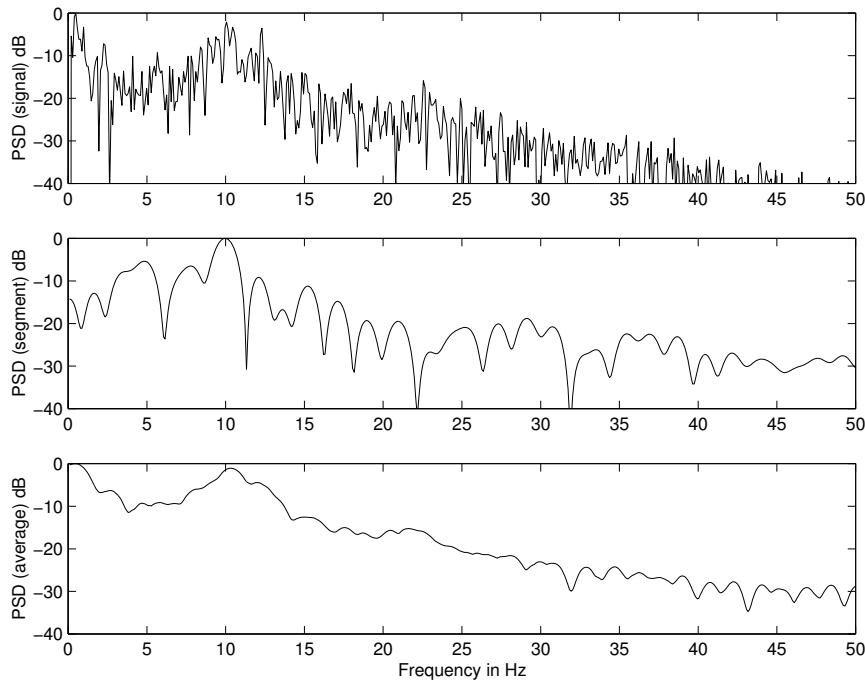


Figure 6.6 Bartlett PSD estimate of the o2 channel of the EEG signal in Figure 1.41. Top trace: PSD of the entire signal. Middle trace: PSD of the 11th segment. Bottom trace: Averaged PSD using $K = 11$ segments of the signal. The rectangular window was (implicitly) used in all cases. Number of samples in the entire signal: $N = 750$. Number of samples in each segment: $M = 64$. All FFT arrays were computed with $L = 1,024$ samples. Sampling frequency $f_s = 100 \text{ Hz}$. See also Figure 6.7.

the bottom trace of Figure 6.7. Regardless, the averaged PSD with the Hann window appears to be smoother and more amenable to analysis than the corresponding result with the rectangular window.

It should be noted that the segments of EEG signals used in the illustrations in the present section are of short duration. See Section 6.7 for an example of analysis of longer EEG records.

6.3.5 Estimation of the ACF from the PSD

Good estimates of the ACF are required in applications such as the design of the Wiener filter and estimation of the statistics of stochastic processes. Once a PSD estimate has been obtained by a method such as the Bartlett or Welch procedure, we may take the inverse Fourier transform of the result and use the result as an estimate of the ACF. We may also fit a smooth curve or a parametric model to the PSD or to the equivalent ACF model (such as an exponential or a Laplacian function).

Let us consider again the expression

$$\phi_2(m) = \frac{1}{N} \sum_{n=0}^{N-|m|-1} x(n)x(n+m). \quad (6.27)$$

As the ACF is an even function, we need to compute it only for positive m . It is evident that the ACF estimate is simply the result of linear convolution of $x(n)$ with $x(-n)$ (with the scale factor $\frac{1}{N}$). If the DFT of $x(n)$ is $X(k)$, the DFT of $x(-n)$ is $X^*(k)$. Since convolution in the time domain is multiplication in the frequency domain, we could compute the DFT $X(k)$ of $x(n)$, obtain $X(k)X^*(k) = |X(k)|^2$, and take its inverse DFT. However, the DFT procedure provides circular

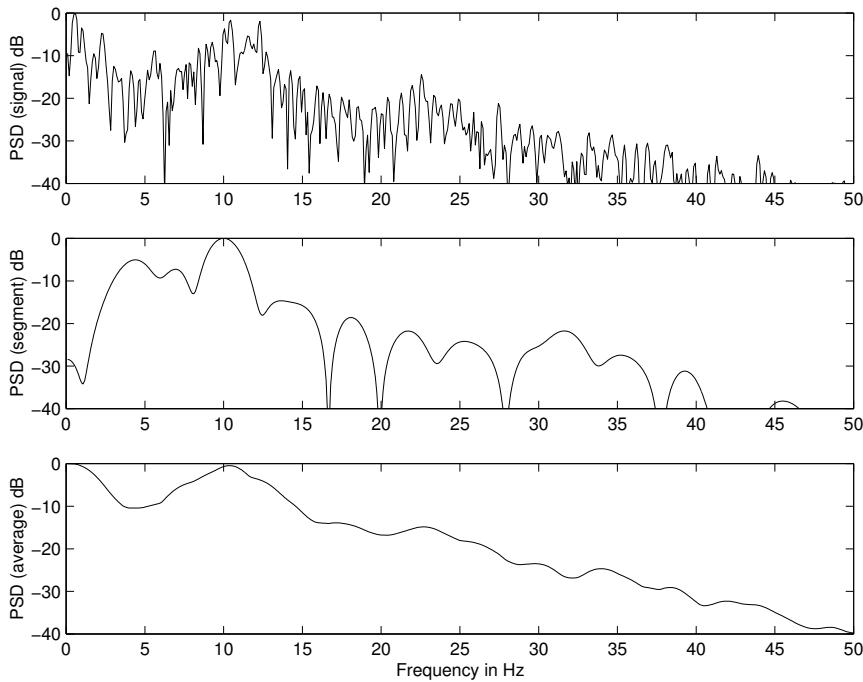


Figure 6.7 Welch PSD estimate of the o2 channel of the EEG signal in Figure 1.41. Top trace: PSD of the entire signal. Middle trace: PSD of the 11th segment. Bottom trace: Averaged PSD using $K = 11$ segments of the signal. The Hann window was used in all cases. Number of samples in the entire signal and the size of the Hann window used in computing the PSD of the entire signal: $N = 750$. Number of samples in each segment and the size of the Hann window used in the averaged periodogram method: $M = 64$. All FFT arrays were computed with $L = 1,024$ samples. Sampling frequency $f_s = 100 \text{ Hz}$. See also Figure 6.6.

convolution and not linear convolution. Therefore, we need to pad $x(n)$ with at least $M - 1$ zeros, where M is the largest lag for which the ACF is desired. The DFT must then be computed with at least $L = N + M - 1$ samples, where N is the number of samples in the original signal. If this requirement is built into the periodogram or averaged periodogram procedure, the inverse DFT of the final PSD estimate may be used as an estimate of the ACF [with the scale factor $\frac{1}{N}$, or division by $\phi(0)$ to get the normalized ACF].

6.3.6 Synchronized averaging of PCG spectra

Every individual is familiar with the comforting *lub-dub* sounds of his or her heart beat; every prospective parent would have taken pleasure in listening to the throbbing heart of the yet-to-be-born baby. Use of the heart sounds is extremely common in clinical practice: the stethoscope is the most common sign and tool of a physician. Yet, behind this common signal lie many sophisticated and potentially complicating characteristics.

The PCG is a nonstationary signal due to the fact that the amount of blood in each cardiac chamber and the state of contraction of the muscles change continuously during each cardiac cycle. S2 usually has more high-frequency content than S1: The PSD of a normal PCG signal changes within about 300 ms. Valve opening or closing sounds, being of short duration of the order of 10 ms, are of a transient and high-frequency character. The presence of murmurs adds another dimension of nonstationarity, with frequency content well beyond that of the normal heart sounds: the PSD of an abnormal PCG could change every 100 ms or less. Individual epochs of S1, S2,

valve snaps, and murmurs are of limited durations of the order of $10 - 300\text{ ms}$. These aspects of the PCG preclude segmented averaging as recommended by the Bartlett or Welch procedures.

Over and above all of the factors mentioned in the preceding paragraph, the transmission characteristics of the chest wall change during breathing. (Living systems are dynamic!) The PCG signals recorded at various locations on the chest are also subject to different transmission-path effects. Whereas adult subjects may cooperate in PCG signal acquisition by holding their breath or performing other maneuvers, these possibilities cannot be considered in the case of infants and young children in poor states of health. The PCG signal presents more challenges in acquisition and analysis than most of the other biomedical signals we have encountered [29].

Problem: Propose a method to obtain averaged PSD estimates of the systolic and diastolic heart sounds.

Solution: The cyclostationarity of the PCG lends itself to a unique approach to averaging PCG segments corresponding to the same phase of the cardiac cycle extracted from multiple heart beats. If the subject were to hold his/her breath during the period of acquisition of the PCG record, the chest-wall transmission characteristics will be stationary over the multiple cardiac cycles in the record. Therefore, we may segment S1, S2, or any portion of the cardiac cycle of interest from as many beats as are available, and then average their PSD estimates in a procedure similar to the Bartlett or Welch procedures. (*Note:* Direct averaging of the PCG signals themselves or of their complex Fourier transforms could lead to undesired cancellation of noise-like murmurs or asynchronous frequency components and their disappearance from the result! Refer to Sections 4.10 and 6.5 for discussions on intentional cancellation of asynchronous components in the PCG via synchronized averaging.)

We have seen in Sections 5.5.2 and 5.5.3 how an envelope or the envelopogram of the PCG may be averaged over several cardiac cycles. However, there was no need to segment parts of a cardiac cycle in envelope analysis: Nonstationarity of the signal within a cardiac cycle was not of concern. In the present application of PSD analysis, there is a need to segment the PCG further.

A procedure was described in Section 4.9 for segmentation of the systolic and diastolic parts of PCG signals based on the detection of the QRS complex in the ECG and the detection of the dicrotic notch in the carotid pulse signal. Further segmentation of the systolic or diastolic parts into S1 and systolic murmur or S2 and diastolic murmur, respectively, would require more sophisticated methods, which are the topics of Chapter 8. For now, let us consider the task of obtaining averaged PSDs of the systolic and diastolic parts of a PCG signal.

Figure 6.8 shows the PCG signal over one cardiac cycle of a normal subject segmented using the procedure described in Section 4.9 and illustrated in Figure 4.30 (see also Figure 5.7). The periodograms of the systolic and diastolic parts of the PCG cycle illustrated are also shown in the figure. In order to obtain better PSD estimates, the periodogram of each systolic or diastolic segment was computed separately and averaged over 16 cardiac cycles. No data window was applied (the rectangular window was used, in effect); therefore, the procedure used is similar to the Bartlett procedure. Individual systolic or diastolic segments could be of different durations; for the present illustration, all periodograms were computed with the same number of samples, which was taken to be the maximum *RR* interval in the ECG record of the subject. The averaged systolic and diastolic PSD estimates are shown in Figure 6.8. (Averaging was performed using the nonnegative PSDs before conversion to *dB*.) The averaging procedure provides smoother estimates of the PSDs by removing beat-to-beat variations that are neither significant nor of interest. Spectral peaks may be clearly observed in the averaged periodograms and may be considered to be more reliable estimates of resonance than the peaks found in individual periodograms.

Figure 6.9 illustrates a PCG signal cycle as well as the individual and averaged systolic and diastolic PSD estimates for a patient with systolic murmur, split S2, and opening snap of the mitral valve (see also Figures 4.31 and 5.8). It is unlikely that the patient held her breath during data acquisition. The presence of increased high-frequency power in the range $120 - 250\text{ Hz}$ due to the

systolic murmur is evident in the averaged systolic PSD. The diastolic PSDs are comparable to the corresponding normal diastolic PSDs in Figure 6.8.

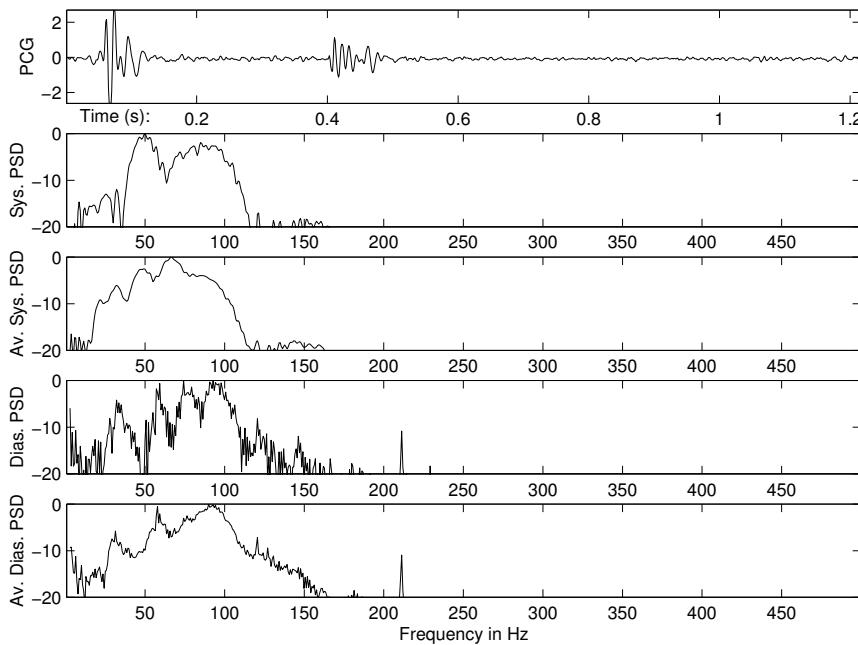


Figure 6.8 Top to bottom: A sample PCG signal over one cardiac cycle of a normal subject (male, 23 years; see also Figures 4.30 and 5.7); periodogram of the systolic portion of the signal (approximately 0 – 0.4 s); averaged periodogram of the systolic parts of 16 cardiac cycles segmented as illustrated in Figure 4.30; periodogram of the diastolic portion of the signal shown in the first plot (approximately 0.4 – 1.2 s); averaged periodogram of the diastolic parts of 16 cardiac cycles. The periodograms are on a log scale (dB). Av. = average. Sys. = systolic. Dias. = diastolic.

6.4 Measures Derived from Power Spectral Density Functions

The Fourier spectrum or PSD provides us with a density function of signal amplitude, power, or energy versus frequency. We would typically have a large number of samples of the PSD over a wide frequency range, which may not lend itself to easy analysis. We may, of course, study the shape of the spectrum graphically and observe its general characteristics. Such an approach is often referred to as *nonparametric* spectral analysis. The spectral models described in Section 7.4 are characterized by a small number of parameters; hence, the procedure is known as *parametric* spectral analysis (or modeling).

Problem: Derive parameters or measures from a Fourier spectrum or PSD that can help in the characterization of the spectral variations or features contained therein.

Solution: Since the PSD is a nonnegative function as well as a density function, we may readily treat it as a PDF, and compute statistics using moments. We may also detect peaks corresponding to resonance, measure their bandwidth or quality factor, and derive measures of concentration of power in specific frequency bands of interest or concern. Although the PSD itself is nonparametric, we may derive several parameters that, while not completely representing the entire PSD, may facilitate the identification and classification of physiological and/or pathological phenomena. We will investigate a few different approaches toward this end in the following sections.

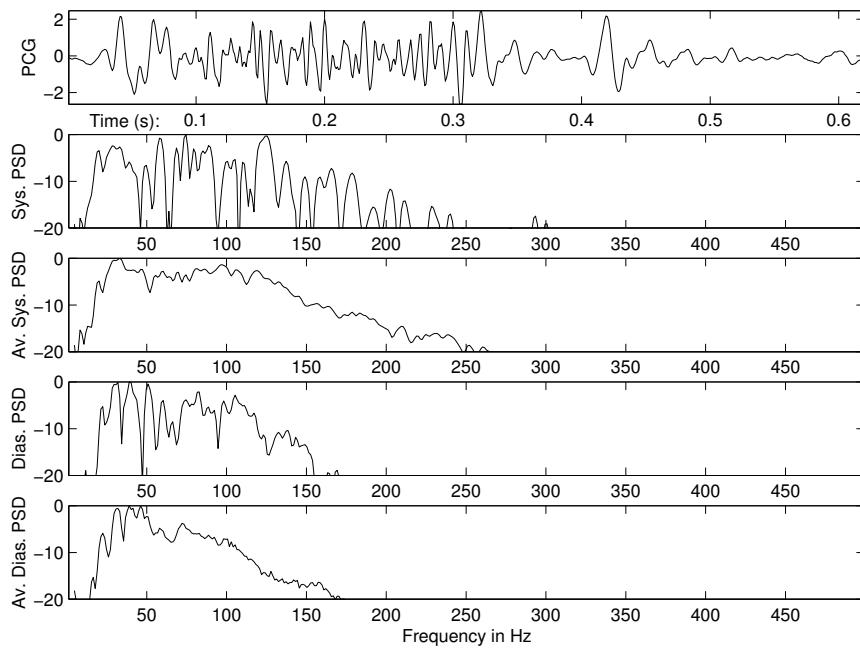


Figure 6.9 Top to bottom: A sample PCG signal over one cardiac cycle of a patient (female, 14 months; see also Figures 4.31 and 5.8) with systolic murmur, split S2, and opening snap of the mitral valve; periodogram of the systolic portion of the signal (approximately 0 – 0.28 s); averaged periodogram of the systolic parts of 26 cardiac cycles segmented as illustrated in Figure 4.31; periodogram of the diastolic portion of the signal shown in the first plot (approximately 0.28 – 0.62 s); averaged periodogram of the diastolic parts of 26 cardiac cycles. The periodograms are on a log scale (dB). Av. = average. Sys. = systolic. Dias. = diastolic.

6.4.1 Moments of PSD functions

Let us recall and reformat the relationships given by Parseval's theorem from Equation 3.91 as follows:

$$\begin{aligned} E_x &= \sum_{n=0}^{N-1} |x(n)|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X(k)|^2 = \sum_{k=0}^{N-1} S_{xx}(k) \\ &= \frac{1}{2\pi} \int_0^{2\pi} |X(\omega)|^2 d\omega = \int_{f_n=0}^1 |X(f_n)|^2 df_n = \int_{f_n=0}^1 S_{xx}(f_n) df_n. \end{aligned} \quad (6.28)$$

The frequency variable f_n above is normalized as $f_n = f/f_s$, such that $0 \leq f_n \leq 1$. $S_{xx}(k)$ is the PSD of $x(n)$, given by

$$S_{xx}(k) = \frac{1}{N} |X(k)|^2, \quad (6.29)$$

where $X(k)$ is the DFT of $x(n)$. Note the presence of the factor $\frac{1}{N}$ in the definition of the PSD. An additional factor of $\frac{1}{N}$ appears in the relationship among the DFT frequency index k , the sampling frequency f_s , and the frequency f in Hz, as

$$f = \frac{f_s}{N} k. \quad (6.30)$$

As evident from the equations given above, the area under the PSD curve represents the total signal power or energy. In order to compute moments, it would be suitable to treat the function

being used as a PDF, with the condition that the area under the curve be unity. PSDs of real-valued signals possess even-symmetry about the folding frequency $f_m = f_s/2$ and include two unique values at DC and f_m ; see Section 3.4.5 and Figure 3.38. Therefore, it would be appropriate to compute moments using the portion of the PSD from DC to f_m or $f_n = 0.5$. This portion of the PSD could be normalized to have unit area by dividing by its integral or sum, given by

$$E_p = \int_{f_n=0}^{0.5} S_{xx}(f_n) df_n \quad (6.31)$$

or

$$E_p = \sum_{k=0}^{N/2} S_{xx}(k). \quad (6.32)$$

To obtain a measure of the concentration or spread of the signal power over its frequency range, we may compute the mean frequency \bar{f} as the first-order moment

$$\bar{f}_n = \frac{1}{E_p} \int_{f_n=0}^{0.5} f_n S_{xx}(f_n) df_n \quad (6.33)$$

or as

$$\bar{f}_{Hz} = \frac{f_s}{N} \frac{1}{E_p} \sum_{k=0}^{N/2} k S_{xx}(k), \quad (6.34)$$

where \bar{f}_n and \bar{f}_{Hz} indicate the mean frequency in its normalized form and in Hz, respectively, and N is the number of samples in the DFT-based representation of the PSD. The upper limit of integration of 0.5 represents integration from DC to the maximum frequency present in the signal, which is one-half of the sampling frequency, the frequency variable having been normalized to the range $0 \leq f_n \leq 1$. Note that the integration or summation is performed over one-half period of the periodic function $S_{xx}(f_n)$ or $S_{xx}(k)$, which possesses even-symmetry about one-half of the sampling frequency for real-valued signals.

To facilitate improved comprehension, two versions of each moment are provided in this section: one using the normalized (continuous) frequency variable $0 \leq f_n \leq 1$, and the other using the DFT index $0 \leq k \leq N - 1$.

The median frequency f_{med} is defined as that frequency which splits the PSD in half:

$$f_{\text{med}} = \frac{m}{N} f_s \text{ with the largest } m \text{ such that} \\ \frac{1}{E_p} \sum_{k=0}^m S_{xx}(k) < \frac{1}{2}; \quad 0 \leq m \leq \frac{N}{2}. \quad (6.35)$$

We may also compute higher-order statistics as follows:

- Variance, f_{m2} , is given by the second-order moment using $(f_n - \bar{f}_n)^2$ in place of f_n [the function of frequency that is multiplied with $S_{xx}(f_n)$] in Equation 6.33 or the equivalent expression in k in Equation 6.34, that is,

$$f_{m2} = \frac{1}{E_p} \int_{f_n=0}^{0.5} (f_n - \bar{f}_n)^2 S_{xx}(f_n) df_n \quad (6.36)$$

or

$$f_{m2} = \left(\frac{f_s}{N}\right)^2 \frac{1}{E_p} \sum_{k=0}^{N/2} (k - \bar{k})^2 S_{xx}(k), \quad (6.37)$$

where $\bar{k} = N \overline{f_{Hz}} / f_s$ is the frequency sample index corresponding to $\overline{f_{Hz}}$.

- Skewness is obtained as

$$\text{skewness} = \frac{f_{m3}}{(f_{m2})^{3/2}}, \quad (6.38)$$

where the third-order moment f_{m3} is computed with $(f_n - \overline{f_n})^3$ in place of f_n in Equation 6.33, that is,

$$f_{m3} = \frac{1}{E_p} \int_{f_n=0}^{0.5} (f_n - \overline{f_n})^3 S_{xx}(f_n) df_n \quad (6.39)$$

or

$$f_{m3} = \left(\frac{f_s}{N} \right)^3 \frac{1}{E_p} \sum_{k=0}^{N/2} (k - \bar{k})^3 S_{xx}(k). \quad (6.40)$$

- Kurtosis is defined as

$$\text{kurtosis} = \frac{f_{m4}}{(f_{m2})^2}, \quad (6.41)$$

where the fourth-order moment f_{m4} is computed with $(f_n - \overline{f_n})^4$ in place of f_n in Equation 6.33, that is,

$$f_{m4} = \frac{1}{E_p} \int_{f_n=0}^{0.5} (f_n - \overline{f_n})^4 S_{xx}(f_n) df_n \quad (6.42)$$

or

$$f_{m4} = \left(\frac{f_s}{N} \right)^4 \frac{1}{E_p} \sum_{k=0}^{N/2} (k - \bar{k})^4 S_{xx}(k). \quad (6.43)$$

In order to avoid notational clutter, the same symbols have been used for the two versions each of f_{m2} , f_{m3} , and f_{m4} ; it is understood that similar versions of the moments are used to compute the ratios for skewness and kurtosis.

The mean frequency is a useful measure of the concentration of signal power, and could indicate the resonance frequency in the case of unimodal distributions. However, a nearly uniform PSD could lead to one-half of the maximum frequency as the mean frequency, which by itself may not be a useful representation of the PSD. The presence of multiple resonance frequencies could also lead to a mean frequency that may not be a useful measure. Multimodal distributions of PSDs may be characterized better by a series of peak frequencies, along with measures of their relative levels and bandwidths or quality factors (described in the following section).

The square root of f_{m2} provides a measure of spectral spread (*SD* about the mean) and an indication of the bandwidth (but not at -3 dB) about the mean frequency. The skewness is zero if the density function is symmetric about the mean frequency; otherwise, it indicates the extent of asymmetry of the distribution. Kurtosis indicates if the PSD is a long-tailed function.

Moments of PSDs may be useful in characterizing the general trends in the distribution of the power of a signal over its bandwidth. The higher-order moments are sensitive to noise or spurious variations in the PSD estimate, and they may not yield reliable measures if the PSD pattern is not simple or if the PSD estimate is poor (has a high variance). The reliability of moments may be improved by smoothing the PSD estimate, or by fitting a smooth parametric curve (such as a Gaussian, a Laplacian, or a spline) as a model of the PSD estimate and computing the moments of the model. Saltzberg and Burch [30] discuss the relationship between moments of PSDs and *ZCR*, along with their application to EEG analysis.

6.4.2 Spectral power ratios

The moments described in the preceding section provide general statistical characterization of the PSD treated as a PDF. In the case of analysis of biomedical signals, it may be more advantageous to define specific measures based on *a priori* information or empirical knowledge about the signals, the systems, and the physiological or pathological processes of concern. For example, in the case of PCG analysis for the detection of murmurs, we could specifically investigate the presence of signal power in the frequency range beyond that of S1 and/or S2. If a specific type of pathology of interest is known to cause a shift in the frequency content within a certain band of frequencies, we may measure spectral power ratios over partitions of the band of interest. We have seen in Sections 6.2.1 and 6.2.2 how such measures have been used for the analysis of ventricular elasticity, diagnosis of myocardial infarction, and detection of murmurs.

The fraction of signal power in a frequency band of interest $f_1 : f_2$ may be computed as

$$E_{f_1:f_2} = \frac{2}{E_x} \int_{f=f_1}^{f_2} |X(f)|^2 df = \frac{2}{NE_x} \sum_{k=k_1}^{k_2} |X(k)|^2, \quad (6.44)$$

where k_1 and k_2 are the DFT indices corresponding to f_1 and f_2 , respectively. Fractions of power as above may be computed for several bands of interest that may or may not span the entire signal bandwidth.

In a variation of the fractional-power measure defined above, Johnson et al. [19] compared the integral of the magnitude spectrum of the systolic murmurs due to aortic stenosis over the band 75 : 150 Hz to that over the band 25 : 75 Hz. They considered the higher-frequency band to represent the *predictive area (PA)* of the spectrum related to aortic stenosis and considered the lower-frequency band to represent a *constant area (CA)* that would be common to all systolic PCG signal segments and serve as a reference. The ratio of *PA* to *CA* was defined as

$$\frac{PA}{CA} = \frac{\int_{f=f_2}^{f_3} |X(f)| df}{\int_{f=f_1}^{f_2} |X(f)| df}, \quad (6.45)$$

with $f_1 = 25$ Hz, $f_2 = 75$ Hz, and $f_3 = 150$ Hz. The $\frac{PA}{CA}$ ratio is provided for the PSDs of systolic murmurs of four patients with aortic stenosis in Figure 6.2; Johnson et al. showed that the ratio correlates well with the severity of aortic stenosis.

Binnie et al. [31, 32] described the application of spectral analysis to EEG for the detection of epilepsy. Their method was based on partitioning or banding of the EEG spectrum into not only the traditional δ , θ , α , and β bands, but also into seven other nonuniform bands specified as 1 – 2, 2 – 4, 4 – 6, 6 – 8, 8 – 11, 11 – 14, and $>$ 14 Hz. Additional features related to *FF* (see Section 5.6.4) were also used. In a study with 275 patients with suspected epilepsy, 90% of the signals of the patients with pathology were classified as abnormal by their methods; conversely, 86% of the patients whose EEGs were classified as abnormal had confirmed pathology.

When analyzing a spectral peak, we may also compute the -3 dB bandwidth of the peak and, furthermore, compute its quality factor as the ratio of the peak frequency to the bandwidth. Such measures may be computed for not only the dominant peak but also several peaks at progressively lower levels of signal power. Essentially, each potential resonance peak is treated and characterized as a bandpass filter. Durand et al. [33] used such measures to characterize the PSDs of sounds produced by prosthetic heart valves (discussed in Section 6.5).

6.5 Application: Evaluation of Prosthetic Heart Valves

Efficient opening and closing actions of cardiac valves are of paramount importance for proper pumping of blood by the heart. When native valves fail, they may be replaced by mechanical

prosthetic valves or by bioprosthetic valves extracted from pigs. Mechanical prosthetic valves are prone to sudden failure due to fracture of their components. Bioprosthetic valves fail gradually due to tissue degeneration and calcification, and have been observed to last 7 – 12 years [33]. Follow-up of the health of patients with prosthetic valves requires periodic, noninvasive assessment of the functional integrity of the valves.

Problem: *Deposition of calcium causes the normally pliant and elastic bioprosthetic valve leaflets to become stiff. Propose a method to assess the functional integrity of bioprosthetic valves.*

Solution: Based on the premise that valve opening and closure contribute directly to heart sounds, analysis of PCG components offers a noninvasive and passive approach to evaluation of prosthetic valves. The increased stiffness is expected to lead to higher-frequency components in the opening or closing sounds of the valve. Durand et al. [33] studied the spectra of the entire S1 signal segment to evaluate the sounds contributed by the closure of porcine (pig) bioprosthetic valves implanted in the mitral position in humans. They demonstrated that, whereas normal S1 spectra were limited in bandwidth to about 100 Hz, degenerated bioprosthetic valves created significant spectral energy in the range 100 – 250 Hz. Figure 6.10 shows the spectra of S1 in the case of a normal bioprosthetic valve and a degenerated bioprosthetic valve; increased high-frequency amplitude is evident in the latter.

Durand et al. derived several parameters from S1 spectra and used them to discriminate normal from degenerated bioprosthetic valves. Some of the parameters used by them are the first and second dominant peak frequencies; the bandwidth and quality factor of the dominant peak; integrated mean area above -20 dB; the highest frequency found at -3 dB; total area and *RMS* value of the spectrum; area and *RMS* value in the 20 – 100 Hz, 100 – 200 Hz, and 200 – 300 Hz bands; and the median frequency. Normal-versus-degenerated valve classification accuracies as high as 98% were achieved.

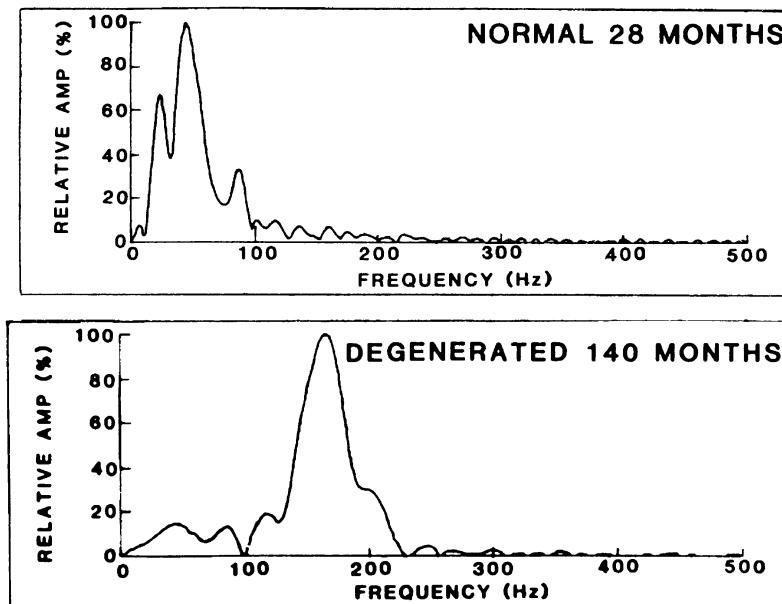


Figure 6.10 First heart sound spectra in the case of normal and degenerated porcine bioprosthetic valves implanted in the mitral position. Reproduced with permission from L.G. Durand, M. Blanchard, G. Cloutier, H.N. Sabbah, and P.D. Stein, Comparison of pattern recognition methods for computer-assisted classification of spectra of heart sounds in patients with a porcine bioprosthetic valve implanted in the mitral position, *IEEE Transactions on Biomedical Engineering*, 37(12):1121–1129, 1990. ©IEEE.

Durand et al. [34] also studied the sounds of bioprosthetic valves in the aortic position. They argued that the aortic and pulmonary components (A2 and P2, respectively) of S2, each lasting about 50 ms, are not temporally correlated during normal breathing. The two components of S2 are separated by 30 – 60 ms during inspiration, but get closer and could overlap during expiration. Furthermore, P2 is weaker than A2 if the PCG is recorded in the aortic area. Thus, P2 may be suppressed, and A2 strengthened by coherent detection and averaging of S2 over several cardiac and breath cycles; see Section 4.10. Durand et al. performed spectral analysis of A2 extracted as above for the purpose of evaluation of bioprosthetic valves in the aortic position. Among a selection of spectral analysis methods including the basic periodogram, Welch's averaged periodogram, all-pole modeling (see Section 7.5), and pole-zero modeling (see Section 7.6), they found the basic periodogram to provide the best compromise for the estimation of both the spectral distribution and the dominant frequency peaks of bioprosthetic valve sounds.

Cloutier et al. [35] studied the bias and variability of several diagnostic spectral parameters computed from simulated closing sounds of bioprosthetic valves in the mitral position. They found that the most-dominant spectral peak frequency and its quality factor were best estimated using an FFT-based PSD estimate with a rectangular window. However, the -3 dB bandwidth of the most-dominant spectral peak, the frequency of the second-dominant peak, and a few other parameters were best estimated by the Steiglitz–McBride method of pole-zero modeling (see Section 7.6.2). Some other parameters were best estimated by all-pole modeling using the covariance method (see Section 7.5). It was concluded that a single method would not provide the best estimates of all possible spectral parameters of interest.

6.6 Application: Fractal Analysis of VAG Signals

Problem: *Knee-joint VAG signals exhibit varying levels of waveform complexity depending upon the state of the cartilage covering the joint surfaces. Given that fractals demonstrate complex patterns depending on certain parameters, can fractal analysis assist in parametric representation and classification of VAG signals?*

Solution: As explained in Sections 1.2.14 and 5.12, normal knee joints with smooth cartilage surfaces produce little sound or vibration. However, when the cartilage surfaces are eroded or damaged due to pathological conditions, additional sounds could be generated. Rangayyan and Oloumi [36] and Rangayyan et al. [37] explored the use of FD as a parameter for screening of VAG signals. Given the nonstationary nature of VAG signals, multiple FD values were derived from the PSDs of segments of VAG signals. The related background and methods are described in the following paragraphs.

6.6.1 Fractals and the $1/f$ model

One of the several definitions of fractals is based on geometric structures that have self-similarity at different scales of length or size [38–40]; see Section 5.13. Fractals lack any single scale of length or size, and have a noninteger (fractional) dimension, referred to as FD . Fractal geometry has been used to represent and synthesize several natural forms such as leaves (especially ferns), trees, mountains, and clouds. Fractal analysis forms a part of advanced techniques for nonlinear analysis of biomedical signals [41–43]; see Section 5.13.

Wiener's geometric model of physical Brownian motion has been used as the basis for algorithms to generate fractal signals [40]. In this model, the unpredictable variation of a quantity, V , over time, t , is viewed as a noise process, $V(t)$. The PSD, $P_V(f)$, is used to estimate the power of fluctuations at a given frequency, f , and also of the variations over a time scale of the order of $1/f$.

A time-varying quantity, $V(t)$, with the best-fitting line to its PSD, $P_V(f)$, varying as $1/f^\beta$ on a log-log scale, is referred to as $1/f$ noise; this is known as the inverse power law or the $1/f$ model.

According to Voss [40], most natural phenomena under this model have β in the range of $[0.5, 1.5]$. The PSD of a noise process represented by Brownian motion or random walk varies as $1/f^2$; the power decreases quadratically with frequency. The trace of such a signal is a fractal curve. A direct relationship exists between the *FD* of the signal and the slope, β , of the best-fitting line to its PSD on a log–log scale [40].

The fractional Brownian motion (fBm) model of Mandelbrot [38] has formed the basis of several mathematical models for computer generation of natural fractal scenery. Models based on fBm have been extended to 2D for the synthesis of Brownian surfaces and 3D to generate Brownian clouds [39].

Figure 6.11 shows examples of signals, $V_H(t)$, generated as functions of an arbitrary time variable, t , based on the fBm model. The scaling of the traces is characterized by the scaling parameter H , known as the Hurst coefficient, in the range $0 \leq H \leq 1$. A high value of H close to 1 results in a relatively smooth signal. A lower value of H produces a rougher trace. The variable H relates the changes in V , $\Delta V = V(t_2) - V(t_1)$, to differences in the time variable, $\Delta t = t_2 - t_1$, by the scaling law [40]

$$\Delta V \propto (\Delta t)^H. \quad (6.46)$$

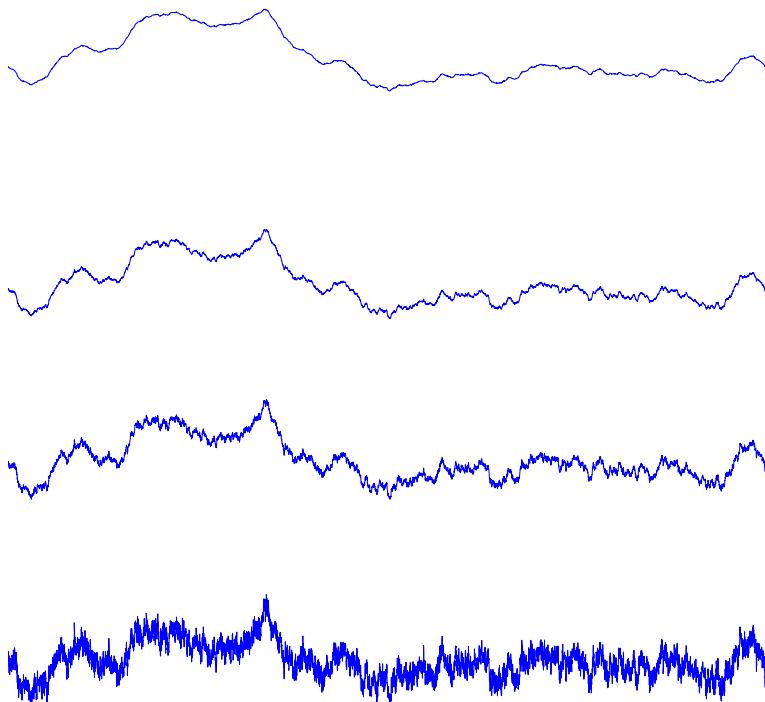


Figure 6.11 Examples of signals generated based on the fBm model for different values of H and *FD*. Top to bottom: $H = 0.9, 0.6, 0.4, 0.1$; model *FD* = 1.1, 1.4, 1.6, 1.9; estimated *FD* = 1.103, 1.401, 1.600, 1.898. Reproduced with permission from R.M. Rangayyan and F. Oloumi, Fractal analysis of knee-joint vibroarthrographic signals, Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB 2010), Corfu, Greece, November 2010, pp 1–4. ©IEEE.

Self-similar patterns repeat themselves at various levels of magnification. However, fBm traces repeat statistically only when t and V are magnified by different amounts. If t is magnified by a factor r as rt , the value of V is magnified by the factor r^H and becomes $r^H V$. Nonuniform scaling of this nature is known as self-affinity [38]. The zero-set of an fBm signal is given by the

intersections of the signal with the horizontal axis. The zero-set is a set of disconnected points with a topological dimension of zero and fractal dimension, D_0 , given by [40]

$$D_0 = 1 - H. \quad (6.47)$$

The zero-set of a self-affine signal is self-similar; different estimates of D_0 will yield the same result. The FD of the signal is related to D_0 as

$$FD = D_0 + 1 \quad (6.48)$$

and to the scaling parameter, H , as

$$FD = 2 - H. \quad (6.49)$$

6.6.2 FD via power spectral analysis

The best method available to estimate the FD of a self-affine signal is power spectral analysis (PSA). As explained in Section 6.6.1, an fBm signal has a PSD that follows the $1/f^\beta$ model. A high value of β indicates a rapid decrease in the high-frequency content of the signal. A self-affine fBm function in an E -dimensional Euclidean space has its PSD $P_V(f) \propto 1/f^\beta$, with [40]

$$FD = E + 1 - H, \quad (6.50)$$

where

$$H = \frac{\beta - 1}{2}. \quad (6.51)$$

The FD , in terms of the spectral component, β , for a 1D signal with $E = 1$, is

$$FD = \frac{5 - \beta}{2}. \quad (6.52)$$

In the studies of Rangayyan and Oloumi [36] and Rangayyan et al. [37], to estimate β , a Hann window was first applied to the given VAG signal; then, the DFT of the windowed signal was calculated, and the squared magnitude of the result was used to obtain the PSD of the signal. The slope of the best-fitting line to the log-log plot of PSD was then determined. See Chang et al. [44] for related discussions.

When applying the PSA method, it is important to specify an appropriate frequency range to obtain the linear fit. Low-frequency details related to any slow drift present in the signal as well as high-frequency content dominated by noise or artifacts in the signal should be disregarded in order to obtain accurate estimates of β and FD . In the works of Rangayyan and Oloumi [36] and Rangayyan et al. [37], when applying the PSA method to the synthesized fBm signals, PSD components spanning 1% of the total frequency range were removed at both the low and high ends of the positive frequency axis. For VAG signals, the frequency range to obtain the linear fit was varied between [10, 300] Hz and [10, 380] Hz to study the effect of the frequency range used on classification accuracy. This was based on experimentation and analysis of the band of frequencies over which appreciable differences were observed between normal and abnormal VAG signals. (The VAG signals had been prefiltered to the range 10 Hz to 1 kHz at the time of data acquisition; therefore, there is no useful information below 10 Hz.)

6.6.3 Examples of synthesized fractal signals

Two types of methods are available to generate fBm signals with known FD : approximation by spatial methods and approximation by spectral synthesis. The most reliable method is spectral

synthesis by means of inverse Fourier filtering [45]. If a_k is the k^{th} complex coefficient of the DFT, to obtain $P(f) \propto 1/f^\beta$, the condition is

$$E[|a_k|^2] \propto \frac{1}{k^\beta}, \quad (6.53)$$

where k denotes the frequency index corresponding to f . By randomly selecting coefficients that satisfy the stated condition and then taking their inverse DFT, we can obtain the corresponding signal in the time domain. In the works of Rangayyan and Oloumi [36] and Rangayyan et al. [37], a direct implementation of the algorithm given by Saupe [45] was used to generate 110 synthetic signals, 10 for each value of H , with $0 \leq H \leq 1$ in intervals of 0.1. Figure 6.11 illustrates four examples of the synthesized signals. It is evident from the illustration that the variability and complexity of the signals increase as H decreases (or as FD increases).

The estimated FD values for the four synthesized signals shown in Figure 6.11 are given in the caption of the figure. The RMS error between the known and estimated values of FD for the 110 synthesized signals was computed to be 0.0198, which indicates accurate estimation of FD by the PSA method. The PSA method performed well with the synthesized signals for FD in the range of [1.1, 1.8], but overestimated FD values in the range [1.9, 2.0].

Estimates of FD were also obtained for the synthesized fBm signals using the well-known and popular box-counting and ruler methods [38, 46, 47] (see Section 5.13); the two methods led to slightly poorer results, with higher RMS errors of 0.1387 and 0.2243, respectively. The results indicate that the PSA method is the best suited method for the estimation of FD of fBm and self-affine signals, among the three methods mentioned.

6.6.4 Fractal analysis of segments of VAG signals

The dataset used in the studies of Rangayyan and Oloumi [36] and Rangayyan et al. [37] consists of 89 VAG signals; see Sections 1.2.14 and 5.12.1 for details. Examples of a normal signal and an abnormal signal are shown in Figures 6.12 and 6.13. Each signal was normalized to the amplitude range [0, 1] and resampled to the length of 8,192 samples by linear interpolation.

It is known that the VAG signal is nonstationary, and that it is appropriate to analyze segments of fixed or adaptive duration [48–52] or apply nonstationary signal processing techniques such as wavelets [53] and TFDs [54]. However, such methods increase the computational load. Furthermore, it is difficult, in practice, to associate parts of the knee joint affected by pathology to segments of VAG signals. The use of VAG signal segments of limited duration may lead to inaccuracies in the estimation of the PSD in low-frequency ranges. On the other hand, it has been observed that parts of VAG signals during certain portions of the swing cycle, especially extension, can provide features with higher discriminant capability than other parts [48, 50–52]. In the works of Rangayyan and Oloumi [36] and Rangayyan et al. [37], to facilitate separate analysis of parts of VAG signals during extension and flexion and to derive multiple features, the PSA method was applied not only to get a parameter for the full VAG signal in each case ($FD1$), but also to the first and second halves of the normalized duration ($FD1h$ and $FD2h$, corresponding to extension and flexion), as well as to four segments with each spanning one quarter of the total duration ($FD1q$, $FD2q$, $FD3q$, and $FD4q$).

Figures 6.12 and 6.13 show the spectra and the linear fits derived to estimate the $FD1$ values of the normal and abnormal VAG signals shown in the same figures. The estimated $FD1$ values for the two signals are given in the captions of Figures 6.12 and 6.13. The average and SD values of $FD1$ for the 51 normal signals were 1.8061 ± 0.2398 ; those for the 38 abnormal VAG signals were 1.6695 ± 0.2226 (using the frequency range [10, 300] Hz). The discriminant capability of the FD values was assessed in terms of the area, A_z , under the receiver operating characteristic (ROC) curve using ROCKIT [55]. The FD values obtained by the PSA method applied to the full VAG signals led to $A_z = 0.6872$.

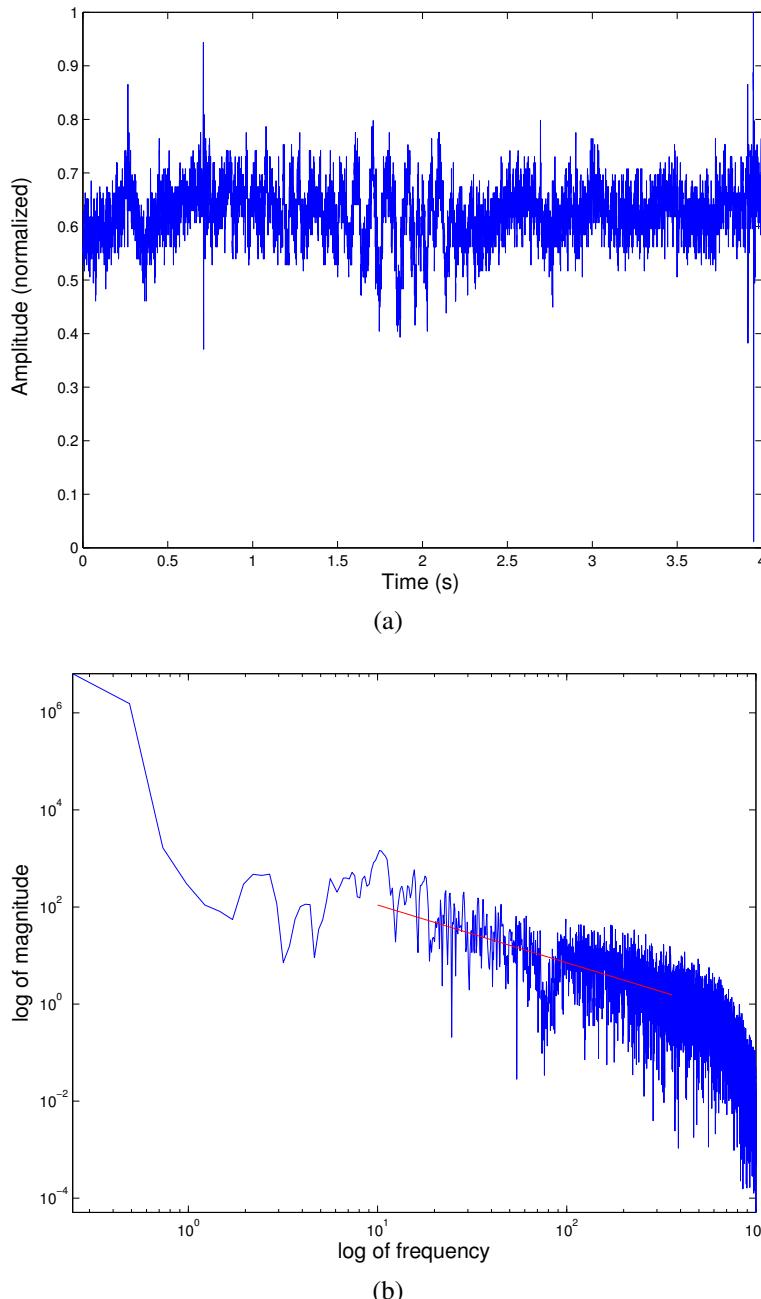


Figure 6.12 (a) Example of a normal VAG signal. (b) Spectrum of the signal with the straight-line fit to the range $[10, 360]$ Hz. $FD_1 = 1.905$. Reproduced with permission from R.M. Rangayyan and F. Oloumi, Fractal analysis of knee-joint vibroarthrographic signals, Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB 2010), Corfu, Greece, November 2010, pp 1–4. ©IEEE.

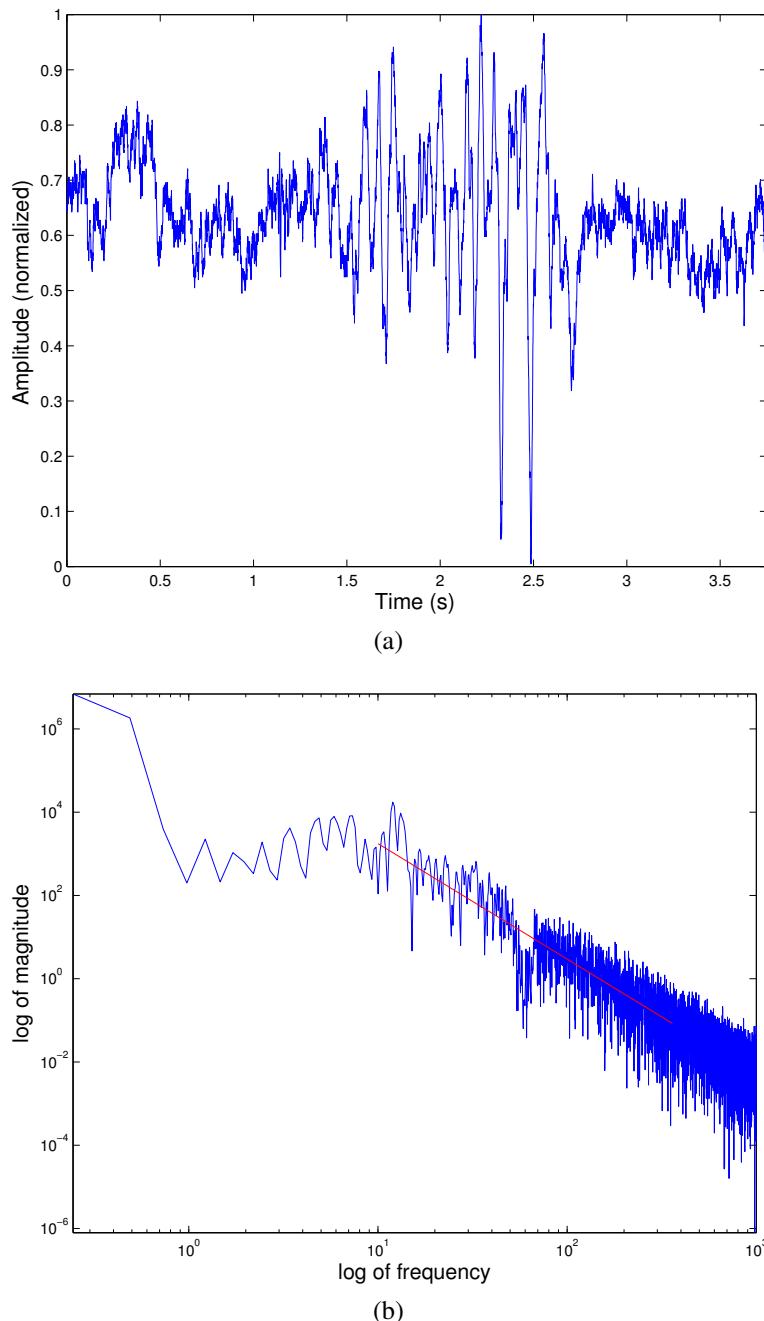


Figure 6.13 (a) Example of an abnormal VAG signal. (b) Spectrum of the signal with the straight-line fit to the range [10, 360] Hz. $FD_1 = 1.113$. Reproduced with permission from R.M. Rangayyan and F. Oloumi, Fractal analysis of knee-joint vibroarthrographic signals, Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB 2010), Corfu, Greece, November 2010, pp 1–4. ©IEEE.

The values of FD estimated for the first half (extension) and second half (flexion) of each signal led to poorer A_z values of 0.6133 and 0.5916, respectively. With the FD values derived using quarter portions of the VAG signals, the four parts, in order, led to $A_z = 0.6546, 0.7394, 0.5959$, and 0.7023 (using the frequency range [10, 380] Hz). Based on the p -values obtained via the t -test, it was noted that the differences between the values of $FD1$, $FD2q$, and $FD4q$ were statistically highly significant ($p < 0.01$), whereas the differences in the $FD1q$ values for the normal and abnormal categories were statistically significant ($0.01 < p < 0.05$). Thus, the usefulness of some of the FD -based parameters in discriminating between normal and abnormal VAG signals was demonstrated.

The FD values for abnormal VAG signals were observed to be, on the average, lower than those for normal signals, which was interpreted as follows. The presence of defects in the articular surfaces of the knee joint due to loss or shedding of cartilage leads to structured vibration components that disrupt or replace the random, fBm-like signals associated with the normal friction-free cartilage surfaces. Although the VAG signals of some normal knees could possibly contain nonrandom components, such as clicks and crepitus, they would be fewer and occur less often than the grinding noise components found in abnormal knee joints. The disruption of the fBm-like characteristics of normal VAG signals by pathological conditions may be expected to result in lower FD values for abnormal VAG signals via the PSA approach and the $1/f$ model. The range of the FD values of the abnormal VAG signals obtained in the study of Rangayyan et al. [37] agrees with the ranges of FD of acceleration signals obtained from finger joints of patients with calcium pyrophosphate deposition disease, rheumatoid arthritis, or spondyloarthropathy in the study of Shah et al. [56] (see Section 5.13.3).

The classification performance of the features derived via fractal analysis, with the highest A_z value of about 0.74 in ROC analysis, was noted to be comparable to the performance provided by several features obtained by predictive modeling, cepstral analysis, measures of variability of power, turns count, probabilistic models, wavelets, and TFDs [49–54]. Selected combinations of FD with some of the other features mentioned above led to better results with A_z values in the range [0.92, 0.96]. See Rangayyan et al. [37] for further details and additional results with various feature selection and pattern classification procedures.

6.7 Application: Spectral Analysis of EEG Signals

Problem: It is known that the EEG signal exhibits shifts towards lower frequencies as the subject goes into deeper stages of sleep [57, 58]. Propose methods to obtain quantitative measures to characterize the spectral content of EEG signals and relate them to stages of sleep.

Solution: As shown in Section 1.2.6, EEG signals exhibit rhythms in certain frequency bands. A subject with eyes closed and awake will typically demonstrate alpha waves in the EEG. As the subject goes to sleep, the EEG rhythms are expected to shift to lower bands of theta and delta rhythms (see Figure 1.40). Episodic EEG activities related to rapid eye movement (REM) and sleep spindles (see Figure 4.1) could contribute power in frequency bands that are higher than the theta and delta bands. Overnight PSG studies commonly include several channels of the EEG signal that are used to establish the stage of sleep by either manual (visual) analysis or computer-aided analysis [57]. The amount of low-frequency power in the delta and theta bands in relation to the power in the alpha band is a useful feature in sleep staging [57].

Figure 6.14 illustrates three traces of the C3–A2 EEG channel of a subject during sleep in stages 0, 1, and 2. (Stage 0 indicates that the subject is awake.) The EEG signal was analyzed by an expert who assigned sleep stages to 5,988-sample segments, corresponding to intervals of 29.94 s with $f_s = 200$ Hz [57]. Regardless of the noise present in the signals, it is evident that, in general, the signals shift to increasing power in lower frequencies as the sleep stage progresses from 0 to 1 to 2.

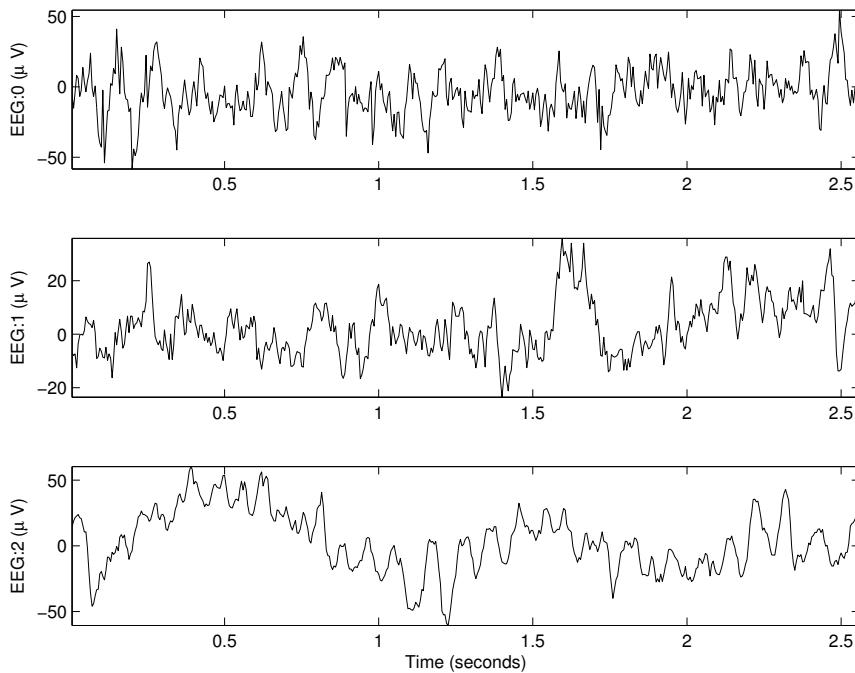


Figure 6.14 Top to bottom: Segments of the C3–A2 EEG channel of a subject during sleep in stages 0, 1, and 2. See also Figure 6.15. EEG data courtesy of R. Agarwal [57].

In order to obtain estimates of the PSDs of the EEG signals, the Welch method of windowing signal segments and averaging their PSDs was applied to 5,988-sample segments of the EEG signal. Nonoverlapping segments of length 512 samples were obtained and the Hann window was applied before computing the DFT. The magnitude spectra of 10 segments were averaged. Figure 6.15 illustrates the average log-magnitude PSDs of three EEG segments corresponding to sleep stages 0, 1, and 2 (top to bottom). Sample traces of the same EEG segments with durations of 512 samples are shown in Figure 6.14. The PSDs clearly demonstrate a reduction in the power of the EEG signals in the higher parts of the range of 0 – 25 Hz plotted in the figure. The EEG signal for stage 0, which represents a wakeful state with the eyes closed (presumably), shows a clear peak at about 8 Hz corresponding to the alpha rhythm. No peak in the alpha range is seen in the remaining two PSDs, which also show a reduction in high-frequency power. Peaks at lower frequencies (in the delta and theta bands) are evident in the PSDs of the EEG for sleep stages 1 and 2.

The mean frequencies of the PSDs were computed according to Equation 6.34 over the band of [0, 100] Hz. For the PSDs illustrated in Figure 6.15, the mean frequencies are 20.21, 10.09, and 3.98 Hz, respectively, indicating a clear downward trend over the sleep stages of 0, 1, and 2. Lowpass filtering the signals to remove noise and reducing the frequency range used to compute the mean frequency could lead to lower values.

Instead of computing a single mean frequency value over an entire PSD, one could also compute fractions of the total power in several frequency bands, as illustrated in Figure 6.2 and in the discussion in Section 6.4.2. One of the features used by Agarwal and Gotman [57] is the ratio of power in the alpha band (defined by them as 8.0 to 11 Hz) to the combined power in the delta (0.5 to 3.5 Hz) and theta (3.5 to 8.0 Hz) bands; this was referred to as the alpha-to-slow-wave index. The values of this index for the PSDs illustrated in Figure 6.15 are 1.68, 0.21, and 0.09, which indicate the expected reducing trend from stage 0 to 1 to 2, representative of the fact that alpha waves are replaced by slower theta and delta waves.

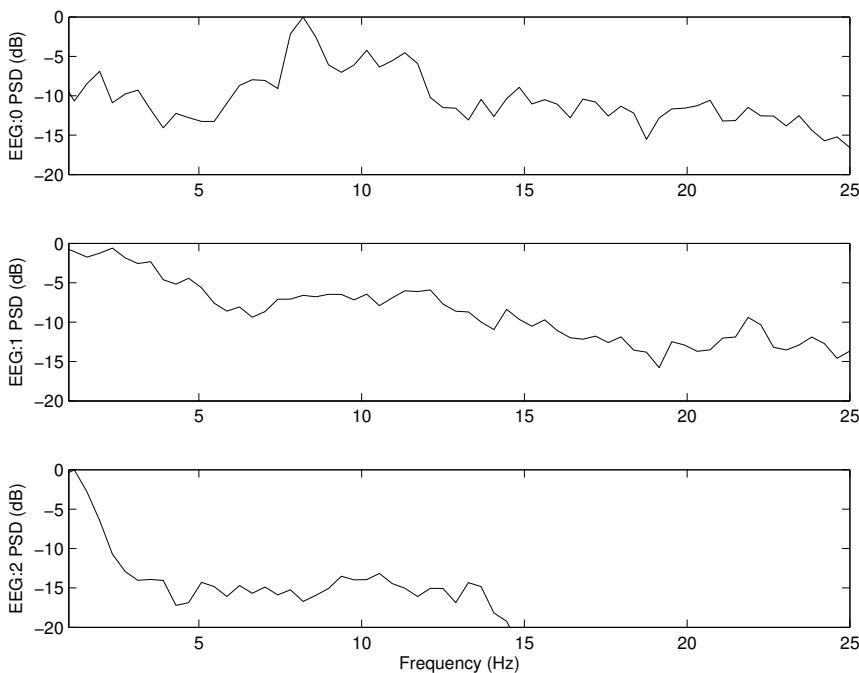


Figure 6.15 Top to bottom: Average PSDs of segments of the C3–A2 EEG channel of a subject during sleep in stages 0, 1, and 2. Figure 6.14 shows sample segments used in the derivation of the PSDs.

Figures 6.16 and 6.17 show cluster plots of the mean frequency and alpha-to-slow-wave index values for 702 segments of an overnight sleep EEG record with sleep stages of 0, 1, and 2. (Each segment has 5,988 samples, corresponding to an interval of 29.94 s with $f_s = 200 \text{ Hz}$. The total EEG record has a duration of about 6 h and 50 min.) Each “ \times ” mark in the plots represents the value of the parameter and the corresponding sleep stage for one EEG segment. In general, the two parameters show trends of decreasing values as the sleep stage varies from 0 to 1 to 2. However, both parameters show substantial variance and overlap in their ranges for the three stages of sleep. It should be noted that EEG signals of a subject who is awake could contain substantial spectral variations. The presence of alpha waves in a stage-0 segment could cause the alpha-to-slow-wave index to be higher than that for a segment without alpha: This is indicated by the presence of multiple clusters of the index values for stage 0 in Figure 6.17.

Figure 6.18 illustrates a cluster plot of the mean frequency and alpha-to-slow-wave index values combined into a 2D vector for each EEG segment. The diamond marks for the feature vectors of stage 2 segments have formed a tight cluster with low values for both features, as expected. However, the “ \circ ” and “ $+$ ” marks representing the features for EEG segments corresponding to stages 0 and 1 overlap substantially. As mentioned in the preceding paragraph, stage-0 EEG segments could have a broad spread of the parameters depending on the presence or absence of alpha; Figure 6.18 indicates the presence of two broad clusters for stage 0. Stage-1 EEG segments are also known to possess various types of waves and a broad spread of spectral features, as seen in Figure 6.18. These characteristics of EEG signals create difficulties in manual labeling of sleep stages and in automatic staging with a small number of parameters. The results shown in the present section indicate the need for additional features or measures to separate EEG segments of sleep stages 0 and 1.

EEG segments corresponding to sleep stages 3 and 4 were not present in the entire sleep record of the subject in the present case. Such segments could be expected to possess lower values for the two spectral parameters described in the present section.

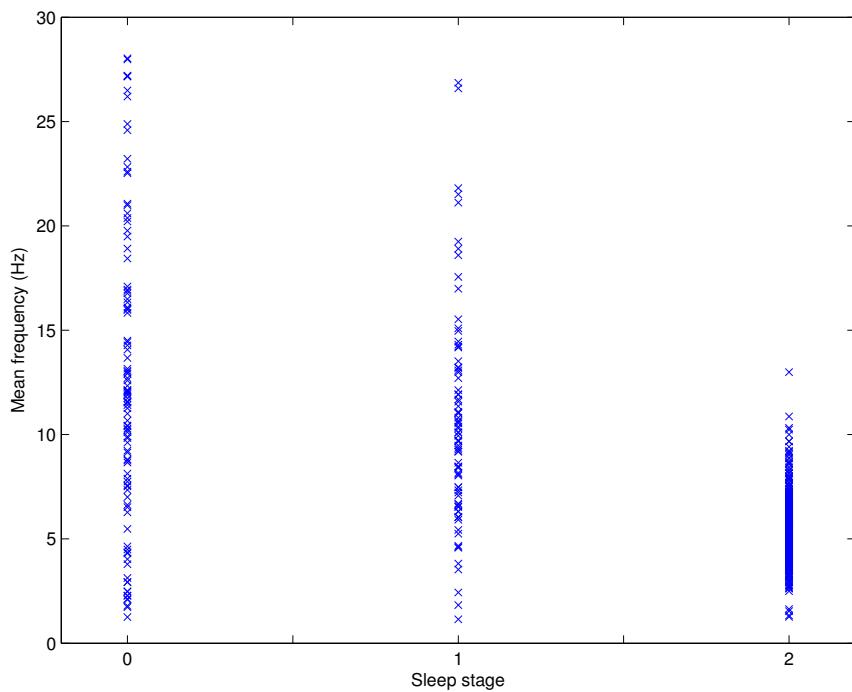


Figure 6.16 Cluster plot of mean frequency values for 702 EEG segments with sleep stages of 0, 1, and 2.

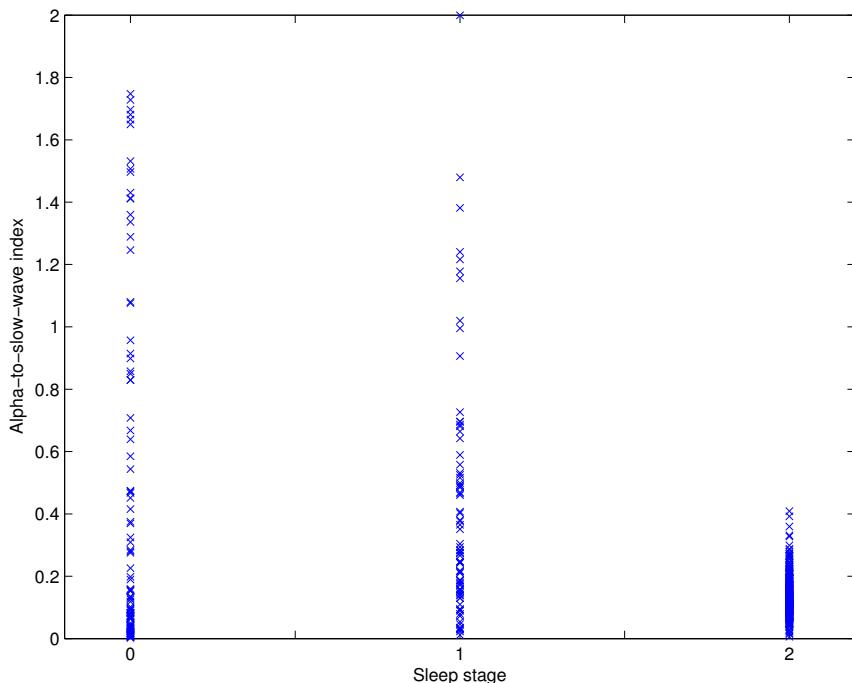


Figure 6.17 Cluster plot of the alpha-to-slow-wave index for 702 EEG segments with sleep stages of 0, 1, and 2.

Notwithstanding the results demonstrated in this section, it should be noted that staging of sleep requires several parameters derived from not only multiple channels of the EEG signal but also additional channels of the EOG and EMG signals; see Agarwal and Gotman [57] for descriptions of several measures for this application.

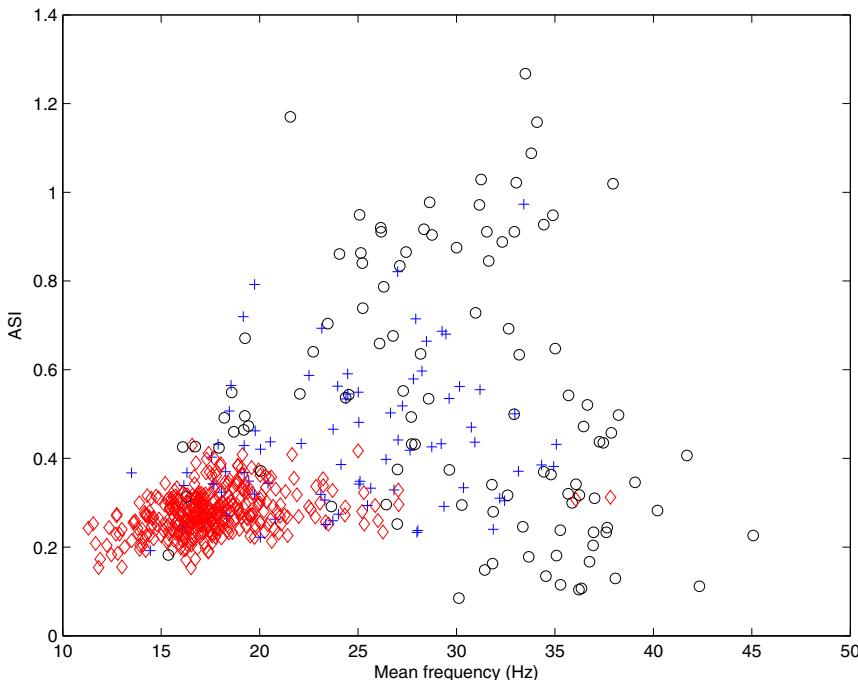


Figure 6.18 Cluster plot of the mean frequency and alpha-to-slow-wave index (ASI) for 702 EEG segments with sleep stages of 0 (“o” mark), 1 (“+” mark), and 2 (diamond mark).

6.8 Remarks

We have investigated the frequency-domain characteristics of a few biomedical signals and the corresponding physiological systems, with particular attention to the PCG and the cardiovascular system. Frequency-domain analysis via PSDs and parameters derived from PSDs can enable us to view the signal from a different perspective than the time domain. Certain signals such as the PCG and EEG may not lend themselves to easy interpretation in the time domain and, therefore, may benefit from a move to the frequency domain.

PSDs and their parameters facilitate investigation of the behavior of physiological systems in terms of rhythms, resonance, and parameters that could be related to the physical characteristics of anatomical entities (for example, the loss of elasticity of the myocardial muscles due to ischemia or infarction, the extent of aortic valvular stenosis, or the extent of calcification and stiffness of bioprosthetic valves). Pathological states may also be derived or simulated by modifying the spectral parameters or representations of the corresponding normal physiological states and signals.

It is worthwhile to pause at this stage of our study, and recognize the importance of the topics presented in the preceding chapters. A good understanding of the physiological systems that produce the biomedical signals we deal with, as well as of the pathological processes that alter their characteristics, is of paramount importance before we may process the signals. Preprocessing the signals to remove artifacts and detect events is essential before we may derive parameters to facili-

tate their analysis in the time and/or frequency domains. The design of biomedical signal analysis techniques requires a thorough understanding of the characteristics and properties of the biomedical systems behind the signals, in addition to detailed knowledge of mathematical principles, statistical analysis, computer techniques, and DSP algorithms.

The following chapters present advanced techniques for modeling, nonstationary analysis, adaptive analysis, adaptive decomposition, and pattern classification of biomedical signals.

6.9 Study Questions and Problems

1. The impulse response of a filter is specified by the series of sample values $\{3, 1, -1, 0\}$. (a) What will be the response of the filter to the input whose sample values are $\{4, 4, 2, 1\}$? (b) Is the filter response obtained by linear convolution or circular convolution of the input with the impulse response? (c) What will be the response with the type of convolution other than the one you indicated as the answer to the questions above? (d) How would you implement convolution of the two signals listed above using the FFT? Which type of convolution will this procedure provide? How would you get the other type of convolution for the signals in this problem via the FFT-based procedure?
2. A conjugate-symmetric (even) signal $x_e(n)$ is defined as a signal with the property $x_e(n) = x_e^*(-n)$. A conjugate-antisymmetric (odd) signal $x_o(n)$ is defined as a signal with the property $x_o(n) = -x_o^*(-n)$. An arbitrary signal $x(n)$ may be expressed as the sum of its conjugate-symmetric and conjugate antisymmetric parts as $x(n) = x_e(n) + x_o(n)$, where $x_e(n) = \frac{1}{2}[x(n) + x^*(-n)]$ and $x_o(n) = \frac{1}{2}[x(n) - x^*(-n)]$. Prove that $FT[x_e(n)] = \text{real}[X(\omega)]$, and $FT[x_o(n)] = j\text{imag}[X(\omega)]$, where $FT[x(n)] = X(\omega)$, and FT stands for the Fourier transform [21].
3. A signal $x(t)$ is transmitted through a channel. The received signal $y(t)$ is a scaled, shifted, and noisy version of $x(t)$ given as $y(t) = \alpha x(t - t_0) + \eta(t)$, where α is a scale factor, t_0 is the time delay, and $\eta(t)$ is noise. Assume that the noise process has zero mean and is statistically independent of the signal process, and that all processes are stationary. Derive expressions for the PSD of $y(t)$ in terms of the PSDs of x and η [59, 60].
4. Consider a continuous-time sinusoidal signal of frequency 10 Hz. (a) Derive an analytical expression for the ACF of the signal. (b) Draw a schematic plot of the ACF, including detailed labeling of the time axis. (c) State the relationship of the PSD to the ACF. (d) Derive an analytical expression for the PSD of the given signal. (e) Draw a schematic plot of the PSD, including detailed labeling of the frequency axis.
5. Two real signals $x_1(n)$ and $x_2(n)$ are combined to form a complex signal defined as $y(n) = x_1(n) + jx_2(n)$. Derive a procedure to extract the DFTs $X_1(k)$ and $X_2(k)$ of $x_1(n)$ and $x_2(n)$, respectively, from the DFT $Y(k)$ of $y(n)$.
6. Distinguish between ensemble averages and temporal (time) averages. Identify applications of first-order and second-order averages of both types in analysis of PCG signals in the time and frequency domains.
7. Propose a procedure to process PCG signals to identify the possible presence of murmurs due to aortic stenosis.
8. Propose algorithms in the time and frequency domains to detect the presence of the alpha rhythm in an EEG signal. Propose extensions to the algorithms to detect the joint presence of the same rhythm in four simultaneously recorded EEG channels.
9. Give an equation to define the mean frequency (centroidal frequency) of a PSD function. Explain the role of each item of your equation. Explain how you would implement the procedure to obtain the mean frequency of a signal. Draw schematic sketches of two PSD functions and indicate their approximate mean frequencies. Explain the difference between the two examples.
10. Using continuous-time and continuous-frequency representation, the fraction of the energy of a signal $x(t)$ in the frequency band $[f_1 : f_2]$ Hz is given by

$$E_{f_1:f_2} = \frac{\int_{f_1}^{f_2} |X(f)|^2 df}{\int_0^\infty |X(f)|^2 df}.$$

Here, $X(f)$ is the Fourier transform of the signal $x(t)$ being processed. A researcher obtains a sampled version $x(n)$ of the signal $x(t)$ with the sampling frequency $f_s = 1,000 \text{ Hz}$. The number of samples in the signal is $N = 1,900$. The researcher then pads the signal with zeros to increase the length to $M = 2,048$ samples and applies the FFT routine to obtain the DFT $X(k)$ of the signal. The researcher wants to compute the fractions of the energy of the signal in the frequency bands $[0 : 50] \text{ Hz}$, $[51 : 100] \text{ Hz}$, and $[101 : 400] \text{ Hz}$. Help the researcher by writing an algorithm to compute the three fractions. Ensure that you give the index of the DFT array for each frequency listed above. Sketch a schematic representation of the spectrum $X(k)$ over the $M = 2,048$ samples in the DFT array. Indicate all of the frequency bands required to compute the three fractions of energy, in terms of the frequency in Hz and the index of the DFT array.

11. A researcher in biomedical signal processing wishes to compute averaged estimates of the PSD for the systolic parts and diastolic parts of PCG signals. The related ECG and carotid pulse (CP) signals are available. The researcher also wants to derive quantitative measures from the averaged PSD. Help the researcher with the following: (a) Give a step-by-step algorithm to identify the beginning and end points of the systolic part and the diastolic part for each cardiac cycle in a given PCG signal. Explain how the ECG and CP signals may be used for this purpose. No equation is required for this part. (b) Give a step-by-step algorithm to compute the PSD of each systolic or diastolic segment and to derive the averaged PSD for systole and diastole over an entire PCG signal including several cardiac cycles. (c) Explain the notions of cyclostationary signals and synchronized averaging. Relate these concepts to the PCG, ECG, and CP signals and the procedure to obtain averaged PSDs as above. (d) Given the averaged PSD, $S(k)$, $k = 1, 2, \dots, N$, where N is the number of the samples used in the computation of the DFT, explain how the mean frequency of the PSD may be computed. Give an equation and explain each term or variable used.
12. A student new to the area of biomedical signal processing wishes to develop methods to detect the presence of murmurs in PCG signals. The student wishes to derive quantitative measures from segments of the PCG signal to facilitate characterization of murmurs. Help the student with the following: (a) Explain the notion of waveform complexity. Give equations to compute the mean, the variance, and the form factor (FF) of a signal $x(n)$ with N samples. Explain how FF may be expected to differ between normal heart sounds and murmurs. (b) Give a step-by-step algorithm to compute the PSD of a signal. Include mathematical formulas to compute the DFT of the signal and to obtain the PSD from the DFT. (c) Give typical bandwidths of normal heart sounds and murmurs. Draw schematic diagrams of the PSDs of normal heart sounds and murmurs, and explain the differences between them. (d) Give a step-by-step procedure to compute the ratio of the power of a given signal in a particular band of frequencies $[f_1, f_2]$ to the total power of the signal. (e) Assume that a signal segment has $N = 500$ samples, that the DFT is computed using the same number of samples, and that the sampling rate is $1,000 \text{ Hz}$. For the purpose of detecting murmurs in PCG signals, give a typical range, in Hz , for the band $[f_1, f_2]$ mentioned above. For the same range of frequencies, give the range of the DFT samples to be used to compute the ratio of power mentioned above. Explain how this measure could be used to detect murmurs and to distinguish murmurs from normal heart sounds.

6.10 Laboratory Exercises and Projects

Note: Data files related to the exercises are available at the site

<https://github.com/srikrishnan1972/Biomedical-Signal-Analysis>

1. Using MATLAB®, prepare a signal that contains the sum of two cosine waves of equal amplitude at 40 Hz and 45 Hz . Let the sampling rate be 1 kHz . (a) Compute the power spectrum of the signal with a rectangular window of duration 2 s . (b) Compute the power spectrum of the signal with a Hamming window of duration 2 s . (c) Compute the power spectrum of the signal with a rectangular window of duration 0.5 s . (d) Compute the power spectrum of the signal with a Hamming window of duration 0.5 s . To obtain the power spectrum, you may take the FFT and square the absolute value of the result. Compare the spectra obtained in parts (a)–(d) and comment upon their similarities and/or differences. In order to

visualize the differences clearly, use 2,048-point FFTs and plot the logarithm of the magnitude-squared spectra with an expanded scale from 0 to 100 Hz only. Be sure to label the frequency axis in Hz !

What should the ideal spectrum look like?

2. Two VAG signals are given in the files vag1.dat and vag2.dat (see also the file vag.m). The sampling rate is 2 kHz . Obtain and plot their PSDs using MATLAB®. Label the frequency axis in Hz !

Compute the mean frequency as the first moment of the PSD for each signal. Compute also the variance frequency (second central moment) of each PSD. What are the units of these parameters?

Compare the spectra and the parameters derived and give your evaluation of the frequency content of the signals.

3. The file safety.wav contains the speech signal for the word “safety” uttered by a male speaker, sampled at 8 kHz (see also the file safety.m). The signal has a significant amount of background noise (as it was recorded in a computer laboratory). Develop procedures to segment the signal into voiced, unvoiced, and silence (background noise) portions using short-time RMS , turns count, or ZCR measures. Compute the PSD for each segment that you obtain and study its characteristics.

4. The files pec1.dat, pec33.dat, and pec52.dat give three-channel recordings of the PCG, ECG, and carotid pulse signals (sampled at 1,000 Hz ; you may read the signals using the program in the file plotpec.m). The signals in pec1.dat and pec52.dat are normal; the PCG signal in pec33.dat has systolic murmur, and is of a patient suspected to have pulmonary stenosis, ventricular septal defect, and pulmonary hypertension.

Apply the Pan–Tompkins method for QRS detection to the ECG channel and the Lehner and Rangayyan method to detect the dicrotic notch in the carotid pulse channel. Extrapolate the timing information from the ECG and carotid pulse channels to segment the PCG signal into two parts: the systolic part from the onset of an S1 and to the onset of the following S2, and the diastolic part from the onset of an S2 to the onset of the following S1. Compute the PSD of each segment.

Extend the procedure to average the systolic and diastolic PSDs over several cardiac cycles. Compare the PSDs obtained for the three cases.

5. Compute the mean frequency and the ratio of the energy in the range 100 – 300 Hz to the total energy for each PSD derived in the previous problem. What can you infer from these measures?

6. Compute the PSDs of a few channels of the EEG in the file eeg1-xx.dat using Welch’s procedure (see also the file eeg1.m). Study the changes in the PSDs derived with variations in the window width, the number of segments averaged, and the type of the window used. Compare the results with the PSDs computed using the entire signal in each channel. Discuss the results in terms of the effects of the procedures and parameters on spectral resolution and leakage.

7. Compute the PSD of an ECG signal over a long duration including several cardiac cycles or beats. It would be good to select a segment with a steady heart rate (almost the same RR interval for each beat) and without any PVCs. Extract a segment over one cardiac cycle from the same signal and compute its PSD. Compare the two PSDs and explain the similarities and/or differences between them.

8. It is desired to obtain averaged PSDs of the systolic and diastolic segments of a PCG signal over several cardiac cycles. Let $x_m(n)$ represent one such segment, where n is the time index, and m is the segment number.

A researcher develops the following procedure (P1): add the segments $x_m(n)$ for all m and get the average $x_a(n)$; compute the DFT $X_a(k)$ of $x_a(n)$; compute the PSD as $|X_a(k)|^2$.

Another researcher’s procedure is as follows (P2): compute the DFT $X_m(k)$ for each $x_m(n)$; add $X_m(k)$ for all m and get the average $X_a(k)$; compute the PSD as $|X_a(k)|^2$.

Yet another researcher comes up with the following procedure (P3): compute the DFT $X_m(k)$ for each $x_m(n)$; obtain $|X_m(k)|^2$ for each m ; add $|X_m(k)|^2$ for all m and get the average.

Develop a program to implement all of the three procedures given above and process a PCG signal over several cardiac cycles. Study the various versions of the averaged PSD and compare them with the PSDs of a few individual systolic and diastolic segments. Which procedure gives the correct result? What are the errors and their effects in the other two procedures?

References

- [1] Sakai A, Feigen LP, and Luisada AA. Frequency distribution of heart sounds in normal man. *Cardiovascular Research*, 5:358–363, 1971.
- [2] Gerbarg DS, Holcomb Jr. FW, Hofler JJ, Bading CE, Schultz GL, and Sears RE. Analysis of phonocardiogram by a digital computer. *Circulation Research*, 11:569–576, 1962.
- [3] Gerbarg DS, Taranta A, Spagnuolo M, and Hofler JJ. Computer analysis of phonocardiograms. *Progress in Cardiovascular Diseases*, 5(4):393–405, 1963.
- [4] Frome EL and Frederickson EL. Digital spectrum analysis of the first and second heart sounds. *Computers and Biomedical Research*, 7:421–431, 1974.
- [5] Yoganathan AP, Gupta R, Udwadia FE, Miller JW, Corcoran WH, Sarma R, Johnson JL, and Bing RJ. Use of the fast Fourier transform for frequency analysis of the first heart sound in normal man. *Medical and Biological Engineering*, 14:69–73, 1976.
- [6] Yoganathan AP, Gupta R, Udwadia FE, Corcoran WH, Sarma R, and Bing RJ. Use of the fast Fourier transform in the frequency analysis of the second heart sound in normal man. *Medical and Biological Engineering*, 14:455–459, 1976.
- [7] Adolph RJ, Stephens JF, and Tanaka K. The clinical value of frequency analysis of the first heart sound in myocardial infarction. *Circulation*, 41:1003–1014, 1970.
- [8] Clarke WB, Austin SM, Shah PM, Griffen PM, Dove JT, McCullough J, and Schreiner BF. Spectral energy of the first heart sound in acute myocardial ischemia. *Circulation*, 57(3):593–598, 1978.
- [9] Jacobs JE, Horikoshi K, and Petrovick MA. Feasibility of automated analysis of phonocardiograms. *Journal of the Audio Engineering Society*, 17(1):49–54, 1969.
- [10] Yokoi M, Uozumi Z, Okamoto N, Mizuno Y, Iwatsuka T, Takahashi H, Watanabe Y, and Yasui S. Clinical evaluation on 5 years' experience of automated phonocardiographic analysis. *Japanese Heart Journal*, 18(4):482–490, 1977.
- [11] Geckeler GD, Likoff W, Mason D, Ries RR, and Wirth CH. Cardiospectrograms: A preliminary report. *American Heart Journal*, 48:189–196, 1954.
- [12] McKusick VA, Talbot SA, and Webb GN. Spectral phonocardiography: Problems and prospects in the application of the Bell sound spectrograph to phonocardiography. *Bulletin of the Johns Hopkins Hospital*, 94:187–198, 1954.
- [13] McKusick VA, Webb GN, Humphries JO, and Reid JA. On cardiovascular sound: Further observations by means of spectral phonocardiography. *Circulation*, 11:849–870, 1955.
- [14] Winer DE, Perry LW, and Caceres CA. Heart sound analysis: A three dimensional approach. Contour plotting of sound for study of cardiovascular acoustics. *American Journal of Cardiology*, 16:547–551, 1965.
- [15] Rushmer RF. *Cardiovascular Dynamics*. WB Saunders, Philadelphia, PA, 4th edition, 1976.
- [16] Yoshimura S. Principle and practice of phonocardiography in reference to frequency intensity characteristics of heart sounds and murmurs. *Japanese Circulation Journal*, 24:921–931, 1960.
- [17] van Vollenhoven E, van Rotterdam A, Dorenbos T, and Schlesinger FG. Frequency analysis of heart murmurs. *Medical and Biological Engineering*, 7:227–231, 1969.
- [18] Sarkady AA, Clark RR, and Williams R. Computer analysis techniques for phonocardiogram diagnosis. *Computers and Biomedical Research*, 9:349–363, 1976.
- [19] Johnson GR, Adolph RJ, and Campbell DJ. Estimation of the severity of aortic valve stenosis by frequency analysis of the murmur. *Journal of the American College of Cardiology*, 1(5):1315–1323, 1983.
- [20] Johnson GR, Myers GS, and Lees RS. Evaluation of aortic stenosis by spectral analysis of the murmur. *Journal of the American College of Cardiology*, 6(1):55–63, 1985.
- [21] Oppenheim AV and Schafer RW. *Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1975.

- [22] Welch PD. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15:70–73, 1967.
- [23] Harris FJ. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.
- [24] Bingham C, Godfrey M, and Tukey JW. Modern techniques of power spectrum estimation. *IEEE Transactions on Audio and Electroacoustics*, 15:56–66, 1967.
- [25] Kay SM and Marple, Jr. SL. Spectrum analysis — A modern perspective. *Proceedings of the IEEE*, 69(11):1380–1419, 1981.
- [26] Thomson DJ. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096, 1982.
- [27] Robinson EA. A historical perspective of spectrum estimation. *Proceedings of the IEEE*, 70(9):885–907, 1982.
- [28] Childers DG, Aunon JI, and McGillem CD. Spectral analysis: Prediction and extrapolation. *CRC Critical Reviews in Biomedical Engineering*, 6(2):133–175, 1981.
- [29] Rangayyan RM and Lehner RJ. Phonocardiogram signal processing: A review. *CRC Critical Reviews in Biomedical Engineering*, 15(3):211–236, 1988.
- [30] Saltzberg B and Burch NR. Period analytic estimates of moments of the power spectrum: A simplified EEG time domain procedure. *Electroencephalography and Clinical Neurophysiology*, 30:568–570, 1971.
- [31] Binnie CD, Batchelor BG, Bowring PA, Darby CE, Herbert L, Lloyd DSL, Smith DM, Smith GF, and Smith M. Computer-assisted interpretation of clinical EEGs. *Electroencephalography and Clinical Neurophysiology*, 44:575–585, 1978.
- [32] Binnie CD, Batchelor BG, Gainsborough AJ, Lloyd DSL, Smith DM, and Smith GF. Visual and computer-assisted assessment of the EEG in epilepsy of late onset. *Electroencephalography and Clinical Neurophysiology*, 47:102–107, 1979.
- [33] Durand LG, Blanchard M, Cloutier G, Sabbah HN, and Stein PD. Comparison of pattern recognition methods for computer-assisted classification of spectra of heart sounds in patients with a porcine bioprosthetic valve implanted in the mitral position. *IEEE Transactions on Biomedical Engineering*, 37(12):1121–1129, 1990.
- [34] Durand LG, de Guise J, Cloutier G, Guardo R, and Brais M. Evaluation of FFT-based and modern parametric methods for the spectral analysis of bioprosthetic valve sounds. *IEEE Transactions on Biomedical Engineering*, 33(6):572–578, 1986.
- [35] Cloutier G, Durand LG, Guardo R, Sabbah HN, and Stein PD. Bias and variability of diagnostic spectral parameters extracted from closing sounds produced by bioprosthetic valves implanted in the mitral position. *IEEE Transactions on Biomedical Engineering*, 36(8):815–825, 1989.
- [36] Rangayyan RM and Oloumi F. Fractal analysis of knee-joint vibroarthrographic signals. In *Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine (ITAB 2010)*, pages 1–4, Corfu, Greece, November, 2010. IEEE.
- [37] Rangayyan RM, Oloumi F, Wu YF, and Cai SX. Fractal analysis of knee-joint vibroarthrographic signals via power spectral analysis. *Biomedical Signal Processing and Control*, 8(1):23–29, 2013.
- [38] Mandelbrot BB. *Fractal Geometry of Nature*. WH Freeman, San Francisco, CA, 1983.
- [39] Voss RF. Random fractal forgeries. In Earnshaw RA, editor, *Fundamental Algorithms for Computer Graphics*. Springer-Verlag, New York, NY, 1985.
- [40] Voss RF. Fractals in nature: From characterization to simulation. In Peitgen HO and Saupe D, editors, *The Science of Fractal Images*. Springer-Verlag, New York, NY, 1988.
- [41] Akay M, editor. *Nonlinear Biomedical Signal Processing: Volume I, Fuzzy Logic, Neural Networks, and New Algorithms*. IEEE, New York, NY, 2000.
- [42] Akay M, editor. *Nonlinear Biomedical Signal Processing: Volume II, Dynamic Analysis and Modeling*. IEEE, New York, NY, 2001.

- [43] Cerutti S and Marchesi C, editors. *Advanced Methods of Biomedical Signal Processing*. IEEE and Wiley, New York, NY, 2011.
- [44] Chang S, Li SJ, Chiang MJ, Hu SJ, and Hsyu MC. Fractal dimension estimation via spectral distribution function and its application to physiological signals. *IEEE Transactions on Biomedical Engineering*, 54(10):1895–1898, 2007.
- [45] Saupe D. Algorithms for random fractals. In Peitgen HO and Saupe D, editors, *The Science of Fractal Images*. Springer, New York, NY, 1988.
- [46] Peitgen HO, Jürgens H, and Saupe D. *Chaos and Fractals: New Frontiers of Science*. Springer, New York, NY, 2004.
- [47] Cabral TM and Rangayyan RM. *Fractal Analysis of Breast Masses in Mammograms*. Morgan & Claypool and Springer, 2012.
- [48] Ladly KO, Frank CB, Bell GD, Zhang YT, and Rangayyan RM. The effect of external loads and cyclic loading on normal patellofemoral joint signals. *Special Issue on Biomedical Engineering, Defence Science Journal (India)*, 43:201–210, July 1993.
- [49] Rangayyan RM, Krishnan S, Bell GD, Frank CB, and Ladly KO. Parametric representation and screening of knee joint vibroarthrographic signals. *IEEE Transactions on Biomedical Engineering*, 44(11):1068–1074, 1997.
- [50] Rangayyan RM and Wu YF. Screening of knee-joint vibroarthrographic signals using statistical parameters and radial basis functions. *Medical and Biological Engineering and Computing*, 46(3):223–232, 2008.
- [51] Rangayyan RM and Wu YF. Analysis of vibroarthrographic signals with features related to signal variability and radial basis functions. *Annals of Biomedical Engineering*, 37(1):156–163, 2009.
- [52] Rangayyan RM and Wu YF. Screening of knee-joint vibroarthrographic signals using probability density functions estimated with Parzen windows. *Biomedical Signal Processing and Control*, 5(1):53–58, 2010.
- [53] Umapathy K and Krishnan S. Modified local discriminant bases algorithm and its application in analysis of human knee joint vibration signals. *IEEE Transactions on Biomedical Engineering*, 53(3):517–523, March 2006.
- [54] Krishnan S, Rangayyan RM, Bell GD, and Frank CB. Adaptive time-frequency analysis of knee joint vibroarthrographic signals for noninvasive screening of articular cartilage pathology. *IEEE Transactions on Biomedical Engineering*, 47(6):773–783, June 2000.
- [55] The University of Chicago, Chicago, IL, <http://metz-roc.uchicago.edu/MetzROC/software>, accessed on 2023-04-26. *Metz ROC Software*.
- [56] Shah EN, Reddy NP, and Rothschild BM. Fractal analysis of acceleration signals from patients with CPPD, rheumatoid arthritis, and spondyloarthropathy of the finger joint. *Computer Methods and Programs in Biomedicine*, 77(3):233–239, 2005.
- [57] Agarwal R and Gotman J. Computer-assisted sleep staging. *IEEE Transactions on Biomedical Engineering*, 48(12):1412–1423, 2001.
- [58] Johnson L, Lubin A, Naitoh P, Nute C, and Austin M. Spectral analysis of the EEG of dominant and non-dominant alpha subjects during waking and sleeping. *Electroencephalography and Clinical Neurophysiology*, 26(4):361–370, 1969.
- [59] Schwartz M. *Information Transmission, Modulation, and Noise*. McGraw-Hill, New York, NY, 3rd edition, 1980.
- [60] Bendat JS and Piersol AG. *Random Data: Analysis and Measurement Procedures*. Wiley, New York, NY, 2nd edition, 1986.

CHAPTER 7

MODELING OF BIOMEDICAL SIGNAL-GENERATING PROCESSES AND SYSTEMS

Up to this point in the book, we have concentrated on processing and analysis of biomedical signals. The signals were treated in their own right as conveyors of diagnostic information. While it was emphasized that the design and application of signal analysis procedures require an understanding of the physiological and pathological processes and systems that generate the signals, no specific mathematical model was used to represent the genesis of the signals in the methods we studied in the preceding chapters.

We now consider the modeling approach, where an explicit mathematical model is used to represent the process or the system that generates the signal of interest. The parameters of the model are then investigated for use in signal analysis, pattern recognition, and decision-making. As we will see, the model parameters may also be related to the physical or physiological aspects of the related systems. The *parametric modeling* approach often leads to succinct and efficient representation of signals and systems. Regardless of the emphasis on modeling, the final aim of the methods described in this chapter is analysis or classification of the signal of interest.

7.1 Problem Statement

Propose mathematical models to represent the generation of biomedical signals. Identify the possible relationships between the mathematical models and the physiological and pathological processes and systems that generate the signals. Explore the potential use of the model parameters in signal analysis, pattern recognition, and classification.

Given the diversity of the biomedical signals that we have already encountered and the many others that exist, a generic model cannot be expected to represent a large number of signals. Indeed, a very specific model is often required for each signal. Bioelectric signals such as the ECG and

EMG may be modeled using the basic action potential or SMUAP as the building block. Sound and vibration signals such as the PCG and speech may be modeled using fluid-filled resonating chambers, turbulent flow across a baffle or through a constriction, vibrating pipes, and acoustic or vibrational excitation of a tract of variable shape. We investigate a few representative signals and models in the following sections and then study a few modeling techniques that facilitate signal analysis based on the parameters extracted.

7.2 Illustration of the Problem with Case Studies

7.2.1 Motor-unit firing patterns

We have seen in Section 1.2.4 that the surface EMG of an active skeletal muscle is the spatiotemporal summation of the action potentials of a large number of motor units that have been recruited into action (see Figure 1.15). If we consider the EMG of a single motor unit, we have a train of SMUAPs; the same basic wave (spike, pulse, or wavelet) is repeated in a quasiperiodic sequence. For the sake of generality, we may represent the intervals between the SMUAPs by a random variable: Although an overall periodicity exists and is represented by the firing rate in *pps*, the intervals between the pulses, known as the *interpulse interval* or IPI, may not precisely be the same from one SMUAP to another.

Agarwal and Gottlieb [1] modeled the single-motor-unit EMG as the convolution of a series of unit impulses or Dirac delta functions — known as a *point process* [2–5] — with the basic SMUAP wave. The SMUAP train $y(t)$ may then be modeled as the output of a linear system whose impulse response $h(t)$ is the SMUAP, and the input is a point process $x(t)$:

$$y(t) = \int_0^t h(t - \tau) x(\tau) d\tau. \quad (7.1)$$

Physiological conditions dictate that successive action potentials of the same motor unit cannot overlap: The interval between any two pulses should be greater than the SMUAP duration. In normal muscle activation, SMUAP durations are of the order of 3 – 15 ms and motor unit firing rates are in the range 6 – 30 pps; the IPI is, therefore, in the range of about 30 – 170 ms, which is significantly higher than the SMUAP duration. An SMUAP train, therefore, consists of discrete (distinct and separated) events or waves.

The model as above permits independent analysis of SMUAP waveshape and firing pattern: The two are, indeed, physiologically separate entities. The SMUAP waveshape depends on the spatial arrangement and characteristics of the muscle fibers that constitute the motor unit, whereas the firing pattern is determined by the motor neuron that stimulates the muscle fibers. Statistics of the point process representing the IPI may be used to study the muscle activation process independently of the SMUAP waveshape. Details on point processes and their application to EMG modeling are presented in Section 7.3.

7.2.2 Cardiac rhythm

The ECG is a quasiperiodic signal that is also cyclostationary in the normal case (see Section 1.2.5). Each beat is triggered by a pulse from the SA node. The P wave is the combined result of the action potentials of the atrial muscle cells, whereas the QRS and T waves are formed by the spatiotemporal summation of the action potentials of the ventricular muscle cells.

In rhythm analysis, one is more interested in the timing of the beats than in their individual waveshape (with the exception of PVCs). Diseases that affect the SA node could disturb the normal rhythm, and lead to abnormal variability in the *RR* intervals. Disregarding the details of atrial and ventricular ECG waves, an ECG rhythm may be modeled by a point process representing the

firing pattern of the SA node. Sinus arrhythmia and HRV may then be investigated by studying the distribution and statistics of the *RR* interval.

Figure 7.1 illustrates the representation of ECG complexes in terms of the instantaneous heart rate values defined as the inverse of the *RR* interval of each beat, in terms of a series of *RR* interval values, and as a train of delta functions at the SA or AV node firing instants [6]. A discrete-time signal may be derived by sampling the signal in Figure 7.1 (b) at equidistant points; the result, however, may not be continuous or differentiable [6]. The signal in Figure 7.1 (c), known as the interval series, has values $I_k = t_k - t_{k-1}$, where the instants t_k represent the time instants at which the QRS complexes occur in the ECG signal. The I_k series is defined as a function of interval number and not of time, and hence may pose difficulties regarding interpretation in the frequency domain. Finally, the signal in Figure 7.1 (d) is defined as a train of Dirac delta functions $s(t) = \sum \delta(t - t_k)$. The series of impulses represents a point process that may be analyzed and interpreted with relative ease, as shown in Section 7.3. The last two representations may be used to analyze cardiac rhythm and HRV, as demonstrated in Section 7.9 (see also Section 8.12).

See Section 7.8 for other electrophysiological models of the heart.

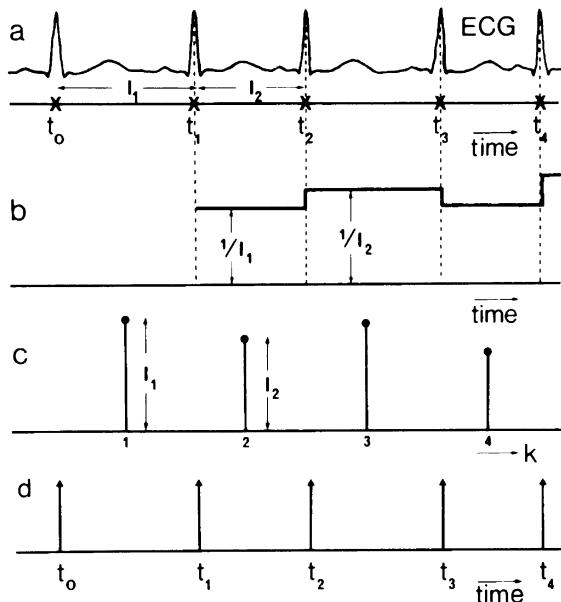


Figure 7.1 The train of ECG complexes in (a) is represented in terms of: (b) the instantaneous heart rate values defined as the inverse of the *RR* interval of each beat; (c) a series of *RR* interval values (known as the interval series); and (d) a train of delta functions at the SA or AV node firing instants. Reproduced with permission from R.W. DeBoer, J.M. Karemaker, and J. Strackee, Comparing spectra of a series of point events particularly for heart rate variability studies, *IEEE Transactions on Biomedical Engineering*, 31(4):384–387, 1984. ©IEEE.

7.2.3 Formants and pitch in speech

Speech signals are formed by exciting the vocal tract with either a pulse train or a random signal produced at the glottis, and possibly their combination as well (see Section 1.2.13). The shape of the vocal tract is varied according to the nature of the sound or phoneme to be produced; the system is, therefore, a time-variant system. We may model the output as the convolution of the (time-variant) impulse response of the vocal tract with the input glottal waveform. The input may be modeled by

a random process for unvoiced speech and as a point process for voiced speech; see Figures 1.51, 1.52, 1.53, and 7.2. Clearly, the speech signal is a nonstationary signal; however, the signal may be considered to be quasistationary over short intervals of time during which the same phoneme is being produced.

Figure 7.2 illustrates the commonly used model for speech production [7]; see also Figures 1.51, 1.52, and 1.53. The speech signal may be modeled using the same convolutional relationship as in Equation 7.1, with the limitation that the expression is valid over durations of time when the vocal-tract shape is held fixed, and the same glottal excitation is applied. Then, $h(t)$ represents the impulse response of the vocal-tract system (filter) for the time interval considered, and $x(t)$ represents the glottal waveform that is applied as input to the system. In the case of voiced speech, the IPI statistics of the point-process input, in particular its mean, are related to the pitch. Furthermore, the frequency response of the filter $H(\omega)$ representing the vocal tract determines the spectral content of the speech signal: The dominant frequencies or peaks are known as formants in the case of voiced speech.



Figure 7.2 Model for production of speech, treating the vocal tract as a time-variant linear system. A point-process input generates quasiperiodic voiced speech, whereas a random-noise input generates unvoiced speech. See also Figures 1.51, 1.52, and 1.53.

Point processes are described in Section 7.3. Parametric spectral modeling and analysis techniques suitable for formant extraction are described in Sections 7.4, 7.5, and 7.6.

7.2.4 Patellofemoral crepitus

Among the various types of VAG signals produced by the knee joint (see Section 1.2.14), the most common is a signal known as physiological patellofemoral crepitus (PPC) [8–12]. The PPC signal is a random sequence of vibrational pulses generated between the surfaces of the patella and the femur, typically observed during slow movement of the knee joint. The PPC signal may carry information on the state and lubrication of the knee joint. A mechanical model of the knee-joint surfaces that generate PPC, as proposed by Beverland et al. [11], is described in Section 7.7.3.

Zhang et al. [8] proposed a model for generation of the PPC signal based on point processes, similar to that for the SMUAP train described in Section 7.2.1. The effects of the repetition rate (or IPI) and the basic patellofemoral pulse (PFP) waveform on the spectrum of the PPC signal were analyzed separately. It was suggested that the model could represent the relationships between physiological parameters, such as the mean and *SD* of the IPI as well as the PFP waveshape, and parameters that could be measured from the PPC signal, such as its mean, *RMS*, and PSD-based features. Illustrations related to this application are provided at the end of Section 7.3.

7.3 Point Processes

Problem: Formulate a mathematical model representing the generation of a train of SMUAPs and derive an expression for the PSD of the signal.

Solution: In the model for EMG generation proposed by Agarwal and Gottlieb [1], a point process is used to represent the motor neuron firing sequence, and the SMUAP train is modeled by the convolution integral as in Equation 7.1. The IPI is treated as a sequence of independent random variables with identical normal (Gaussian) PDFs.

Let the interval between the i^{th} SMUAP and the preceding one be τ_i , and let the origin be set at the instant of appearance of the first SMUAP at $i = 0$ with $\tau_0 = 0$. The time of arrival of the i^{th} SMUAP is then given by $t_i = \tau_1 + \tau_2 + \dots + \tau_i$. The variable t_i is the sum of i independent random variables; note that $\tau_i > 0$. It is assumed that the mean μ and variance σ^2 of the random variable representing each IPI are the same. Then, the mean of t_i is $i\mu$, and its variance is $i\sigma^2$. Furthermore, t_i is also a random variable with the Gaussian PDF

$$p_{t_i}(t_i) = \frac{1}{\sqrt{2\pi i} \sigma} \exp\left[-\frac{(t_i - i\mu)^2}{2i\sigma^2}\right]. \quad (7.2)$$

If the SMUAP train has $N + 1$ SMUAPs labeled as $i = 0, 1, 2, \dots, N$, the motor neuron firing sequence is represented by the point process

$$x(t) = \sum_{i=0}^N \delta(t - t_i). \quad (7.3)$$

The Fourier transform of the point process is

$$\begin{aligned} X(\omega) &= \int_{-\infty}^{\infty} \sum_{i=0}^N \delta(t - t_i) \exp(-j\omega t) dt \\ &= \sum_{i=0}^N \exp(-j\omega t_i). \end{aligned} \quad (7.4)$$

$X(\omega)$ is a function of the random variable t_i , which is, in turn, a function of i random variables $\tau_1, \tau_2, \dots, \tau_i$. Therefore, $X(\omega)$ is random. The ensemble average of $X(\omega)$ may be obtained by computing its expectation, taking into account the PDF of t_i , as follows [1]:

$$\bar{X}(\omega) = E[X(\omega)] = \sum_{i=0}^N E[\exp(-j\omega t_i)]. \quad (7.5)$$

$$E[\exp(-j\omega t_i)] = \int_{-\infty}^{\infty} \exp(-j\omega t_i) p_{t_i}(t_i) dt_i. \quad (7.6)$$

Using the expression for $p_{t_i}(t_i)$ in Equation 7.2, we get

$$E[\exp(-j\omega t_i)] = \frac{1}{\sqrt{2\pi i} \sigma} \int_{-\infty}^{\infty} \exp(-j\omega t_i) \exp\left[-\frac{(t_i - i\mu)^2}{2i\sigma^2}\right] dt_i. \quad (7.7)$$

Substituting $t_i - i\mu = r$, where r is a temporary variable, we get

$$E[\exp(-j\omega t_i)] = \frac{1}{\sqrt{2\pi i} \sigma} \exp(-j\omega i\mu) \int_{-\infty}^{\infty} \exp\left[-\frac{r^2}{2i\sigma^2}\right] \exp(-j\omega r) dr. \quad (7.8)$$

Given that the Fourier transform of $\exp(-\frac{r^2}{2\sigma^2})$ is $\sigma\sqrt{2\pi} \exp(-\frac{\sigma^2\omega^2}{2})$ [13], it follows that

$$E[\exp(-j\omega t_i)] = \exp(-j\omega i\mu) \exp\left[-\frac{i\sigma^2\omega^2}{2}\right]. \quad (7.9)$$

Finally, we have

$$\bar{X}(\omega) = \sum_{i=0}^N \exp(-j\omega i\mu) \exp\left[-\frac{i\sigma^2\omega^2}{2}\right]. \quad (7.10)$$

(In the equations in the present context, i is an index and $j = \sqrt{-1}$.) The ensemble-averaged Fourier transform of the SMUAP train is given by

$$\bar{Y}(\omega) = \bar{X}(\omega)H(\omega), \quad (7.11)$$

where $H(\omega)$ is the Fourier transform of an individual SMUAP. The Fourier transform of an SMUAP train is, therefore, a multiplicative combination of the Fourier transform of the point process representing the motor neuron firing sequence and the Fourier transform of an individual SMUAP.

Illustration of application to EMG: Figure 7.3 illustrates EMG signals synthesized using the point-process model as above using 1, 20, 40, and 60 motor units, all with the same biphasic SMUAP of 8 ms duration and IPI statistics $\mu = 50$ ms and $\sigma = 6.27$ ms [1]. It is seen that the EMG signal complexity increases as more motor units are activated. The interference patterns resemble real EMG signals but obscure the shape of the SMUAP used to generate the signals.

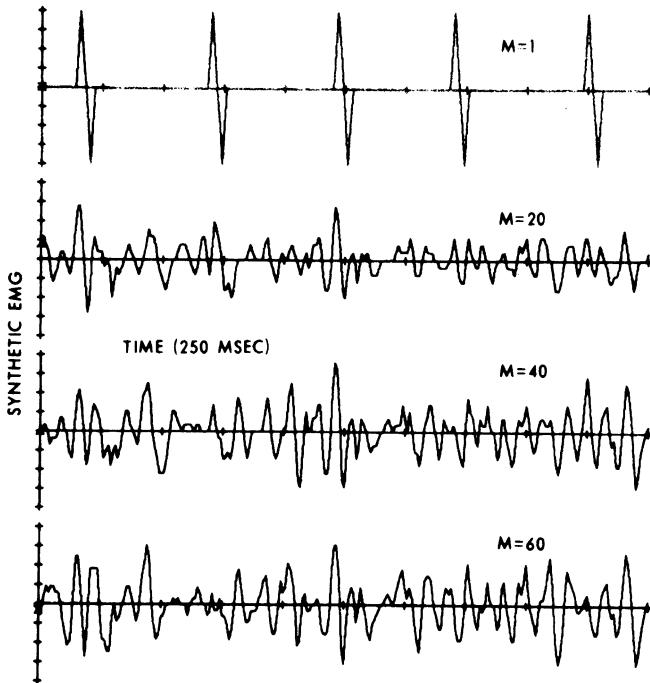


Figure 7.3 Synthesis of an SMUAP train and EMG interference pattern using the point-process model. Top to bottom: SMUAP train of a single motor unit, and interference patterns of the activities of 20, 40, and 60 motor units. SMUAP duration = 8 ms. IPI statistics $\mu = 50$ ms and $\sigma = 6.27$ ms. The duration of each signal is 250 ms. Reproduced with permission from G.C. Agarwal and G.L. Gottlieb, An analysis of the electromyogram by Fourier, simulation and experimental techniques, *IEEE Transactions on Biomedical Engineering*, 22(3):225–229, 1975. ©IEEE.

Figure 7.4 shows the magnitude spectra of two synthesized EMG signals, the first one with one active motor unit, and the other with 15 active motor units, with biphasic SMUAP duration of 8 ms, $\mu = 20$ ms, and $\sigma = 4.36$ ms [1]. The smooth curve superimposed on the second spectrum in Figure 7.4 was derived from the mathematical model described in the preceding paragraphs. An important point to observe from the spectra is that the average magnitude spectrum of several identical motor units approaches the spectrum of a single MUAP. The spectral envelope of an SMUAP train or that of an interference pattern of several SMUAP trains with identical SMUAP waveshape is determined by the shape of an individual SMUAP.

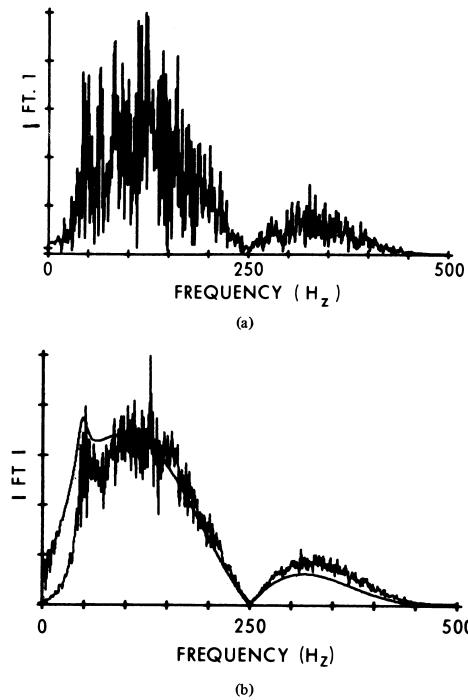


Figure 7.4 Magnitude spectra of synthesized EMG signals with (a) one motor unit, and (b) 15 motor units, with biphasic SMUAP duration of 8 ms, $\mu = 20$ ms, and $\sigma = 4.36$ ms. The smooth curve superimposed on the spectrum in (b) was derived from the point-process model with 10 SMUAPs. Reproduced with permission from G.C. Agarwal and G.L. Gottlieb, An analysis of the electromyogram by Fourier, simulation and experimental techniques, *IEEE Transactions on Biomedical Engineering*, 22(3):225–229, 1975. ©IEEE.

Figure 7.5 shows the magnitude spectra of surface EMG signals recorded from the gastrocnemius–soleus muscle, averaged over 1, 5, and 15 signal records [1]. The spectra in Figures 7.4 and 7.5 demonstrate comparable features. If all of the motor units active in a composite EMG record were to have similar or identical MUAPs, the spectral envelope of the signal could provide information on the MUAP waveshape (via an inverse Fourier transform). As we have noted earlier in Section 1.2.4, MUAP shape could be useful in the diagnosis of neuromuscular diseases. In reality, however, many motor units of different MUAP shapes could be contributing to the EMG signal even at low levels of effort, and analysis as above may have limited applicability. Regardless, the point-process model provides an interesting approach to model EMG signals. The same model is applicable to the generation of voiced-speech signals, as illustrated in Figure 7.2. For other models to represent the characteristics of the EMG signal, refer to the papers by Parker et al. [14], Lindström and Magnusson [15], Zhang et al. [16], Parker and Scott [17], Shwedyk et al. [18], Person and Libkind [19], Person and Kudina [20], de Luca [21, 22], Lawrence and de Luca [23], and de Luca and van Dyk [24].

Illustration of application to PPC: Zhang et al. [8] proposed a point-process model to represent knee-joint PPC signals, which they called PFP trains or signals (see Section 7.2.4). Figure 7.6 illustrates the PSDs of two point processes simulated with mean repetition rate $\mu_r = 21$ pps and coefficient of variation $CV_r = \sigma_r/\mu_r = 0.1$ and 0.05, where σ_r is the SD of the repetition rate. A Gaussian distribution was used to model the IPI statistics. The spectra clearly show the most-dominant peak at the mean repetition rate of the point process, followed by smaller peaks at its harmonics. The higher-order harmonics are better defined in the case with the lower CV_r ; in the

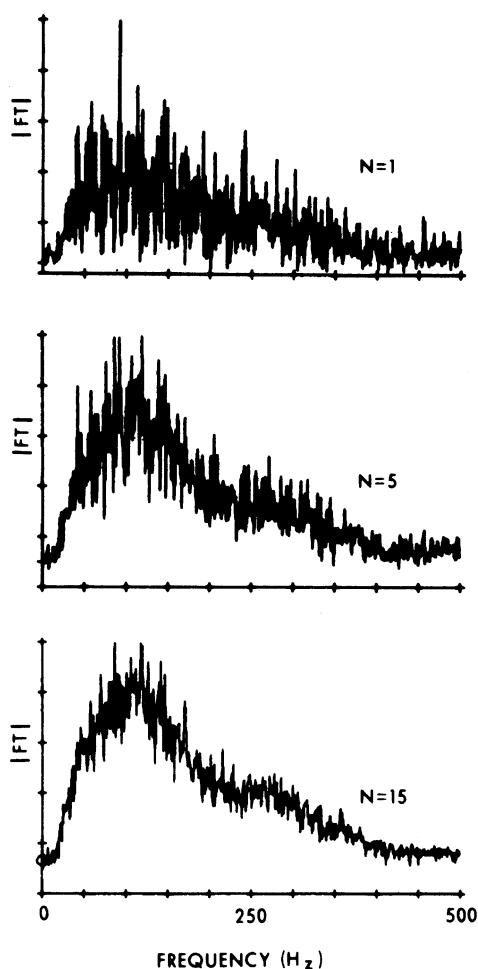


Figure 7.5 Magnitude spectra of surface EMG signals recorded from the gastrocnemius-soleus muscle, averaged over 1, 5, and 15 signal records. Reproduced with permission from G.C. Agarwal and G.L. Gottlieb, An analysis of the electromyogram by Fourier, simulation and experimental techniques, *IEEE Transactions on Biomedical Engineering*, 22(3):225–229, 1975. ©IEEE.

limit, the PSD will be a periodic impulse train with all impulses of equal strength when the point process is exactly periodic ($\sigma_r = 0, CV_r = 0$).

Zhang et al. [8] simulated PFP trains for different IPI statistics using a sample PFP waveform from a real VAG signal recorded at the patella of a normal subject using an accelerometer. The duration of the PFP waveform was 21 ms, and the IPI statistics μ_r and CV_r were limited such that the PFP trains synthesized would have nonoverlapping PFP waveforms and resemble real PFP signals. Figures 7.7 and 7.8 illustrate the PSDs of synthesized PFP signals for different μ_r but with the same CV_r , and for the same μ_r but with different CV_r , respectively. The PSDs clearly illustrate the influence of IPI statistics on the spectral features of signals generated by point processes. Some important observations to be made are:

- The PSD envelope of the PFP train remains the same, regardless of the IPI statistics.
- The PSD envelope of the PFP train is determined by the PSD of an individual PFP waveform.

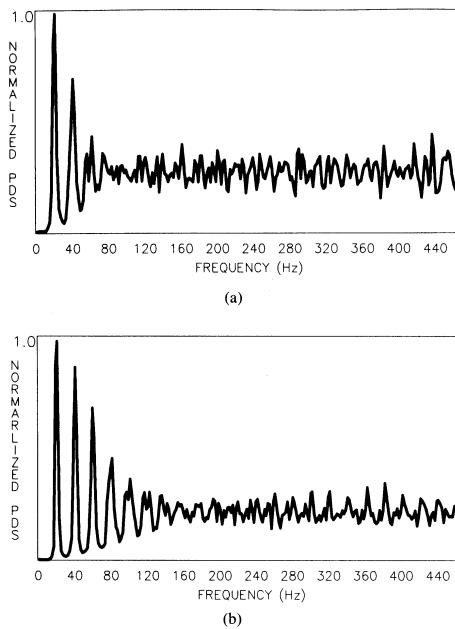


Figure 7.6 Normalized PSDs of synthesized point processes with (a) $\mu_r = 21$ pps and $CV_r = 0.1$, and (b) $\mu_r = 21$ pps and $CV_r = 0.05$. Note: PDS represents the power density spectrum, which is the same as the PSD. Reproduced with permission from Y.T. Zhang, C.B. Frank, R.M. Rangayyan, and G.D. Bell, Mathematical modelling and spectrum analysis of the physiological patello-femoral pulse train produced by slow knee movement, *IEEE Transactions on Biomedical Engineering*, 39(9):971–979, 1992. ©IEEE.

- The PSD envelope of the PFP train is modulated by a series of impulses with characteristics determined by the IPI statistics. The first impulse indicates the mean repetition rate.
- The point process has a highpass effect: low-frequency components of the PSD of the basic PFP are suppressed due to multiplication with the PSD of the point process.
- Physiological signals rarely exhibit precise periodicity. The CV_r value will be reasonably large, thereby limiting the effect of repetition to low frequencies in the PSD of the PFP train.

The observations made above are, in general, valid for all signals generated by point processes, including SMUAP trains and voiced-speech signals.

Zhang et al. [8] verified the point-process model for PFP signals by computing the IPI statistics and PSDs of real PFP signals recorded from normal subjects. Figure 7.9 shows the IPI histograms computed from the PFP signals of two normal subjects. The IPI statistics computed for the two cases are $\mu_r = 25.2$ pps and $CV_r = 0.07$ for the first signal, and $\mu_r = 16.1$ pps and $CV_r = 0.25$ for the second signal. While the IPI histogram for the first signal appears to be close to a Gaussian distribution, the second is not. The PSDs of the two signals are shown in Figure 7.10. The PSDs of the real signals demonstrate features that are comparable to those observed from the PSDs of the synthesized signals and agree with the observations listed above. The envelopes of the two PSDs demonstrate minor variations: The basic PFP waveforms in the two cases were not identical.

7.4 Parametric System Modeling

The importance of spectral analysis of biomedical signals is demonstrated in Chapter 6. The methods described are based on the computation and use of the Fourier spectrum; while this approach

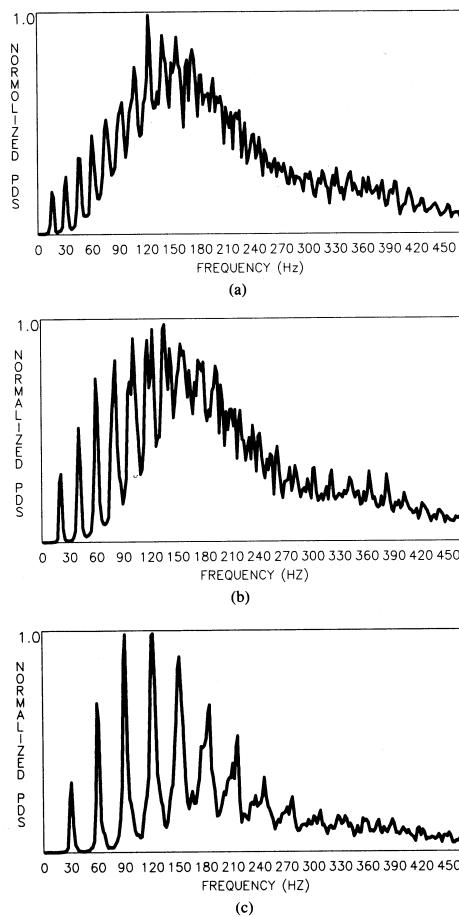


Figure 7.7 Normalized PSDs of synthesized PFP trains using a real PFP waveform with a duration of 21 ms, $CV_r = 0.05$, and (a) $\mu_r = 16$ pps, (b) $\mu_r = 21$ pps, and (c) $\mu_r = 31$ pps. Note: PDS represents the power density spectrum, which is the same as the PSD. Reproduced with permission from Y.T. Zhang, C.B. Frank, R.M. Rangayyan, and G.D. Bell, Mathematical modelling and spectrum analysis of the physiological patello-femoral pulse train produced by slow knee movement, *IEEE Transactions on Biomedical Engineering*, 39(9):971–979, 1992. ©IEEE.

is, to begin with, nonparametric, a few parameters can be computed from Fourier spectra. The limitations of such an approach are also discussed in Chapter 6. We now study methods for parametric modeling and analysis that, although based on time-domain data and models at the outset, can facilitate parametric characterization of the spectral properties of signals and systems.

Problem: Explore the possibility of parametric modeling of signal characteristics using the general linear system model.

Solution: The difference equation that gives the output of a general discrete-time LSI system is

$$y(n) = - \sum_{k=1}^P a_k y(n-k) + G \sum_{l=0}^Q b_l x(n-l), \quad (7.12)$$

with $b_0 = 1$. (Note: The advantage of the negative sign before the summation with a_k will become apparent later in this section; some model formulations use a positive sign, which does not make any significant difference in the rest of the derivation.) The input to the system is $x(n)$; the output

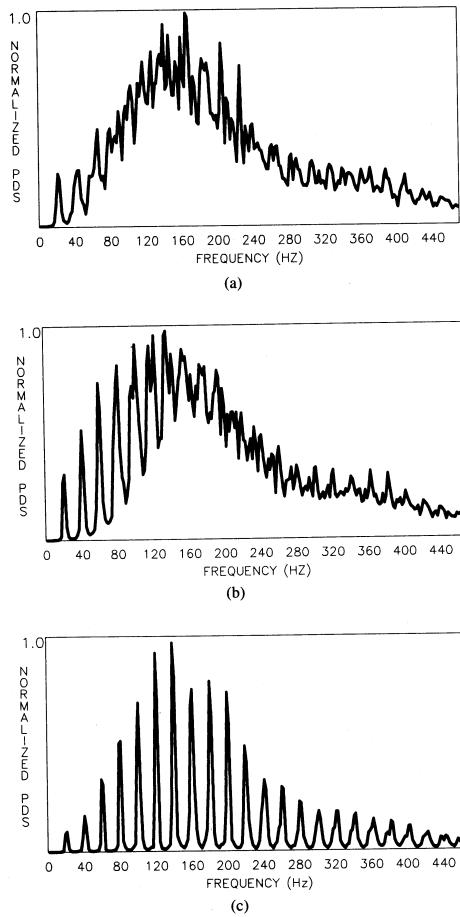


Figure 7.8 Normalized PSDs of synthesized PFP trains using a real PFP waveform with a duration of 21 ms, $\mu_r = 21$ pps, and (a) $CV_r = 0.1$, (b) $CV_r = 0.05$, and (c) $CV_r = 0.01$. Note: PDS represents the power density spectrum, which is the same as the PSD. Reproduced with permission from Y.T. Zhang, C.B. Frank, R.M. Rangayyan, and G.D. Bell, Mathematical modelling and spectrum analysis of the physiological patello-femoral pulse train produced by slow knee movement, *IEEE Transactions on Biomedical Engineering*, 39(9):971–979, 1992. ©IEEE.

is $y(n)$; the parameters $b_l, l = 0, 1, 2, \dots, Q$, indicate how the present and Q past samples of the input are combined, in a linear manner, to generate the present output sample; the parameters $a_k, k = 1, 2, \dots, P$, indicate how the past P samples of the output are linearly combined (in a feedback loop) to produce the current output; G is a gain factor; and P and Q determine the order of the system. The summation over x represents the *moving-average* or MA part of the system; the summation over y represents the *autoregressive* or AR part of the system; the entire system may be viewed as a combined *autoregressive, moving-average*, or ARMA system. The feedback part typically makes the impulse response of the system infinitely long; the system may then be viewed as an IIR filter (see Figures 3.62 and 3.63).

Equation 7.12 indicates that the output of the system is simply a linear combination of the present input sample, a few past input samples, and a few past output samples. The use of the past input and output samples in computing the present output sample represents the memory of the system. The model also indicates that the present output sample may be of the

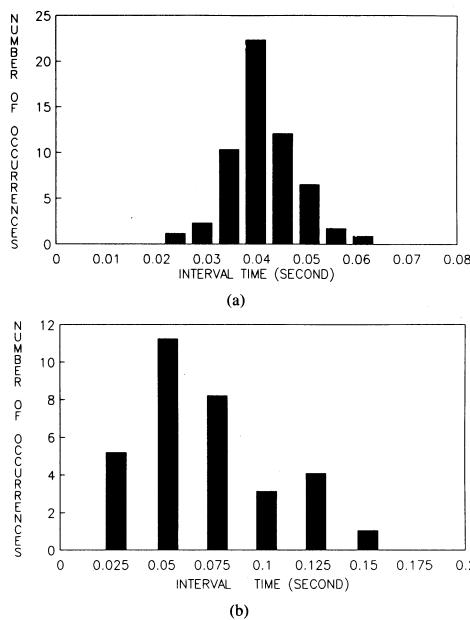


Figure 7.9 IPI histograms computed from real PFP trains recorded from two normal subjects. The statistics computed were (a) $\mu_r = 25.2 \text{ pps}$ and $CV_r = 0.07$, and (b) $\mu_r = 16.1 \text{ pps}$ and $CV_r = 0.25$. See also Figure 7.10. Reproduced with permission from Y.T. Zhang, C.B. Frank, R.M. Rangayyan, and G.D. Bell, Mathematical modelling and spectrum analysis of the physiological patello-femoral pulse train produced by slow knee movement, *IEEE Transactions on Biomedical Engineering*, 39(9):971–979, 1992. ©IEEE.

present and a few past input samples, and a few past output samples. For this reason, the model is also known as the *linear prediction* or LP model [7, 25–27].

Applying the z -transform to Equation 7.12, we can obtain the transfer function of the system as

$$H(z) = \frac{Y(z)}{X(z)} = G \frac{1 + \sum_{l=1}^Q b_l z^{-l}}{1 + \sum_{k=1}^P a_k z^{-k}}. \quad (7.13)$$

(The advantage of the negative sign before the summation with a_k in Equation 7.12 is now apparent in the numerator – denominator symmetry of Equation 7.13.) The system is completely characterized by the parameters $a_k, k = 1, 2, \dots, P$; $b_l, l = 1, 2, \dots, Q$; and G . In most applications, the gain factor G is not important; the system is, therefore, completely characterized by the a and b parameters, with the exception of a gain factor. Furthermore, we may factorize the numerator and denominator polynomials in Equation 7.13 and express the transfer function as

$$H(z) = G^1 \frac{\prod_{l=1}^Q (1 - z_l z^{-1})}{\prod_{k=1}^P (1 - p_k z^{-1})}, \quad (7.14)$$

where $z_l, l = 1, 2, \dots, Q$, are the zeros of the system, and $p_k, k = 1, 2, \dots, P$, are the poles of the system. The model may now be referred to as a *pole-zero model*. It is evident from Equation 7.14 that the system is completely characterized by its poles and zeros, but for a gain factor G^1 .

Equations 7.12, 7.13, and 7.14 demonstrate the applicability of the same conceptual model in the time and frequency domains. The a and b parameters are directly applicable in both the time and the frequency domains in expressing the input–output relationship or the system transfer function. The poles and zeros are more specific to the frequency domain, although the contribution of each pole or zero to the time-domain impulse response of the system may be derived directly from its coordinates in the z -plane [13].

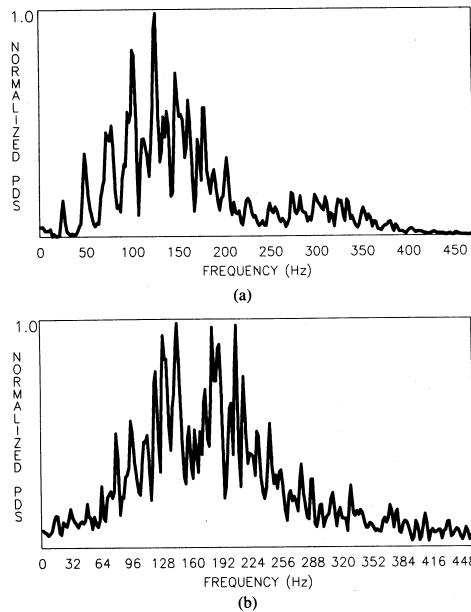


Figure 7.10 Normalized PSDs of the real PFP trains recorded from two normal subjects whose IPI histograms are shown in Figure 7.9. The IPI statistics of the two cases are (a) $\mu_r = 25.2 \text{ pps}$ and $CV_r = 0.07$, and (b) $\mu_r = 16.1 \text{ pps}$ and $CV_r = 0.25$. Note: PDS represents the power density spectrum, which is the same as the PSD. Reproduced with permission from Y.T. Zhang, C.B. Frank, R.M. Rangayyan, and G.D. Bell, Mathematical modelling and spectrum analysis of the physiological patello-femoral pulse train produced by slow knee movement, *IEEE Transactions on Biomedical Engineering*, 39(9):971–979, 1992. ©IEEE.

Given a particular input signal $x(n)$ and the corresponding output of the system $y(n)$, we could derive their z -transforms $X(z)$ and $Y(z)$, and subsequently obtain the system transfer function $H(z)$ in some form. Difficulties arise at values of z for which $X(z) = 0$; as the system is linear and $Y(z) = H(z)X(z)$, we have $Y(z) = 0$ at such points as well. Then, $H(z)$ cannot be determined at the corresponding values of z . [The ideal test signal is the unit-impulse function $x(n) = \delta(n)$, for which $X(z) = 1$ for all z : The response of an LSI system to an impulse completely characterizes the system with the corresponding $y(n) = h(n)$ or its z -domain equivalent $H(z)$.] Methods to determine an AR or ARMA model for a given signal for which the corresponding input to the system is not known (or is assumed to be a point process or a random process) are described in the following sections.

7.5 Autoregressive or All-pole Modeling

Problem: How can we obtain an AR (or LP) model when the input to the system that caused the given signal as its output is unknown?

Solution: In the AR or all-pole model [7, 25], the output is modeled as a linear combination of P past values of the output and the present input sample as

$$y(n) = - \sum_{k=1}^P a_k y(n-k) + G x(n). \quad (7.15)$$

(The discussion on AR modeling here closely follows that by Makhoul [25], with permission.) Some model formulations use a positive sign in place of the negative sign before the summation in

the equation given above. It should be noted that the model in Equation 7.15 does not account for the presence of noise.

The all-pole transfer function corresponding to Equation 7.15 is

$$H(z) = \frac{G}{1 + \sum_{k=1}^P a_k z^{-k}}. \quad (7.16)$$

In the case of biomedical signals such as the EEG or the PCG, the input to the system is totally unknown. Then, we can only approximately predict the current sample of the output signal using its past values as

$$\tilde{y}(n) = - \sum_{k=1}^P a_k y(n-k), \quad (7.17)$$

where the \sim symbol indicates that the predicted value is only approximate. The error in the predicted value (also known as the residual) is

$$e(n) = y(n) - \tilde{y}(n) = y(n) + \sum_{k=1}^P a_k y(n-k). \quad (7.18)$$

The general signal-flow diagram of the AR model viewed as a prediction or error filter is illustrated in Figure 7.11.

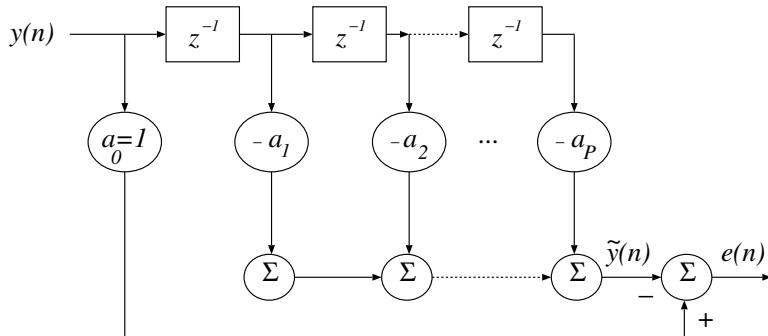


Figure 7.11 Signal-flow diagram of the AR model.

The least-squares method: In the least-squares method, the parameters a_k are obtained by minimizing the MSE with respect to all of the parameters. The procedure is similar to that used to derive the optimal Wiener filter (see Section 3.9 and Haykin [26]).

Given an observed signal $y(n)$, the following procedure is applicable for minimization of the MSE [25]. The total squared error (TSE) ε is given by

$$\varepsilon = \sum_n e^2(n) = \sum_n \left(y(n) + \sum_{k=1}^P a_k y(n-k) \right)^2. \quad (7.19)$$

(Note: The TSE is the same as the MSE except for a scale factor.) Although the range of the summation in Equation 7.19 is important, we may minimize ε without specifying the range for the time being. Minimization of ε is performed by applying the conditions

$$\frac{\partial \varepsilon}{\partial a_k} = 0, \quad 1 \leq k \leq P \quad (7.20)$$

to Equation 7.19, which yields

$$\sum_{k=1}^P a_k \sum_n y(n-k) y(n-i) = - \sum_n y(n) y(n-i), \quad 1 \leq i \leq P. \quad (7.21)$$

For a given signal $y(n)$, Equation 7.21 provides a set of P equations in the P unknowns $a_k, k = 1, 2, \dots, P$, known as the *normal equations*; the similarities between the normal equations here and those in the case of the Wiener filter (see Section 3.9) will become apparent later.

By expanding Equation 7.19 and using the relationship in Equation 7.21, the minimum TSE ε_P for the model of order P is obtained as

$$\varepsilon_P = \sum_n y^2(n) + \sum_{k=1}^P a_k \sum_n y(n) y(n-k). \quad (7.22)$$

The expression for TSE will be simplified in the following paragraphs.

The autocorrelation method: If the range of summation in Equations 7.19 and 7.21 is specified to be $-\infty < n < \infty$, the error is minimized over an infinite duration. We have

$$\phi_y(i) = \sum_{n=-\infty}^{\infty} y(n) y(n-i), \quad (7.23)$$

where $\phi_y(i)$ is the ACF of $y(n)$. In practice, the signal $y(n)$ will be available only over a finite interval, such as $0 \leq n \leq N-1$; the given signal may then be assumed to be zero outside this range and treated as a windowed version of the true signal, as we have seen in Section 6.3. Then, the ACF may be expressed as

$$\phi_y(i) = \sum_{n=i}^{N-1-i} y(n) y(n-i), \quad i \geq 0, \quad (7.24)$$

where the scale factor $\frac{1}{N}$ is omitted. (It will become apparent later that the scale factor is immaterial in the derivation of the model coefficients.) The normal equations then become

$$\sum_{k=1}^P a_k \phi_y(i-k) = -\phi_y(i), \quad 1 \leq i \leq P. \quad (7.25)$$

We now see that an AR model may be derived for a signal with the knowledge of only its ACF; the signal samples themselves are not required. It is also obvious now that the scale factor $\frac{1}{N}$ that was omitted in defining the ACF is of no consequence in deriving the model coefficients. It may be advantageous to use the normalized ACF values given as $\bar{\phi}_y(i) = \phi_y(i)/\phi_y(0)$, which have the property that $|\bar{\phi}_y(i)| \leq 1$.

The minimum TSE is given by

$$\varepsilon_P = \phi_y(0) + \sum_{k=1}^P a_k \phi_y(k). \quad (7.26)$$

Application to random signals: If the signal $y(n)$ is a sample of a random process, the error $e(n)$ is also a sample of a random process. We then have to use the expectation operation to obtain the *MSE* as follows:

$$\varepsilon = E[e^2(n)] = E \left[\left(y(n) + \sum_{k=1}^P a_k y(n-k) \right)^2 \right]. \quad (7.27)$$

Applying the condition for minimum error as in Equation 7.20, we get the normal equations as

$$\sum_{k=1}^P a_k E[y(n-k) y(n-i)] = -E[y(n) y(n-i)], \quad 1 \leq i \leq P. \quad (7.28)$$

The minimum MSE is

$$\varepsilon_P = E[y^2(n)] + \sum_{k=1}^P a_k E[y(n) y(n-k)]. \quad (7.29)$$

If the signal is a sample of a stationary random process, we have $E[y(n-k) y(n-i)] = \phi_y(i-k)$. This leads to the same normal equations as in Equation 7.25. If the process is ergodic, the ACF may be computed as a time average as in Equation 7.24.

If the signal is a sample of a nonstationary random process, $E[y(n-k) y(n-i)] = \phi_y(n-k, n-i)$; the ACF is a function of not only the shift but also time. We would then have to compute the model parameters for every instant of time n ; we then have a time-variant model. Methods for modeling and analysis of nonstationary signals are described in Chapter 8.

Computation of the gain factor G : Since we assumed earlier that the input to the system being modeled is unknown, the gain parameter G is not important. Regardless, the derivation of G demonstrates a few important points. Equation 7.18 may be rewritten as

$$y(n) = -\sum_{k=1}^P a_k y(n-k) + e(n). \quad (7.30)$$

Comparing this with Equation 7.15, we see that the only input signal $x(n)$ which can result in $y(n)$ at the output is given by the condition $Gx(n) = e(n)$. This condition indicates that the input signal is proportional to the error of prediction when the estimated model parameters are equal to the real system parameters. Regardless of the input, a condition that could be applied is that the energy of the output be equal to that of the signal $y(n)$ being modeled. Since the transfer function $H(z)$ is fixed, we then have the condition that the total energy of the input signal $Gx(n)$ be equal to the total energy of the error ε_P .

As illustrated in the model for speech generation in Figure 7.2, two types of input that are of interest are the impulse function and a random process that is stationary white noise. In the case when $x(n) = \delta(n)$, we have the impulse response $h(n)$ at the output, and

$$h(n) = -\sum_{k=1}^P a_k h(n-k) + G \delta(n). \quad (7.31)$$

Multiplying both sides of the expression above with $h(n-i)$ and summing over all n , we get expressions in terms of the ACF $\phi_h(i)$ of $h(n)$ as

$$\phi_h(i) = -\sum_{k=1}^P a_k \phi_h(i-k), \quad 1 \leq |i| \leq \infty \quad (7.32)$$

and

$$\phi_h(0) = -\sum_{k=1}^P a_k \phi_h(k) + G^2. \quad (7.33)$$

Due to the condition that the total energy of $h(n)$ be equal to that of $y(n)$, the condition $\phi_h(0) = \phi_y(0)$ must be satisfied. Comparing Equations 7.25 and 7.32, we then have

$$\phi_h(i) = \phi_y(i), \quad 0 \leq i \leq P. \quad (7.34)$$

Therefore, for a model of order P , the first $(P + 1)$ ACF terms of the impulse response $h(n)$ must be equal to the corresponding ACF terms of the signal $y(n)$ being modeled. It follows from Equations 7.26, 7.33, and 7.34 that

$$G^2 = \varepsilon_P = \phi_y(0) + \sum_{k=1}^P a_k \phi_y(k). \quad (7.35)$$

In the case when the input is a sequence of uncorrelated samples of a random process (white noise) with zero mean and unit variance, we could use the same procedure as for the impulse-input case, with the difference being that expectations are taken instead of summing over all n . {The conditions to be noted in this case are $E[x(n)] = 0$ and $E[x(n)x(n-i)] = \delta(i)$.} The same relations as above for the impulse-input case are obtained. The identical nature of the results for the two cases follows from the fact that the two types of input have identical ACFs and PSDs. These characteristics are relevant in the speech model shown in Figure 7.2.

Computation of the model parameters: For low orders of the model, Equation 7.25 may be solved directly. However, direct methods may not be feasible when P is large.

The normal equations in Equation 7.25 may be written in matrix form as

$$\begin{bmatrix} \phi_y(0) & \phi_y(1) & \cdots & \phi_y(P-1) \\ \phi_y(1) & \phi_y(0) & \cdots & \phi_y(P-2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_y(P-1) & \phi_y(P-2) & \cdots & \phi_y(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = - \begin{bmatrix} \phi_y(1) \\ \phi_y(2) \\ \vdots \\ \phi_y(P) \end{bmatrix}. \quad (7.36)$$

For real signals, the $P \times P$ ACF matrix is symmetric, and the elements along any diagonal are identical, that is, it is a Toeplitz matrix.

It is worth noting the following similarities and differences between the normal equations in the case of the Wiener filter as given in Equation 3.170 and those given above in Equation 7.36:

- The filter vector on the LHS contains the coefficients of the filter being designed in both cases — the Wiener filter or the prediction filter.
- The vector on the RHS is the CCF between the input and the desired response in the case of the Wiener filter, whereas it is the ACF of the input signal starting from a time lag of 1 in the present case. (Note that the desired response in the case of a prediction filter is the input signal itself but advanced by one time sample.)

Haykin [26] provides a more detailed correspondence between the AR model and the Wiener filter.

A procedure known as Durbin's method [28,29] or the Levinson–Durbin algorithm (see Makhoul [25], Rabiner and Schafer [7], or Haykin [26]) provides a recursive method to solve the normal equations in Equation 7.36. The procedure starts with a model order of 1; computes the model parameters, the error, and a secondary set of parameters known as the reflection coefficients; updates the model order and the parameters; and repeats the procedure until the model of the desired order is obtained. The Levinson–Durbin algorithm is summarized below.

Initialize model order $i = 0$ and error $\varepsilon_0 = \phi_y(0)$. Perform the following steps recursively for $i = 1, 2, \dots, P$.

1. Increment model order i and compute the i^{th} reflection coefficient γ_i as

$$\gamma_i = -\frac{1}{\varepsilon_{i-1}} \left[\phi_y(i) + \sum_{j=1}^{i-1} a_{i-1,j} \phi_y(i-j) \right], \quad (7.37)$$

where $a_{i-1,j}$ denotes the j^{th} model coefficient at iteration $(i - 1)$; the iteration index is also the recursively updated model order.

2. Let $a_{i,i} = \gamma_i$.
3. Update the predictor coefficients as

$$a_{i,j} = a_{i-1,j} + \gamma_i a_{i-1,i-j}, \quad 1 \leq j \leq i-1. \quad (7.38)$$

4. Compute the error value as

$$\varepsilon_i = (1 - \gamma_i^2) \varepsilon_{i-1}. \quad (7.39)$$

The final model parameters are given as $a_k = a_{P,k}$, $1 \leq k \leq P$. The Levinson–Durbin algorithm computes the model parameters for all orders up to the desired order P . As the order of the model is increased, the TSE reduces, and hence we have $0 \leq \varepsilon_i \leq \varepsilon_{i-1}$. The reflection coefficients may also be used to test the stability of the model (filter) being designed: $|\gamma_i| < 1$, $i = 1, 2, \dots, P$, is the required condition for stability of the model of order P .

The covariance method: In deriving the autocorrelation method, the range of summation of the prediction error in Equations 7.19 and 7.21 was specified to be $-\infty < n < \infty$. If, instead, we specify the range of summation to be a finite interval, such as $0 \leq n \leq N-1$, we get

$$\sum_{k=1}^P a_k C(k, i) = -C(0, i), \quad 1 \leq i \leq P \quad (7.40)$$

instead of Equation 7.25 based on the ACF, and the minimum TSE is given by

$$\varepsilon_P = C(0, 0) + \sum_{k=1}^P a_k C(0, k) \quad (7.41)$$

instead of Equation 7.26, where

$$C(i, k) = \sum_{n=0}^{N-1} y(n-i) y(n-k) \quad (7.42)$$

is the covariance of the signal $y(n)$ in the specified interval. The matrix formed by the covariance function is symmetric as $C(i, k) = C(k, i)$, similar to the ACF matrix in Equation 7.36; however, the elements along each diagonal will not be equal, as $C(i+1, k+1) = C(i, k) + y(-i-1) y(-k-1) - y(N-1-i) y(N-1-k)$. Computation of the covariance coefficients also requires $y(n)$ to be known for $-P \leq n \leq N-1$. The distinctions disappear as the specified interval of summation (error minimization) tends to infinity.

7.5.1 Spectral matching and parameterization

The AR model was derived in the preceding section based on time-domain formulations in the autocorrelation and covariance methods. We now see that equivalent formulations can be derived in the frequency domain, which can lead to a different interpretation of the model. Applying the z -transform to Equation 7.18, we get

$$E(z) = \left[1 + \sum_{k=1}^P a_k z^{-k} \right] Y(z) = A(z) Y(z) \quad (7.43)$$

and

$$H(z) = \frac{G}{A(z)}, \quad (7.44)$$

where

$$A(z) = 1 + \sum_{k=1}^P a_k z^{-k}, \quad (7.45)$$

and $E(z)$ is the z -transform of $e(n)$. We can now view the error $e(n)$ as the result of passing the signal being modeled $y(n)$ through the filter $A(z)$, which may be considered to be an *inverse filter*. In the case of $y(n)$ being a deterministic signal, applying Parseval's theorem, the TSE to be minimized may be written as

$$\varepsilon = \sum_{n=-\infty}^{\infty} e^2(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(\omega)|^2 d\omega, \quad (7.46)$$

where $E(\omega)$ is obtained by evaluating $E(z)$ on the unit circle in the z -plane. Using $S_y(\omega)$ to represent the PSD of $y(n)$, we have

$$\varepsilon = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_y(\omega) |A(\omega)|^2 d\omega, \quad (7.47)$$

where $A(\omega)$ is the frequency response of the inverse filter and is given by evaluating $A(z)$ on the unit circle in the z -plane.

From Equations 7.16 and 7.44, we get

$$\tilde{S}_y(\omega) = |H(\omega)|^2 = \frac{G^2}{|A(\omega)|^2} = \frac{G^2}{\left| 1 + \sum_{k=1}^P a_k \exp(-jk\omega) \right|^2}. \quad (7.48)$$

Here, $\tilde{S}_y(\omega)$ represents the PSD of the modeled signal $\tilde{y}(n)$ that is an approximation of $y(n)$ as in Equation 7.17. From Equation 7.43 we have

$$S_y(\omega) = \frac{|E(\omega)|^2}{|A(\omega)|^2}. \quad (7.49)$$

Now, $\tilde{S}_y(\omega)$ is the model's approximation of $S_y(\omega)$. Comparing Equations 7.48 and 7.49, we see that the error PSD $|E(\omega)|^2$ is modeled by a uniform (or "flat" or "white") PSD equal to G^2 . For this reason, the filter $A(z)$ is also known as a "whitening" filter.

From Equations 7.48 and 7.49, we get the TSE as

$$\varepsilon = \frac{G^2}{2\pi} \int_{-\pi}^{\pi} \frac{S_y(\omega)}{\tilde{S}_y(\omega)} d\omega. \quad (7.50)$$

As the model is derived by minimizing the TSE ε , we see now that the model is effectively minimizing the integrated ratio of the signal PSD $S_y(\omega)$ to its approximation $\tilde{S}_y(\omega)$. Makhoul [25] describes the equivalence of the model in the following terms:

- As the model order $P \rightarrow \infty$, the TSE is minimized, that is, $\varepsilon_P \rightarrow 0$.
- For a model of order P , the first $(P+1)$ ACF values of its impulse response are equal to those of the signal being modeled. Increasing P increases the range of the delay parameter (time) over which the model ACF is equal to the signal ACF.
- Given that the PSD and the ACF are Fourier-transform pairs, the preceding point leads to the frequency-domain statement that increasing P leads to a better fit of $\tilde{S}_y(\omega)$ to $S_y(\omega)$. As $P \rightarrow \infty$, the model ACF and PSD become identical to the signal ACF and PSD, respectively.

Thus, *any spectrum may be approximated by an all-pole model of an appropriate order* (see Section 7.5.2 for a discussion on the optimal model order).

Noting from Equation 7.35 that $G^2 = \varepsilon_P$, Equation 7.50 yields another important property of the model as

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{S_y(\omega)}{\tilde{S}_y(\omega)} d\omega = 1. \quad (7.51)$$

Equations 7.50 and 7.51 lead to the following spectral-matching properties of the AR model [25]:

- Due to the fact that the TSE is determined by the ratio of the true PSD to the model PSD, the spectral-matching process performs uniformly over the entire frequency range irrespective of the spectral shape. (Had the error measure been dependent on the difference between the true PSD and the model PSD, the spectral match would have been better at higher-energy frequency coordinates than at lower-energy frequency coordinates.)
- $S_y(\omega)$ will be greater than $\tilde{S}_y(\omega)$ at some frequencies and lower at others, while satisfying Equation 7.51 on the whole; the contribution to the TSE is more significant when $S_y(\omega) > \tilde{S}_y(\omega)$ than in the opposite case. Thus, when the error is minimized, the fitting of $\tilde{S}_y(\omega)$ to $S_y(\omega)$ is better at frequencies where $S_y(\omega) > \tilde{S}_y(\omega)$. Thus, the model PSD fits better at the *peaks* of the signal PSD.
- The preceding point leads to another interpretation: the AR-model spectrum $\tilde{S}_y(\omega)$ is a good estimate of the *spectral envelope* of the signal PSD. This is particularly useful when modeling quasiperiodic signals such as voiced speech, PCG, and other signals that have strong peaks in their spectra representing harmonics, formants, or resonance. Furthermore, by following the envelope, the effects of repetition, that is, the effects of the point-process excitation function (see Section 7.3), are removed.

Since the model PSD is entirely specified by the model parameters (as in Equation 7.48), we now have a *parametric representation* of the PSD of the given signal (subject to the error in the model). The TSE may be related to the signal PSD as follows [25]:

$$\hat{y}(0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log[\tilde{S}_y(\omega)] d\omega \quad (7.52)$$

represents the zeroth coefficient of the power cepstrum of $\tilde{y}(n)$ (see Sections 4.7.3 and 5.4.2). Using the relationship in Equation 7.48, we get

$$\hat{y}(0) = \log G^2 - \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |A(\omega)|^2 d\omega. \quad (7.53)$$

As all of the roots (zeros) of $A(z)$ are inside the unit circle in the z -plane (for the AR model to be stable), the integral in the equation given above is zero [25]. We also have $G^2 = \varepsilon_P$. Therefore,

$$\varepsilon_P = \exp[\hat{y}(0)]. \quad (7.54)$$

The minimum of ε_P is reached as $P \rightarrow \infty$, and is given by

$$\varepsilon_{\min} = \varepsilon_{\infty} = \exp[\hat{y}(0)]. \quad (7.55)$$

This relationship means that the TSE ε_P is the geometric mean of the model PSD $\tilde{S}_y(\omega)$, which is always positive for a positive-definite PSD. The quantity ε_P represents that portion of the signal's information content that is not predictable by a model of order P .

7.5.2 Optimal model order

Given that the AR model performs better and better as the order P is increased, where do we stop? Makhoul [25] shows that if the given signal is the output of a P -pole system, then an AR model of order P would be the optimal model with the minimum error. But how would one find in practice if the given signal was indeed produced by a P -pole system?

One possibility to determine the optimal order to model a given signal is to follow the trend in the TSE as the model order P is increased. This is feasible in a recursive procedure such as the Levinson–Durbin algorithm, where models of all lower orders are computed in deriving a model of order P , and the error at each order is readily available. The procedure could be stopped when there is no significant reduction in the error as the model order is incremented.

Makhoul [25] describes the use of a normalized error measure $\bar{\varepsilon}_P$ defined as

$$\bar{\varepsilon}_P = \frac{\varepsilon_P}{\phi_y(0)} = \frac{\exp[\hat{y}(0)]}{\phi_y(0)}. \quad (7.56)$$

As the model order $P \rightarrow \infty$,

$$\bar{\varepsilon}_{\min} = \bar{\varepsilon}_{\infty} = \frac{\exp[\hat{y}(0)]}{\phi_y(0)}. \quad (7.57)$$

$\bar{\varepsilon}_{\min}$ is a monotonically decreasing function of P , with $\bar{\varepsilon}_0 = 1$ and $\bar{\varepsilon}_{\infty} = \bar{\varepsilon}_{\min}$; furthermore, it can be expressed as a function of the model PSD as

$$\bar{\varepsilon}_P = \frac{\exp\left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \log \tilde{S}_y(\omega) d\omega\right]}{\frac{1}{2\pi} \int_{-\pi}^{\pi} \tilde{S}_y(\omega) d\omega}. \quad (7.58)$$

It is evident that $\bar{\varepsilon}_P$ depends only on the shape of the model PSD, and that $\bar{\varepsilon}_{\min}$ is determined solely by the signal PSD. The quantity $\bar{\varepsilon}_P$ may be viewed as the ratio of the geometric mean of the model PSD to its arithmetic mean, which is a measure of the spread of the PSD: The smaller the spread, the closer is the ratio to unity; the larger the spread, the closer is the ratio to zero. If the signal is the result of an all-pole system with P_0 poles, $\bar{\varepsilon}_P = \bar{\varepsilon}_{P_0}$ for $P \geq P_0$, that is, the curve remains flat. In practice, the incremental reduction in the normalized error may be checked with a condition such as

$$1 - \frac{\bar{\varepsilon}_{P+1}}{\bar{\varepsilon}_P} < \Delta, \quad (7.59)$$

where Δ is a small threshold. The optimal order may be considered to have been reached if the condition is satisfied for several consecutive model orders.

A measure based on an information-theoretic criterion proposed by Akaike [30] is expressed as [25]

$$I(P) = \log \bar{\varepsilon}_P + \frac{2P}{N_e}, \quad (7.60)$$

where N_e is the effective number of data points in the signal taking into account windowing (for example, $N_e = 0.4N$ for a Hamming window, where N is the number of data samples). As P is increased, the first term in the equation given above decreases, while the second term increases. Akaike's measure $I(P)$ may be computed up to the maximum order P of interest or the maximum that is feasible, and then the model of the order for which $I(P)$ is at its minimum could be taken as the optimal model.

Model parameters: The AR (all-pole) model $H(z)$ and its inverse $A(z)$ are uniquely characterized by any one set of the following sets of parameters [25]:

- The model parameters $a_k, k = 1, 2, \dots, P$. The series of a_k parameters is also equal to the impulse response of the inverse filter.

- The impulse response $h(n)$ of the AR model.
- The poles of $H(z)$, which are also the roots (zeros) of $A(z)$.
- The reflection coefficients $\gamma_i, i = 1, 2, \dots, P$.
- The ACF (or PSD) of the a_k coefficients.
- The ACF (or PSD) of $h(n)$.
- The cepstrum of a_k or $h(n)$.

With the inclusion of the gain factor G as required, all of the above sets have a total of $(P + 1)$ values and are equivalent in the sense that one set may be derived from another. Any particular set of parameters may be used, depending on its relevance, interpretability, or relationship to the real-world system being modeled.

Illustration of application to EEG signals: Identification of the existence of rhythms of specific frequencies is an important aspect of EEG analysis. The direct relationship between the poles of an AR model and resonance frequencies makes this technique an attractive tool for the analysis of EEG signals.

Figure 7.12 shows the Fourier spectrum and AR-model spectra with $P = 6$ and $P = 10$ for the o1 channel of the EEG signal shown in Figure 1.41. The Fourier spectrum in the lowest trace of Figure 7.12 includes many spurious variations that make its interpretation difficult. On the other hand, the AR spectra indicate distinct peaks at about 10 Hz corresponding to an alpha rhythm; a peak at 10 Hz is clearly evident even with a low model order of $P = 6$ (the middle trace in Figure 7.12).

The poles of the AR model with order $P = 10$ are also shown in Figure 7.12. The dominant pole (closest to the unit circle in the z -plane) appears at 10 Hz, corresponding to the peak observed in the spectrum in the topmost plot. The radius of the dominant pole is $|z| = 0.95$; the other complex-conjugate pole pairs have $|z| \leq 0.76$. The model with $P = 6$ resulted in two complex-conjugate pole pairs and two real poles, with the dominant pair at 10.5 Hz with $|z| = 0.91$; the magnitude of the other pole pair was 0.74. A simple search for the dominant (complex) pole can provide an indication of the prevalent EEG rhythm with fairly low AR model orders.

Illustration of application to PCG signals: Application of AR modeling is an attractive possibility for the analysis of PCG signals due to the need to identify significant frequencies of resonance in the presence of multiple components, artifacts, and noise. Although the model coefficients themselves do not carry any physical correlates or significance, the poles may be related directly to the physical or physiological characteristics of hearts sounds and murmurs.

Figure 7.13 illustrates the Fourier spectrum of a segment containing S1 and the subsequent systolic portion of the PCG signal of a normal subject, the AR-model spectra for order $P = 10$ and $P = 20$, and the poles of the model of order $P = 20$ (see also Figures 4.30, 5.7, and 6.8). Figure 7.14 illustrates the same items for a segment containing S2 and the subsequent diastolic portion of the same subject. It is evident that the AR spectra follow the dominant peaks in the spectra of the original signals. The spectra for the models of order $P = 20$ provide closer fits than those for $P = 10$; peaks in the $P = 10$ spectra gloss over multiple peaks in the original spectra. Observe the presence of poles close to the unit circle in the z -plane at frequencies corresponding to the peaks in the spectra of the signals. The AR-model spectra are smoother and easier to interpret than the periodogram-based spectra illustrated in Figure 6.8 for the same subject. The spectra for the diastolic portion indicate more medium-frequency energy than those for the systolic portion, as expected. The model coefficients or poles provide a compact parametric representation of the signals and their spectra.

Figure 7.15 illustrates the Fourier spectrum of a segment containing S1 and the subsequent systolic portion of the PCG signal of a subject with systolic murmur, split S2, and opening snap of

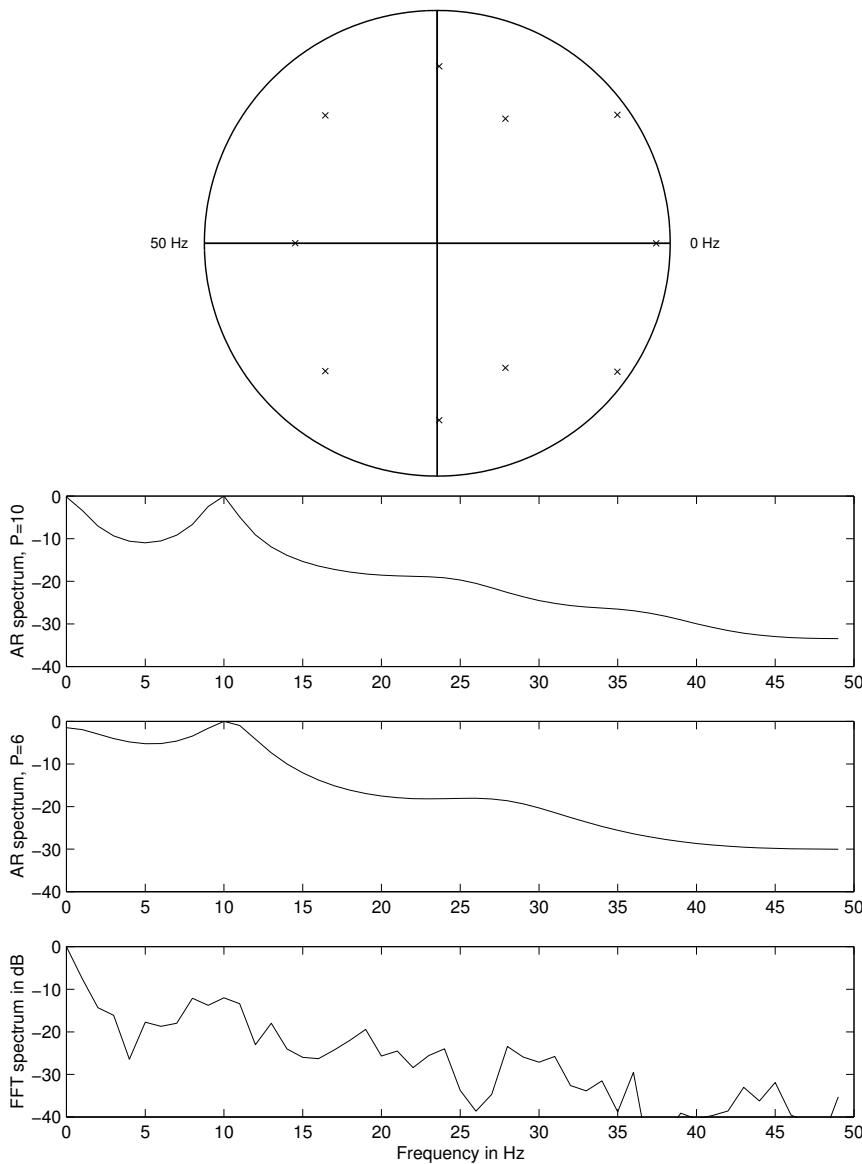


Figure 7.12 From bottom to top: Fourier spectrum; AR-model spectra with $P = 6$ and $P = 10$; and poles of the model with $P = 10$ for the o1 channel of the EEG signal shown in Figure 1.41.

the mitral valve (see also Figures 4.31, 5.8, and 6.9); the AR-model spectra for order $P = 10$ and $P = 20$; and the poles of the model of order $P = 20$. Figure 7.16 illustrates the same items for a segment containing S2 and the subsequent diastolic portion of the same subject. The systolic murmur has given rise to more medium-frequency components than in the case of the normal subject in Figure 7.13. The AR-model spectra clearly indicate additional and stronger peaks at 140 Hz and 250 Hz, which are confirmed by poles close to the unit circle at the corresponding frequencies in Figure 7.15.

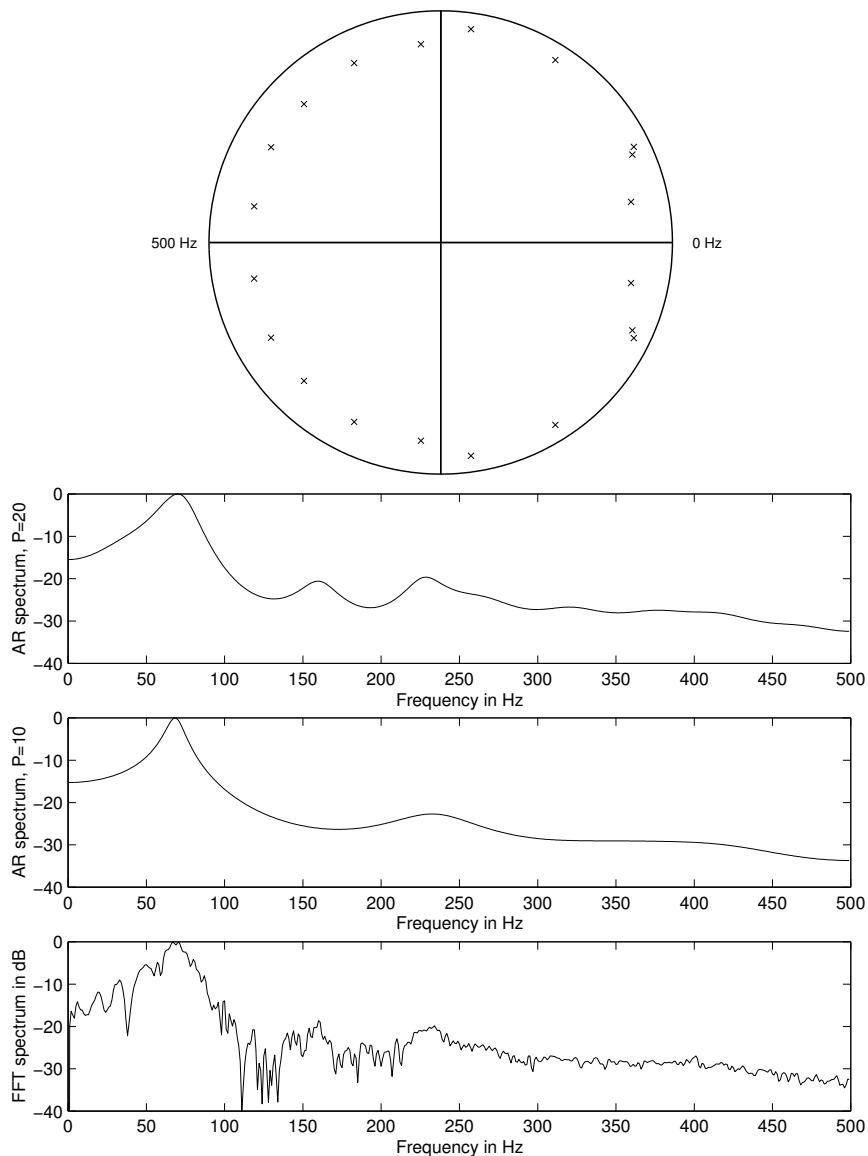


Figure 7.13 Bottom to top: Fourier spectrum of S1 and the systolic portion of the PCG of a normal subject (male, 23 years); AR-model spectrum with order $P = 10$; AR-model spectrum with order $P = 20$; and poles of the AR model with order $P = 20$. (See also Figures 4.30, 5.7, and 6.8.)

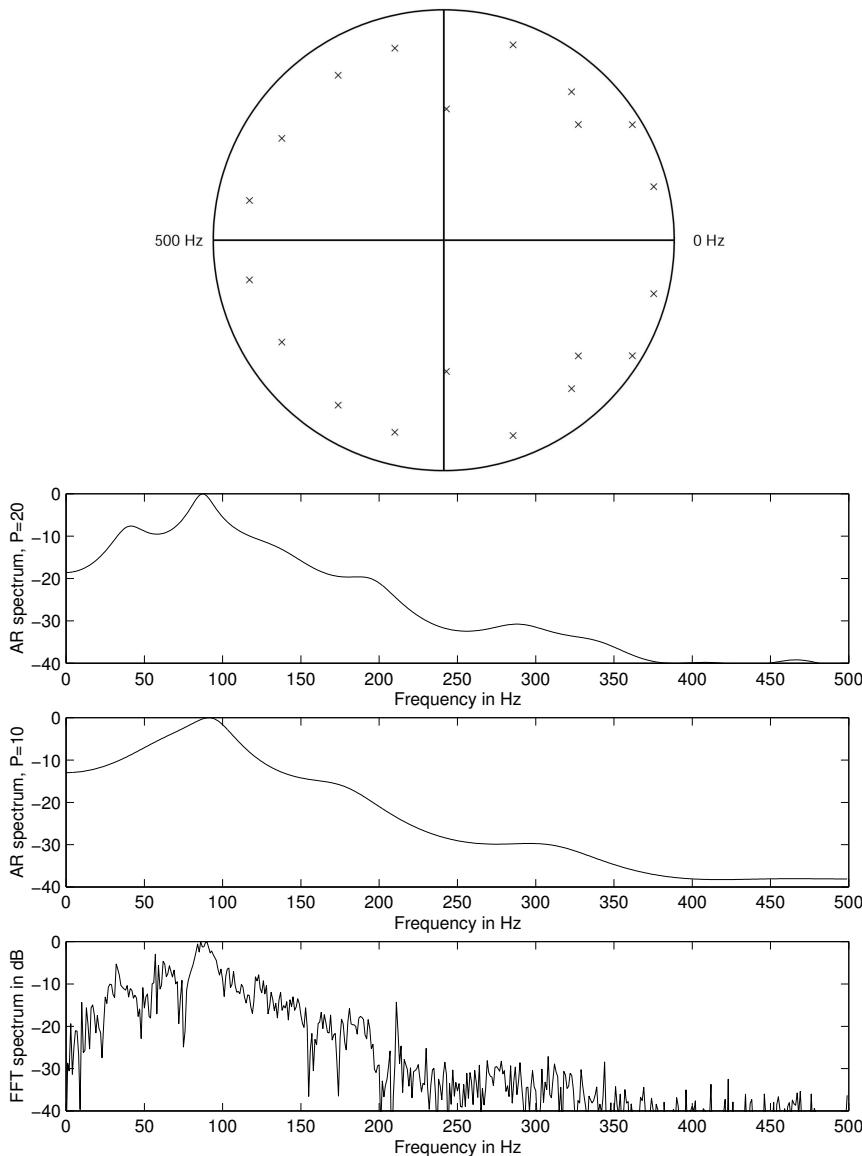


Figure 7.14 Bottom to top: Fourier spectrum of S2 and the diastolic portion of the PCG of a normal subject (male, 23 years); AR-model spectrum with order $P = 10$; AR-model spectrum with order $P = 20$; and poles of the AR model with order $P = 20$. (See also Figures 4.30, 5.7, and 6.8.)

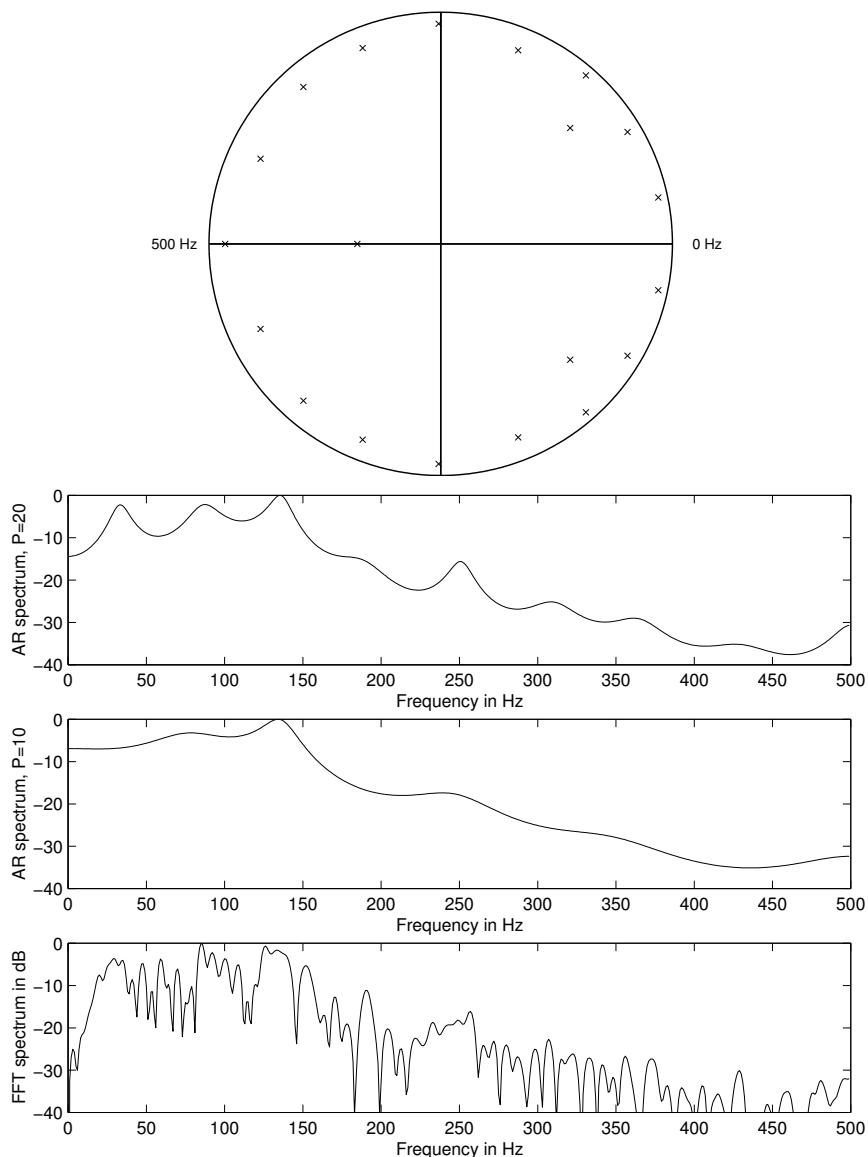


Figure 7.15 Bottom to top: Fourier spectrum of S1 and the systolic portion of the PCG of a subject (female, 14 months) with systolic murmur, split S2, and opening snap of the mitral valve; AR-model spectrum with order $P = 10$; AR-model spectrum with order $P = 20$; and poles of the AR model with order $P = 20$. (See also Figures 4.31, 5.8, and 6.9.)

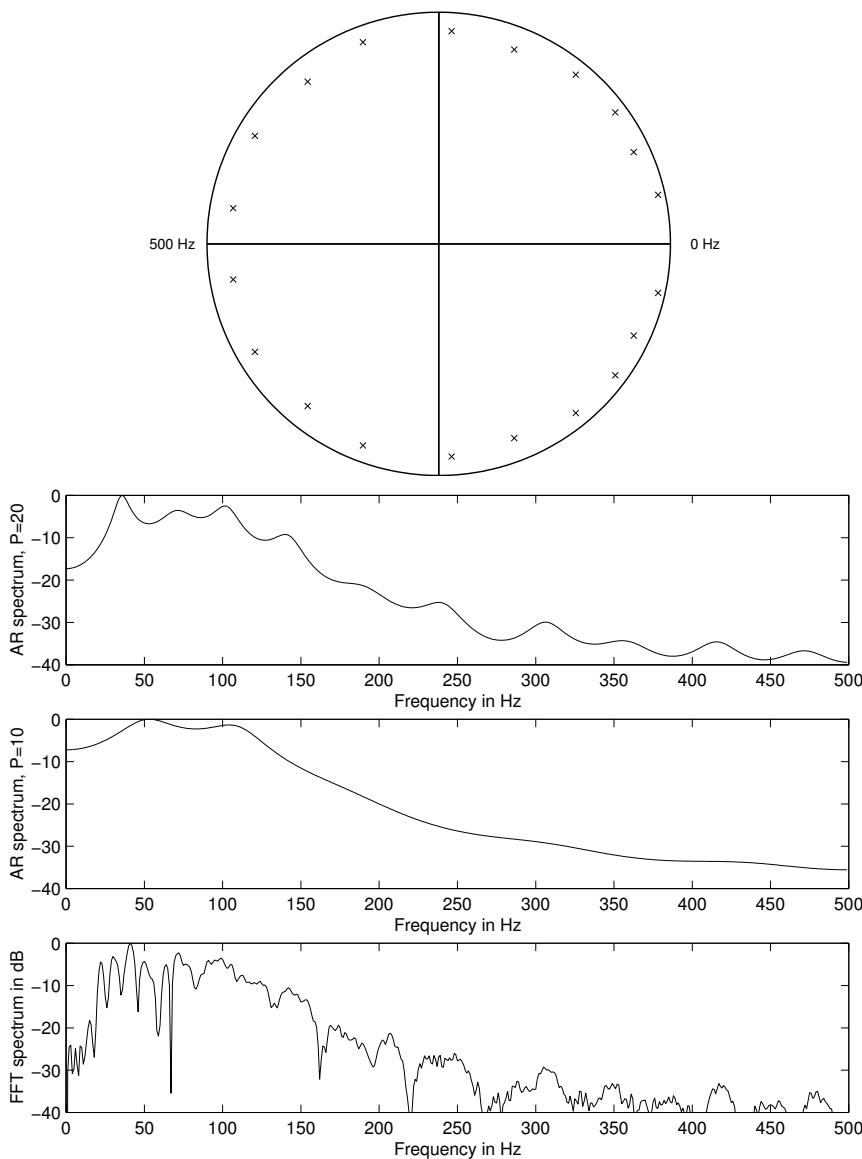


Figure 7.16 Bottom to top: Fourier spectrum of S2 and the diastolic portion of the PCG of a subject (female, 14 months) with systolic murmur, split S2, and opening snap of the mitral valve; AR-model spectrum with order $P = 10$; AR-model spectrum with order $P = 20$; and poles of the AR model with order $P = 20$. (See also Figures 4.31, 5.8, and 6.9.)

7.5.3 AR and cepstral coefficients

If the poles of $H(z)$ are inside the unit circle in the complex z -plane, from the theory of complex variables, $\ln H(z)$ can be expanded into a Laurent series as

$$\ln H(z) = \sum_{n=1}^{\infty} \hat{h}(n) z^{-n}. \quad (7.61)$$

Considering the definition of the complex cepstrum as the inverse z -transform of the logarithm of the z -transform of the signal, and the fact that the LHS of the equation given above represents the z -transform of $\hat{h}(n)$, it is clear that the coefficients of the series $\hat{h}(n)$ are the cepstral coefficients of $h(n)$; see Section 4.7. If $H(z)$ has been approximated by an AR model with coefficients a_k , $1 \leq k \leq P$, we have

$$\ln \left(\frac{1}{1 + \sum_{k=1}^P a_k z^{-k}} \right) = \sum_{n=1}^{\infty} \hat{h}(n) z^{-n}. \quad (7.62)$$

Differentiating both sides of the equation given above with respect to z^{-1} , we get

$$\frac{- \left(\sum_{k=1}^P k a_k z^{-k+1} \right)}{1 + \sum_{k=1}^P a_k z^{-k}} = \sum_{n=1}^{\infty} n \hat{h}(n) z^{-n+1}, \quad (7.63)$$

or

$$- \sum_{k=1}^P k a_k z^{-k+1} = \left(1 + \sum_{k=1}^P a_k z^{-k} \right) \sum_{n=1}^{\infty} n \hat{h}(n) z^{-n+1}. \quad (7.64)$$

By equating the constant term and the like powers of z^{-1} on both sides, the following relationship can be obtained [31]:

$$\begin{aligned} \hat{h}(1) &= -a_1, \\ \hat{h}(n) &= -a_n - \sum_{k=1}^{n-1} \left(1 - \frac{k}{n} \right) a_k \hat{h}(n-k), \quad 1 < n \leq P. \end{aligned} \quad (7.65)$$

As noted in Section 4.7.3, phase unwrapping could cause difficulties in estimating the cepstral coefficients using the inverse Fourier transform of the logarithm of the Fourier transform of a given signal [32]. Estimation of the cepstral coefficients using the AR coefficients has the advantage that it does not require phase unwrapping. Although the cepstral coefficients are deduced from the AR coefficients, it is expected that the nonlinear characteristics of the transformation could lead to an improvement in signal classification using the former than the latter set of coefficients. Cepstral coefficients have provided better classification than AR coefficients in speech [31], EMG [33], and VAG [34] signal analysis.

See Section 7.10 for a discussion on the application of AR modeling for the analysis of PCG signals. See Chisci et al. [35] for details on the application of AR modeling for the extraction of features from intracranial EEG signals and prediction of epileptic seizures. See Sections 8.5 and 8.10 for discussions on the use of AR modeling for segmentation of nonstationary signals.

7.6 Pole-Zero Modeling

Although AR or all-pole modeling can provide good spectral models for any kind of spectra with appropriately high model orders, it has a few limitations. The AR model essentially follows the peaks in the PSD of the signal being modeled, and thus resonance characteristics are represented

well. However, if the signal has spectral nulls or valleys (antiresonance), the AR-model spectrum will not provide a good fit in such spectral segments. Spectral zeros are important in modeling certain signals, such as nasal speech signals [36]. Furthermore, an all-pole model assumes the signal to be a minimum-phase signal or a maximum-phase signal, and does not allow mixed-phase signals [37].

The main conceptual difficulty posed by pole-zero modeling is that it is inherently nonunique, because a zero can be approximated by a large number of poles, and vice versa [25]. However, if the system being modeled has a number of influential zeros, the number of poles required for an all-pole model can become large. For these reasons, ARMA or pole-zero modeling [25, 36–40] is important in certain applications.

The ARMA normal equations: From the ARMA model represented by Equation 7.13, we can write the model PSD as [25]

$$\tilde{S}_y(\omega) = |H(\omega)|^2 = G^2 \frac{|B(\omega)|^2}{|A(\omega)|^2} = G^2 \frac{S_b(\omega)}{S_a(\omega)}, \quad (7.66)$$

where

$$S_a(\omega) = \left| 1 + \sum_{k=1}^P a_k \exp(-jk\omega) \right|^2 \quad (7.67)$$

and

$$S_b(\omega) = \left| 1 + \sum_{l=1}^Q b_l \exp(-jl\omega) \right|^2. \quad (7.68)$$

The total spectral-matching error is given by

$$\varepsilon = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_y(\omega) \frac{S_a(\omega)}{S_b(\omega)} d\omega, \quad (7.69)$$

which may be viewed as the residual energy after passing the modeled signal through the inverse filter $\frac{A(z)}{B(z)}$. In order to obtain the optimal pole-zero model, we need to determine the coefficients a_k and b_l such that the error ε is minimized.

Before taking the derivatives of ε with respect to a_k and b_l , the following relationships are worth noting. Taking the partial derivative of $S_a(\omega)$ in Equation 7.67 with respect to a_i , we get

$$\frac{\partial S_a(\omega)}{\partial a_i} = 2 \sum_{k=0}^P a_k \cos[(i-k)\omega], \quad (7.70)$$

with $a_0 = 1$. Similarly, from Equation 7.68 we get

$$\frac{\partial S_b(\omega)}{\partial b_i} = 2 \sum_{l=0}^Q b_l \cos[(i-l)\omega]. \quad (7.71)$$

Let

$$\phi_{y\alpha\beta}(i) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_y(\omega) \frac{[S_a(\omega)]^\beta}{[S_b(\omega)]^\alpha} \cos(i\omega) d\omega. \quad (7.72)$$

$\phi_{y00}(i)$ is the inverse Fourier transform of $S_y(\omega)$ and hence simply $\phi_y(i)$.

Now, we can take the partial derivative of ε in Equation 7.69 with respect to a_i as

$$\begin{aligned}\frac{\partial \varepsilon}{\partial a_i} &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{S_y(\omega)}{S_b(\omega)} \frac{\partial}{\partial a_i} S_a(\omega) d\omega \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{S_y(\omega)}{S_b(\omega)} 2 \sum_{k=0}^P a_k \cos[(i-k)\omega] d\omega \\ &= 2 \sum_{k=0}^P a_k \phi_{y10}(i-k), \quad 1 \leq i \leq P.\end{aligned}\quad (7.73)$$

In the same manner, we can obtain

$$\frac{\partial \varepsilon}{\partial b_i} = -2 \sum_{l=0}^Q b_l \phi_{y21}(i-l), \quad 1 \leq i \leq Q. \quad (7.74)$$

Because $\phi_{y10}(i-k)$ in Equation 7.73 is not a function of a_k , we obtain a set of linear equations by setting the final expression in Equation 7.73 to zero, which could be solved to obtain the a_k coefficients in a manner similar to the procedures used in AR modeling. However, $\phi_{y21}(i-l)$ in Equation 7.74 is a function of the b_l coefficients, which leads to a set of nonlinear equations that must be solved to obtain the b_l coefficients; the zeros of the model may then be derived from the b_l coefficients. Obtaining the ARMA model, therefore, requires solving P linear equations and Q nonlinear equations.

Iterative solution of the ARMA normal equations: Makhoul [25] describes the following iterative procedure to solve the $(P+Q)$ ARMA model normal equations based on the Newton-Raphson procedure:

Let $\mathbf{a} = [a_1, a_2, \dots, a_P]^T$, $\mathbf{b} = [b_1, b_2, \dots, b_Q]^T$, and $\mathbf{c} = [a_1, a_2, \dots, a_P, b_1, b_2, \dots, b_Q]^T$ represent the model coefficients to be derived in vectorial form. The vector at iteration $(m+1)$ is derived from that at iteration m as

$$\mathbf{c}_{m+1} = \mathbf{c}_m - \mathbf{J}^{-1} \left. \frac{\partial \varepsilon}{\partial \mathbf{c}} \right|_{\mathbf{c}=\mathbf{c}_m}, \quad (7.75)$$

where \mathbf{J} is the $(P+Q) \times (P+Q)$ symmetric Hessian matrix defined as $\mathbf{J} = \frac{\partial^2 \varepsilon}{\partial \mathbf{c} \partial \mathbf{c}^T}$. The vector \mathbf{c} may be partitioned as $\mathbf{c}^T = [\mathbf{a}^T, \mathbf{b}^T]$, and the iterative procedure may be expressed as

$$\begin{bmatrix} \mathbf{a}_{m+1} \\ \mathbf{b}_{m+1} \end{bmatrix} = \begin{bmatrix} \mathbf{a}_m \\ \mathbf{b}_m \end{bmatrix} - \begin{bmatrix} \frac{\partial^2 \varepsilon}{\partial \mathbf{a} \partial \mathbf{a}^T} & \frac{\partial^2 \varepsilon}{\partial \mathbf{a} \partial \mathbf{b}^T} \\ \frac{\partial^2 \varepsilon}{\partial \mathbf{b} \partial \mathbf{a}^T} & \frac{\partial^2 \varepsilon}{\partial \mathbf{b} \partial \mathbf{b}^T} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial \varepsilon}{\partial \mathbf{a}} \\ \frac{\partial \varepsilon}{\partial \mathbf{b}} \end{bmatrix} \quad \begin{array}{ll} \mathbf{a} = \mathbf{a}_m & \mathbf{a} = \mathbf{a}_m \\ \mathbf{b} = \mathbf{b}_m & \mathbf{b} = \mathbf{b}_m \end{array}. \quad (7.76)$$

Equations 7.73 and 7.74 give the first-order partial derivatives required above. The second-order partial derivatives are given as follows [25]:

$$\frac{\partial^2 \varepsilon}{\partial a_i \partial a_j} = 2\phi_{y10}(i-j), \quad (7.77)$$

$$\frac{\partial^2 \varepsilon}{\partial a_i \partial b_j} = -2 \sum_{k=0}^P \sum_{l=0}^Q a_k b_l [\phi_{y20}(j+i-l-k) + \phi_{y20}(j-i-l+k)], \quad (7.78)$$

and

$$\frac{\partial^2 \varepsilon}{\partial b_i \partial b_j} = -2\phi_{y21}(i-j) + 4 \sum_{k=0}^P \sum_{l=0}^Q b_k b_l [\phi_{y31}(j+i-l-k) + \phi_{y31}(j-i-l+k)]. \quad (7.79)$$

(In the present context, i and j are both indices with integral values.) The iterative procedure works well if the initial estimate is close to the optimal model; otherwise, one of the noniterative methods described in the following sections may be considered.

7.6.1 Sequential estimation of poles and zeros

Given the difficulties with the nonlinear nature of direct pole-zero modeling, a few methods have been proposed to split the problem into two parts: identify the poles first by AR modeling, and then treat the residual error in some manner to estimate the zeros [25, 36–40]. (Note: Several notational differences exist between the various references cited here. The following derivations use notations consistent with those used so far in the present chapter.)

Shanks' method: Let us consider a slightly modified version of Equation 7.13 as

$$H(z) = \frac{Y(z)}{X(z)} = \frac{B(z)}{A(z)} = \frac{1 + \sum_{l=1}^Q b_l z^{-l}}{1 + \sum_{k=1}^P a_k z^{-k}}, \quad (7.80)$$

where the gain factor G has been set to be unity: $G = 1$. The difference equation relating the output to the input is given by a small change to Equation 7.12 as

$$y(n) = - \sum_{k=1}^P a_k y(n-k) + \sum_{l=0}^Q b_l x(n-l). \quad (7.81)$$

The effect of the numerator and denominator polynomials in Equation 7.80 may be separated by considering $Y(z) = V(z)B(z)$, where $V(z) = \frac{X(z)}{A(z)}$. This leads to the all-zero or MA part of the system

$$y(n) = \sum_{l=0}^Q b_l v(n-l), \quad (7.82)$$

with $v(n)$ given by the all-pole or AR part of the model as

$$v(n) = - \sum_{k=1}^P a_k v(n-k) + x(n). \quad (7.83)$$

Let us consider the case of determining the a_k and b_l coefficients (equivalently, the poles and zeros) of the system $H(z)$ given its impulse response. Recollect that $y(n) = h(n)$ when $x(n) = \delta(n)$; consequently, we have $X(z) = 1$, and $Y(z) = H(z)$. The impulse response of the system is given by

$$h(n) = - \sum_{k=1}^P a_k h(n-k) + \sum_{l=0}^Q b_l \delta(n-l), \quad (7.84)$$

which simplifies to

$$h(n) = - \sum_{k=1}^P a_k h(n-k), \quad n > Q. \quad (7.85)$$

The effect of the impulse input does not last beyond the number of zeros in the system: the system output is then perfectly predictable from the preceding P samples, and hence an AR or all-pole model is adequate to model $h(n)$ for $n > Q$. As a consequence, Equation 7.32 is modified to

$$\phi_h(i) = - \sum_{k=1}^P a_k \phi_h(i-k), \quad i > Q. \quad (7.86)$$

This system of equations may be solved by considering P equations with $Q+1 \leq i \leq Q+P$. Thus, the a_k coefficients, and hence the poles of the system, may be computed independently of the b_l coefficients or the zeros by restricting the AR error analysis to $n > Q$.

In a practical application, the error of prediction needs to be considered, as the model order P will not be known or some noise will be present in the estimation. Kopec et al. [36] recommend that the covariance method described in Section 7.5 be used to derive the AR model by considering the error of prediction as

$$e(n) = h(n) + \sum_{k=1}^P a_k h(n-k) \quad (7.87)$$

and minimizing the TSE defined as

$$\varepsilon = \sum_{n=Q+1}^{\infty} |e(n)|^2. \quad (7.88)$$

The first Q points are left out as they are not predictable with an all-pole model.

Let us assume that the AR modeling part has been successfully performed by the procedure described above. Let

$$\tilde{A}(z) = 1 + \sum_{k=1}^P \tilde{a}_k z^{-k} \quad (7.89)$$

represent the denominator polynomial of the system that has been estimated. The TSE in modeling $h(n)$ is given by

$$\varepsilon = \sum_{n=0}^{\infty} \left| h(n) - \sum_{l=0}^Q b_l \tilde{v}(n-l) \right|^2, \quad (7.90)$$

where $\tilde{v}(n)$ is the impulse response of the AR model derived, with $\tilde{V}(z) = \frac{1}{\tilde{A}(z)}$. Minimization of ε as above leads to the set of linear equations

$$\sum_{l=0}^Q b_l \phi_{\tilde{v}\tilde{v}}(l, j) = \phi_{h\tilde{v}}(0, j), \quad 0 \leq j \leq Q, \quad (7.91)$$

where

$$\phi_{h\tilde{v}}(l, j) = \sum_{n=0}^{\infty} h(n-l) \tilde{v}(n-j) \quad (7.92)$$

is the CCF between $h(n)$ and $\tilde{v}(n)$, and $\phi_{\tilde{v}\tilde{v}}$ is the ACF of $\tilde{v}(n)$.

The frequency-domain equivalents of the steps above may be analyzed as follows. The TSE is

$$\begin{aligned} \varepsilon &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| H(\omega) \tilde{A}(\omega) - \sum_{l=0}^Q b_l \exp(-jl\omega) \right|^2 |\tilde{V}(\omega)|^2 d\omega \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} |E_h(\omega) - B(\omega)|^2 |\tilde{V}(\omega)|^2 d\omega, \end{aligned} \quad (7.93)$$

where $E_h(z) = H(z)\tilde{A}(z)$ is the AR-model error in the z -domain. [Recall that the Fourier spectrum of a signal is obtained by evaluating the corresponding function of z with $z = \exp(j\omega)$.] The method described above, which is originally attributed to Shanks [38] and has been described as above by Kopec et al. [36], therefore estimates the numerator polynomial of the model by fitting a polynomial (spectral function) to the z -transform of the error of the AR or all-pole model.

Makhoul [25] and Kopec et al. [36] suggest another method labeled as *inverse LP* modeling, where the inverse of the AR-model error $e_h^{-1}(n)$ given as the inverse z -transform of $E_h^{-1}(z)$ is subjected to all-pole modeling. The poles so obtained are the zeros of the original system being modeled.

7.6.2 Iterative system identification

Problem: Given a noisy observation of the output of a linear system in response to a certain input, develop a method to estimate the numerator and denominator polynomials of a rational z -domain model of the system.

Solution: In consideration of the difficulty in solving the nonlinear problem inherent in ARMA modeling or pole-zero estimation, Steiglitz and McBride [39] proposed an iterative procedure based upon an initial estimate of the denominator (AR) polynomial. Since their approach to system identification is slightly different from the LP approach we have been using so far in this chapter, it is appropriate to restate the problem.

The Steiglitz-McBride method: Figure 7.17 shows a block diagram illustrating the problem of system identification. The system is represented by its transfer function $H(z)$, input $x(n)$, output $y(n) = h(n) * x(n)$, and the noisy observation $w(n) = y(n) + \eta(n)$, where $\eta(n)$ is a noise process that is statistically independent of the signals being considered. $H(z)$ is represented as a rational function of z , as

$$H(z) = \frac{Y(z)}{X(z)} = \frac{B(z)}{A(z)} = \frac{\sum_{l=0}^Q b_l z^{-l}}{\sum_{k=0}^P a_k z^{-k}}. \quad (7.94)$$

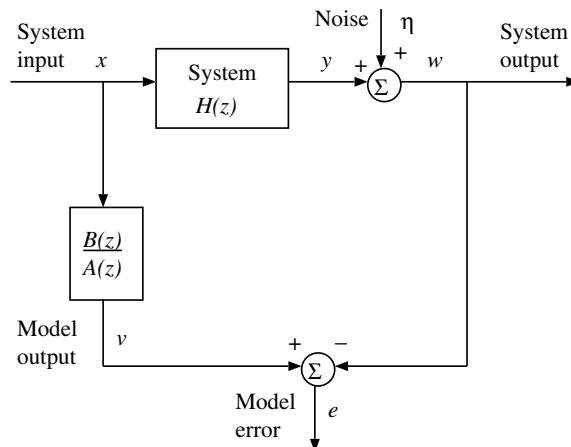


Figure 7.17 Schematic representation of system identification. Adapted from Steiglitz and McBride [39].

The error to be minimized may be written as [39]

$$\sum_{n=0}^{N-1} e^2(n) = \frac{1}{2\pi j} \oint \left| X(z) \frac{B(z)}{A(z)} - W(z) \right|^2 \frac{dz}{z}, \quad (7.95)$$

where the RHS represents the inverse z -transform of the function of z involved, $W(z)$ is the z -transform of $w(n)$, and N is the number of data samples available. The functions of z within the integral essentially compare the predicted model output with the observed output of the physical system.

As seen earlier, this approach leads to a nonlinear problem. The problem may be simplified (linearized) by taking the approach of separate identification of the numerator and denominator polynomials: the estimation problem illustrated in Figure 7.18 treats $A(z)$ and $B(z)$ as separate systems. The error to be minimized may then be written as [39]

$$\sum_{n=0}^{N-1} e^2(n) = \frac{1}{2\pi j} \oint |X(z)B(z) - W(z)A(z)|^2 \frac{dz}{z}. \quad (7.96)$$

(This approach was originally proposed by Kalman [41].)

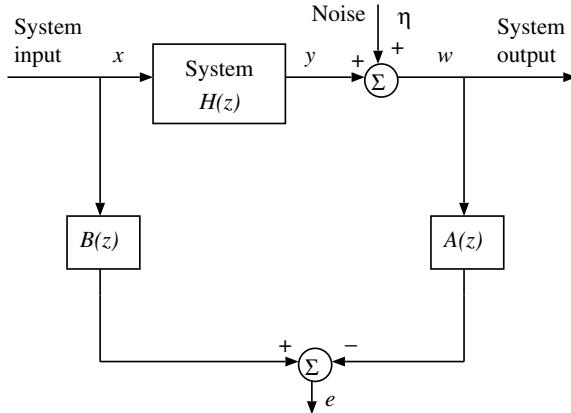


Figure 7.18 Schematic representation of system identification with separate estimation of $A(z)$ and $B(z)$. Adapted from Steiglitz and McBride [39].

The sample-model error is given by

$$e(n) = \sum_{l=0}^Q b_l x(n-l) - \sum_{k=1}^P a_k w(n-k) - w(n). \quad (7.97)$$

The model coefficients and the input-output data samples may be written in vectorial form as

$$\mathbf{c} = [b_0, b_1, \dots, b_Q, -a_1, -a_2, \dots, -a_P]^T \quad (7.98)$$

and

$$\mathbf{d}(n) = [x(n), x(n-1), \dots, x(n-Q), w(n-1), w(n-2), \dots, w(n-P)]^T, \quad (7.99)$$

with the vectors being of size $(P+Q+1)$. Then, the error is given by

$$e(n) = \mathbf{d}^T(n) \mathbf{c} - w(n). \quad (7.100)$$

The condition for minimum TSE is given by

$$\frac{\partial}{\partial \mathbf{c}} \sum_{n=0}^{N-1} e^2(n) = 2 \sum_{n=0}^{N-1} \frac{\partial e(n)}{\partial \mathbf{c}} e(n) = 2 \sum_{n=0}^{N-1} \mathbf{d}(n) e(n) = 0. \quad (7.101)$$

Substitution of the expression for the error in Equation 7.100 in the condition stated above gives

$$\left(\sum_{n=0}^{N-1} \mathbf{d}(n) \mathbf{d}^T(n) \right) \mathbf{c} = \sum_{n=0}^{N-1} w(n) \mathbf{d}(n). \quad (7.102)$$

If we let

$$\Phi = \sum_{n=0}^{N-1} \mathbf{d}(n) \mathbf{d}^T(n) \quad (7.103)$$

represent the $(P+Q+1) \times (P+Q+1)$ correlation matrix of the combined string of input-output data samples $\mathbf{d}(n)$, and let

$$\Theta = \sum_{n=0}^{N-1} w(n) \mathbf{d}(n) \quad (7.104)$$

represent the correlation between the signal $w(n)$ and the data vector $\mathbf{d}(n)$ of size $(P+Q+1)$, we get the solution to the estimation problem as

$$\mathbf{c} = \Phi^{-1} \Theta. \quad (7.105)$$

Although the vectors and matrices related to the filter coefficients and the signal correlation functions are defined in a different manner, the solution obtained above is comparable to that of the optimal Wiener filter (see Section 3.9 and Equation 3.169).

The limitation of the approach described above is that the error used has no physical significance. The separation of the numerator and denominator functions as in Figure 7.18, while simplifying the estimation problem, has led to a situation that is far from reality.

To improve upon the match between the real physical situation and the estimation problem, Steiglitz and McBride [39] proposed an iterative procedure which is schematically illustrated in Figure 7.19. The basic approach is to treat the system identified using the simplified procedure described above as an initial estimate, labeled as $A_1(z)$ and $B_1(z)$; filter the original signals $x(n)$ and $w(n)$ with the system $\frac{1}{A_1(z)}$; use the filtered signals to obtain new estimates $A_2(z)$ and $B_2(z)$; and iterate the procedure until convergence is achieved.

The error to be minimized may be written as [39]

$$\begin{aligned} \sum_{n=0}^{N-1} e^2(n) &= \frac{1}{2\pi j} \oint \left| X(z) \frac{B_i(z)}{A_{i-1}(z)} - W(z) \frac{A_i(z)}{A_{i-1}(z)} \right|^2 \frac{dz}{z} \\ &= \frac{1}{2\pi j} \oint \left| X(z) \frac{B_i(z)}{A_i(z)} - W(z) \right|^2 \left| \frac{A_i(z)}{A_{i-1}(z)} \right|^2 \frac{dz}{z}, \end{aligned} \quad (7.106)$$

with $A_0(z) = 1$. It is obvious that, upon convergence, when $A_i(z) = A_{i-1}(z)$, the minimization problem stated above reduces to the ideal (albeit nonlinear) situation expressed in Equation 7.95 and illustrated in Figure 7.17.

Steiglitz and McBride [39] suggest a modified iterative procedure to improve the estimate further, by imposing the condition that the partial derivatives of the true error criterion with respect to the model coefficients be equal to zero at convergence. The error of the true (ideal) model is given in the z -domain as (refer to Figure 7.17)

$$E(z) = X(z) \frac{B(z)}{A(z)} - W(z) = V(z) - W(z). \quad (7.107)$$

$V(z)$ is the output of the model for the input $x(n)$. The derivatives of $E(z)$ with respect to the model coefficients are given by

$$\frac{\partial \varepsilon}{\partial a_i} = -\frac{X(z)B(z)}{A^2(z)} z^{-i} = -\frac{V(z)}{A(z)} z^{-i} = -\tilde{V}(z)z^{-i} \quad (7.108)$$

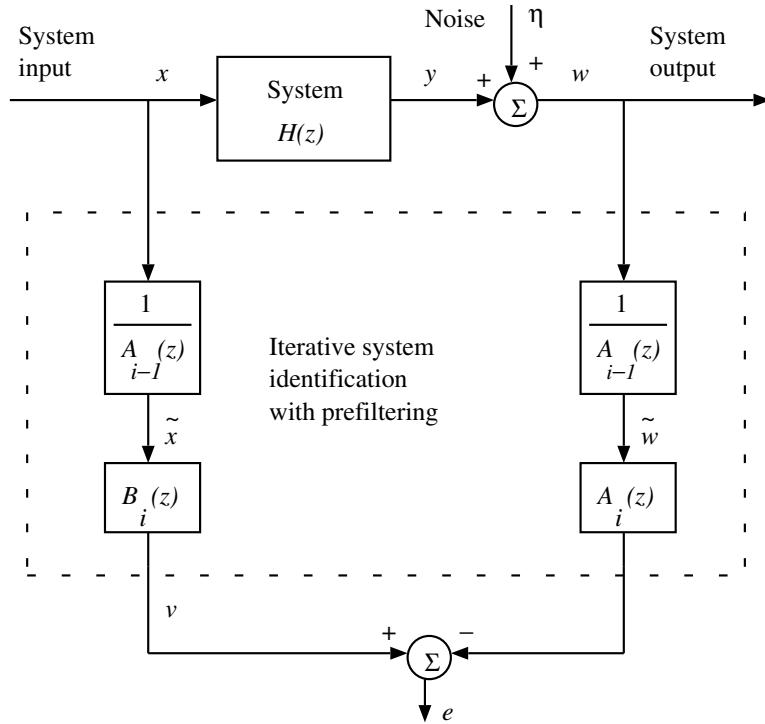


Figure 7.19 Schematic representation of system identification via iterative prefiltering. Adapted from Steiglitz and McBride [39].

and

$$\frac{\partial \varepsilon}{\partial b_i} = \frac{X(z)}{A(z)} z^{-i} = \tilde{X}(z) z^{-i}, \quad (7.109)$$

where the superscript \sim represents a filtered version of the corresponding signal, the filter transfer function being $\frac{1}{A(z)}$. A new data vector is defined as

$$\mathbf{d}_1(n) = [\tilde{x}(n), \tilde{x}(n-1), \dots, \tilde{x}(n-Q), \tilde{v}(n-1), \tilde{v}(n-2), \dots, \tilde{v}(n-P)]^T. \quad (7.110)$$

The error gradient in Equation 7.101 is modified to

$$\begin{aligned} \frac{\partial}{\partial \mathbf{c}} \sum_{n=0}^{N-1} e^2(n) &= 2 \sum_{n=0}^{N-1} \frac{\partial e(n)}{\partial \mathbf{c}} e(n) = 2 \sum_{n=0}^{N-1} \mathbf{d}_1(n) e(n) \\ &= 2 \sum_{n=0}^{N-1} [\mathbf{d}_1(n) \mathbf{d}_1^T(n) \mathbf{c} - w(n) \mathbf{d}_1(n)], \end{aligned} \quad (7.111)$$

where the last equality is true only at convergence. The rest of the procedure remains the same as before, but with the correlation functions defined as

$$\Phi_1 = \sum_{n=0}^{N-1} \mathbf{d}_1(n) \mathbf{d}_1^T(n) \quad (7.112)$$

and

$$\Theta_1 = \sum_{n=0}^{N-1} w(n) \mathbf{d}_1(n). \quad (7.113)$$

Once the a_k and b_l coefficients are obtained, the related polynomials may be solved to obtain the poles and zeros of the system being modeled. Furthermore, the polynomials may be used to derive spectral models of the system or the signal of interest. Note that the procedures given above are applicable to the special case of system identification when the impulse response $h(n)$ is given: We just need to change $x(n) = \delta(n)$ and $X(z) = 1$. Steiglitz and McBride [39] did not provide any proof of convergence of their methods; however, it was indicated that the method performed successfully in many practical applications.

The Steiglitz–McBride method was applied to the modeling and classification of PCG signals by Joo et al. [42]. The first and second peak frequencies were detected from the model spectra and used to analyze porcine prosthetic valve function. Murthy and Prasad [43] applied the Steiglitz–McBride method to ECG signals. Pole–zero models derived from ECG signals including a few cardiac cycles were used to reconstruct and identify the ECG waveform over a single cycle, and also to reconstruct separately (that is, to segment) the P, QRS, and T waves.

7.6.3 Homomorphic prediction and modeling

Problem: Given the relative ease of all-pole modeling, is it possible to convert the zeros of a system to poles?

Solution: As mentioned earlier, an all-pole model assumes the signal being modeled to be a minimum-phase signal or a maximum-phase signal, and does not allow mixed-phase signals [37]. We have seen in Sections 4.7.3 and 5.4.2 that homomorphic filtering can facilitate the separation of the minimum-phase and maximum-phase components of a mixed-phase signal, and further facilitate the derivation of a minimum-phase version or correspondent (MPC) of a mixed-phase signal. Makhoul [25], Oppenheim et al. [37], and Kopec et al. [36] suggested methods to combine homomorphic filtering and LP into a procedure that has been labeled *homomorphic prediction* or *cepstral prediction*.

An intriguing property that arises in homomorphic prediction is that if a signal $x(n)$ has a rational z -transform, then $n\hat{x}(n)$ [where $\hat{x}(n)$ is the complex cepstrum of $x(n)$] has a rational z -transform whose poles correspond to the poles and zeros of $x(n)$. The basic property of the z -transform we need to recollect here is that if $X(z)$ is the z -transform of $x(n)$, then the z -transform of $nx(n)$ is $-z \frac{dX(z)}{dz}$. Now, the complex cepstrum $\hat{x}(n)$ of $x(n)$ is defined as the inverse z -transform of $\hat{X}(z) = \log X(z)$. Therefore, we have

$$ZT[n\hat{x}(n)] = -z \frac{d\hat{X}(z)}{dz} = -z \frac{1}{X(z)} \frac{dX(z)}{dz}, \quad (7.114)$$

where $ZT[\cdot]$ represents the z -transform operator. If $X(z) = \frac{B(z)}{A(z)}$, we get

$$ZT[n\hat{x}(n)] = -z \frac{A(z)B'(z) - B(z)A'(z)}{A(z)B(z)}, \quad (7.115)$$

where the prime ' denotes the derivative of the associated function with respect to z . A general representation of a rational function of z (which represents an exponential signal in the z -domain) in terms of its poles and zeros is given by [37]

$$X(z) = A z^r \frac{\prod_{i=1}^{Q_i} (1 - z_{il} z^{-1})}{\prod_{k=1}^{P_i} (1 - p_{ik} z^{-1})} \frac{\prod_{n=1}^{Q_o} (1 - z_{on} z)}{\prod_{m=1}^{P_o} (1 - p_{om} z)}, \quad (7.116)$$

with the magnitudes of all of the z_i , z_o , p_i , and p_o coefficients being less than unity (see also Sections 3.4.3, 4.7.3, and 5.4.2). The p_i and z_i values above give the P_i poles and Q_i zeros, respectively, of the system that are inside the unit circle in the z -plane; $\frac{1}{p_o}$ and $\frac{1}{z_o}$ give the P_o poles and Q_o zeros,

respectively, that lie outside the unit circle. A causal and stable system will not have any poles outside the unit circle; regardless, the general representation given above permits the analysis and modeling of maximum-phase signals that are anticausal. Computation of the complex cepstrum requires the removal of any linear phase component that may be present, and hence we could impose the condition $r = 0$. We then have

$$\begin{aligned}\hat{X}(z) &= \log X(z) = \log A \\ &+ \sum_{l=1}^{Q_i} \log(1 - z_{il} z^{-1}) + \sum_{n=1}^{Q_o} \log(1 - z_{on} z) \\ &- \sum_{k=1}^{P_i} \log(1 - p_{ik} z^{-1}) - \sum_{m=1}^{P_o} \log(1 - p_{om} z),\end{aligned}\quad (7.117)$$

and, furthermore,

$$\begin{aligned}-z \frac{d\hat{X}(z)}{dz} &= - \sum_{l=1}^{Q_i} \frac{z_{il} z^{-1}}{(1 - z_{il} z^{-1})} + \sum_{n=1}^{Q_o} \frac{z_{on} z}{(1 - z_{on} z)} \\ &+ \sum_{k=1}^{P_i} \frac{p_{ik} z^{-1}}{(1 - p_{ik} z^{-1})} - \sum_{m=1}^{P_o} \frac{p_{om} z}{(1 - p_{om} z)}.\end{aligned}\quad (7.118)$$

From the expression given above, it is evident that $n\hat{x}(n)$ has simple (first-order) poles at every pole as well as every zero of $x(n)$. Therefore, we could apply an all-pole modeling procedure to $n\hat{x}(n)$, and then separate the poles so obtained into the desired poles and zeros of $x(n)$. An initial all-pole model of $x(n)$ can assist in the task of separating the poles from the zeros. Oppenheim et al. [37] show further that even if $X(z)$ is irrational, $n\hat{x}(n)$ has a rational z -transform with first-order poles corresponding to each irrational factor in $X(z)$.

Illustration of application to a synthetic speech signal: Figures 7.20 and 7.21 show examples of the application of pole-zero and all-pole modeling techniques to a synthetic speech signal [37]. The impulse response of a synthetic system with two poles at 292 Hz and 3,500 Hz with bandwidth equal to 79 Hz and 100 Hz, respectively, and one zero at 2,000 Hz with bandwidth equal to 200 Hz is shown in Figure 7.20 (a), with its log-magnitude spectrum in Figure 7.21 (a). The formant or resonance structure of the signal is evident in the spectral peaks. (The sampling rate is 12 kHz.) Excitation of the system with a pulse train with the repetition rate of 120 Hz resulted in the signal in Figure 7.20 (b), whose spectrum is shown in Figure 7.21 (b); the spectrum clearly shows the effect of repetition of the basic wavelet in the series of waves that are superimposed on the basic spectrum of the wavelet. Application of homomorphic filtering to the signal in Figure 7.20 (b) provided an estimate of the basic wavelet as shown in Figure 7.20 (c), with the corresponding spectrum in Figure 7.21 (c).

The pole-zero modeling method of Shanks was applied to the result of homomorphic filtering in Figure 7.20 (c) with four poles and two zeros. The impulse response of the model and the corresponding spectrum are shown in Figure 7.20 (d) and Figure 7.21 (d), respectively. It is seen that the two peaks and the valley in the original spectrum are faithfully reproduced in the modeled spectrum. The frequencies of the poles (and their bandwidths) given by the model are 291 Hz (118 Hz) and 3,498 Hz (128 Hz), and those of the zero are 2,004 Hz (242 Hz), which compare well with those of the synthesized system listed in the preceding paragraph.

Application of the autocorrelation method of LP modeling with six poles to the original signal in Figure 7.20 (a) resulted in the model impulse response and spectrum illustrated in Figures 7.20 (e) and 7.21 (e). While the all-pole model spectrum has followed the spectral peaks well, it has failed to represent the valley or null related to the zero.

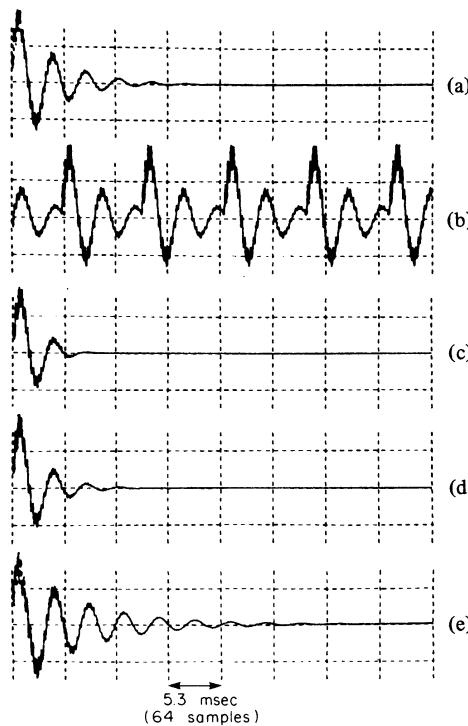


Figure 7.20 Time-domain signals: (a) impulse response of a 4-pole, 2-zero synthetic system; (b) synthesized voiced-speech signal obtained by triggering the system with an impulse train; (c) result of basic wavelet extraction via application of homomorphic filtering to the signal in (b); (d) impulse response of a 4-pole, 2-zero model of the signal in (c) obtained by Shanks' method; and (e) impulse response of a 6-pole AR model. msec = ms. See also Figure 7.21. Reproduced with permission from A.V. Oppenheim, G.E. Kopec, and J.M. Tribble, Signal analysis by homomorphic prediction, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):327–332, 1976. ©IEEE.

Illustration of application to a real speech signal: Figure 7.22 (a) shows the log-magnitude spectrum of a real speech signal (preemphasized) of the nasalized vowel /U/ in the word “moon” [36]. Part (b) of the same figure shows the spectrum after homomorphic filtering to remove the effects of repetition of the basic wavelet. Parts (c) and (d) show 10-pole, 6-zero model spectra obtained using Shanks' method and inverse LP modeling, respectively. The spectra of the models have successfully followed the peaks and valleys in the signal spectrum.

Shanks' method was applied to the minimum-phase and maximum-phase components of ECG signals obtained via homomorphic filtering by Murthy et al. [44]. Akay et al. [45] used ARMA techniques to model diastolic heart sounds for the detection of coronary heart disease; however, only the dominant poles of the model were used in pattern analysis (see Section 7.11 for details of this application).

7.7 Electromechanical Models of Signal Generation

While purely mathematical models of signal generation, such as point processes and linear system models, provide the advantage of theoretical elegance and convenience, they may not be able to represent directly certain physical and physiological aspects of the systems that generate the signals. For example, the models described in the preceding sections cannot directly accommodate the

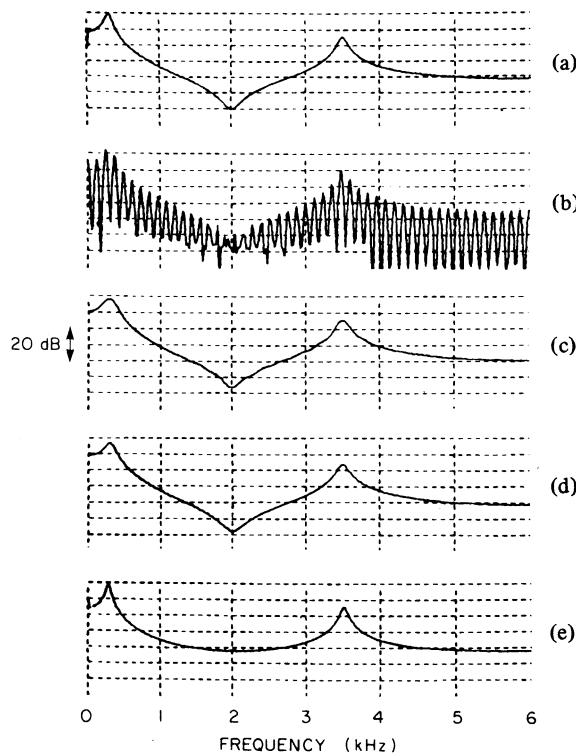


Figure 7.21 Log-magnitude spectra of the time-domain signals in Figure 7.20: (a) actual spectral response of the 4-pole, 2-zero synthetic system; (b) spectrum of the synthesized voiced-speech signal obtained by triggering the system with an impulse train; (c) spectrum of the basic wavelet extracted via application of homomorphic filtering to the signal corresponding to (b); (d) spectral response of a 4-pole, 2-zero model of the signal in (c) obtained by Shanks' method; and (e) spectral response of a 6-pole AR model. Reproduced with permission from A.V. Oppenheim, G.E. Kopec, and J.M. Tribble, Signal analysis by homomorphic prediction, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):327–332, 1976. ©IEEE.

physical dimensions of blood vessels or valves, the loss in the compliance of a valve leaflet due to stenosis, or the lubrication (or the lack thereof) or friction between joint surfaces.

Sikarskie et al. [46] proposed a model to characterize aortic valve vibration for the analysis of its contribution to S2; in addition to mathematical relationships, they included physical factors, such as the valve forcing function, valve mass, and valve stiffness. It was shown that the amplitude and frequency content of A2 depend strongly on the valve forcing function and valve stiffness. Valve mass was shown to have little effect on the amplitude and frequency content of A2; blood density was shown to have no effect on the same parameters.

We now study three representative applications of electromechanical modeling, where mechanical models and their electrical counterparts are used to represent the generation and altered characteristics of sounds in the respiratory system, coronary arteries, and knee joints.

7.7.1 Modeling of respiratory sounds

Problem: Given that parts of the respiratory system are composed of pipe-like segments, propose an electromechanical model to characterize the sounds produced due to respiration.

Solution: Moussavi [47] proposed mechanical as well as electrical circuit models to characterize parts of the respiratory system. The models were used in the analysis of sounds related to breathing.

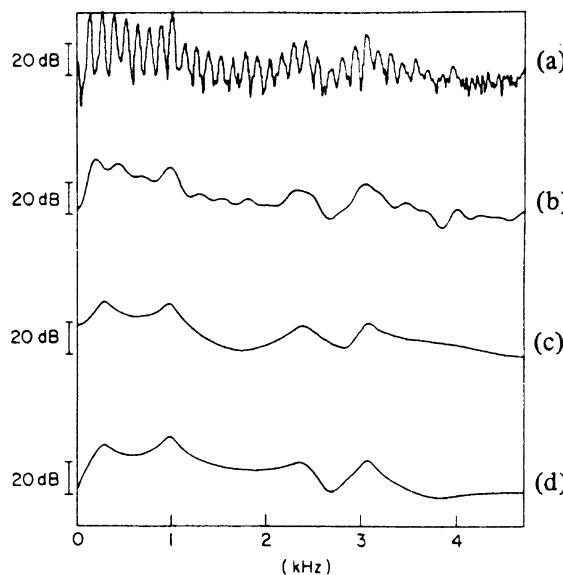


Figure 7.22 (a) Log-magnitude spectrum of the preemphasized, real speech signal of the nasalized vowel /U/ in the word “moon”; (b) spectrum after homomorphic filtering to remove the effects of repetition of the basic wavelet; (c) spectral response of a 10-pole, 6-zero model obtained by Shanks’ method; (d) spectral response of a 10-pole, 6-zero model obtained by inverse LP modeling. Reproduced with permission from G.E. Kopec, A.V. Oppenheim, and J.M. Tribble, Speech analysis by homomorphic prediction, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(1):40–49, 1977. ©IEEE.

Regardless of the complex structure of the human vocal and respiratory tracts, their main characteristics can be represented by mechanical models with tubes or pipes and their equivalent electrical circuit models [47, 48]; see also Section 1.2.13. A simple model of the vocal tract is given by a pipe closed at one end by the glottis and open at the other end at the lips. A pipe with length L has resonance frequencies given by $f_n = \frac{nv}{4L}$, $n = 1, 3, 5, \dots$, where n is the harmonic number, and v is the velocity of air through the pipe. In order to take into account the variations of the respiratory tract along its length, the model could be modified to include multiple short segments of pipes or tubes with different diameters and lengths. Due to the mass of the air in each segment of the pipe model, the related inertance that opposes acceleration needs to be taken into account. Furthermore, the compliance related to the compressibility of air and the elasticity of the pipes’ walls need to be represented. Overall, these considerations lead to a lossy transmission line or circuit model [47, 48]. Figure 7.23 shows a mechanical model and an electrical circuit model for a segment of the respiratory system given by Moussavi [47]; see also Figures 1.51 and 1.52 as well as Flanagan [48].

If a tube is smooth and its walls are hard, loss of energy occurs via viscous friction at the walls of the tube and heat conduction. The viscous loss is proportional to the square of the particle velocity and the loss due to heat conduction is proportional to the square of the pressure [48]. Some of the similarities that can be drawn between mechanical and electrical models for the system described above are as follows [47, 48]: The sound pressure, P , is analogous to voltage, V ; the acoustic volume velocity, U , is comparable to current, I ; viscous loss can be represented by I^2R loss, where R is resistance; heat conduction loss can be related to V^2/R or V^2G loss, where G is conductance; inertance of air mass is analogous to inductance; and compliance can be modeled by capacitance.

Moussavi [47] gives the following derivations for various parameters of the models; see also Flanagan [48]. Starting with Newton’s second law, we have force $F = ma$, where m is the mass and a is the acceleration of an object. In a tube with cross-sectional area A and pressure P , we have $F = PA$; $m = \rho A l$, where ρ is the density of air, and l is the length of the pipe; and $a = \frac{du}{dt}$,

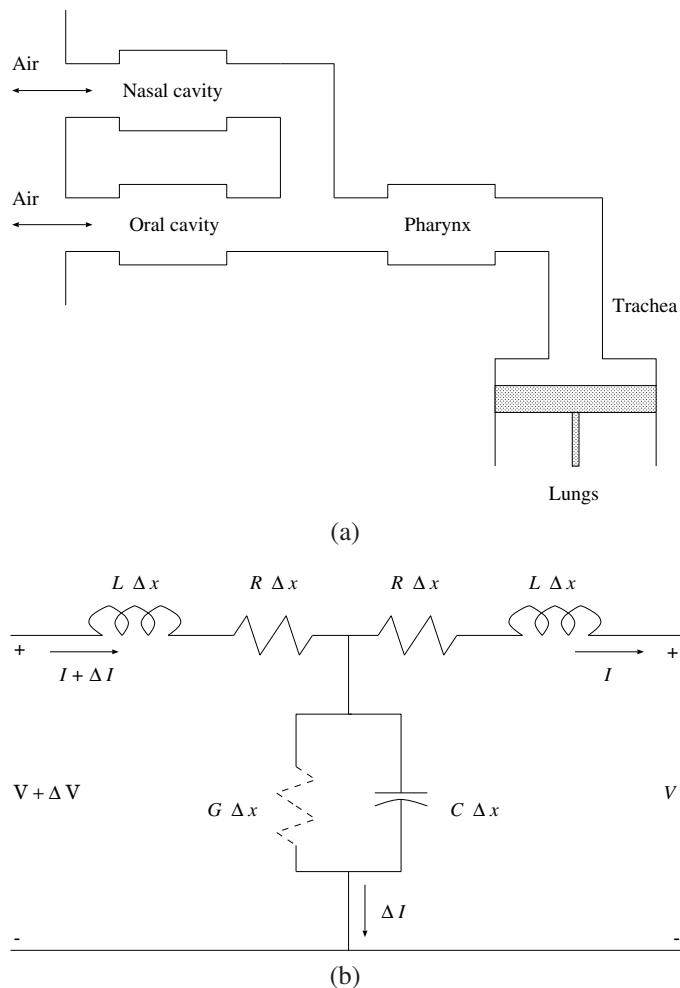


Figure 7.23 (a) Mechanical and (b) electrical circuit models of parts of the respiratory system. Adapted with permission from Z. Moussavi, *Fundamentals of Respiratory Sounds and Analysis*, Morgan & Claypool, San Francisco, CA, 2006, ©Morgan & Claypool.

where u is the particle velocity, and t is time. The volume velocity U is given by $U = Au$. Then, we have

$$PA = \rho A l \frac{du}{dt} = \rho l \frac{dU}{dt}, \quad (7.119)$$

which leads to

$$P = \rho \frac{l}{A} \frac{dU}{dt}. \quad (7.120)$$

Comparing the equation given above with that for the voltage across an inductance L given by

$$V = L \frac{dI}{dt}, \quad (7.121)$$

we have the acoustic equivalent of inductance as

$$L_a = \rho \frac{l}{A}. \quad (7.122)$$

Now, consider the adiabatic gas law $PV_a^\eta = \text{a constant}$, where V_a is the volume of the gas under pressure P , and η is the adiabatic constant. Differentiating this equation with respect to time, we have

$$P \eta V_a^{\eta-1} \frac{dV_a}{dt} + V_a^\eta \frac{dP}{dt} = 0, \quad (7.123)$$

which leads to

$$\frac{1}{P} \frac{dP}{dt} = -\frac{\eta}{V_a} \frac{dV_a}{dt} = \frac{\eta}{V_a} U \quad (7.124)$$

and

$$U = \frac{V_a}{P \eta} \frac{dP}{dt}. \quad (7.125)$$

Comparing the last equation given above with the equation for the current through a capacitor C , expressed as

$$I = C \frac{dV}{dt}, \quad (7.126)$$

we have the acoustic equivalent of capacitance or compliance as

$$C_a = \frac{V_a}{P \eta}. \quad (7.127)$$

Acoustic resistance is given by

$$R_a = \frac{l S}{A^2} \sqrt{\frac{\omega \rho \mu}{2}}, \quad (7.128)$$

where ω is the frequency component of air flow, μ is the viscosity coefficient, and S is the circumference of the pipe, and acoustic conductance is given by

$$G_a = S l \frac{\eta - l}{\rho c^2} \sqrt{\frac{\lambda \omega}{2c_p \rho}}, \quad (7.129)$$

where λ is the coefficient of heat conduction, and c_p is the specific heat of air at constant pressure; see Flanagan [48] and Mousavi [47] for detailed derivations of these expressions.

Parameters derived as above may be used to model segments of tubes or pipes that can be combined to form more elaborate models of parts of the respiratory system. The effects of various segments could be analyzed as filtering of a certain input signal (such as an impulse, a pulse, or random noise) as it passes through the segments. Various respiratory diseases cause narrowing of parts of the respiratory tract as well changes in their smoothness, stiffness, and other mechanical properties. If such pathological changes can be included in the mechanical or electrical model, it becomes possible to analyze the related respiratory sounds and their characteristics. See Mousavi [47] for examples of sounds related to various diseases that affect the respiratory system and their characteristics.

7.7.2 Modeling sound generation in coronary arteries

Problem: Propose an electromechanical model to characterize the sounds produced due to blood flow in stenosed arteries.

Solution: Blood vessels are normally flexible, elastic, and pliant, with smooth internal surfaces. When a segment of a blood vessel is hardened due to the deposition of calcium and other minerals, the segment becomes rigid. Furthermore, the development of plaque inside the vessel causes narrowing or constriction of the vessel, which impedes the flow of blood. The result is turbulent flow of blood, with accompanying high-frequency sounds.

Wang et al. [49,50] proposed a sound-source model combining an incremental-network model of the left coronary artery tree with a transfer-function model describing the resonance characteristics of arterial chambers. The network model, illustrated in Figure 7.24, predicts flow in normal and stenosed arteries. It was noted that stenotic branches may require division into multiple segments in the model due to greater geometric variations. Furthermore, it was observed that a stenotic segment may exhibit poststenotic dilation as illustrated in Figure 7.25, due to increased pressure fluctuations caused by turbulence at the point of stenosis.

The resonance frequency of a segment of an artery depends on the length and diameter of the segment, as well as the distal (away from the heart) hydraulic pressure loading the segment. The physical parameters required for the model were obtained from arteriograms of the patient being examined. The terminal resistances, labeled Z in Figure 7.24, represent loading of the resistive arteriolar beds, assumed to be directly related to the areas that the terminal branches serve.

Wang et al. related the network elements (resistance R , inertance or inductance L , and capacitance C) to the physical parameters of the artery segments as

$$R = 8\pi\nu \frac{l}{A^2}, \quad (7.130)$$

$$L = \rho \frac{l}{A}, \quad (7.131)$$

and

$$C = Alh \frac{D}{E}, \quad (7.132)$$

where $\nu = 0.04 \text{ g cm}^{-1} \text{ s}$ is the viscosity of blood, $\rho = 1.0 \text{ g cm}^{-3}$ is the density of blood, $E = 2 \times 10^6 \text{ g cm}^{-1} \text{ s}^2$ is the Young's modulus of the blood vessel, D is the diameter of the segment, $A = \pi \frac{D^2}{4}$ is the cross-sectional area of the segment, $h \approx 0.08D$ is the wall thickness of the segment, and l is the length of the segment (see also Section 7.7.1). Wang et al. [50] remarked that, while the network elements may be assumed to be approximately constant during diastole, the assumption would not be valid during systole due to variations in the parameters of the segments.

In analyzing the artery–network model, voltage is analogous to pressure (P), and current is analogous to blood flow (Q). State-variable differential equations were used by Wang et al. [50] to derive the flow through the artery tree model for various pressure waveforms. It was hypothesized that turbulence at the point of stenosis would provide the excitation power, and that the stenotic segment and the dilated segment distal to the point of stenosis (see Figure 7.25) would act as resonance chambers.

Wang et al. [49] used the following relationships to compute the RMS pressure fluctuation (see also Fredberg [51]):

$$\langle P^2 \rangle_{\max} = 10^{-4} \rho u^2 f(x), \quad (7.133)$$

$$f(x) = 25.1 - 37.1x + 15.5x^2 - 0.08x^3 - 0.89x^4 + 0.12x^5, \quad (7.134)$$

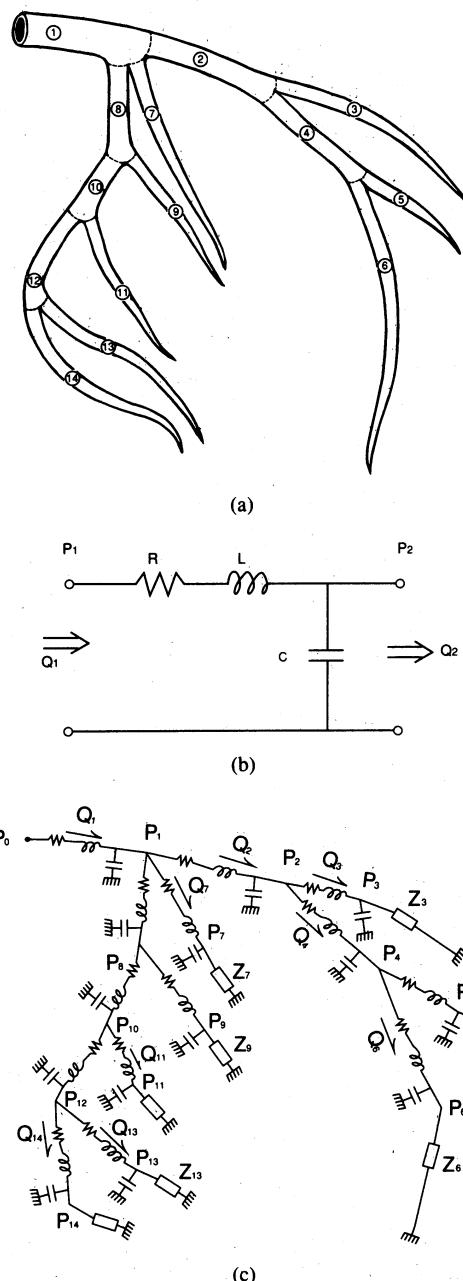
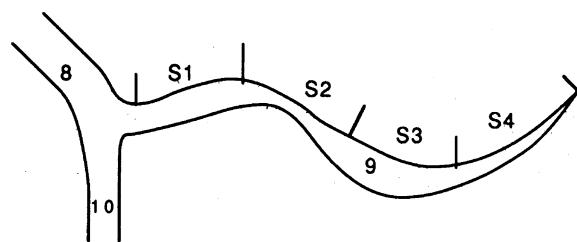


Figure 7.24 Electromechanical model of a coronary artery tree: (a) the left coronary artery tree is divided into 14 branches; (b) circuit model of a segment and (c) circuit model of the artery tree. Reproduced with permission from J.Z. Wang, B. Tie, W. Welkowitz, J.L. Semmlow, and J.B. Kostis, Modeling sound generation in stenosed coronary arteries, *IEEE Transactions on Biomedical Engineering*, 37(11):1087–1094, 1990, ©IEEE; and J.Z. Wang, B. Tie, W. Welkowitz, J. Kostis, and J. Semmlow, Incremental network analogue model of the coronary artery, *Medical & Biological Engineering & Computing*, 27:416–422, 1989. ©IFMBE.



- S1: Proximal to Stenosis Segment**
- S2: Stenotic Segment**
- S3: Poststenotic Dilation Segment**
- S4: Distal to Dilation Segment**

Figure 7.25 Hypothetical example of stenosis in coronary artery branch number 9. Reproduced with permission from J.Z. Wang, B. Tie, W. Welkowitz, J.L. Semmlow, and J.B. Kostis, Modeling sound generation in stenosed coronary arteries, *IEEE Transactions on Biomedical Engineering*, 37(11):1087–1094, 1990. ©IEEE.

and

$$x = 10^{-3} \frac{ud}{\nu} \left(\frac{D}{d} \right)^{0.75}, \quad (7.135)$$

where u is the blood velocity in the stenotic segment, and d is the diameter of the stenotic segment. The incremental network model was used to estimate the blood velocity in each segment.

The wideband spectrum of the sound associated with turbulent flow was modeled as (see also Fredberg [51])

$$S(f) = \frac{0.7 \frac{d}{U} \langle P^2 \rangle_{\max}}{1 + 0.5 [f \frac{d}{U}]^{\frac{10}{3}}}, \quad (7.136)$$

where U is the velocity of blood in a normal segment, and f is frequency in Hz. Wang et al. used the function $S(f)$ given above as the source of excitation power to derive the response of their network model. It was observed that the model spectra indicated two resonance frequencies, the magnitude and frequency of which depended on the geometry and loading of the segments. Wang et al. cautioned that the results of the model are sensitive to errors in the estimation of the required parameters from arteriograms or other sources.

Figure 7.26 illustrates the model spectra for segment number 12 of the artery tree model in Figure 7.24 with no stenosis and with stenosis of two grades. Narrowing of the segment with increasing stenosis is seen to shift the second peak in the spectrum to higher frequencies, while the magnitude and frequency of the first peak are both reduced. The results were confirmed by comparing the model spectra with spectra of signals recorded from a few patients with stenosed coronary arteries. Examples of spectral analysis of signals recorded from patients before and after angioplasty to correct for stenosis are presented in Section 7.11.

7.7.3 Modeling sound generation in knee joints

Problem: Develop a mechanical analog of the knee joint to model the generation of the pulse train related to PPC.

Solution: Beverland et al. [11] studied the PPC signals produced during very slow movement of the leg (at about $4^\circ/s$). The signals were recorded by taping accelerometers to the skin above the upper pole and/or the lower pole of the patella. Reproducible series of bursts of vibration were

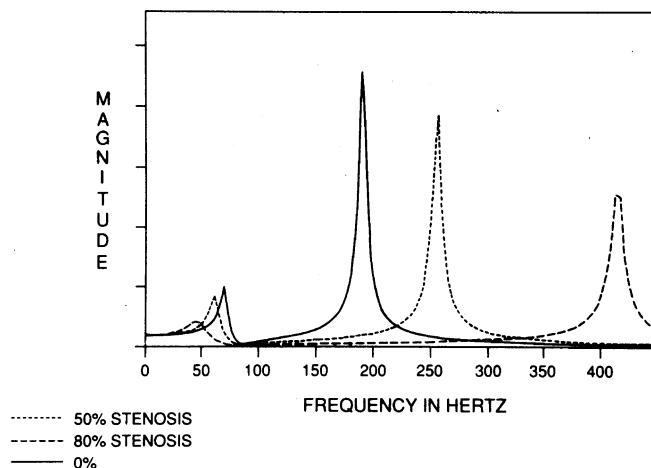


Figure 7.26 Shift in frequency components predicted by the transfer-function model for the case of stenosis in element number 12 in the model of the coronary artery in Figure 7.24. Reproduced with permission from J.Z. Wang, B. Tie, W. Welkowitz, J.L. Semmlow, and J.B. Kostis, Modeling sound generation in stenosed coronary arteries, *IEEE Transactions on Biomedical Engineering*, 37(11):1087–1094, 1990. ©IEEE.

recorded in their experiments. Figure 7.27 illustrates two channels of simultaneously recorded PPC signals from the upper and lower poles of the patella during extension and flexion of the leg. The signals display reversed similarity when extension versus flexion or upper-pole versus lower-pole recordings are compared.

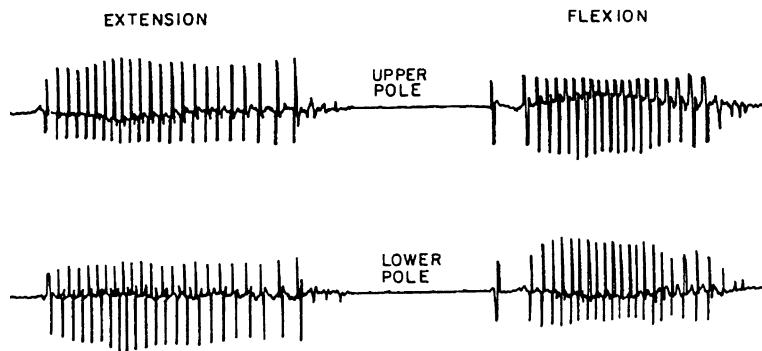


Figure 7.27 Simultaneously recorded PPC signals from the upper and lower poles of the patella during extension and flexion of the leg. The duration of the signal was not specified. Reproduced with permission from D.E. Beverland, W.G. Kernohan, G.F. McCoy, and R.A.B. Mollan, What is physiological patellofemoral crepitus?, *Proceedings of XIV International Conference on Medical and Biological Engineering and VII International Conference on Medical Physics*, pp 1249–1250, Espoo, Finland, 1985. ©IFMBE

Beverland et al. proposed a mechanical model to explain the generation of the PPC signals. The patella was considered to behave like a seesaw in the model, which was supported by the observation that a pivot point exists at the midpoint of the patella. The apparatus constructed, as illustrated in Figure 7.28, included a rubber wheel to represent the trochlear surface of the femur, on top of which was tensioned a rectangular piece of hardboard to represent the patella.

It was argued that, as the wheel in the model is slowly rotated clockwise (representing extension), it would initially stick to the overlying patella (hardboard) due to static friction. This would tend to

impart an anticlockwise rotatory motion, as a rotating cogwheel would impart an opposite rotation to a cog in contact with it (as illustrated in the upper right-hand corner of Figure 7.28). The upper end of the patella would then move toward the wheel. A point would be reached where the static friction would be overcome, when the patella would slip and the rotation is suddenly reversed, with the upper pole jerking outward, and the lower pole jerking inward. The actions would be the opposite to those described above in the case of flexion. The mechanical model was shown to generate signals similar to those recorded from subjects, thereby confirming the stick-slip frictional model for the generation of PPC signals.

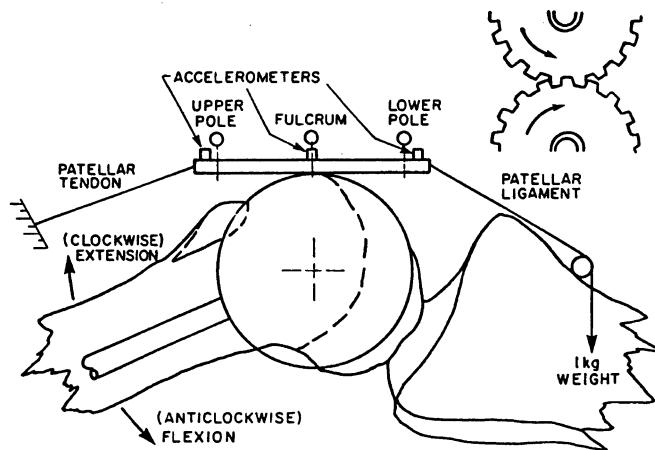


Figure 7.28 Apparatus to mimic the generation of PPC signals via a stick-slip frictional model. Reproduced with permission from D.E. Beverland, W.G. Kernohan, G.F. McCoy, and R.A.B. Mollan, What is physiological patellofemoral crepitus?, *Proceedings of XIV International Conference on Medical and Biological Engineering and VII International Conference on Medical Physics*, pp 1249–1250, Espoo, Finland, 1985. ©IFMBE

7.8 Electrophysiological Models of the Heart

Electrophysiology is the area of study of the electrical characteristics of cells, tissues, and organs. Electrophysiological studies can be conducted on living organisms, excised tissue, synthetic tissues, or their combination; see Section 1.2.1. In the case of the heart, electrical activities can be recorded as intracardiac electrograms by placing electrodes directly on the tissues of interest, as well as ECGs by placing electrodes directly on the chest surface.

Electrophysiological modeling of the heart is a rapidly growing research topic due to the availability of tools and mechanisms to understand the heart's function at the subcellular, cellular, tissue, organ, and system levels. Recent advancements present an intriguing opportunity to gain a better understanding of electrical pathways at all levels, from the genome to the proteome to the physiome. Although substantial details are known about these subsystems' underlying physiological activity, much remains unknown regarding their interoperability, which relates to understanding the healthy and diseased nature of the heart. Integration of cellular models and understanding of the electromechanical interactions between many subsystems have been made possible by 3D modeling of the heart. Models have been developed of blood circulation in the heart's chambers and coronary arteries [52]. Researchers are investigating the biological factors underlying various types of arrhythmia in which regular cardiac excitation fragments into many wavelets during the ventricular fibrillation process [52]. In the following sections, we look into electrophysiological modeling at the cellular, tissue, and organ levels.

7.8.1 Electrophysiological modeling at the cellular level

A cardiac muscle cell or myocyte is composed of a membrane and many ion channels and exchangers. Upon stimulation, a series of fast changes in the membrane potential results in an electrical pulse known as the action potential; see Section 1.2.1. Cardiac action potentials are generated as a result of electrical stimulation beginning at the heart's SA node and spreading across the cardiac muscle tissues; see Section 1.2.5.

The electrical activity begins at the cellular level with the exchange of the four primary ion types Na^+ , K^+ , Ca^+ , and Cl^- , as well as additional ions across the membrane that separates the extracellular and intracellular domains. A membrane transport protein, alternatively referred to as a transporter, is a protein that facilitates the flow of ions, small molecules, and macromolecules across a biological membrane [53]. The two major classes of proteins involved in this form of transport are channels and carriers [52]. A carrier cannot be open to both the external and intracellular environments concurrently. In comparison, a channel can be open to both environments simultaneously, letting molecules to diffuse unhindered. In physical terms, an ion channel is a membrane protein that forms a continuous conduit for the diffusion of ions across the membrane.

Within the heart, the cardiac action potential propagates through three distinct cell types: epicardial, myocardial, and endocardial cells. The three types of cells have slightly varied resting ion concentrations, resulting in slightly different action potentials. Na^+ , Cl^- , and Ca^+ exist in larger external concentrations than K^+ , which has a higher intracellular concentration and is maintained by ion pumps [53]. Without external stimulation, a cardiac cell is electrically at rest, which means there is no change in its transmembrane potential, which is defined as the intracellular voltage minus the extracellular voltage, and is typically between -85 and -90 mV for most cardiac cells. When a stimulus that is stronger than a threshold is applied, the transmembrane voltage increases above its resting state, causing depolarization. Ion-channel gating processes are dependent on the transmembrane voltage and its variations as a result of depolarization, which results in a reduction in transmembrane voltage and return to the resting potential; this is referred to as repolarization. An action potential is formed when a cell undergoes depolarization and repolarization in response to a stimulus; see Section 1.2.1.

Between the extracellular and intracellular compartments of a cardiac cell, there are four basic types of channels: voltage-gated channels, stretch-activated channels, ligand-gated channels, and sodium–calcium exchangers.

A voltage-gated channel is a type of transmembrane protein that functions as an ion channel when the electrical membrane potential near the channel changes. The membrane potential modifies the structure of channel proteins, and thereby controls their opening and closing. The channels open and close in response to changes in the ion concentration, and thus the charge gradient, between the two sides of the cell membrane.

Ligand-gated channels are normally closed and are triggered in response to aberrant chemical concentrations. Additionally, they are referred to as ionotropic receptors. This type of channel is triggered when the level of acetylcholine exceeds typical levels.

Stretch-activated channels are ordinarily closed but open when mechanical pressure is applied. Their enlargement or amplification has an effect on the transmembrane voltage. The existence of these channels is critical for mechanosensitive feedback to occur. Mechanical sensitivity, similar to voltage or ligand sensitivity, is a characteristic shared by channels and other proteins.

Sodium–calcium exchangers are critical for the maintenance of the gradient of ion concentration between intracellular and extracellular regions. To maintain balance, sodium–calcium exchangers and sodium–potassium pumps send sodium out of and potassium into the cell.

The following paragraphs provide a brief historical perspective on the development of neuron and cardiac models, and their role in helping to understand the electrophysiological processes of the heart.

The Hodgkin–Huxley model: The Hodgkin–Huxley model is a series of mathematical equations simulating the current flowing across the surface membrane of a large nerve fiber, specifically, a squid nerve fiber [54]. Hodgkin and Huxley were jointly awarded the 1963 Nobel Prize in physiology and medicine for their ground-breaking research. The Hodgkin–Huxley model quantitatively describes the effects of current flow into and out of a cell, as well as how current flow affects the generated action potential and, consequently, the currents generated by each channel protein in the cell. Additionally, the model incorporates the propagation of action potential in 1D, 2D, and 3D structures [54]. (See also Section 1.2.2.)

Figure 7.29 depicts the Hodgkin–Huxley cell membrane circuit model. Three elements comprise the Hodgkin–Huxley model: channel protein currents, membrane/action potential voltage, and action potential propagation. Capacitors are used to simulate the cell membrane’s charge storage capacity, resistors are used to simulate the various types of ion channels embedded in the membrane, and voltage sources are used to simulate the electrochemical potentials caused by varying intra- and extracellular ion concentrations [54]. Hodgkin and Huxley created a detailed mathematical model of the voltage- and time-dependent characteristics of sodium and potassium conductances based on a series of voltage-clamp studies [54]. Ion-channel currents are modeled as [52]

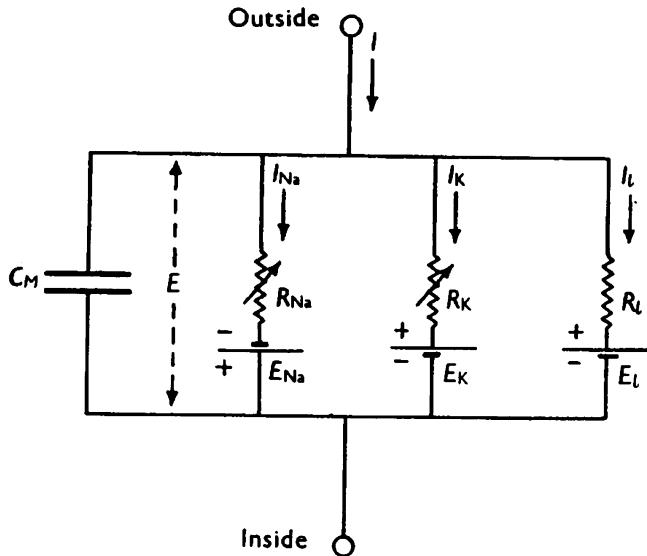


Figure 7.29 Hodgkin–Huxley cell membrane circuit model. $R_{Na} = 1/g_{Na}$, $R_K = 1/g_K$, $R_l = 1/g_l$. R_{Na} and R_K vary with time and membrane potential; the other components are constant [54]. Reproduced with permission from A.L. Hodgkin and A.F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve, *The Journal of Physiology*, 117(4):500–544, 1952. ©The Physiological Society.

$$I_i = g_i N (V_m - E_i) P(o), \quad (7.137)$$

where I_i is the channel current during the open state, g_i is the single-channel conductance, N is the number of channels in the membrane, $P(o)$ is the voltage- and time-dependent probability that the channel is in the open state, V_m is the membrane or action potential, and E_i is the reverse potential (membrane potential) at which there is no net flow of current. Any ion channel with an open gate can be modeled by this generalized equation. With reference to the model in Figure 7.29, i could be one of Na , K , or l , with l representing leakage.

A general equation governing the membrane or action potential generated by a cell is [54]

$$\frac{\partial V}{\partial t} = -\frac{I_{\text{ion}} + I_{\text{stim}}}{C_m}, \quad (7.138)$$

where $\frac{\partial V}{\partial t}$ denotes the change in membrane potential as a function of time, I_{ion} denotes the summation of all channel currents, I_{stim} is the stimulus current, and C_m is the membrane capacitance.

Although Equation 7.138 was designed to model the generation of neuronal action potentials, it may be readily adapted to action potentials of cardiac myocytes with the incorporation of appropriate ion channels.

A general formulation of the reverse potential generated by a cell is given by [55]

$$E_x = \frac{RT}{zF} \log \frac{X_{\text{out}}}{X_{\text{in}}}, \quad (7.139)$$

where R is the gas constant, T is the temperature (in K), z represents ion charge, F is Faraday's constant, $\frac{X_{\text{out}}}{X_{\text{in}}}$ denotes the ratio of the ion concentration outside the cell to that inside, and x could be Na^+ , K^+ , or Ca^{2+} .

The general equation to represent the propagation of the action potentials generated by a collection of cells along a conducting fiber is [54]

$$C_m \frac{\partial V_m}{\partial t} + I_{\text{ion}} = \left(\frac{a}{2r} \right) \left(\frac{\partial^2 V_m}{\partial x^2} \right), \quad (7.140)$$

where V_m denotes the membrane potential, which is now a function of time t and space x ; a denotes the fiber radius; and r is the axial resistance per unit length of the conducting fiber. Note that Equation 7.140 is an expansion of the membrane potential defined in Equation 7.138, as V_m is now a function of both time and space. Additionally, I_{stim} is no longer relevant.

Figure 7.30 shows the simulated action potential of a neuron in the upper plot, and an experimental recording from an axon in the lower plot. Differences in the temperature used in the model and of the experimental setup caused variations in the action potential's temporal characteristics. Notwithstanding a few minor variations in certain details, Hodgkin and Huxley [54] reported good general agreement between the simulated and experimental recordings in terms of amplitude, form, and temporal characteristics.

The Luo–Rudy model: Luo and Rudy [56] proposed a cardiac-specific model of the action potential of ventricular cells based on an extension of the Hodgkin–Huxley model for neuronal action potentials. The model was successful in representing the action potential's depolarization and repolarization phase as well as various electrophysiological phenomena, such as cardiac cell excitability, periodicity, and aperiodic patterns in response to periodic stimulation. However, the model was considered to be passive as it did not account for dynamic changes of intracellular ion concentrations.

To simulate real-life conditions, the Luo–Rudy dynamic (LRd) model was proposed as an extension of the passive model [57]. The LRd model has been widely utilized to simulate action potential propagation in cardiac tissue as it is reasonably easy to change both the duration of the action potential and its rate dependence.

As shown in Figure 7.31, the LRd model includes six currents: a slow inward current I_{SI} , a fast sodium current I_{Na} , a background current I_b , a plateau potassium current I_{Kp} , a time-independent potassium current I_{K1} , and a time-dependent potassium current I_K [56]. The total time-independent potassium current is

$$I_{K1(T)} = I_{K1} + I_{Kp} + I_b. \quad (7.141)$$

Figure 7.32 shows the different parameters computed using the LRd model for fast pacing with a cycle length (CL) of 250 ms and a slow pacing rate with a CL of 1,000 ms. The rate-dependent

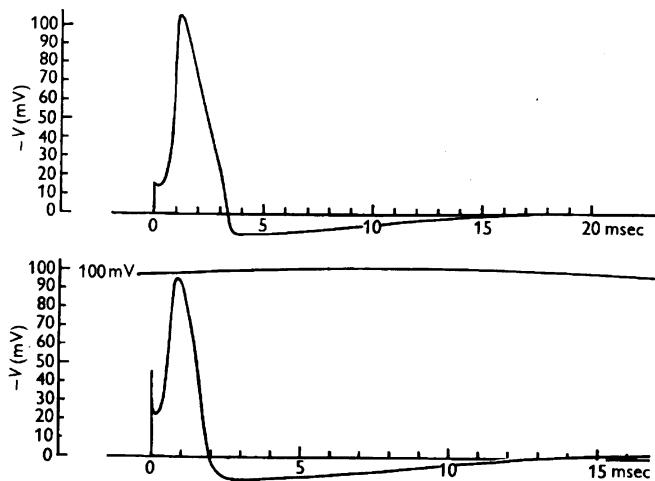


Figure 7.30 Upper plot: Simulated action potential for a stimulus of 15 mV and temperature of 6° C. Lower plot: Experimental recording of an action potential from an axon at 9.1° C. The time scales are different due to the difference in the temperature. Reproduced with permission from A.L. Hodgkin and A.F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve, *The Journal of Physiology*, 117(4): 500–544, 1952. ©The Physiological Society.

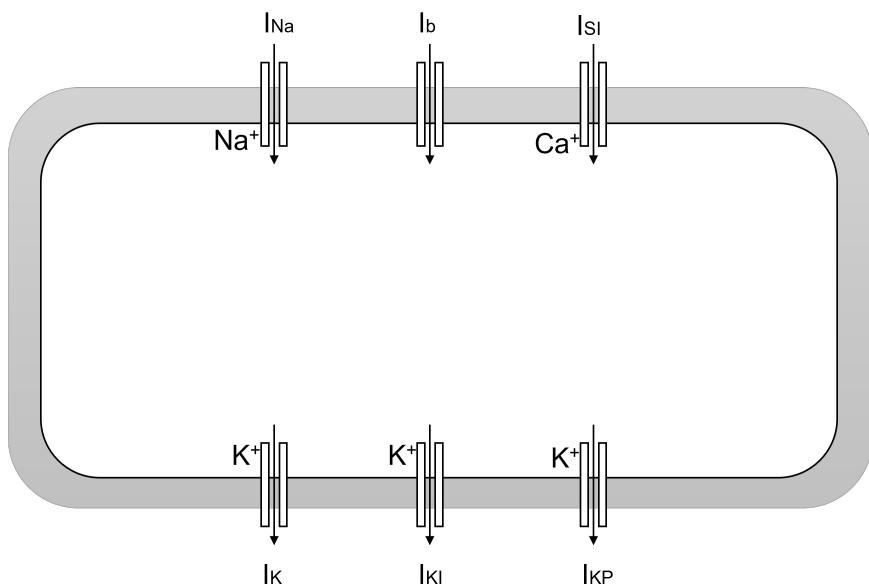


Figure 7.31 The Luo–Rudy electrophysiological cell model. Luo and Rudy [56]/The CellML project.

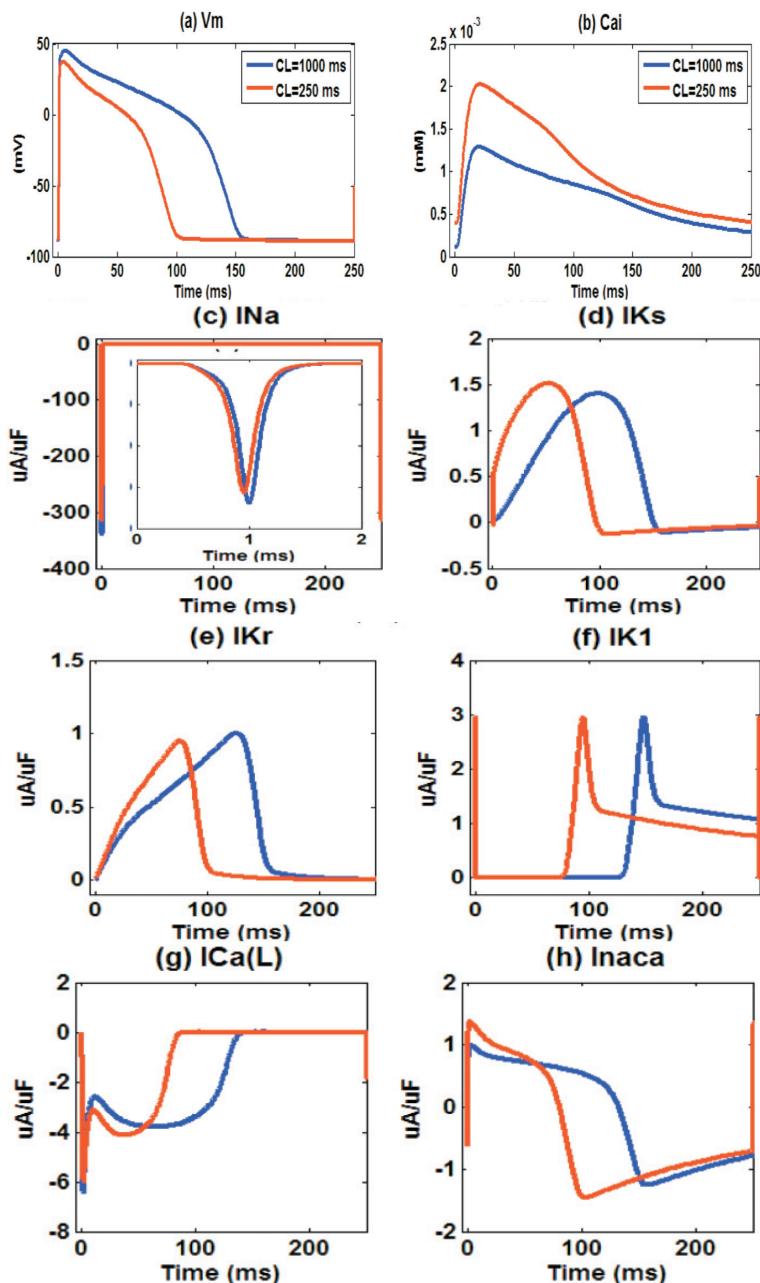


Figure 7.32 LRD model parameters under fast pacing ($CL = 250\text{ ms}$) and slow pacing ($CL = 1,000\text{ ms}$) conditions: (a) action potential, (b) intracellular calcium transient, (c) to (h) major ion-channel currents. Reproduced with permission from Y.H. Chan, C.T. Wu, Y.H. Yeh, and C.T. Kuo. Reappraisal of Luo-Rudy dynamic cell model. *Acta Cardiologica Sinica*, 26:69–80, 2010. ©Taiwan Society of Cardiology.

adaptation of action potential duration may be observed in the morphology of the waveforms, especially in the case of the major ionic currents. Fast pacing causes a higher peak and plateau in L-type Ca^{2+} ($I_{Ca(L)}$) currents, allowing more Ca^{2+} to enter the cytoplasm and boost Ca^{2+} release. Under rapid pacing, the I_{Na} current is smaller, resulting in a lower action potential peak and affecting the activation of other channels, such as the slow K^+ current (I_{Ks}), rapid K^+ current (I_{Kr}), and $I_{Ca(L)}$ that determine action potential duration. Under fast pacing conditions, a quicker increase and larger peak in the I_{Ks} current is observed to efficiently shorten the action potential duration. However, there is no discernible variation in the amplitude of the I_{Kr} peak between slow and rapid pacing, indicating that I_{Kr} is secondary to I_{Ks} in the rate-dependent adaptation of action potential duration.

The total current I_{ion} , which is calculated for a specific time, is given by [56]

$$I_{ion} = I_{Na} + I_K + I_b + I_{Kp} + I_{K1} + I_{SI}. \quad (7.142)$$

These deviate slightly from the governing equation because they incorporate activation and inactivation gates. The open and closed states of each current are modeled using these gates, which simplify and incorporate the states from their corresponding Markov models, which, in turn, provide representations of stochastic processes that undergo transitions from one cell state to another.

The area of electrophysiological modeling at the cellular level has substantially advanced since the first neuronal cell experiments of Hodgkin and Huxley. With the advancements in cellular experimental protocols and mathematical modeling, there has been an improved understanding of cellular signaling, transport phenomena, and cellular responses. However, given the complexity of the cellular mechanisms and the variability associated with different biological organs, the development of robust and explicable models is still an interesting topic for detailed experimental and theoretical investigations.

7.8.2 Electrophysiological modeling at the tissue and organ levels

Individual cardiac cells are interconnected by gap junctions. Cardiac tissue is made up of cardiac cells that are arranged in certain orientations to construct the heart's muscular structure [58]. The gap junctions allow a cardiac cell's action potential to propagate and influence nearby cells. Gap junctions connect cells electrically and chemically throughout the body. The electrical connection has the capability to act swiftly. The functions of cardiac tissues are coordinated via gap junctions, with intercellular signaling occurring in microseconds. The highest conductivity is found in the bundle branches and the Purkinje network, while the lowest is found in the AV node [58, 59]. The gap junctions in cardiac cells are concentrated in the longitudinal direction, resulting in enhanced conductivity in the direction of the fibers. The gap junctions can be disturbed by disease, causing conduction blockages.

Cardiac cells should be connected together to model electrophysiology and describe propagation at the tissue level. This requires formulation of the electrophysiology of the cell through equations that either characterize action potentials generated by cells or represent ion exchange in detail. Because the propagation of an action potential in muscle tissue is a diffusive phenomenon, the diffusion equation from physics may be utilized. The tissue's structure and conductance are expressed as a diffusion tensor in (x, y, z) , the three spatial dimensions. Histology or diffusion tensor imaging can be utilized to obtain this information to represent the diffusion tensor [58, 59].

In terms of tissue representation, cardiac electrophysiological models are classified into two categories: monodomain and bidomain models [58, 59]. The monodomain model of electrical propagation in cardiac tissue is a simplification of the bidomain model. The reduction in complexity is due to the assumption that the anisotropy ratios of the intracellular and extracellular domains are equal. The bidomain model incorporates two interacting compartments that reflect the extracellular and intracellular environments, whereas the monodomain model views the tissue as a single channel

for the diffusion of electrical activity. This distinction between the intracellular and extracellular environments adds detail to the modeling of electrophysiology and makes it more representative of the true physiological situation; however, one must first obtain estimates of intracellular conductivities, which provide the required detailed information.

The monodomain model suggests that cardiac tissue functions as an excitable medium, with voltage diffusion and local stimulation. It is the simplest way to describe action potential propagation [58, 59], with the model

$$\frac{\partial V_m}{\partial t} = \nabla \cdot (\mathbf{D} \nabla V_m) - \frac{I_{\text{ion}} + I_{\text{applied}}}{C_m}, \quad (7.143)$$

where ∇ is the gradient operator, \cdot denotes the dot product that allows the gradient operator to measure scalar field changes in all directions rather than just the direction of the greatest change, V_m denotes the transmembrane voltage, C_m denotes the membrane capacitance, I_{applied} is the external current injected, \mathbf{D} denotes the diffusion tensor, and I_{ion} denotes the ion channel current specified by the cell model utilized. In isotropic instances, the diffusion tensor is a scalar quantity and is related to conductance as

$$D = \frac{G}{S_v C_m}, \quad (7.144)$$

where S_v denotes the cell's surface-to-volume ratio, and G denotes conductance. The model in Equation 7.143 is simple and requires low computational effort to implement. Typically, the boundary condition for the differential equation is defined on the external surface of the heart, assuming no electrical exchanges between the heart and the neighboring organs, and is expressed as [58, 59]

$$\vec{n} \cdot (\mathbf{D} \nabla V_m) = 0, \quad (7.145)$$

where \vec{n} is the unit vector normal to the myocardial surface of the heart. The bidomain model of cardiac tissue views it as a two-phase medium composed of intracellular and extracellular areas [58, 59]. The transmembrane potential is the difference between the intracellular (ϕ_i) and extracellular (ϕ_e) potentials:

$$V_m = \phi_i - \phi_e. \quad (7.146)$$

The following equations contribute to the general form of the bidomain model:

$$\nabla \cdot (\mathbf{D}_i + \mathbf{D}_e) \nabla \phi_e = -\nabla \cdot (\mathbf{D}_i \nabla V_m), \quad (7.147)$$

$$\nabla \cdot (\mathbf{D}_i \nabla V_m) + \nabla \cdot (\mathbf{D}_i \nabla \phi_e) = -S_v I_m, \quad (7.148)$$

where \mathbf{D}_i is the diffusion tensor within the cell, \mathbf{D}_e denotes the diffusion tensor outside the cell, and I_m denotes the transmembrane current. Both V_m and ϕ_e have the same boundary conditions as in the monodomain model [58, 59].

The transmembrane current I_m of the bidomain model consists of a capacitive part and an ionic part, and is given by

$$I_m = C_m \frac{\partial V_m}{\partial t} + I_{\text{ion}}. \quad (7.149)$$

The bidomain model provides additional information as compared to the monodomain model, particularly when an external current is injected, and is useful in defibrillation investigations; however, the bidomain equations are computationally more intensive due to their complicated formulation [58, 59].

Illustration of the generation of the ECG from models of the heart: The ECG can be considered to be a representation of the electrical field generated by the summation of the action potentials of three types of cells in the endocardium, epicardium, and myocardium, and the effects of their propagation [60].

Gima and Rudy [61] established the cellular and ion channel basis for ECG waveforms using a theoretical modeling method. A detailed, physiologically realistic model of the ventricular cell that consists of all critical pumps, exchangers, and ion channels, and accounts for dynamic variations in the concentrations of Na^+ , Ca^{2+} , and K^+ throughout the action potential, was implemented. To represent the three ventricular cell types, transmural heterogeneities were created. The heterogeneities created by the slow and delayed rectifier current (I_{Ks}) and the transient outward current (I_{to}) affect the T wave and the J wave, respectively, in the ECG.

The propagation of action potentials is based on a 1D fiber model as in the LRD model. The theoretical fiber, which measures 1.65 cm in length, is made up of 165 LRD model cells coupled by gap junctions. The density of the ion channels was chosen to reflect the three cell types: endocardial (cells 1–60), midmyocardial (cells 61–105), and epicardial (cells 106–165).

The extracellular ionic values for the simulation performed by Gima and Rudy [61] are $[Na^+]_o = 150 \text{ mmol/l}$, $[K^+]_o = 4 \text{ mmol/l}$, and $[Ca^{2+}]_o = 1.8 \text{ mmol/l}$, where mmol/l stands for millimole per liter. At rest, the intracellular concentrations are $[Na^+]_i = 14.603 \text{ mmol/l}$, $[K^+]_i = 140.516 \text{ mmol/l}$, and $[Ca^{2+}]_i = 0.08278 \text{ mmol/l}$ and alter dynamically during the action potential. The simulation assumed a temperature of 37 °C. Before a stimulus is administered, the fiber is assumed to be in a resting steady-state condition. The duration of an action potential is defined as the time period between the occurrence of the maximum slope and 90% repolarization. The T wave's end is determined by the intersection of a linear fit to the T wave's sharpest component after its peak with the baseline of the ECG.

It is possible to relate patterns in ECG in a deterministic and precise manner to individual membrane ionic currents and action potential features using the simulation method of Gima and Rudy [61]. After establishing the spatial gradient for the propagation of action potentials, the next step is to generate an ECG signal that can be examined and identified as either normal or not. The ECG reconstruction incorporates the wide planar wavefront propagating from the endocardium to the epicardium during ventricular excitation due to the rapid discharge of the Purkinje network. It should be emphasized that, due to the use of a single type of cardiac fiber in the model, the output signal will be a pseudo-ECG. While the model signal shares many of the same features as a real-life ECG and can be related to specific disorders, it is not a true ECG. The resultant pseudo-ECGs appear similar to the examples shown in Figure 7.33. The Gima and Rudy pseudo-ECG generation model enables us to establish a direct link among cellular ionic processes, the action potential, and the shape of external ECG waveforms. The results reveal how altering a specific ionic current alters the gradient potential across the ventricular wall and, hence, the ECG waveforms. The models allow researchers to investigate pathogenic mechanisms related to action potential propagation and the resulting ECG waveforms.

7.8.3 Extensions to the models of the heart

The *Hund–Rudy model* [62] is a canine action potential model that incorporates a large number of the protein channels that are not present in the LRD model. Hund and Rudy [62] developed a model of the action potential of the canine ventricular epicardium and calcium cycling, and applied it to investigate the underlying mechanisms of the rate dependence of Ca^{2+} transients and action potential duration. They presented a dynamic model of the canine ventricular epicardial cell that emulates experimentally obtained action potential and Ca^{2+} transient values accurately over a wide range of pacing frequency. Due to the wide frequency range of arrhythmia and the relationship between cellular electrophysiology and Ca^{2+} cycling during arrhythmogenesis, this model is a useful investigation tool.

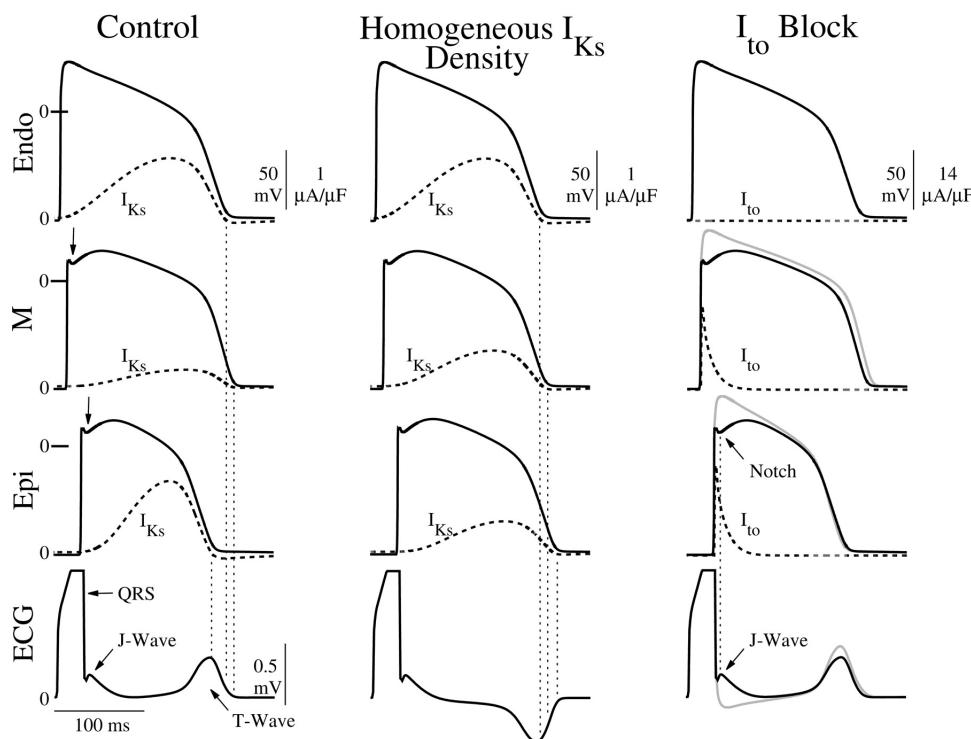


Figure 7.33 Top to bottom rows: Simulated action potentials of endocardial (Endo), myocardial (M), and epicardial (Epi) cells are presented, for control (left), homogeneous transmural I_{Ks} density (center), and in the absence of I_{to} (right), along with the computed ECG. Left: The T wave is inscribed by the transmural voltage gradient caused by the inhomogeneous distribution of I_{Ks} across the ventricular wall. The complete repolarization of the epicardium and midmyocardium occurs at the T wave's peak and end, respectively (vertical dotted lines). The arrows denote the notch formed by I_{to} in epicardial and M cells. The electrocardiographic J wave is inscribed by the transmural gradient caused by the heterogeneity of I_{to} . Center: T-wave inversion occurs when the intrinsic heterogeneity of I_{Ks} is removed, causing the repolarization sequence to follow the activation sequence. Right: I_{to} blocking suppresses the notch and the J wave. Reproduced with permission from K. Gima and Y. Rudy. Ionic current basis of electrocardiographic waveforms: a model study. *Circulation Research*, 90(8):889–896, 2002. ©American Heart Association.

The *Tusscher-Panfilov model* [55] is centered on the formation of human cardiac action potentials and incorporates ion channels that are not included in the LRD Model. ten Tusscher et al. [55] created a mathematical model of human ventricular cell action potentials that, while electrophysiologically detailed, is computationally efficient so as to be useful in large-scale spatial simulations for the study of reentrant arrhythmia. The primary ionic currents used in the model are fitted to data from investigations into human ventricular myocytes and cardiac channel expression, which is an important aspect of the model. The model's restoration of conduction velocity is higher than that of other models. A 2D sheet of human ventricular tissue was used to describe the dynamics of spiral-wave rotation, and the results show that the spiral wave has a convoluted meandering pattern with a period of 265 ms. The model, in general, replicates different electrophysiological features and can be used to study reentrant arrhythmia in human ventricular tissue.

The *Rogers-McCulloch model* [63] is based on the FitzHugh-Nagumo equations [58] for excitable media. It is capable of generating 2D action potential propagation patterns and simulating spiral-pattern action potentials, but lacks the equations and techniques necessary to generate an ECG signal. The computer model presented by Rogers and McCulloch [63] was created to help

researchers better understand how cardiac action potentials propagate through geometrically and structurally complex areas of the heart. In theory, the method can be used in any model of impulse propagation that treats the heart as a continuous medium. The governing finite-element equations were developed using the collocation approach, and the spatial variations of time-dependent excitation and recovery variables were evaluated using cubic Hermite interpolation. The results show that the finite-element method (FEM) is well suited for the assessment of normal and abnormal cardiac activity.

The *Davies et al. model* [64] is a modified version of the Hund–Rudy model that refocused the model to study compounds or drugs affecting midmyocardial tissue. This is the first instance of an ensemble of model variants being used to provide an accurate representation of how effects may vary within a population. The results indicate that *in silico* models have the potential to offer value to preclinical assessment of cardiac risk and to be consolidated into the practices of pharmaceutical industries and regulatory agencies.

The *Geodesic-based Earliest Activation Sites Identification (GEASI) model* [65] is capable of simulating a 3D mesh containing an unlimited number of early activation sites in the Purkinje network. The model is capable of doing so by analyzing a given ECG signal and then generating comparable ECG signals. In essence, it can replicate a patient's ECG using a 3D mesh of early activation sites. This model is complex and computationally costly because it is based on Eikonal equations (nonlinear first-order partial differential equations) related to Hamilton–Jacobi formulations.

A *simplified 3D heart model* has been proposed by Sovilj et al. [66]. Their model offers a balance between model complexity and computational efficiency. The model could be implemented as a research tool to simulate cell, tissue, and whole-heart properties under various pathological conditions, and to synthesize the corresponding ECG signals.

To understand cardiac mechanisms connected to rhythm and morphology of surface ECGs, the aforementioned mathematical and computer models of the heart are required due to the limitations in clinical and experimental work with human hearts. Given that the models are primarily based on linear and nonlinear differential equations, advances in numerical implementation of the equations will allow us to implement and observe certain complex events related to arrhythmia and other cardiac disorders.

7.8.4 Challenges and future considerations in modeling the heart

Animal hearts are employed in the majority of scientific investigations, but human hearts are used in rare circumstances, such as after heart transplantation (see Section 8.16). Models are inherently variable for a variety of reasons, including inconsistency in experimental design and biological variation; such differences are inadequately described, impeding continued improvement in experimental and theoretical methodologies [58].

There are still issues remaining in collecting adequate data to generate a complete action potential model from single-cell models. The variety of physiological phenomena that mathematical models may capture is enormous, and newer models attempt to address more and more of the related complexities. Often, determining the set of virtual experiments that should be run on a model to assess whether it is the appropriate model for a new application can be challenging and time consuming [58].

Apart from the issue of obtaining the required amount of data, the nonlinearity of mathematical models and the accompanying numerical optimization problems complicate the task of finding an effective solution. Additionally, as model complexity increases to incorporate dozens of variables and hundreds of parameters, the number of local minima often grows larger and causes more problems [58]. Realizable models contain fewer equations; a primary advantage is greater computational tractability, which enables larger, longer, and more spatially detailed 3D simulations [58]. In terms of implementation, when combined with cell models, both monodomain and bidomain models in-

clude a system of differential equations that lacks a closed-form solution [58]; as a consequence, numerical approaches need to be used to resolve them. Euler integration is the simplest method to solve first-order ordinary differential equations. In reality, the Euler method is extremely fast, yet susceptible to instability and inaccuracy. The modified Euler technique is similar to the forward Euler method, except that it finds the solution using the trapezoidal rule. The trapezoidal rule approximates the integral as a trapezoid regardless of the path traveled. Rather than looking backward to identify the next point, this solution determines the result by combining the past and current points.

The most well-known member of the Runge–Kutta family, which involves a sequence of iterations, is typically referred to as RK4. When paired with an iterative adaptive step-size technique, RK4 is an excellent option for the numerical solution of differential equations. The disadvantages of Runge–Kutta methods are that they consume much more computer time than multistep approaches with comparable accuracy, and that they do not necessarily provide accurate truncation error estimates.

The electrophysiological system is decomposed into simpler components using mesh-generation techniques from a computational standpoint. FEM and the finite-difference method (FDM) are the two primary numerical approaches used for this purpose. Keener and Bogar [67] provide a viable FDM-based solution for complex systems of differential equations in bidomain models. FEM is more extensively employed because it performs better with complicated domains and is more accurate in approximating values between nodes; however, it is less stable than FDM [68]. Austin et al. [69], dos Santos et al. [70], and Vigmond et al. [71] developed several FEM-based techniques for bidomain models. Generally, the choice of one method over another is determined by the nature of the problem. FEM is preferred for monodomain and bidomain models of cardiac electrophysiology.

Personalized medicine and digital twins: Personalized medicine is an emerging trend in health-care that focuses on developing optimal treatment and intervention procedures for a particular patient at a particular time through the use of personalized mathematical models [72]. Personalized medicine is advancing as a result of the continuing development of sophisticated mathematical models and the application of advanced computing resources. Engineering and computational models are used by medical practitioners in attempts to recreate physiological systems. The models are used to mimic the following: the outcomes of diseases; the short- and long-term effects of therapy (drug-based or not); and the overall functionality of a system from the protein level to the organ level. The models, termed digital twins, provide computer representations of individuals that dynamically reflect their chemical, physiological, and lifestyle states through time [73]. Digital twins enable tailored treatment without compromising the patient's safety [72, 73]. The models are also used to replicate the following: harmful short- and long-term negative effects associated with drug testing, with no guarantee of a cure for the problem, and long-term negative consequences of surgical treatment that may not improve the situation.

Digital twins provide for rapid evaluation of a variety of therapeutic options for a specific illness in a constrained time frame. They enable the identification of effective treatments for diseases or conditions that would be difficult or impossible to study in clinical trials due to interindividual heterogeneity. Digital twins circumvent the limitations of a one-size-fits-all strategy by modeling distinct patient-specific effects of diseases using individualized data. This approach addresses the restrictions and limitations associated with physical models or physical replications of circumstances utilizing animal models and cellular cultures [74]; the demand for genetics specialists to build 2D and 3D cell culture models; the environment required to improve physiological system function; and the inability of *in vivo* and *ex vivo* animal and cell culture models to reproduce accurately critical characteristics of human systems, such as the heart [74].

Specifically, in the instance of cardiac digital twins, mathematical equations are used to recreate the heart's functionality under normal and unusual settings. The methods produce either specific cardiac activities or a plethora of interrelated cardiac functions from the protein level to the out-

put ECG signal. Single-function models may be used to model only a single parameter, such as the action potential or the ECG signal, using a set of equations. Complex models may attempt to represent processes ranging from the currents of protein channels and their interactions with one another that generate action potentials, to action potential propagation, and the model's ECG signal. The integrated knowledge that could be gained by analyzing and understanding the complex mechanisms of cardiac function could help the research community and industry practitioners in a variety of applications and in understanding the pathophysiological processes of cardiac disorders.

The domain of electrophysiological modeling and simulation at multiple levels holds significant promise as computer implementation of complex nonlinear mathematical equations continues to advance, and could provide new paths in the analysis of cause and effect through signal analysis applied to a specific human organ, such as the heart, the brain, or the lung.

7.9 Application: Analysis of Heart-rate Variability

Problem: Explore the applicability of Fourier spectral analysis methods to study HRV.

Solution: DeBoer et al. [6] applied Fourier analysis techniques to two types of measures derived from heart-rate data (see also Akselrod et al. [75]). They noted that the standard Fourier analysis methods cannot be applied directly to a series of point events. Therefore, they derived three types of signals from trains of ECG beats as illustrated in Figure 7.1.

The *interval spectrum* was derived by computing the Fourier spectrum of the interval series, normalized as $\tilde{I}_k = (I_k - \bar{I}) / \bar{I}$, where \bar{I} is the mean interval length. The frequency axis was scaled by considering the time-domain data to be spaced at distances equal to the mean interval length \bar{I} , that is, the effective sampling frequency is $1 / \bar{I}$.

The *spectrum of counts* was derived by taking the Fourier transform of the impulse-train representation, derived from *RR* interval data as shown in Figure 7.1. The signal was normalized and scaled as $\tilde{s}(t) = \sum [\bar{I} \delta(t - t_k)] - N$, where N is the number of data samples, and the Fourier transform was computed. The spectra computed were smoothed with a 27-point rectangular window. DeBoer et al. demonstrated that the two spectra exhibit similar characteristics under certain conditions of slow or slight modulation of the data about the mean heart rate.

The *RR* interval data of a subject breathing freely and the two spectra derived from the data are shown in Figure 7.34. Three peaks are seen in both the spectra, which were explained as follows [6]:

- the effect of respiration at about 0.3 Hz;
- the peak at 0.1 Hz related to 10 s waves seen in the BP; and
- a peak at a frequency lower than 0.1 Hz caused by the thermoregulatory system.

Figure 7.35 shows the *RR* interval data and spectra for a subject breathing at a fixed rate of 0.16 Hz. The spectra display well-defined peaks at both the average heart rate (1.06 Hz) and at the breathing rate, as well as their harmonics. The spectra clearly illustrate the effect of respiration on the heart rate, and may be used to analyze the coupling between the cardiovascular and respiratory systems (see Section 2.2.4).

Note that direct Fourier analysis of a stream of ECG signals will not provide the same information as above. The reduced representation (model) of the *RR* interval data, as illustrated in Figure 7.1, has permitted Fourier analysis of the heart rate and its relationship with respiration. The methods have application in studies on HRV [76–82].

As described in Section 2.4.3, Mendez et al. [83] proposed a method for the detection of OSA based on the ECG recorded during sleep. Features that gave good results in the detection of OSA included coherence between the *RR* interval data and the QRS area, as well as measures of HRV.

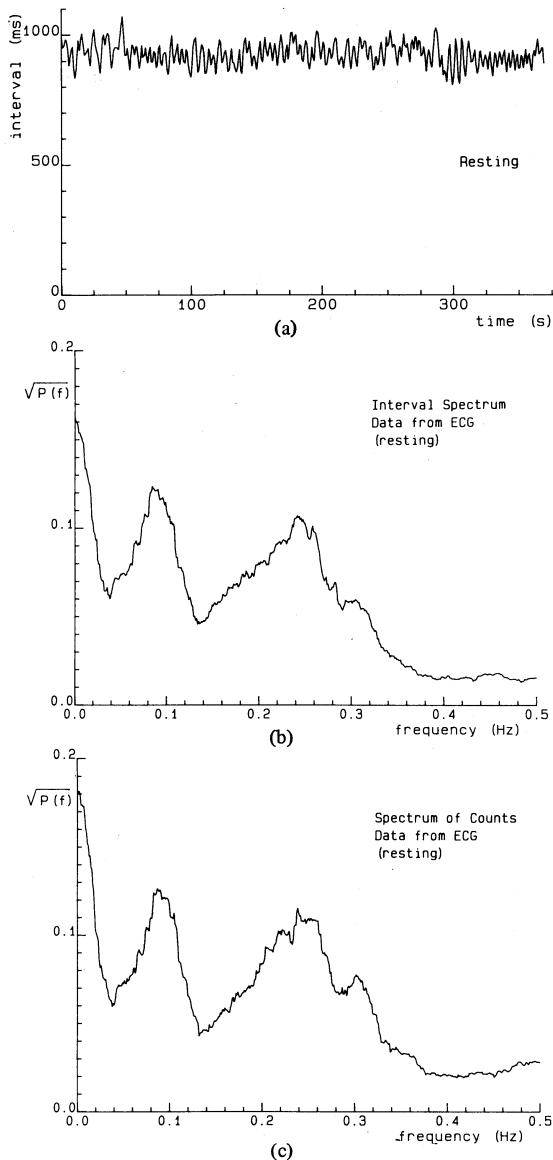


Figure 7.34 (a) 400 RR interval values from a healthy subject breathing freely. (b) Interval spectrum computed from a total of 940 intervals, including the 400 shown in (a) at the beginning. (c) Spectrum of counts. The spectra are shown for the range $0 - 0.5 \text{ Hz}$ only. Reproduced with permission from R.W. DeBoer, J.M. Karemaker, and J. Strackee, Comparing spectra of a series of point events, particularly for heart rate variability studies, *IEEE Transactions on Biomedical Engineering*, 31(4):384–387, 1984. ©IEEE.

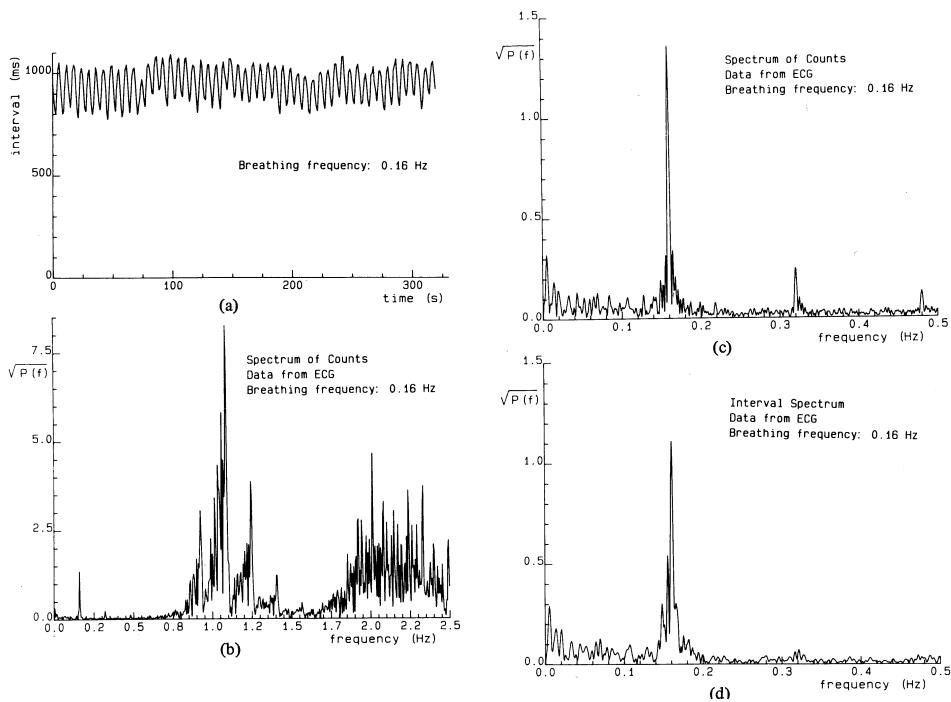


Figure 7.35 (a) 340 *RR* interval values from a healthy subject breathing at a fixed rate of 0.16 Hz. (b) Spectrum of counts for the range 0 – 2.5 Hz. (c) Spectrum of counts for the range 0 – 0.5 Hz. (d) Interval spectrum. Reproduced with permission from R.W. DeBoer, J.M. Karemaker, and J. Strackee, Comparing spectra of a series of point events particularly for heart rate variability studies, *IEEE Transactions on Biomedical Engineering*, 31(4):384–387, 1984. ©IEEE.

7.10 Application: Spectral Modeling and Analysis of PCG Signals

Iwata et al. [84,85] applied AR modeling and parametric spectral analysis techniques to PCG signals for the detection of murmurs as well as the detection of the onset of S1 and S2. Their techniques included AR modeling, extraction of the dominant poles for pattern classification, and spectral tracking, which are explained in the following paragraphs.

Dominant poles: After the a_k , $k = 1, 2, \dots, P$, coefficients of an all-pole or AR model of order P have been computed, the polynomial $A(z)$ may be factorized and solved to obtain the locations of the poles p_k , $k = 1, 2, \dots, P$, of the system. The closer a pole is to the unit circle in the z -plane, the narrower is its bandwidth and the stronger is its contribution to the impulse response of the system. Poles that are close to the unit circle may be related to the resonance frequencies of the system, and could be used in system identification and pattern recognition.

In view of the nonstationary nature of the PCG signal, Iwata et al. [84] computed a new model with order $P = 8$ for every window or frame of duration 25 ms, allowing an overlap of 12.5 ms between adjacent frames (with the sampling rate $f_s = 2$ kHz). The frequency of a pole p_k was calculated as

$$f_k = \frac{\angle p_k}{2\pi} f_s, \quad (7.150)$$

and its bandwidth was calculated as

$$bw_k = \frac{\log |p_k|}{\pi} f_s. \quad (7.151)$$

Conditions based upon the difference in the spectral power estimate of the model from one frame to the next, and the existence of poles with $f_k < 300 \text{ Hz}$ with the minimal bandwidth for the model considered, were used to segment each PCG signal into four phases: S1, a systolic phase spanning the S1 – S2 interval, S2, and a diastolic phase spanning the interval from one S2 to the following S1. (See also Section 4.9.)

Figures 7.36 and 7.37 show the PCG signals, spectral contours, the spectral power estimate, and the dominant poles for a normal subject and a patient with murmur due to patent ductus arteriosus. Most of the dominant poles of the model for the normal subject are below 300 Hz. The model for the patient with murmur indicates many dominant poles above 300 Hz.

The mean and SD of the poles with $bw_k < 80 \text{ Hz}$ of the model of each PCG phase were computed and used for pattern classification. The five coefficients of a fourth-order polynomial fitted to the series of spectral power estimates of the models for each phase were also used as features. Twenty-six out of 29 design samples and 14 out of 19 test samples were correctly classified. However, the number of cases was low compared to the number of features used in most of the six categories present in the samples (normal, aortic insufficiency, mitral insufficiency, mitral stenosis, aortic stenosis combined with insufficiency, and atrial septal defect).

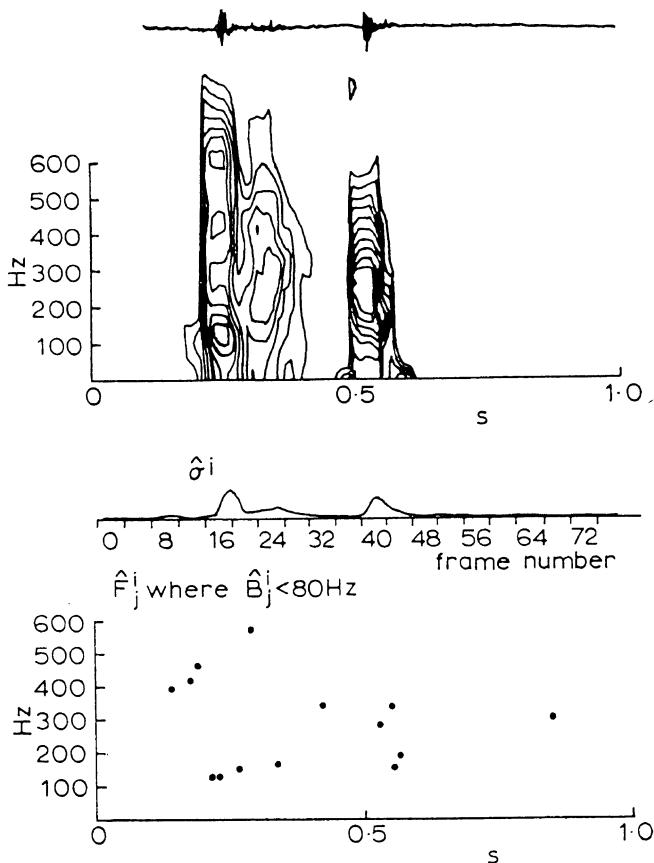


Figure 7.36 Illustration of feature extraction based upon all-pole modeling of a normal PCG signal. From top to bottom: PCG signal; model spectrum in the form of isointensity contours; model spectral power estimate $\hat{\sigma}^i$, where i refers to the frame number; the frequencies \hat{F}_j^i of the dominant poles with bandwidth $\hat{B}_j^i < 80 \text{ Hz}$. Reproduced with permission from A. Iwata, N. Suzumara, and K. Ikegaya, Pattern classification of the phonocardiogram using linear prediction analysis, *Medical & Biological Engineering & Computing*, 15:407–412, 1977. ©IFMBE.

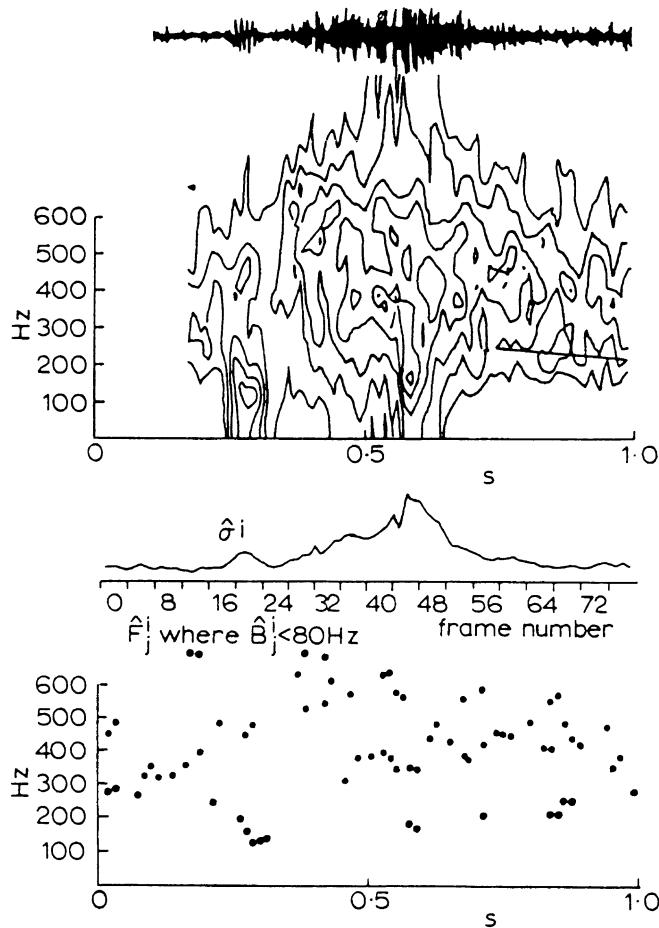


Figure 7.37 Illustration of feature extraction based upon all-pole modeling of the PCG signal of a patient with murmur due to patent ductus arteriosus. From top to bottom: PCG signal; model spectrum in the form of iso-intensity contours; model spectral power estimate $\hat{\sigma}^i$, where i refers to the frame number; the frequencies \hat{F}_j^i of the dominant poles with bandwidth $\hat{B}_j^i < 80$ Hz. Reproduced with permission from A. Iwata, N. Suzumura, and K. Ikegaya, Pattern classification of the phonocardiogram using linear prediction analysis, *Medical & Biological Engineering & Computing*, 15:407–412, 1977. ©IFMBE.

Spectral tracking: In another application of AR modeling for the analysis of PCG signals, Iwata et al. [85] proposed a spectral-tracking procedure based upon AR modeling to detect S1 and S2. PCG signals were recorded at the apex with a highpass filter that, at 100 Hz, had a gain -40 dB below the peak gain at 300 Hz (labeled Ap-H). The signals were lowpass filtered with a gain of -20 dB at 1,000 Hz and sampled at 2 kHz. The AR model was computed with order $P = 8$ for frames of length 25 ms; the frame-advance interval was only 5 ms. The model PSD was computed as

$$\tilde{S}(\omega) = \frac{\sigma_r^2}{2\pi} \frac{1}{\sum_{k=0}^P \phi_a(k) \cos(\omega T)}, \quad (7.152)$$

where

$$\phi_a(k) = \sum_{j=0}^{P-k} a_j a_{j+k}, \quad (7.153)$$

with a_k being the AR-model coefficients, $P = 8$, $T = 0.5 \text{ ms}$, σ_r^2 being the model residual energy (error), and $a_0 = 1$.

Based on a study of the spectra of 69 normal and abnormal PCG signals, Iwata et al. [85] found the mean peak frequency of S1 to be 127 Hz, and that of S2 to be 170 Hz; it should be noted that the PCG signals were highpass filtered (as described in the preceding paragraph) at the time of data acquisition. The model spectral power at 100 Hz was used as the tracking function to detect S1: The peak in the tracking function after the location t_R of the R wave in the ECG was taken to be the position of S1. The tracking function to detect S2 was based on the spectral power at 150 Hz; the peak in the interval $t_R + 0.25RR \leq t \leq t_R + 0.6RR$, where RR is the interbeat interval, was treated as the position of S2. The use of a normalized spectral density function based on the AR-model coefficients but without the σ_r^2 factor in Equation 7.152 was recommended, in order to overcome problems due to the occurrence of murmurs close to S2.

Figure 7.38 illustrates the performance of the tracking procedure with a normal PCG signal. The peaks in the 100 Hz and 150 Hz spectral-tracking functions (lowest traces) coincide well with the timing instants of S1 and S2, respectively. Figure 7.39 illustrates the application of the tracking procedure to the PCG signal of a patient with mitral insufficiency. The systolic murmur completely fills the interval between S1 and S2, and no separation is seen between the sounds and the murmur. While the 150 Hz spectral-tracking function labeled (b) in the figure does not demonstrate a clear peak related to S2, the normalized spectral-tracking function labeled (c) shows a clear peak corresponding to S2. The two additional PCG traces shown at the bottom of the figure (labeled Ap-L for the apex channel including more low-frequency components with a gain of -20 dB at 40 Hz, and 3L-H for a channel recorded at the third left intercostal space with the same bandwidth as the Ap-H signal) illustrate S2 more distinctly than the Ap-H signal, confirming the peak location of the spectral-tracking function labeled (c) in the figure.

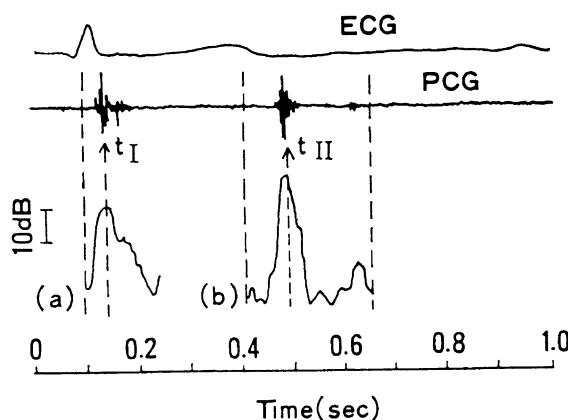


Figure 7.38 Illustration of the detection of S1 and S2 via spectral tracking based upon all-pole modeling of a normal PCG signal. From top to bottom: ECG signal; PCG signal; spectral-tracking functions at (a) 100 Hz for S1 and (b) 150 Hz for S2. The S1 and S2 locations detected are marked as t_1 and t_{II} , respectively. Reproduced with permission from A. Iwata, N. Ishii, N. Suzumura, and K. Ikegaya, Algorithm for detecting the first and the second heart sounds by spectral tracking, *Medical & Biological Engineering & Computing*, 18:19–26, 1980. ©IFMBE.

7.11 Application: Detection of Coronary Artery Disease

The diastolic segment of a normal PCG signal after S2 is typically silent; in particular, the central portion of the diastolic segment after the possible occurrence of AV valve-opening snaps is silent.

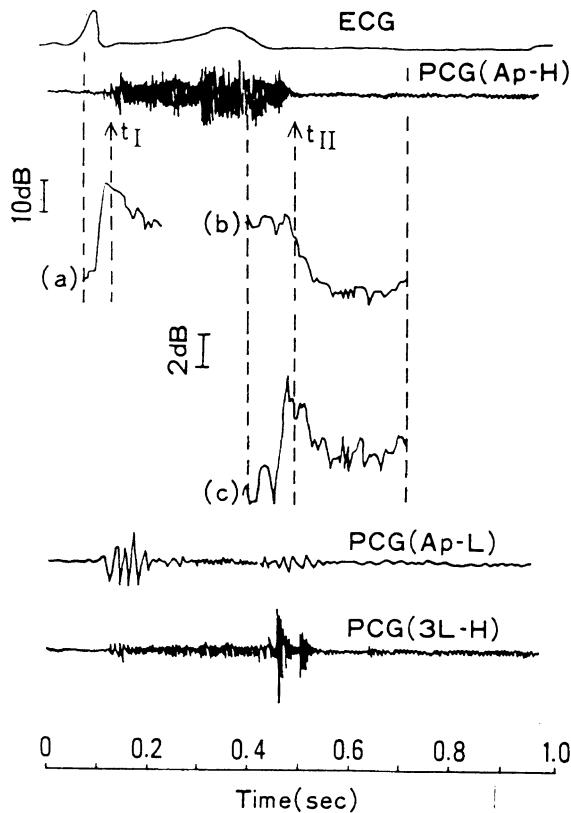


Figure 7.39 Illustration of the detection of S1 and S2 via spectral tracking based upon all-pole modeling of a PCG signal with systolic murmur due to mitral insufficiency. From top to bottom: ECG signal; PCG (Ap-H) signal; spectral-tracking functions at (a) 100 Hz for S1, (b) 150 Hz for S2, and (c) normalized spectral-tracking function at 150 Hz for S2; PCG (Ap-L) signal from the apex with more low-frequency components included; and PCG (3L-H) signal from the third left intercostal space with the same filters as for Ap-H. The S1 and S2 locations detected are marked as t_1 and t_{II} , respectively. Reproduced with permission from A. Iwata, N. Ishii, N. Suzumura, and K. Ikegaya, Algorithm for detecting the first and the second heart sounds by spectral tracking, *Medical & Biological Engineering & Computing*, 18:19–26, 1980. ©IFMBE.

Akay et al. [86] conjectured that blood flow in the coronary arteries is at its maximum during middiastole, and further that the effects of coronary artery disease, such as occlusion or stenosis, could present high-frequency sounds in this period due to turbulent blood flow (see Section 7.7.2).

Akay et al. [86] studied the spectra of middiastolic segments of the PCGs, averaged over 20 – 30 beats, of normal subjects and patients with coronary artery disease confirmed by angiography. It was found that the PCG signals in the case of coronary artery disease exhibited greater portions of their energy above 300 Hz than the normal signals.

Figure 7.40 illustrates the AR-model spectra of two normal subjects and two patients with coronary artery disease. The signals related to coronary artery disease are seen to possess a high-frequency peak in the range 400 – 600 Hz that is not evident in the normal cases.

Akay et al. [87] further found that the relatively high power levels of resonance frequencies in the range of 400 – 600 Hz that were evident in patients with coronary artery disease were reduced after angioplasty. Figure 7.41 shows the spectra of the diastolic heart sounds of a patient before and after coronary artery occlusion was corrected by angioplasty. It may be readily observed that the high-frequency components that were present before surgery (“preang.”) are not present after the

treatment (“postang.”). (The minimum-norm method of PSD estimation used by Akay et al. [87] — labeled as “MINORM” in the figure — is not discussed in this book.)

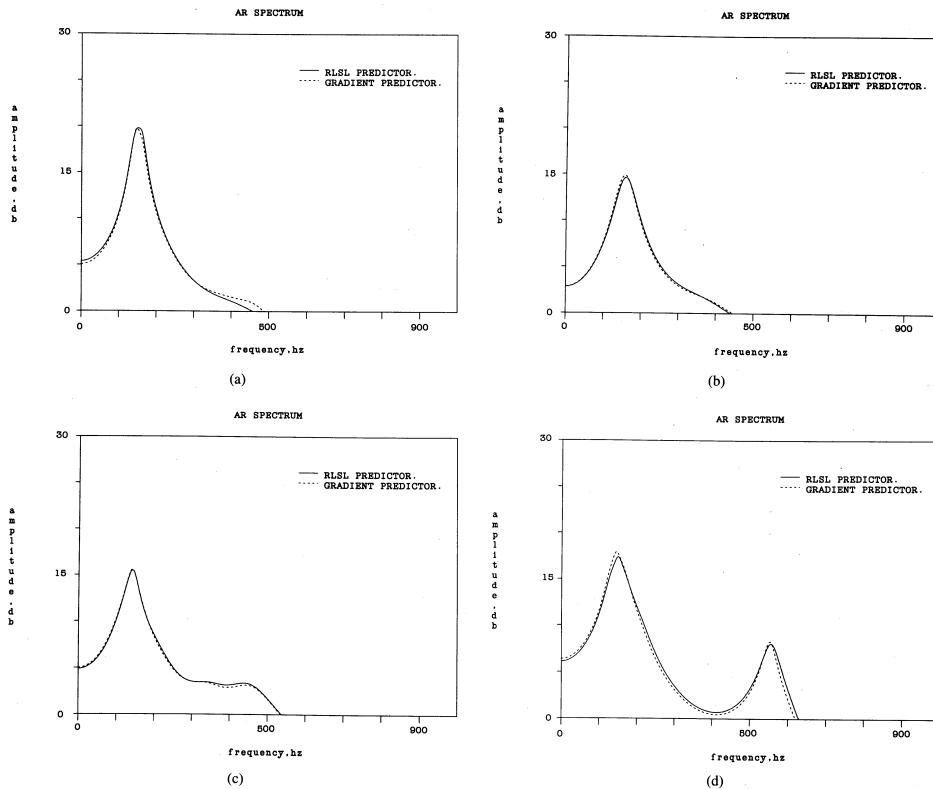


Figure 7.40 Diastolic heart sound spectra of (a, b) two normal subjects and (c, d) two patients with coronary artery disease. The method of estimating AR models identified in the figure as “RLSL” is described in Section 8.6.2; the gradient predictor method is not discussed in this book. Reproduced with permission from A.M. Akay, J.L. Semmlow, W. Welkowitz, M.D. Bauer, and J.B. Kostis, Detection of coronary occlusions using autoregressive modeling of diastolic heart sounds, *IEEE Transactions on Biomedical Engineering*, 37(4):366–373, 1990. ©IEEE.

7.12 Remarks

In this chapter, we have studied how mathematical models may be derived to represent physiological processes that generate biomedical signals, and furthermore, how the models may be related to changes in signal characteristics due to functional and pathological processes. The important point to note in the modeling approach is that the models provide a small number of *parameters* that characterize the signal and/or system of interest; the modeling approach is, therefore, useful in *parametric analysis* of signals and systems. As the number of parameters derived is usually much smaller than the number of signal samples, the modeling approach could also assist in data compression and compact representation of signals and related information.

Pole-zero models could be used to view physiological systems as control systems. Pathological states may be derived or simulated by modifying the parameters of the related models. Models of signals and systems are also useful in the design and control of prostheses.

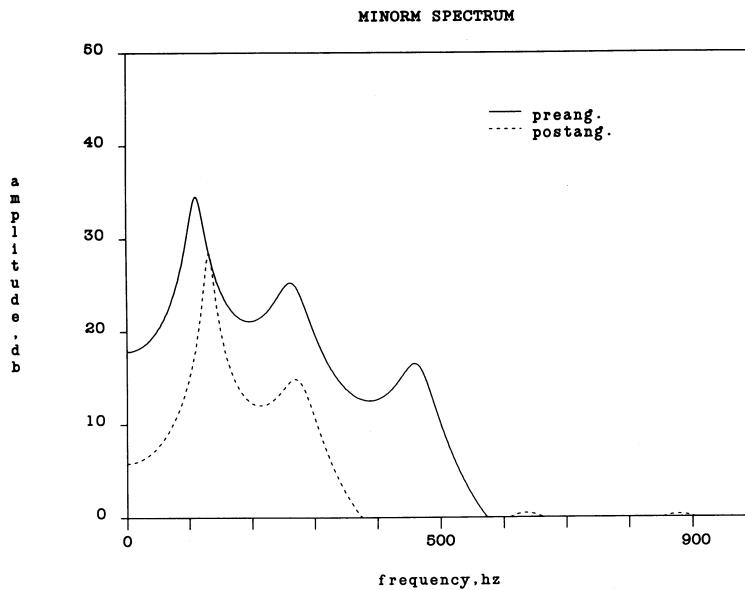


Figure 7.41 Diastolic heart sound spectra before (preang.) and after angioplasty (postang.) of a patient for whom coronary artery occlusion was corrected. (The minimum-norm method of PSD estimation used by Akay et al. [87] — labeled as “MINORM” in the figure — is not discussed in this book.) Reproduced with permission from A.M. Akay, J.L. Semmlow, W. Welkowitz, M.D. Bauer, and J.B. Kostis, Noninvasive detection of coronary stenoses before and after angioplasty using eigenvector methods, *IEEE Transactions on Biomedical Engineering*, 37(11):1095–1104, 1990. ©IEEE.

A combination of mathematical modeling with electromechanical modeling can allow the inclusion of physical parameters, such as the diameter of a blood vessel, constriction due to plaque, stiffness due to stenosis, and friction coefficient. Although accurate estimation of such parameters for human subjects may not always be possible, the models could lead to a better understanding of the related biomedical signals and systems.

Electrophysiological modeling at multiple scales related to cellular, tissue, and organ levels of detail could help understand the generation and propagation of action potentials. By combining electrophysiological, mathematical, and computational models, a detailed understanding of the characteristics of a signal related to its genesis and observed properties may be obtained for diagnostic and modeling applications.

7.13 Study Questions and Problems

1. Consider the LP model given by $\tilde{y}(n) = a y(n - 1)$. Define the prediction error, and derive the optimal value for a by minimizing the TSE.
2. The AR-model coefficients of a signal are $a_0 = 1$, $a_1 = 1$, and $a_2 = 0.5$. What is the transfer function of the model? Draw the pole-zero diagram of the model. What are the resonance frequencies of the system?
3. The AR-model coefficient vectors of a number of signals are made available to you. Propose two measures to compare the signals for (a) similarity, and (b) dissimilarity.
4. In AR modeling of signals, show why setting the derivative of the TSE with respect to any coefficient to zero will always lead to the minimum error (and not the maximum).
5. What type of a filter can convert the autocorrelation matrix of a signal to a diagonal matrix?

6. A biomedical signal is sampled at 500 Hz and subjected to AR modeling. The poles of the model are determined to be at $0.4 \pm j0.5$ and $-0.7 \pm j0.6$. (a) Derive the transfer function of the model. (b) Derive the difference equation in the time domain. (c) What are the resonance frequencies of the system that is producing the signal?
7. A model is described by the relationship $y(n) = x(n) + 0.5x(n-1) + 0.25x(n-2)$, where $x(n)$ is the input, and $y(n)$ is the output. What is the type of this system among AR, MA, and ARMA systems? What is the model order? What is its transfer function? Draw the pole-zero diagram of the system. Comment upon the stability of the system.
8. A model is described by the relationship $y(n) = -0.5y(n-1) - y(n-2) + x(n) + 0.5x(n-1) - x(n-2)$, where $x(n)$ is the input, and $y(n)$ is the output. What is the type of this system among AR, MA, and ARMA systems? What is the model order? What is its transfer function? Draw the pole-zero diagram of the system. Comment upon the stability of the system.

7.14 Laboratory Exercises and Projects

Note: Data files related to the exercises are available at the site

<https://github.com/srikrishnan1972/Biomedical-Signal-Analysis>

1. The file safety.wav contains the speech signal for the word “safety” uttered by a male speaker, sampled at 8 kHz (see the file safety.m). The signal has a significant amount of background noise (as it was recorded in a computer laboratory). Develop procedures to segment the signal into voiced, unvoiced, and silence (background noise) portions using short-time RMS , turns count, or ZCR measures.

Apply the AR modeling procedure to each segment; the command *lpc* in MATLAB® may be used for this purpose. Compute the model PSD for each segment. Compare the model PSD with the Fourier PSD for each segment. What are the advantages and disadvantages of the model-based PSD in the case of voiced and unvoiced sounds?

2. Derive the poles of the models you obtained in the preceding problem. Express each pole in terms of not only its z -plane coordinates but also its frequency and bandwidth. Study the variations in the pole positions as the type of the sound varies from one segment to the next over the duration of the signal.
3. The files pec1.dat, pec33.dat, and pec52.dat give three-channel recordings of the PCG, ECG, and carotid pulse signals (sampled at $1,000 \text{ Hz}$; you may read the signals using the program in the file plotpec.m). The signals in pec1.dat and pec52.dat are normal; the PCG signal in pec33.dat has systolic murmur, and is of a patient suspected to have pulmonary stenosis, ventricular septal defect, and pulmonary hypertension.

Segment each signal into its systolic and diastolic parts. Apply the AR modeling procedure to each segment and derive the model PSD. Compare the result with the corresponding PSDs obtained using Welch’s procedure.

4. Derive the poles of the models you obtained in the preceding problem. Express each pole in terms of not only its z -plane coordinates but also its frequency and bandwidth. Study the variations in the pole positions from the systolic part to the diastolic part of each signal. What are the major differences between the pole plots for the normal cases and the case with murmur?
5. The files ECG3, ECG4, ECG5, and ECG6 contain ECG signals sampled at the rate of 200 Hz (see the file ECGS.m). Apply the Pan-Tompkins method for QRS detection to each signal. Create impulse sequences including a delta function at every QRS location for the four signals. Create also the interval series for each signal as illustrated in Figure 7.1. Compute the spectra corresponding to the two representations of cardiac rhythm and study their relationship to the heart rate and its variability in each case.
6. Using one of the numerical methods (for example, Euler’s forward method) to solve ordinary and partial differential equations, implement a MATLAB® program to solve the Hodgkin–Huxley differential equation. Plot the voltage versus time function of the model output.

References

- [1] Agarwal GC and Gottlieb GL. An analysis of the electromyogram by Fourier, simulation and experimental techniques. *IEEE Transactions on Biomedical Engineering*, 22(3):225–229, 1975.
- [2] Abeles M and Goldstein Jr. MH. Multispike train analysis. *Proceedings of the IEEE*, 65(5):762–773, 1977.
- [3] Landolt JP and Correia MJ. Neuromathematical concepts of point process theory. *IEEE Transactions on Biomedical Engineering*, 25(1):1–12, 1978.
- [4] Anderson DJ and Correia MJ. The detection and analysis of point processes in biological signals. *Proceedings of the IEEE*, 65(5):773–780, 1977.
- [5] Cohen A. *Biomedical Signal Processing*. CRC Press, Boca Raton, FL, 1986.
- [6] deBoer RW, Karemaker JM, and Strackee J. Comparing spectra of a series of point events particularly for heart rate variability studies. *IEEE Transactions on Biomedical Engineering*, 31(4):384–387, 1984.
- [7] Rabiner LR and Schafer RW. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, 1978.
- [8] Zhang YT, Frank CB, Rangayyan RM, and Bell GD. Mathematical modeling and spectrum analysis of the physiological patello-femoral pulse train produced by slow knee movement. *IEEE Transactions on Biomedical Engineering*, 39(9):971–979, 1992.
- [9] Kernohan WG, Beverland DE, McCoy GF, Hamilton A, Watson P, and Mollan RAB. Vibration arthrometry. *Acta Orthopædica Scandinavica*, 61(1):70–79, 1990.
- [10] Beverland DE, Kernohan WG, and Mollan RAB. Analysis of physiological patello-femoral crepitus. In Byford GH, editor, *Technology in Health Care*, pages 137–138. Biological Engineering Society, London, UK, 1985.
- [11] Beverland DE, Kernohan WG, McCoy GF, and Mollan RAB. What is physiological patellofemoral crepitus? In *Proceedings of the XIV International Conference on Medical and Biological Engineering and VII International Conference on Medical Physics*, pages 1249–1250. IFMBE, Espoo, Finland, 1985.
- [12] Beverland DE, McCoy GF, Kernohan WG, and Mollan RAB. What is patellofemoral crepitus? *Journal of Bone and Joint Surgery*, 68-B:496, 1986.
- [13] Lathi BP. *Signal Processing and Linear Systems*. Berkeley-Cambridge, Carmichael, CA, 1998.
- [14] Parker PA, Stuller JA, and Scott RN. Signal processing for the multistate myoelectric channel. *Proceedings of the IEEE*, 65(5):662–674, 1977.
- [15] Lindström LH and Magnusson RI. Interpretation of myoelectric power spectra: A model and its applications. *Proceedings of the IEEE*, 65(5):653–662, 1977.
- [16] Zhang YT, Parker PA, and Scott RN. Study of the effects of motor unit recruitment and firing statistics on the signal-to-noise ratio of a myoelectric control channel. *Medical and Biological Engineering and Computing*, 28:225–231, 1990.
- [17] Parker PA and Scott RN. Statistics of the myoelectric signal from monopolar and bipolar electrodes. *Medical and Biological Engineering*, 11:591–596, 1973.
- [18] Shwedyk E, Balasubramanian R, and Scott RN. A nonstationary model for the electromyogram. *IEEE Transactions on Biomedical Engineering*, 24(5):417–424, 1977.
- [19] Person RS and Libkind MS. Simulation of electromyograms showing interference patterns. *Electroencephalography and Clinical Neurophysiology*, 28:625–632, 1970.
- [20] Person RS and Kudina LP. Cross-correlation of electromyograms showing interference patterns. *Electroencephalography and Clinical Neurophysiology*, 25:58–68, 1968.
- [21] de Luca CJ. A model for a motor unit train recorded during constant force isometric contractions. *Biological Cybernetics*, 19:159–167, 1975.
- [22] de Luca CJ. Physiology and mathematics of myoelectric signals. *IEEE Transactions on Biomedical Engineering*, 26:313–325, 1979.

- [23] Lawrence JH and de Luca CJ. Myoelectric signal versus force relationship in different human muscles. *Journal of Applied Physiology*, 54(6):1653–1659, 1983.
- [24] de Luca CJ and van Dyk EJ. Derivation of some parameters of myoelectric signals recorded during sustained constant force isometric contractions. *Biophysical Journal*, 15:1167–1180, 1975.
- [25] Makhoul J. Linear prediction: A tutorial. *Proceedings of the IEEE*, 63(4):561–580, 1975.
- [26] Haykin S. *Adaptive Filter Theory*. Prentice-Hall, Upper Saddle River, NJ, 3rd edition, 1996.
- [27] Vaidyanathan PP. *The Theory of Linear Prediction*. Morgan & Claypool and Springer, Switzerland AG, 2018.
- [28] Durbin J. *The Fitting of Time-series Models*. Mimeograph Series No. 244, Institute of Statistics, University of North Carolina, Chapel Hill, NC, 1959.
- [29] Durbin J. Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society, Series B (Methodological)*, 22(1):139–153, 1960.
- [30] Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.
- [31] Atal BS. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55(6):1304–1313, June 1974.
- [32] Childers DG, Skinner DP, and Kemerait RC. The cepstrum: A guide to processing. *Proceedings of the IEEE*, 65(10):1428–1443, 1977.
- [33] Kang WJ, Shiu JR, Cheng CK, Lai JS, Tsao HW, and Kuo TS. The application of cepstral coefficients and maximum likelihood method in EMG pattern recognition. *IEEE Transactions on Biomedical Engineering*, 42(8):777–785, 1995.
- [34] Rangayyan RM, Krishnan S, Bell GD, Frank CB, and Ladly KO. Parametric representation and screening of knee joint vibroarthrographic signals. *IEEE Transactions on Biomedical Engineering*, 44(11):1068–1074, 1997.
- [35] Chisci L, Mavino A, Perferi G, Sciandrone M, Anile C, Colicchio G, and Fuggetta F. Real-time epileptic seizure prediction using AR models and support vector machines. *IEEE Transactions on Biomedical Engineering*, 57(5):1124–1132, 2010.
- [36] Kopec GE, Oppenheim AV, and Trbolet JM. Speech analysis by homomorphic prediction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(1):40–49, 1977.
- [37] Oppenheim AV, Kopec GE, and Trbolet JM. Signal analysis by homomorphic prediction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):327–332, 1976.
- [38] Shanks JL. Recursion filters for digital processing. *Geophysics*, 32(1):33–51, 1967.
- [39] Steiglitz K and McBride LE. A technique for the identification of linear systems. *IEEE Transactions on Automatic Control*, 10:461–464, 1965.
- [40] Steiglitz K. On the simultaneous estimation of poles and zeros in speech analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3):229–234, 1977.
- [41] Kalman RE. Design of a self-optimizing control system. *Transactions of the ASME*, 80:468–478, 1958.
- [42] Joo TH, McClellan JH, Foale RA, Myers GS, and Lees RA. Pole-zero modeling and classification of phonocardiograms. *IEEE Transactions on Biomedical Engineering*, 30(2):110–118, 1983.
- [43] Murthy ISN and Prasad GSSD. Analysis of ECG from pole-zero models. *IEEE Transactions on Biomedical Engineering*, 39(7):741–751, 1992.
- [44] Murthy ISN, Rangaraj MR, Udupa KJ, and Goyal AK. Homomorphic analysis and modeling of ECG signals. *IEEE Transactions on Biomedical Engineering*, 26(6):330–344, 1979.
- [45] Akay AM, Welkowitz W, Semmlow JL, and Kostis JB. Application of the ARMA method to acoustic detection of coronary artery disease. *Medical and Biological Engineering and Computing*, 29:365–372, 1991.

- [46] Sikarskie DL, Stein PD, and Vable M. A mathematical model of aortic valve vibration. *Journal of Biomechanics*, 17(11):831–837, 1984.
- [47] Moussavi Z. *Fundamentals of Respiratory Sounds and Analysis*. Morgan & Claypool and Springer, San Rafael, CA, 2006.
- [48] Flanagan JL. *Speech Analysis Synthesis and Perception*. Springer, New York, NY, 2nd edition, 1972.
- [49] Wang JZ, Tie B, Welkowitz W, Semmlow JL, and Kostis JB. Modeling sound generation in stenosed coronary arteries. *IEEE Transactions on Biomedical Engineering*, 37(11):1087–1094, 1990.
- [50] Wang JZ, Tie B, Welkowitz W, Kostis J, and Semmlow J. Incremental network analogue model of the coronary artery. *Medical and Biological Engineering and Computing*, 27:416–422, 1989.
- [51] Fredberg JJ. Origin and character of vascular murmurs: Model studies. *Journal of the Acoustical Society of America*, 61(4):1077–1085, 1977.
- [52] Fink M, Niederer SA, Cherry EM, Fenton FH, Koivumäki JT, Seemann G, Thul R, Zhang H, Sachse FB, Beard D, Crampin EJ, and Smith NP. Cardiac cell modelling: Observations from the heart of the cardiac physiome project. *Progress in Biophysics and Molecular Biology*, 104(1–3):2–21, 2011.
- [53] Wang Y and Rudy Y. Action potential propagation in inhomogeneous cardiac tissue: Safety factor considerations and ionic mechanism. *American Journal of Physiology — Heart and Circulatory Physiology*, 278(4):H1019–H1029, 2000.
- [54] Hodgkin AL and Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544, 1952.
- [55] ten Tusscher KHWJ, Noble D, Noble PJ, and Panfilov AV. A model for human ventricular tissue. *American Journal of Physiology — Heart and Circulatory Physiology*, 286(4):H1573–H1589, 2004.
- [56] Luo C and Rudy Y. A model of the ventricular cardiac action potential. Depolarization, repolarization, and their interaction. *Circulation Research*, 68(6):1501–1526, 1991.
- [57] Luo C and Rudy Y. A dynamic model of the cardiac ventricular action potential. I. Simulations of ionic currents and concentration changes. *Circulation Research*, 74(6):1071–1096, 1994.
- [58] Clayton RH and Panfilov AV. A guide to modelling cardiac electrical activity in anatomically detailed ventricles. *Progress in Biophysics and Molecular Biology*, 96(1–3):19–43, 2008.
- [59] Beheshti MA, Umapathy K, and Krishnan S. Electrophysiological cardiac modeling: a review. *Critical Reviews™ in Biomedical Engineering*, 44(1–2):99–122.
- [60] Silva JR, Pan H, Wu D, Nekouzadeh A, Decker KF, Cui J, Baker NA, Sept D, and Rudy Y. A multi-scale model linking ion-channel molecular dynamics and electrostatics to the cardiac action potential. *Proceedings of the National Academy of Sciences*, 106(27):11102–11106, 2009.
- [61] Gima K and Rudy Y. Ionic current basis of electrocardiographic waveforms: A model study. *Circulation Research*, 90(8):889–896, 2002.
- [62] Hund TJ and Rudy Y. Rate dependence and regulation of action potential and calcium transient in a canine cardiac ventricular cell model. *Circulation*, 110(20):3168–3174, 2004.
- [63] Rogers JM and McCulloch AD. A collocation-Galerkin finite element model of cardiac action potential propagation. *IEEE Transactions on Biomedical Engineering*, 41(8):743–757, 1994.
- [64] Davies MR, Mistry HB, Hussein L, Pollard CE, Valentin J-P, Swinton J, and Abi-Gerges N. An in silico canine cardiac midmyocardial action potential duration model as a tool for early drug safety assessment. *American Journal of Physiology — Heart and Circulatory Physiology*, 302(7):H1466–H1480, 2012.
- [65] Grandits T, Effland A, Pock T, Krause R, Plank G, and Pezzuto S. GEASI: geodesic-based earliest activation sites identification in cardiac models. *International Journal for Numerical Methods in Biomedical Engineering*, 37(8):e3505, 2021.
- [66] Sovilj S, Magjarević R, Lovell NH, and Dokos S. A simplified 3D model of whole heart electrical activity and 12-lead ECG generation. *Computational and Mathematical Methods in Medicine*, 2013:134208, 2013.
- [67] Keener JP and Bogar K. A numerical method for the solution of the bidomain equations in cardiac tissue. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 8(1):234–241, 1998.

- [68] Strang G and Fix GJ. *An Analysis of the Finite Element Method*. Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [69] Austin TM, Hooks DA, Hunter PJ, Nickerson DP, Pullan AJ, Sands GB, Smaill BH, and Trew ML. Modeling cardiac electrical activity at the cell and tissue levels. *Annals of the New York Academy of Sciences*, 1080(1):334–347, 2006.
- [70] dos Santos RW, Plank G, Bauer S, and Vigmond EJ. Parallel multigrid preconditioner for the cardiac bidomain model. *IEEE Transactions on Biomedical Engineering*, 51(11):1960–1968, 2004.
- [71] Vigmond EJ, Aguel F, and Trayanova NA. Computational techniques for solving the bidomain equations in three dimensions. *IEEE Transactions on Biomedical Engineering*, 49(11):1260–1269, 2002.
- [72] Niederer SA, Lumens J, and Trayanova NA. Computational models in cardiology. *Nature Reviews — Cardiology*, 16(2):100–111, 2019.
- [73] Bruynseels K, Santoni de Sio F, and van den Hoven J. Digital twins in health care: Ethical implications of an emerging engineering paradigm. *Frontiers in Genetics*, 9:1–11, 2018.
- [74] Sharma P, Wang X, Ming CLC, Vettori L, Figtree G, Boyle A, and Gentile C. Advanced cardiac models: Considerations for the bioengineering of advanced cardiac in vitro models of myocardial infarction (small 15/2021). *Small*, 17(15):2170067, 2021.
- [75] Akselrod S, Gordon D, Ubel FA, Shannon DC, Barger AC, and Cohen RJ. Power spectrum analysis of heart rate fluctuation: A quantitative probe of beat-to-beat cardiovascular control. *Science*, 213:220–222, 10 July 1981.
- [76] Sayers B.McA. Analysis of heart rate variability. *Ergonomics*, 16(1):17–32, 1973.
- [77] Kobayashi M and Musha T. 1/f fluctuation of heartbeat period. *IEEE Transactions on Biomedical Engineering*, 29(6):456–457, 1982.
- [78] Rompelman O, Snijders JBIM, and van Spronken CJ. The measurement of heart rate variability spectra with the help of a personal computer. *IEEE Transactions on Biomedical Engineering*, 29(7):503–510, 1982.
- [79] Rosenblum MG, Kurths J, Pikovsky A, Schäfer C, Tass P, and Abel HH. Synchronization in noisy systems and cardiorespiratory interaction. *IEEE Engineering in Medicine and Biology Magazine*, 17(6):46–53, 1998.
- [80] Pompe B, Blidh P, Hoyer D, and Eiselt M. Using mutual information to measure coupling in the cardiorespiratory system. *IEEE Engineering in Medicine and Biology Magazine*, 17(6):32–39, 1998.
- [81] Kamath MV, Watanabe M, and Upton A, editors. *Heart Rate Variability (HRV) Signal Analysis: Clinical Applications*. CRC Press, Boca Raton, FL, 2012.
- [82] Kamath MV and Fallen EL. Power spectral analysis of heart rate variability: A noninvasive signature of cardiac autonomic function. *Critical Reviews in Biomedical Engineering*, 21(3):245–311, 1993.
- [83] Mendez MO, Bianchi AM, Matteucci M, Cerutti S, and Penzel T. Sleep apnea screening by autoregressive models from a single ECG lead. *IEEE Transactions on Biomedical Engineering*, 56(12):2838–2850, 2009.
- [84] Iwata A, Suzumara N, and Ikegaya K. Pattern classification of the phonocardiogram using linear prediction analysis. *Medical and Biological Engineering and Computing*, 15:407–412, 1977.
- [85] Iwata A, Ishii N, Suzumara N, and Ikegaya K. Algorithm for detecting the first and the second heart sounds by spectral tracking. *Medical and Biological Engineering and Computing*, 18:19–26, 1980.
- [86] Akay AM, Semmlow JL, Welkowitz W, Bauer MD, and Kostis JB. Detection of coronary occlusions using autoregressive modeling of diastolic heart sounds. *IEEE Transactions on Biomedical Engineering*, 37(4):366–373, 1990.
- [87] Akay AM, Semmlow JL, Welkowitz W, Bauer MD, and Kostis JB. Noninvasive detection of coronary stenoses before and after angioplasty using eigenvector methods. *IEEE Transactions on Biomedical Engineering*, 37(11):1095–1104, 1990.

CHAPTER 8

ADAPTIVE ANALYSIS OF NONSTATIONARY SIGNALS

A stationary signal is one that possesses the same statistical measures for all time, or at least over the duration of observation. We have seen in the preceding chapters that most biomedical signals, being manifestations of dynamic systems and pathophysiological processes, are *nonstationary*: Figure 3.3 shows that the variance of the speech signal used as an example varies with time; Figure 3.4 shows that the spectrum or frequency content of the speech signal also varies considerably over its duration. Figures 6.8 and 6.9 show that the spectrum of a heart sound signal or PCG varies from systole to diastole and could vary in between the two events as well.

When the characteristics of a signal being studied vary considerably over the duration of interest, measures and transforms computed over the entire duration do not carry useful information: they pay no attention to the dynamics of the signal. A single PSD computed from a long EMG, PCG, VAG, or speech record is of no practical value. The PSD does not provide information on time localization of the frequency components of the signal. We addressed this concern in PCG signal analysis in Section 6.3.6 by segmenting the PCG into its systolic and diastolic parts by using the ECG and carotid pulse signals as timing references. But how would we be able to handle the situation when murmurs are present in systole and diastole, and when we need to analyze the spectra of the murmurs without the contributions of S1 and S2? How could one perform segmentation of a PCG cycle into separate portions with S1, S2, and murmurs (if present)?

Furthermore, the EEG signal changes its nature in terms of rhythms, waves, transients, and spindles for which no independent references are available (see Section 1.2.6). In fact, the EEG represents a conglomeration of a number of mental and physiological processes going on in the brain at any given instant of time.

The VAG signal has nonstationary characteristics related to the cartilage surfaces that come into contact depending upon the activity performed, and no other source of information can assist in identifying time instants when the signal's properties change (see Section 1.2.14). Indeed, a VAG

signal contains no specific events that may be identified as such, but is a concatenation of nonspecific vibrations (with, perhaps, the exception of clicks). Would we be able to extend the application of the well-established signal analysis techniques that we have studied so far to such nonstationary signals?

In addition to being nonstationary, several biomedical signals possess multiple components from the same source as well as other sources that may be active at the same time. Cohen [1] defines a multicomponent signal as one that has delineated concentrations of power or energy in the time-frequency (TF) plane, with each such portion possibly related to a component of the signal; see Boashash [2] for related discussions. The speech signal is considered to be an example of a multicomponent signal: The formant structure of voiced speech indicates the presence of multiple resonances that relate to different components (see Section 1.2.13). Each phoneme in a speech signal is also a component that is different from the other phonemes present in the signal. Extending the same interpretation, we could consider EEG, PCG, and VAG signals to be multicomponent signals.

In addition to the presence of multiple delineated components in the TF plane, biomedical signals could also possess multiple components arising from different sources that are active at the same time. A VAG signal could possibly contain simultaneous contributions from multiple sources of sound or vibration in the same knee joint due to different types of cartilage pathology (see Section 8.2.3). The ECG of an expectant mother recorded from the surface of her abdomen contains a mixture of the maternal as well as fetal ECGs (see Section 3.3.5). Such mixtures are also multicomponent signals with multiple sources. Given such composite signals that are also nonstationary, how could we separate them into their components for further detailed analysis?

In this chapter, we shall focus on adaptive analysis of nonstationary signals, with the possible presence of multiple components. Specialized techniques to decompose and analyze multicomponent and multisource signals are presented in Chapter 9.

8.1 Problem Statement

Develop methods to study the dynamic characteristics of nonstationary biomedical signals. Propose schemes to apply the well-established Fourier transform and AR-modeling techniques to analyze and parameterize nonstationary signals.

The case studies presented in the following section provide the motivation for the present study from the perspective of a few representative biomedical signals. Approaches to solve the stated problem are presented in the sections to follow. This chapter presents several techniques for segmentation-based analysis of nonstationary signals. In addition, the chapter includes introductions to adaptive filters, Kalman filtering, wavelets, signal dictionary approaches, and joint TF analysis of nonstationary signals without segmentation.

8.2 Illustration of the Problem with Case Studies

8.2.1 Heart sounds and murmurs

We noted in Section 6.3.6 that the spectral contents of S1 and S2 are different due to the different states of contraction or relaxation of the ventricular muscles and the differences in their blood content during the corresponding cardiac phases. In the normal case, the QRS in the ECG signal and the dicrotic notch in the carotid pulse signal may be used to split a given PCG signal into S1 and S2, and separate PSDs may be obtained for the signal parts, as illustrated in Section 6.3.6. However, when a PCG signal contains murmurs in systole and/or diastole and possibly valve opening snaps (see Figure 6.9), it may be desirable to split the signal further.

Iwata et al. [3] applied AR modeling to PCG signals by breaking the signal into fixed segments of 25 ms duration (see Section 7.10). While this approach may be satisfactory, it raises questions

on optimality. What should be the window duration? Is it necessary to break the intervals between S1 and S2 into multiple segments? Would it not be more efficient to compute a single AR model for the entire durations of each of S1, S2, systolic murmur, and diastolic murmur — that is, a total of only four models over a cardiac cycle? It is conceivable that each model would be more accurate as all available signal samples would be used to estimate the required ACF if the signal were to be segmented adaptively as mentioned above instead of using fixed segments of shorter duration.

8.2.2 EEG rhythms and waves

The scalp EEG represents a combination of the multifarious activities of many small zones of the cortical surface beneath each electrode. The signal changes its characteristics in relation to mental tasks, external stimuli, and physiological processes. As we have noted in Section 1.2.6 and observed in Figure 1.40, a visual stimulus blocks the alpha rhythm; slower waves become prominent as the subject goes to deeper stages of sleep (see Section 6.7); and patients with epilepsy may exhibit sharp spikes and trains of spike-and-wave complexes (see Sections 1.2.6 and 4.4.3). The description of an EEG record, as outlined in Sections 1.2.6, 4.2.4, and 4.4.3, requires the identification of several types of waves and rhythms. This suggests that the signal may first have to be broken into segments, preferably with each segment possessing certain properties that remain the same for the duration of the segment. Each segment may then be described in terms of its characteristic features.

8.2.3 Articular cartilage damage and knee-joint vibration

Movement of the knee joint consists of coupled translation and rotation. The configuration of the patella is such that some portion of the articular surface is in contact with the femur throughout knee flexion and to almost full extension (see Section 1.2.14 and Figure 1.56). Goodfellow et al. [4] demonstrated that initial patellofemoral engagement occurs at approximately 20° of flexion involving both the medial and lateral facets. Figure 8.1 shows the patellar contact areas at different joint angles. As the knee is flexed, the patellofemoral contact area moves progressively upward, involving both the medial and lateral facets. At 90° of flexion, the band of contact engages the upper or superior pole of the patella. The odd facet does not articulate with the lateral margin of the medial femoral condyle until about $120^\circ - 135^\circ$ of knee flexion.

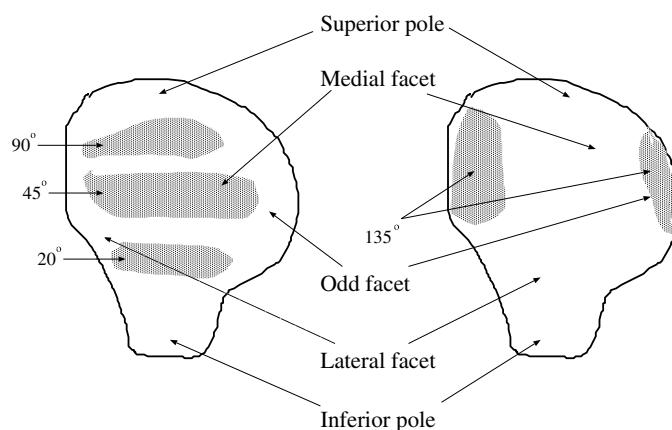


Figure 8.1 Contact areas of the patella with the femur during patellofemoral articulation [5].

Articular cartilage is composed of a solid matrix and synovial fluid [6]; it has no nerves, blood vessels, or lymphatics and is nourished by the synovial fluid covering its free surface. During articulation, friction between the bones is reduced as a result of the lubrication provided by the viscous synovial fluid [7,8]. The material properties of articular cartilage and cartilage thickness are variable not only from joint to joint but also within the same joint. In cases of abnormal cartilage, alterations in the matrix structure, such as increased hydration, disruption of the collagen fibrillar network, and disaggregation or loss of proteoglycans occur. As the compositional and biomechanical properties of abnormal articular cartilage continue to deteriorate, substance loss eventually occurs. This may be focal or diffuse, restricted to superficial fraying and fibrillation, or partial-thickness loss to full-thickness loss. In some cases, focal swelling or blistering of the cartilage may be seen before there is fraying of the articular surface [9].

Chondromalacia patella (soft cartilage of the patella) is a condition in which there is degeneration of patellar cartilage, often associated with anterior knee pain. Exposed subchondral bone and surface fibrillation of the articular cartilage are evident on the posterior patellar surface in chondromalacia patella [10]. Chondromalacia patella is usually graded in terms of the severity of the lesions [11, 12] as follows:

- *Grade I:* Softening, cracking, and blistering, but no loss of articular cartilage.
- *Grade II:* Damage is moderate and there is some loss of cartilage.
- *Grade III:* Severe damage of fibrocartilage has occurred but bone is not exposed.
- *Grade IV:* The cartilage is eroded and the subchondral bone is exposed.

Osteoarthritis is a degenerative joint disease that involves specific changes to bone in addition to cartilage. In the late stages of osteoarthritis, there is full-thickness articular cartilage degeneration and exposed bone. Other structural changes include fibrous changes to the synovium, joint capsule thickening, and further alterations to the bone, such as osteophyte formation [13]. Higher-grade chondromalacia may be categorized as osteoarthritis.

The menisci are subject to vertical compression, horizontal distraction, and rotary and shearing forces of varying degrees in the course of normal activities [14]. Advance of the aging process in both articular cartilage and fibrocartilage causes progressive liability to horizontal cleavage lesion [14].

The semiinvasive procedure of *arthroscopy* (fiber-optic inspection of joint surfaces, usually under general anesthesia) is often used for diagnosis of cartilage pathology. Through an arthroscope, the surgeon can usually see the patellofemoral joint, the femoral condyles, the tibial plateau (menisci), the anterior cruciate ligament, and the medial and lateral synovial spaces. Arthroscopy has emerged as the “gold standard” for relatively low-risk assessment of joint surfaces in order to determine the prognosis and treatment for a variety of conditions. Figure 8.2 shows the different stages of chondromalacia patella as viewed during arthroscopy.

Abnormal structures and surfaces in the knee joint are more likely to generate sound during extension and flexion movements than normal structures. Softened articular cartilage in chondromalacia patella — as well as cracks, fissures, or thickened areas in osteoarthritis — almost certainly increase the friction between the articular surfaces and are, therefore, likely to increase the sounds emitted during normal joint movement [15, 16]. Injury to the menisci in the form of tearing causes irregularity in shape and disruption to normal joint movement, and may produce sharp clicking sounds during normal knee movement [16–18].

It is obvious from this discussion (and the related introduction in Section 1.2.14) that the VAG signal is a nonstationary signal. Different aspects of the articulating surfaces come into contact at different joint angles; their quality in terms of lubrication and functional integrity could vary from one position to another. Inspection of the VAG signals and their spectrograms illustrated in Sections 3.10.3 and 3.15 reveals that the nature of a VAG signal changes significantly over the

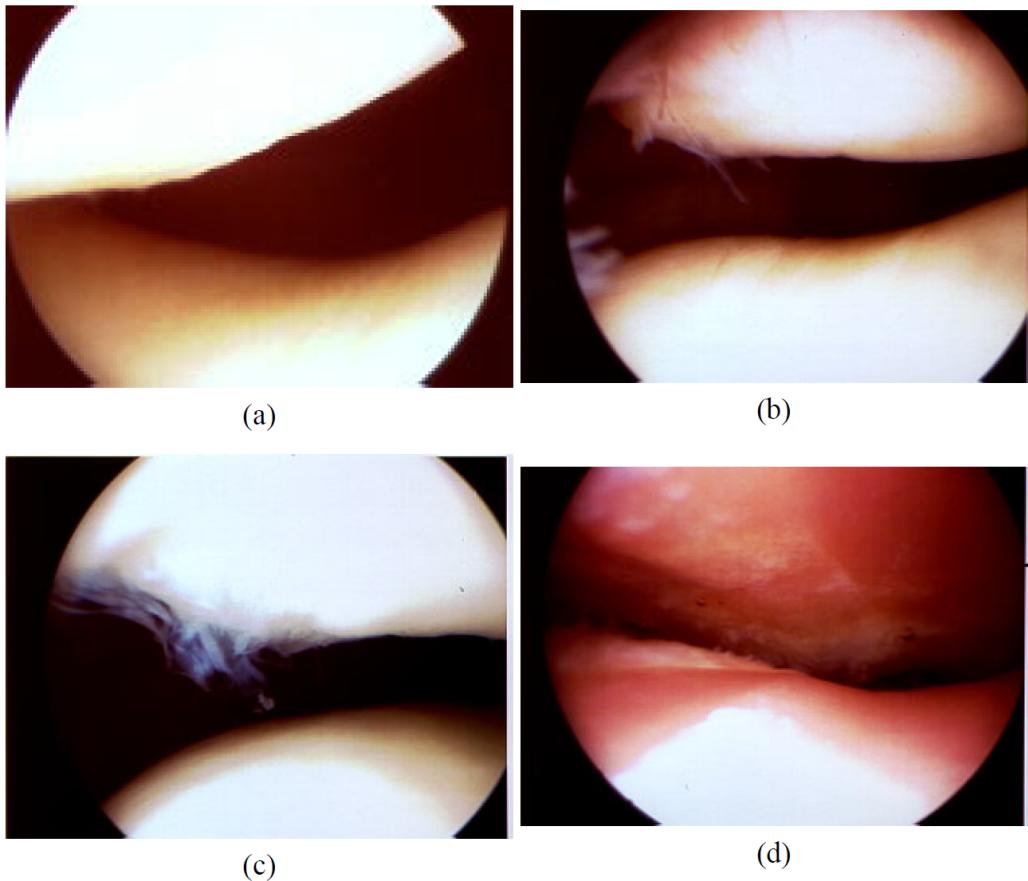


Figure 8.2 Arthroscopic views of the patellofemoral joint. (a) Normal cartilage surfaces. (b) Chondromalacia Grade II at the patella. (c) Chondromalacia Grade III at the patella. (d) Chondromalacia Grade IV at the patella and the femur; the bones are exposed. The under-surface of patella is at the top and the femoral condyle is at the bottom. Figure courtesy: G.D. Bell, Sport Medicine Centre, University of Calgary.

duration of the signal. As no prior or independent information is available about changes in the knee-joint structures that could lead to vibrations, adaptive segmentation of the VAG signal is desirable before it may be analyzed using the methods we have studied to this point in the book. Illustrations of adaptive segmentation of VAG signals are provided in Sections 8.6.1 and 8.6.2. In addition, methods for analysis of VAG signals without segmentation are described in Section 9.9.

8.3 Time-variant Systems

The linear system model represented by Equation 7.12 is a time-invariant system: The coefficients a_k and b_l of the system do not change with time, and consequently, the poles and zeros of the system stay fixed for all time. A nonstationary or dynamic system will possess coefficients that do vary with time: We saw in Sections 3.10.2 and 3.10.3 that the coefficient or tap-weight vectors of the adaptive LMS and RLS filters are expressed as functions of time. (Note: The Wiener filter described in

Section 3.9, once optimized for a given set of signal and noise statistics, is a time-invariant filter.) Since the coefficients of an LMS or RLS filter vary with time, so do the transfer function and the frequency response of the filter. It follows that the impulse response of such a system also varies with time.

Let us consider an all-pole filter for the sake of simplicity; the filter characteristics are determined by the positions of the poles (except for a gain factor). If the poles are expressed in terms of their polar coordinates, their angles correspond to (resonance) frequencies and their radii are related to the associated bandwidths. We may, therefore, characterize time-variant or nonstationary systems and signals by describing their pole positions in the complex z -plane — or, equivalently, the related frequencies and bandwidths — as functions of time. A description of the variation or the modulation of the pole parameters over time can thus capture the nonstationary or dynamic nature of a time-variant system or signal. Variations in the gain factor also lead to nonstationarities in the signal produced by the system. Appel and v. Brandt [19, 20] describe the simulation of different types of nonstationary behavior of signals and systems.

In the general case of a nonstationary system that is an AR process, we may modify Equation 7.17 to indicate that the model coefficients are functions of time:

$$\tilde{y}(n) = - \sum_{k=1}^P a_k(n) y(n-k). \quad (8.1)$$

Methods related to the Kalman filter or the least-squares approach may be used to analyze such a system [21–27]; see Section 8.7. Time-varying AR and ARMA modeling techniques have been applied to analyze the EEG [28], EGG [29], PCG [30], and HRV [31] signals; the application of time-varying analysis techniques to HRV signals is discussed in Section 8.12.

8.3.1 Characterization of nonstationary signals and dynamic systems

The output of a time-variant or dynamic system will be a nonstationary signal. The system may be characterized in terms of its time-variant model coefficients, transfer function, or related parameters derived thereof. Various short-time statistical measures computed over moving windows may be used to characterize a nonstationary signal; the measures may also be used to test for stationarity, or lack thereof, of a signal.

- **Mean:** The short-time mean represents the average or DC level of the signal in the analysis window. Variation of the mean from one window to another is usually an indication of the presence of a wandering baseline or low-frequency artifact, as in the case of the ECG signal in Figure 3.6. Clearly, the signal in Figure 3.6 is nonstationary in the mean. However, the mean is not an important measure in most signals, and it is typically blocked at the data-acquisition stage via capacitive coupling and/or a highpass filter. Furthermore, since a DC level carries no sound or vibration information, its presence or removal is of no consequence in the analysis of signals such as the heart sounds, speech, VAG, and VMG.
- **Variance:** Figure 3.3 illustrates the short-time variance of a speech signal. It is evident that the variance is high in regions of high signal variability (swings or excursions) about the mean, as in the case of the vowels or voiced-speech segments in the signal. The variance is low in the regions related to the fricatives or unvoiced-speech segments in the signal where the amplitude swing is small, in spite of their high-frequency nature. Since the mean of the signal is zero, the variance is equal to the MS value and represents the average power level in the corresponding signal windows. Although variations in the power level of speech signals may be useful in making voiced/ unvoiced/ silence decision, the parameter does not bear much information and provides a limited representation of the general statistical variability of signal characteristics. Regardless of the interpretation of the parameter, it is seen that the speech signal in Figure 3.3

is nonstationary in its variance (and the related measures of *SD*, *MS*, and *RMS*). From the discussion in Section 1.2.13, it is also clear that the vocal-tract system producing the speech signal is a dynamic system with time-varying configuration and filtering characteristics.

- **Measures of activity:** We have studied several measures of activity that indicate the “busyness” of the given signal, such as turning points, *ZCR*, and turns count, in Chapters 3 and 5. (The term “activity” has several connotations in the literature; see Chapter 5.) The short-time count of turning points is plotted in Figure 3.1 for a speech signal: It is evident that the signal is more active or busy in the periods related to the fricatives than those related to the vowels (a trend that is the opposite of that in the short-time variance of the same signal shown in Figure 3.3). The short-time turns count plot of the EMG signal in Figure 5.10 indicates the rising level of complexity of the signal with the level of breathing (inspiration). Although turning points, *ZCR*, and turns count are not among the traditional statistical measures derived from PDFs, they characterize signal variability and complexity in different ways. Both of the examples cited above illustrate variation of the parameters measured over the duration of the corresponding signals: The signals are, therefore, nonstationary in terms of the number of turning points or the turns count.
- **ACF:** The ACF was defined in Section 3.2.1 in general as $\phi_{xx}(t_1, t_1 + \tau) = E[x(t_1)x(t_1 + \tau)]$. In Section 3.2.4, one of the conditions for (wide-sense or second-order) stationarity was defined as the ACF being independent of a shift in the time origin or reference point, that is, $\phi_{xx}(t_1, t_1 + \tau) = \phi_{xx}(\tau)$. A nonstationary signal will not satisfy this condition and will have an ACF that varies with time. Since the ACF is based on the expectation of pairs of signal samples separated by a certain time difference or shift, it is a more general measure of signal variability than the variance and related measures. Note that the ACF for zero lag is the *MS* value of the signal.

One faces limitations in computing the ACF of short-time segments of a signal to investigate (non)stationarity: The shorter the analysis window, the shorter the maximum lag up to which the ACF may be reliably estimated. Regardless, the short-time ACF may be used to track nonstationarities in a signal. If the signal is the result of a dynamic AR system, the system parameters may be derived from the ACF (see Section 7.5).

- **PSD:** The PSD and ACF of a signal are interrelated by the Fourier transform. Therefore, a signal that is (non)stationary in its ACF is also (non)stationary in its PSD. However, the PSD is easier to interpret than the ACF, as we have seen in Chapter 6. The spectrogram of the speech signal in Figure 3.4 indicates significant variations in the short-time PSD of the signal: The speech signal is clearly nonstationary in its PSD (and ACF). The spectrograms of VAG signals in Sections 3.10.3 and 3.15 illustrate their nonstationary nature.
- **Higher-order statistics:** A major limitation of signal analysis using the ACF (or equivalently the PSD) is that the phase information is lost. The importance of phase in signals is discussed by Oppenheim and Lim [32]. Various conditions under which a signal may be reconstructed from its magnitude spectrum only or from its phase spectrum only are described by Hayes et al. [33] and Oppenheim and Lim [32]. Analysis based only upon the ACF cannot be applied to signals that are of mixed phase (that is, not minimum phase; see Section 5.4.2), that are the result of nonlinear systems, or that follow a PDF other than a Gaussian [34].

The general n^{th} -order moment of a random signal $x(t)$ at the instant t_1 is defined as [24,34,35]

$$\begin{aligned} m_x^n(t_1, t_1 + \tau_1, t_1 + \tau_2, \dots, t_1 + \tau_{n-1}) &= \\ E[x(t_1)x(t_1 + \tau_1)x(t_1 + \tau_2)\cdots x(t_1 + \tau_{n-1})], \end{aligned} \quad (8.2)$$

where $\tau_1, \tau_2, \dots, \tau_{n-1}$ are various shifts in time. It is evident that the ACF is a special case of the above with $n = 2$, that is, the ACF is the second-order moment.

A set of parameters known as cumulants may be related to the moments as follows: The second-order and third-order cumulants are equal to the corresponding moments. The fourth-order cumulant is related to the fourth-order moment as [24, 34, 35]

$$\begin{aligned} c_x^4(t_1, t_1 + \tau_1, t_1 + \tau_2, t_1 + \tau_3) &= m_x^4(t_1, t_1 + \tau_1, t_1 + \tau_2, t_1 + \tau_3) \\ &- m_x^2(t_1, t_1 + \tau_1) m_x^2(t_1 + \tau_2, t_1 + \tau_3) \\ &- m_x^2(t_1, t_1 + \tau_2) m_x^2(t_1 + \tau_3, t_1 + \tau_1) \\ &- m_x^2(t_1, t_1 + \tau_3) m_x^2(t_1 + \tau_1, t_1 + \tau_2). \end{aligned} \quad (8.3)$$

The Fourier transforms of the cumulants provide the corresponding higher-order spectra or polyspectra (with as many frequency variables as the order minus one). The Fourier transforms of the second-order, third-order, and fourth-order cumulants are known as the power spectrum (PSD), bispectrum, and trispectrum, respectively. A Gaussian process possesses only first-order and second-order statistics; moments and spectra of order higher than two are zero. Higher-order moments, cumulants, and spectra may be used to characterize nonlinear, mixed-phase, and non-Gaussian signals [24, 34, 35]. Variations over time of such measures may be used to detect the related types of nonstationarity.

- **System parameters:** When a time-varying model of the system producing the given signal is available in terms of its coefficients, such as $a_k(n)$ in Equation 8.1, we may follow or track changes in the coefficients over time. Significant changes in the model parameters indicate corresponding changes in the output signal.

The following sections present several techniques and examples of application related to the concepts and notions listed above.

8.4 Fixed Segmentation

Given a nonstationary signal, the simplest approach to break it into quasistationary segments would be to consider small windows of fixed duration. Given a signal $x(i)$ for $i = 0, 1, 2, \dots, N - 1$, we could consider a fixed segment duration of M samples, with $M \ll N$, and break the signal into $K = N/M$ parts as

$$x_k(n) = x[n + (k - 1)M], \quad 0 \leq n \leq M - 1, \quad 1 \leq k \leq K. \quad (8.4)$$

With the assumption that the signal does not change its characteristics to any significant extent within the duration corresponding to M samples (or $\frac{M}{f_s}$ s), each segment may be considered to be quasistationary.

Note that the segmentation procedure shown here is similar to that in the Bartlett and Welch procedures described in Sections 6.3.3 and 6.3.4. However, we will not be averaging the spectra over the segments now, but will be treating them as separate entities. The signal processing techniques we have studied so far may then be applied to analyze each segment separately.

8.4.1 The short-time Fourier transform

Once a given signal has been segmented into quasistationary parts $x_k(n)$ as in Equation 8.4, we may compute the Fourier spectrum for each segment as

$$X_k(\omega) = \sum_{n=0}^{M-1} x_k(n) \exp(-j\omega n). \quad (8.5)$$

The array of spectra $X_k(\omega)$ for $k = 1, 2, \dots, K$ will describe the time-varying spectral characteristics of the signal.

Segmentation of the given signal as above may be interpreted as the application of a moving window to the signal. The k^{th} segment $x_k(n)$ may be expressed as the multiplication of the signal $x(n)$ with a window function $w(n)$ positioned at the beginning of the segment as

$$x_k(n) = x(n) w[n - (k - 1)M], \quad 1 \leq k \leq K, \quad (8.6)$$

where

$$w(n) = \begin{cases} 1 & \text{for } 0 \leq n \leq M - 1, \\ 0 & \text{otherwise.} \end{cases} \quad (8.7)$$

Figure 8.3 (a) illustrates the PCG of a patient with systolic murmur and opening snap of the mitral valve, with a moving rectangular analysis window of duration 64 ms superimposed on the signal at three different instants of time. The duration of each window is 64 samples, equal to 64 ms with $f_s = 1 \text{ kHz}$. The three windows have been positioned approximately over the S1, systolic murmur, and S2 events in the signal. Figure 8.3 (b) shows the log PSDs of the signal segments extracted by the three analysis windows. It is seen that the PSDs differ significantly, with the second window displaying the largest amount and extent of high-frequency power due to the murmur. The third window displays more medium-frequency content than the first. It is clear that the PCG signal is nonstationary in the PSD.

In general, the window may be positioned at any time instant m , and the resulting segment may be expressed as $x(n) w(n - m)$. We need to state how the window is moved or advanced from one segment to another; in the extreme situation, we may advance the window one sample at a time, in which case adjacent windows would have an overlap of $(M - 1)$ samples. We may then compute the Fourier transform of every segment as

$$X(m, \omega) = \sum_{n=0}^{M-1} [x(n) w(n - m)] \exp(-j\omega n). \quad (8.8)$$

In the case when both the time and frequency variables are continuous, we may write the expression above in a more readily understandable form as

$$X(\tau, \omega) = \int_{-\infty}^{\infty} [x(t) w(t - \tau)] \exp(-j\omega t) dt. \quad (8.9)$$

The spectrum is now expressed not only as a function of frequency ω , but also as a function of time τ . Although the limits of the integral have been stated as $(-\infty, \infty)$, the finite duration of the window placed at time τ performs segmentation of the signal as desired.

The spectral representation of a signal as a function of time in Equations 8.8 and 8.9 is known as a *time-frequency distribution* or TFD [36–39]. In practice, only the squared magnitudes of the expressions in Equations 8.8 and 8.9 are used for the display of TFDs. Because the Fourier transform is applied, in the procedure above, to short windows of the signal in time, the result is known as the *short-time Fourier transform* or STFT of the signal. The method of analysis of a nonstationary signal in short windows is, in general, known as *short-time analysis*. The squared magnitude of the STFT is known as the *spectrogram* of the signal.

Figure 8.4 illustrates the spectrogram of the PCG signal of a patient with systolic murmur and opening snap of the mitral valve: The signal and the window parameters are the same as in Figure 8.3, but now the STFT magnitude spectrum is plotted for every window position with a displacement of 32 ms. The relatively high-frequency nature of the murmur as compared to S1 and S2 is clearly evident in the spectrogram.

We have previously encountered spectrograms of speech and VAG signals: Refer to Figure 3.4 and Sections 3.10.3 and 3.15. More examples of spectrograms are provided at the end of this section and later in this chapter.

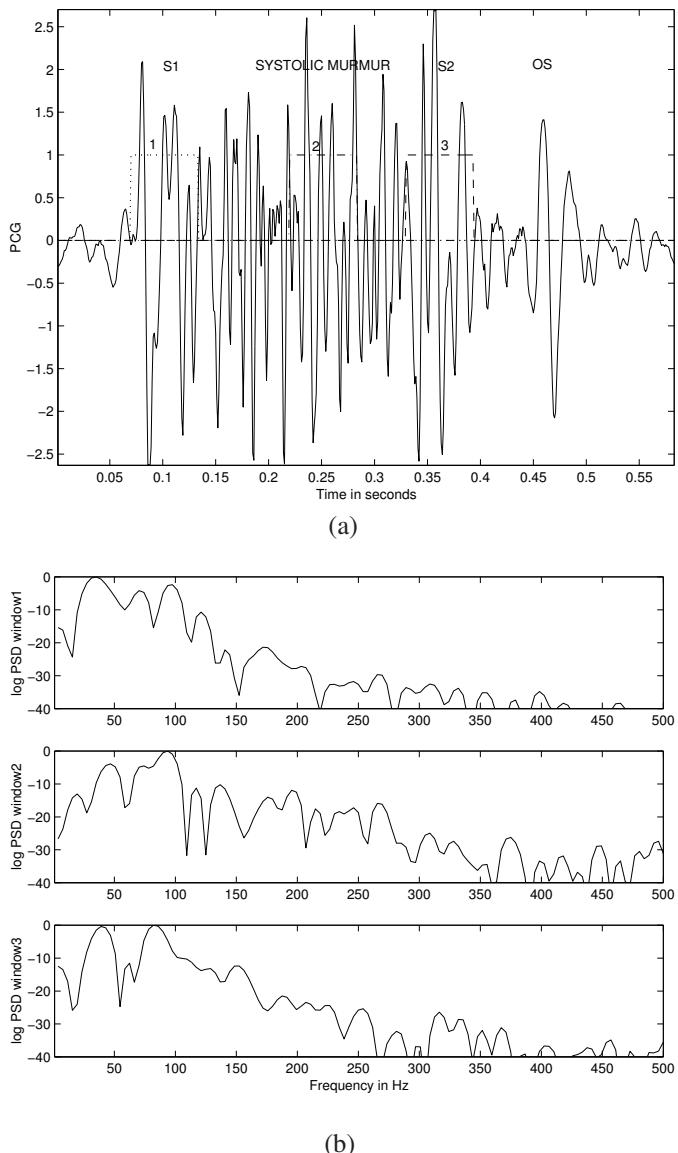


Figure 8.3 (a) PCG signal of a patient (female, 14 months) with systolic murmur and opening snap (OS) of the mitral valve. Three short-time analysis windows are superimposed, each one being a rectangular window of duration 64 ms. (b) Log PSDs of the three windowed signal segments. Each DFT was computed with zero padding to a total length 256 samples. See also Figures 6.9 and 8.4.

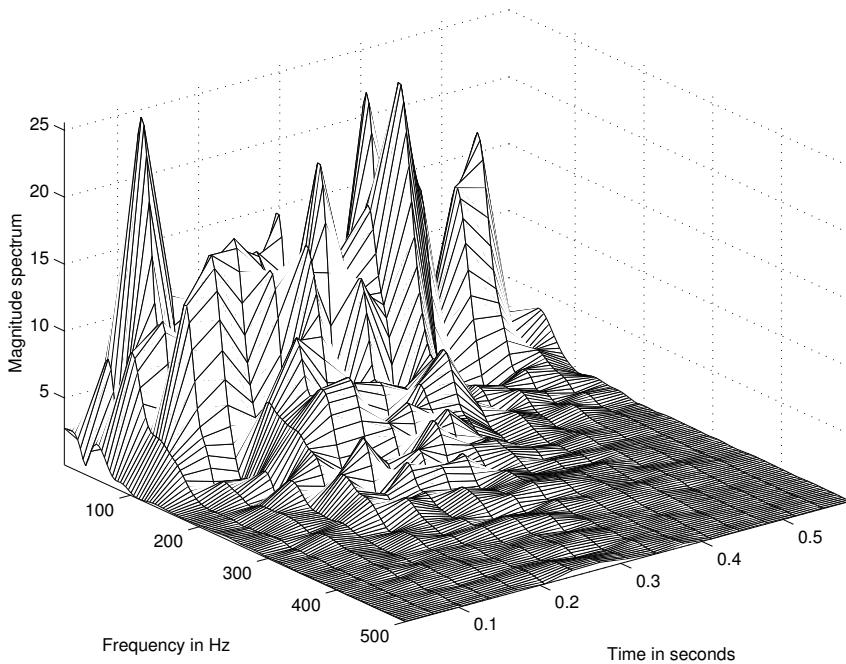


Figure 8.4 Spectrogram of the PCG signal of a patient (female, 14 months) with systolic murmur and opening snap of the mitral valve, computed with a moving short-time analysis window of duration 64 samples (64 ms with $f_s = 1 \text{ kHz}$), with the window advance interval being 32 samples. Each DFT was computed with zero padding to a total length 256 samples. $f_s = 1 \text{ kHz}$. See also Figures 6.9 and 8.3.

8.4.2 Considerations in short-time analysis

Short-time analysis of signals could be computationally expensive. In the case of the STFT, the Fourier transform has to be computed for each segment of the signal. In practice, there should be no need to compute the Fourier transform for every possible window position, that is, for every m in Equation 8.8. We could advance the analysis window by M samples, in which case adjacent windows will not overlap. It is common practice to advance the analysis window by $\frac{M}{2}$ samples, in which case adjacent windows will overlap for $\frac{M}{2}$ samples; some overlap is desirable in order to maintain continuity in the STFT or TFD computed.

An important question arises regarding the duration of the analysis window M to be used in a given application. The window should be short enough to ensure that each segment is stationary, but long enough to permit meaningful analysis and adequate representation of low-frequency components. We have seen in Section 6.3.4 that a short window possesses a wide main lobe in its frequency response. Since the given signal is multiplied in the time domain with the analysis window, the spectrum of the signal gets convolved with the spectral response of the window in the frequency domain. Convolution in the frequency domain with a function having a large main lobe leads to significant blurring and loss of spectral resolution.

The limitation imposed by the use of a window is related to the Heisenberg's uncertainty principle or the time–bandwidth product, expressed as [37]

$$\Delta t \Delta \omega \geq \frac{1}{2}, \quad (8.10)$$

where

$$(\Delta t)^2 = \int_{-\infty}^{\infty} (t - \bar{t})^2 |x(t)|^2 dt, \quad (8.11)$$

$$\bar{t} = \int_{-\infty}^{\infty} t |x(t)|^2 dt, \quad (8.12)$$

$$(\Delta\omega)^2 = \int_{-\infty}^{\infty} (\omega - \bar{\omega})^2 |X(\omega)|^2 d\omega, \quad (8.13)$$

$$\bar{\omega} = \int_{-\infty}^{\infty} \omega |X(\omega)|^2 d\omega, \quad (8.14)$$

and Δt and $\Delta\omega$ represent the time extent (duration) and frequency extent (bandwidth) of the signal $x(t)$ and its Fourier transform $X(\omega)$, respectively. The gist of the limitation stated above is that both a signal and its Fourier transform cannot simultaneously be made arbitrarily narrow. Considering the Fourier transform of a rectangular pulse, it is readily seen that the bandwidth of the signal is expanded when its time duration is shrunk; conversely, the bandwidth is reduced when the time duration is increased. In the extreme limits, it is well known that the Fourier transform of an impulse in time has infinite bandwidth, and that of a constant of infinite duration is an impulse at DC. The effect of this limitation on the STFT and TFD-based analysis is that we cannot simultaneously obtain arbitrarily high resolution along both the time and frequency axes.

At the extremes, a continuous-time signal $x(t)$ provides infinite time resolution but no frequency resolution: The value of the signal is known at every instant of time t , but nothing is known about the frequency content of the signal. Conversely, the PSD $S_{xx}(f)$ provides infinite frequency resolution but no time resolution.

In the case of sampled signals and spectra, the sampling intervals in the time domain and in the frequency domain will be finite and will be limited by Heisenberg's inequality as stated above. Furthermore, increasing the time resolution of the STFT by making the analysis window short in duration will compromise frequency resolution; on the other hand, increasing the window duration will lead to a loss in time resolution.

In general, the window function $w(n)$ included in Equation 8.8 need not be a rectangle: any of the window functions listed in Section 6.3.4 may be used. Once a window is chosen, the joint time and frequency resolution is the same over the entire TF plane.

The STFT expression in Equation 8.8 indicated the placement of a causal analysis window beginning at the time instant of reference m in the argument of the STFT. It is also common practice to use a symmetrical noncausal window defined for $-\frac{M}{2} \leq n \leq \frac{M}{2}$, in which case the reference point of the analysis window would be at the center of the window.

See Allen and Rabiner [40], Portnoff [41], and Rabiner and Schafer [42] for discussions on short-time Fourier analysis. See Rabiner and Schafer [42] for applications of short-time analysis to speech signals. Note that the temporal extent of the windowed segments, while fixed in a given application, has no bearing with the term "short"; that is, the duration of the analysis window could be as short or as long as desired or appropriate.

Illustration of application: Spectrograms of the speech signal in Figure 1.54 with different window parameters are provided in Figures 8.5 and 8.6. The spectrograms are shown here as gray-scale images, with the darkness at each point being proportional to the log PSD for the corresponding temporal analysis window position and frequency coordinate. It is evident that increasing the length of the analysis window provides better frequency resolution (the definition or clarity of the frequency components) while at the same time reducing the temporal resolution (that is, causing smearing in the temporal dimension). Decreasing the window length causes the reverse effects. The spectrogram in Figure 8.5 (b) with the analysis window duration being 16 ms clearly illustrates the high-frequency (broadband) nature of the fricatives; the transient and broadband nature of the

plosive /T/ is also clearly shown. The same features are not clearly depicted by the spectrogram in Figure 8.6 (b) where the analysis window is fairly long (128 ms); however, the formant structure of the voiced-speech components (the vowels) is clearly depicted. The formant structure of the voiced-speech components is not clearly visible in the spectrogram in Figure 8.5 (b).

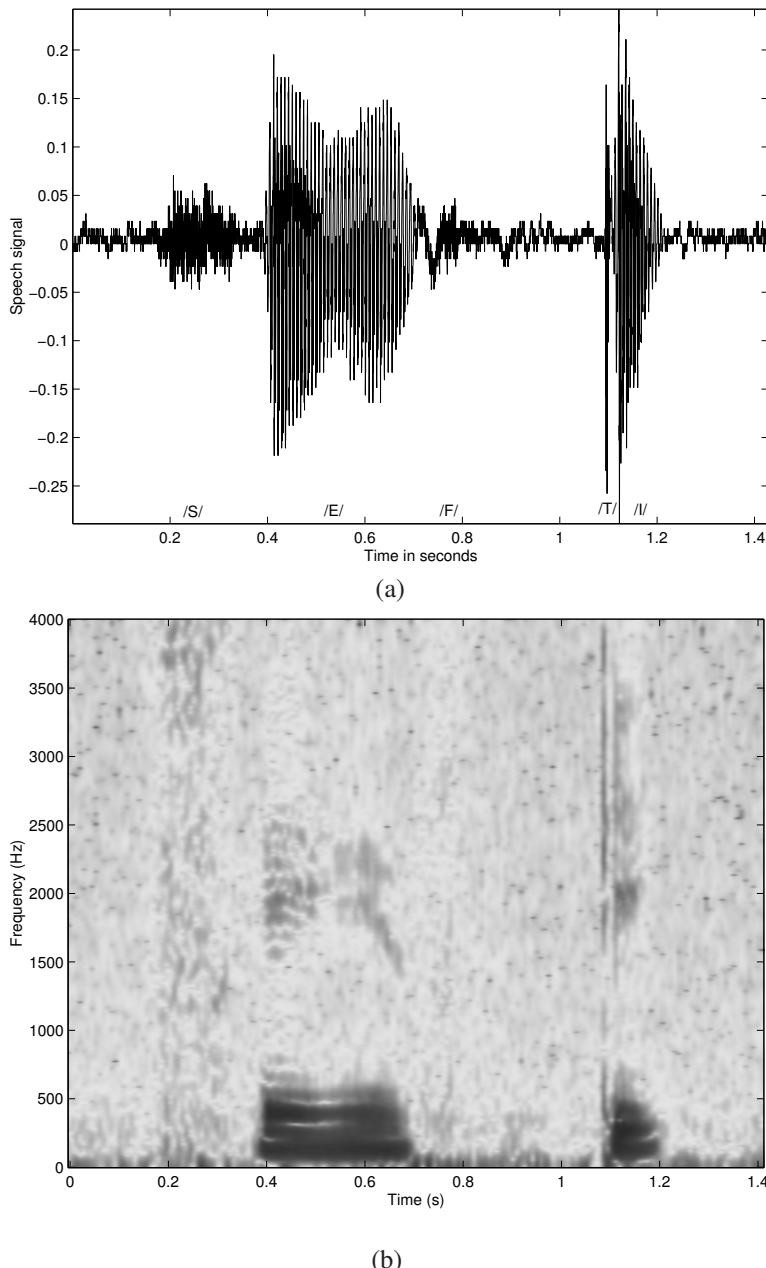
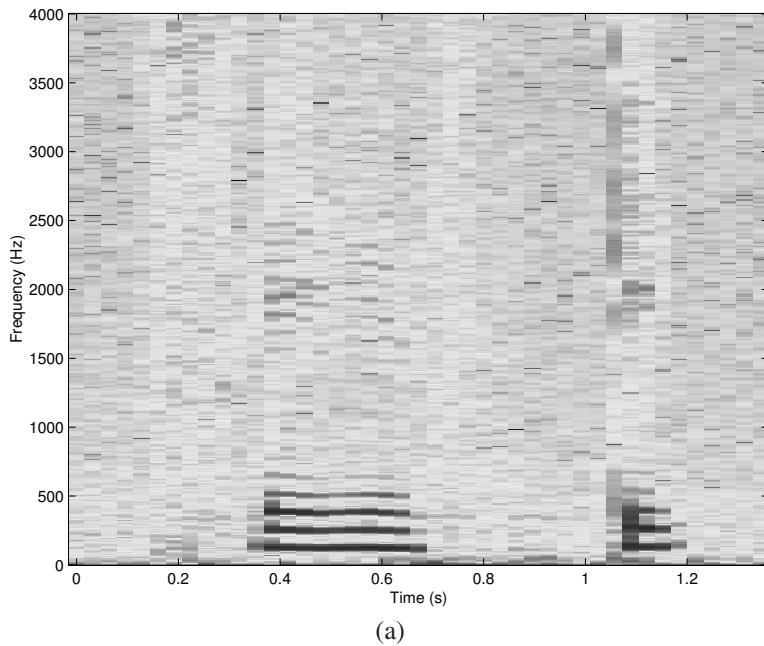
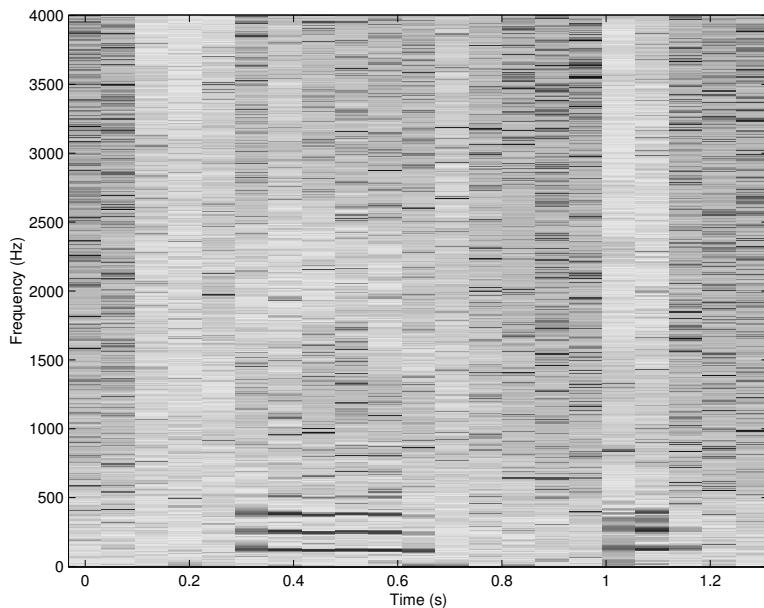


Figure 8.5 (a) Time-domain speech signal of the word “safety” uttered by a male speaker. (The signal is also illustrated in Figures 1.54, 3.1, and 3.3.) (b) Spectrogram (log PSD) of the signal computed with a moving short-time analysis window of duration 16 ms (128 samples with $f_s = 8 \text{ kHz}$), with the window advance interval being 8 ms. See also Figure 8.6.



(a)



(b)

Figure 8.6 Spectrograms (log PSD) of the speech signal in Figure 8.5: (a) with a moving window of duration 64 ms (512 samples with $f_s = 8 \text{ kHz}$) and window advance interval of 32 ms, and (b) with a moving window of duration 128 ms (1,024 samples) and window advance interval of 64 ms.

8.5 Adaptive Segmentation

One of the limitations of short-time analysis lies with the use of a fixed window duration. A signal may remain stationary for a certain duration of time much longer than the window duration chosen, and yet the signal would be broken into many segments over such a duration. Conversely, a signal may change its characteristics within the duration of the fixed window: Short-time analysis cannot guarantee stationarity of the signal over even the relatively short duration of the analysis window used. It would be desirable to adapt the analysis window to changes in the given signal, allowing the window to be as long as possible while the signal remains stationary, and to start a new window at the exact instant when the signal or the related system changes its characteristics.

Problem: *Propose adaptive methods to break a nonstationary signal into quasistationary segments of variable duration.*

Solution: We saw in Section 7.5 that a signal may be represented or modeled as a linear combination of a small number of past values of the signal, subject to a small error of prediction. It then follows that, if a signal were to change its behavior, it would no longer be predictable from its preceding samples as they would correspond to the previous state of the time-variant system generating the nonstationary signal. Therefore, we could expect a large increase or jump in the prediction error at instants of time when the signal changes in its characteristics. Furthermore, the AR-model parameters represent the system generating the signal and provide the poles of the system. If the system were to change in terms of the locations of its poles (related to its resonance frequencies), the same model would no longer hold: A new model would have to be initiated at such instants of change. This suggests that we could estimate AR models on a short-time basis and monitor the model parameters from segment to segment: A significant change in the model parameters would indicate a point of change in the signal. (We have seen in Section 7.10 how a similar approach was used by Iwata et al. [43] to detect S1 and S2 in PCGs.) Adjacent segments that have the same or similar model parameters could be concatenated to form longer segments in a subsequent stage of processing. As the AR model provides several parameters and may be interpreted in several ways (see Section 7.5.2), tracking the behavior of the model over a moving analysis window may be accomplished in many ways. The following sections provide the details of a few approaches for adaptive segmentation based on the notions stated above.

8.5.1 Spectral error measure

Bodenstein and Praetorius [44] and Praetorius et al. [45] used the all-pole LP or AR model (see Section 7.5) for adaptive segmentation of EEG signals into quasistationary segments and also for further feature extraction. They made the following observations about the application of AR modeling to EEG signals:

- *Time domain:* The present value of the prediction error indicates the instantaneous degree of “unexpectedness” in the signal.
- *Autocorrelation domain:* The prediction error is decorrelated.
- *Spectral domain:* The prediction error being white noise, the AR model yields an all-pole representation of the signal’s spectrum, which is particularly suitable for the modeling of resonance.

These properties are useful for

- detection and elimination of transients;
- segmentation of the EEG into quasistationary segments; and
- feature extraction and pattern recognition (diagnosis).

Ferber [46] provided a description of nonstationarities in the EEG and suggested a few approaches to treat the same.

Analysis of spectral change: Let the PSD of the given nonstationary signal be $S(0, \omega)$ at zero time and $S(t, \omega)$ at time t . The *spectral error* of $S(t, \omega)$ with respect to $S(0, \omega)$ may be taken to be dependent upon the difference between the corresponding log PSDs — that is, to be proportional to $\{\log[S(t, \omega)] - \log[S(0, \omega)]\}$ or $\log\left[\frac{S(t, \omega)}{S(0, \omega)}\right]$. Consider the state when an AR model has been adapted to the signal's spectrum $S(0, \omega)$ at zero time. If we pass the signal at time t through the AR model, the prediction error will have an instantaneous spectrum given by

$$S_e(\omega) = \frac{S(t, \omega)}{S(0, \omega)}, \quad (8.15)$$

which is similar to the spectral ratio in Equation 7.50. Thus, the problem of comparing two arbitrary PSDs of a nonstationary signal at two different instants of time may be expressed as testing $S_e(\omega)$ for deviation from a uniform PSD.

Let $a_R(k)$, $k = 1, 2, \dots, P$, represent the reference AR model. When the current signal $y(n)$ is passed through the filter represented by the AR model, we obtain the prediction error

$$e(n) = \sum_{k=0}^P a_R(k) y(n-k), \quad (8.16)$$

with the prediction model defined as in Equation 7.17, the prediction error as given in Equation 7.18, and $a_R(0) = 1$. The error indicates the deviation of the current signal from the previously computed model. Consider the integral

$$\varepsilon = \int_{-\infty}^{\infty} [1 - S_e(\omega)]^2 d\omega, \quad (8.17)$$

where $S_e(\omega)$ is the PSD of the prediction error. Ideally, when the AR model has been optimized for the signal on hand, the prediction error is expected to have a uniform PSD. However, if the signal is nonstationary, some changes would have occurred in the spectral characteristics of the signal, which would be reflected in the PSD of the error. If $\phi_e(k)$ is the ACF corresponding to $S_e(\omega)$, the latter is given by the Fourier transform of the former. However, since both functions are real and even, we have

$$S_e(\omega) = \phi_e(0) + 2 \sum_{k=1}^{\infty} \phi_e(k) \cos(2\pi\omega k). \quad (8.18)$$

Then,

$$\varepsilon = \int_{-\infty}^{\infty} \left[1 - \phi_e(0) - 2 \sum_{k=1}^{\infty} \phi_e(k) \cos(2\pi\omega k) \right]^2 d\omega. \quad (8.19)$$

Due to the orthonormality of the trigonometric functions, we get

$$\varepsilon = [1 - \phi_e(0)]^2 + 2 \sum_{k=1}^{\infty} \phi_e^2(k). \quad (8.20)$$

In practice, the summation may be performed up to some lag, M . Bodenstein and Praetorius [44] recommended normalization of the error measure by division by $\phi_e^2(0)$, leading to the *spectral error measure (SEM)*

$$SEM = \left[\frac{1}{\phi_e(0)} - 1 \right]^2 + 2 \sum_{k=1}^M \left[\frac{\phi_e(k)}{\phi_e(0)} \right]^2. \quad (8.21)$$

Here, the first term represents the change in the total power of the prediction error; the second term depends upon the change in spectral shape only. Note that the prediction error is expected to have a

uniform (flat) PSD as long as the signal remains stationary with respect to the AR model designed. The *SEM* was shown to vary significantly in response to changes in the spectral characteristics of EEG signals and to be useful in breaking the signals into quasistationary parts. Figure 8.7 shows the general scheme of EEG segmentation by using the *SEM*.

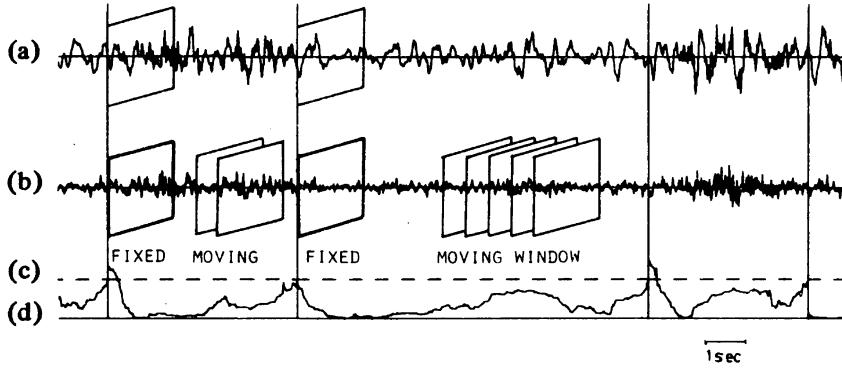


Figure 8.7 Adaptive segmentation of EEG signals via the use of *SEM*. (a) Original EEG signal. The rectangular window at the beginning of each adaptive segment indicates the signal window to which the AR model has been optimized. (b) Prediction error. The initial ACF of the error is computed over the fixed window; the running ACF of the error is computed over the moving window. (c) Segmentation threshold. (d) *SEM*. The vertical lines represent the segmentation boundaries. Reproduced with permission from G. Bodensteiner and H.M. Praetorius, Feature extraction from the electroencephalogram by adaptive segmentation, *Proceedings of the IEEE*, 65(5):642–652, 1977. ©IEEE.

Algorithm for adaptive segmentation [44]: Let $n = 0$ represent the starting point of analysis where the first reference or fixed analysis window is placed for each adaptive segment, as in Figure 8.7 (a). $(N+P)$ samples of the signal $y(n)$ should be available prior to the arbitrarily designated origin at $n = 0$, where $(2N + 1)$ is the size of the analysis window and P is the order of the AR model to be used.

1. Using the signal samples $y(-N)$ to $y(N)$, compute the signal ACF up to lag P .
2. Derive the corresponding AR model of order P .
3. Using the signal values $y(-N - P)$ to $y(n + N)$, compute the prediction error $e(-N)$ to $e(n + N)$ and compute the running short-time ACF $\phi_e(n, m)$ of the prediction error as

$$\phi_e(n, m) = \frac{1}{2N + 1} \sum_{k=-N}^{N-m} e(n+k) e(n+k+m). \quad (8.22)$$

Note that the ACF now has two indices: the first index n indicates the position of the short-time analysis window, and the second index m indicates the lag for which the ACF is computed.

4. Calculate $\phi_e(0, m)$ for $m = 0, 1, \dots, M$. This represents the fixed window at the beginning of each adaptive segment in Figure 8.7 (a). Perform the following three steps for each data point.
5. Compute $\phi_e(n, m)$ for the moving window [see Figure 8.7 (b)] by the recursive relationship

$$\begin{aligned} \phi_e(n, m) &= \phi_e(n-1, m) + e(n+N) e(n+N-m) \\ &\quad - e(n-N-1) e(n-N-1-m). \end{aligned} \quad (8.23)$$

This represents the moving window in Figure 8.7 (b).

6. Compute the *SEM* at time n as

$$SEM(n) = \left[\frac{\phi_e(0,0)}{\phi_e(n,0)} - 1 \right]^2 + 2 \sum_{k=1}^M \left[\frac{\phi_e(n,k)}{\phi_e(n,0)} \right]^2, \quad (8.24)$$

where $\phi_e(0,0)$ accounts for the fact that the signal may have an arbitrary power level.

7. Test if $SEM(n) > Th_1$, where Th_1 is a threshold.

If the condition is not satisfied, increase n by 1 and return to Step 5.

If the condition is satisfied, a segment boundary has been detected at time n , as indicated by the vertical lines in Figure 8.7. Reset the procedure by continuing to the next step.

8. Shift the time axis by substituting $(n+k)$ with $(k-N)$ and start the procedure again with Step 1.

In the investigations of Bodenstein and Praetorius [44], *SEM* demonstrated sharp jumps as transients of duration less than 100 ms entered and left the moving analysis window of duration 2 s ($2N+1 = 101$ samples with $f_s = 50$ Hz). Such jumps could lead to inappropriate segmentation, especially with burst-suppression type EEG episodes as illustrated in Figure 8.8. To overcome this problem, it was suggested that the prediction error $e(n)$ be limited (clipped) by a threshold Th_2 as

$$e(n) = \begin{cases} e(n), & \text{if } |e(n)| < Th_2, \\ \text{sgn}[e(n)] Th_2, & \text{if } |e(n)| \geq Th_2. \end{cases} \quad (8.25)$$

The threshold Th_2 is shown by the dashed lines in Figure 8.8 (c). The *SEM* computed from the clipped $e(n)$ is shown in Figure 8.8 (d), which, when checked against the original threshold Th_1 , yields the correct segmentation boundary. The signal reconstructed from the clipped prediction error is shown in Figure 8.8 (e), which shows that the clipping procedure has suppressed the effect of the transient without affecting the rest of the signal.

In spite of the clipping procedure as in Equation 8.25, it was indicated by Bodenstein and Praetorius [44] that the procedure was more sensitive than desired and caused false alarms. To further limit the effects of random fluctuations in the prediction error, a smoothed version $e_s(n)$ of the squared prediction error was computed as

$$e_s(n) = e^2(n-1) + 2e^2(n) + e^2(n+1) \quad (8.26)$$

for those samples of $e(n)$ that satisfied the condition $|e(n)| > Th_2$. Another threshold Th_3 was applied to $e_s(n)$, and the triplet $\{y(n-1), y(n), y(n+1)\}$ was considered to be a part of a transient only if $e_s(n) > Th_3$. The procedure of Bodenstein and Praetorius combines adaptive segmentation of EEG signals with transient detection as the two tasks are interrelated.

Illustration of application: Figure 8.9 shows the EEG signal of a child in sleep stage I, superimposed with 14 Hz spindles. The *SEM* and its components are also shown in the figure. The vertical lines indicate the segment boundaries detected. Bodenstein et al. [47] and Creutzfeldt et al. [48] describe further extension of the approach to computerized pattern classification of EEG signals including clustering of similar segments and labeling of the types of activity found in an EEG record.

The *SEM* method was applied by Tavathia et al. [49] for adaptive segmentation of VAG signals. It was indicated that each segment could be characterized by the frequency of the most-dominant pole obtained via AR modeling and the spectral power ratio $E_{40:120}$ as per Equation 6.44; however, no classification experiments were performed. More examples of application of the *SEM* technique are presented in Sections 8.5.4 and 8.10.

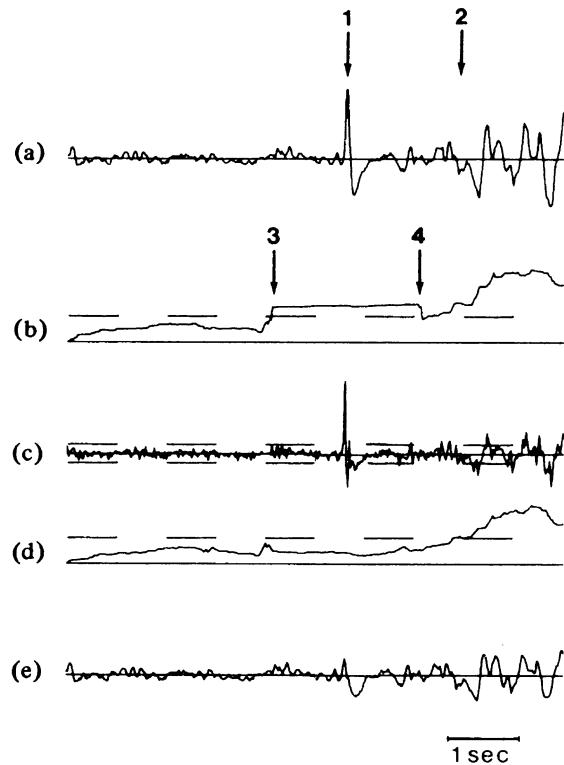


Figure 8.8 Elimination of transients by clipping the prediction error. (a) Original EEG signal of the burst-suppression type. The sharp wave marked by the arrow 1 is followed by the onset of a burst marked by the arrow 2. (b) SEM showing sudden jumps at points indicated by the arrows 3 and 4 as the sharp wave enters and leaves the analysis window. (c) Clipping of the prediction error with threshold Th_2 . (d) SEM after clipping the prediction error. The dashed line represents the threshold Th_1 . (e) Signal reconstructed from the clipped prediction error. Reproduced with permission from G. Bodensteiner and H.M. Praetorius, Feature extraction from the electroencephalogram by adaptive segmentation, *Proceedings of the IEEE*, 65(5):642–652, 1977. ©IEEE.

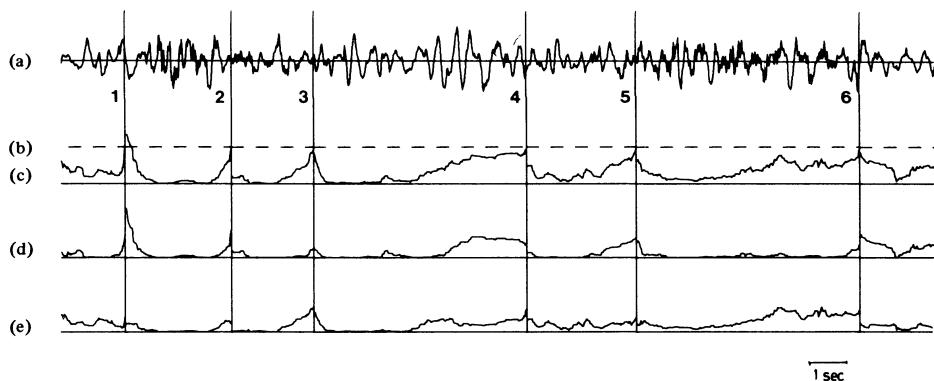


Figure 8.9 Use of SEM to segment an EEG signal. (a) Original EEG signal of a child in sleep stage I with superimposed 14 Hz spindles. (b) Segmentation threshold. (c) SEM. (d) Deviation in prediction error power. (e) Deviation in prediction error spectral shape. The vertical lines represent the segmentation boundaries. Reproduced with permission from G. Bodensteiner and H.M. Praetorius, Feature extraction from the electroencephalogram by adaptive segmentation, *Proceedings of the IEEE*, 65(5):642–652, 1977. ©IEEE.

8.5.2 ACF distance

Michael and Houchin [50] proposed a method comparable to that of Bodenstein and Praetorius [44], but based on a simpler scheme using the ACF. It should be noted that the AR-model coefficients are derived from the ACF and that the spectra used to compute *SEM* are related to the corresponding ACFs by the Fourier transform. However, direct use of the ACF removes the assumption made in AR modeling that the signal is the result of an AR process.

In the method of Michael and Houchin, the ACF is treated as a statistical measure of the given signal, and significant variations in the ACF are used to detect nonstationarity. A reference window is extracted at the beginning of each scan, and the given signal (EEG) is observed through a moving window. The duration of the window has to be chosen such that it is shorter than the shortest expected quasistationary segment of the given signal, but long enough to characterize the lowest frequency present. If the difference between the signal's statistics (ACF) in the moving window and the reference window is significant, a segment boundary is drawn, and the procedure is restarted.

Let $\phi_R(k)$ be the ACF of the reference window at the beginning of a new segmentation step, where k is the lag or delay. Let $\phi_T(n, k)$ be the ACF of the test window positioned at time instant n . Given that the ACF for zero lag is the power of the signal, Michael and Houchin computed a normalized power distance $d_P(n)$ between the ACFs as (see also Appel and v. Brandt [20])

$$d_P(n) = \frac{|\sqrt{\phi_T(n, 0)} - \sqrt{\phi_R(0)}|}{\min\{\sqrt{\phi_T(n, 0)}, \sqrt{\phi_R(0)}\}}. \quad (8.27)$$

A spectral distance $d_F(n)$ was computed using the ACF coefficients only up to lag q as

$$d_F(n) = \frac{\sum_{k=1}^q |\phi_T(n, k) - \phi_R(k)|}{0.5 + \sum_{k=1}^q \min\{\sqrt{\phi_T(n, k)}, \sqrt{\phi_R(k)}\}}. \quad (8.28)$$

The lag limit q was set as the lower value of the lags at which the ACFs changed from positive to negative values for the first time. The net ACF distance $d(n)$ was then computed as

$$d(n) = \frac{d_P(n)}{Th_P} + \frac{d_F(n)}{Th_F}, \quad (8.29)$$

where Th_P and Th_F are thresholds. The condition $d(n) > 1$ was considered to represent a significant change in the ACF, and used to mark a segment boundary.

Due to the use of a moving window of finite size, the true boundary or point of change in the signal characteristics will lie within the last test window before a segment boundary is triggered. Michael and Houchin used a linear interpolation procedure based on the steepness of the ACF distance measure to correct for such a displacement. Barlow et al. [51] provide illustrations of application of the method to clinical EEGs. Their work includes clustering of similar segments based on mean amplitude and mean frequency measures, “dendograms” to illustrate the clustering of segments [52], and labeling of the various states found in an EEG record. Illustrations of application of the ACF method are provided in Section 8.5.4.

8.5.3 The generalized likelihood ratio

Appel and v. Brandt [19] proposed the generalized likelihood ratio (GLR) method, which uses a reference window that is continuously grown as long as no new boundary is marked. The test window is a sliding window of constant duration as in the case of the *SEM* and ACF methods. Figure 8.10 illustrates the windows used. The advantage of the growing reference window is that it contains the maximum amount of information available from the beginning of the new segment to the current instant. Three different datasets are defined: the growing reference window, the sliding

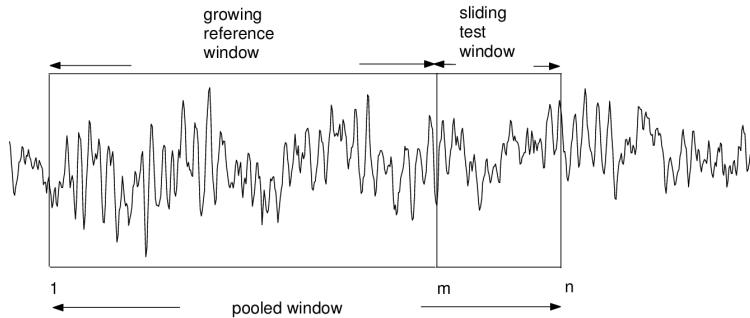


Figure 8.10 The growing reference window, the sliding test window, and the pooled window used in the GLR method for adaptive segmentation.

test window, and a pooled window formed by concatenating the two. Distance measures are then derived using AR-model prediction errors computed for the three datasets.

Let $\varepsilon(m : n)$ represent the prediction error energy (TSE ε as in Equation 7.19) within an arbitrary dataset or window with boundaries m and n . The maximum log likelihood measure $H(m : n)$ for the window is defined as

$$H(m : n) = (n - m + 1) \ln \left[\frac{\varepsilon(m : n)}{(n - m + 1)} \right]. \quad (8.30)$$

Three measures are computed for the three datasets described above as $H(1 : m - 1)$ for the growing reference window, $H(m : n)$ for the test window, and $H(1 : n)$ for the composite or pooled window. Here, the reference window is denoted as commencing from the time instant or sample 1, m is the last sample of the growing reference window, and the current test window spans the duration from m to the current time instant n (see Figure 8.10). The GLR distance measure is defined as

$$d(n) = H(1 : n) - [H(1 : m - 1) + H(m : n)]. \quad (8.31)$$

Here, the first quantity represents the TSE if the test window is appended to the growing reference window; the second quantity represents the TSE of the reference window grown so far; and the third quantity represents the TSE in modeling the test window itself. The measure $d(n)$ answers the question, “How much is the increase in the TSE if we add the test window to the growing reference window?”

Appel and v. Brandt [19] and Cohen [53] provide more details on the GLR. The GLR distance is a measure of the statistical similarity of the reference and test data sequences, with the assumption that their AR-model coefficients have a normal (Gaussian) distribution. The GLR distance is also a measure of the loss of information caused if no segment boundary is drawn at the position of the test window, that is, if it is assumed that the null hypothesis that the two sequences are similar is true.

Appel and v. Brandt [19] discuss issues related to the choice of the parameters involved in the GLR method, including the AR-model order, the test window length, and the threshold, on the GLR distance measure. The GLR method was also used by Willsky and Jones [54] to detect abrupt changes (sporadic anomalies and failures) in the variables of stochastic linear systems, as well as by Basseville and Benveniste [55] for segmentation of nonstationary signals (see also Cohen [53]). Illustrations of application of the GLR method are provided in Section 8.5.4.

8.5.4 Comparative analysis of the ACF, SEM, and GLR methods

Appel and v. Brandt [20] performed a comparative analysis of the performance of the ACF, *SEM*, and GLR methods of adaptive segmentation using synthesized signals as well as EEG signals. A simple two-pole system was used as the basis to simulate nonstationary signals. The gain, pole radius, and pole angle were individually varied back and forth between two sets of values. Several outputs of the dynamic system were computed with random signals (Gaussian-distributed white noise) as input. The signals were processed by the ACF, *SEM*, and GLR methods for adaptive segmentation. The variability of the segment boundaries detected for various realizations of the nonstationary (random) output signals for the same sequences of system parameters was analyzed.

Figure 8.11 shows the results related to variations in the angles of the poles, that is, in the resonance frequency of the system. The angle of the pole in the upper-half of the z -plane was changed from 20° to 40° and back at samples 200 and 400; the conjugate pole was also varied accordingly. The same changes were repeated at samples 700 and 800. The upper panel in the figure shows the pole positions and the related PSDs. The middle panel illustrates one sample of the 200 test signals generated: The higher-frequency characteristics of the signal related to the shifted pole positioned at 40° are evident over the intervals 200 – 400 and 700 – 800 samples. The lower panel illustrates the variability in the detected segment indices (dashed curve) and the estimated segment boundary positions (solid curves) for the three methods over 200 realizations of the test signals. (The true segment indices and boundaries are 1 : 200, 2 : 400, 3 : 700, and 4 : 800; ideally, the curves should exhibit steps at the points of change.) It is evident that the GLR method has provided the most consistent and accurate segmentation results, although at the price of increased computational load. The *SEM* method has performed better than the ACF method, with the latter showing the poorest results.

Figure 8.12 shows the results related to variations in the distance of the poles from the origin, that is, in the bandwidth of the resonance frequency of the system. The distance of the poles from the origin was changed from 0.7 to 0.9 and back at samples 200 and 400. The same changes were repeated at samples 700 and 800. The PSDs display the increased prominence of the spectral peak when the poles are pushed toward the unit circle. The ACF method has not performed well in recognizing the nonstationarities of this type in the test signals. The GLR method has performed better than the ACF method in segmentation.

Figure 8.13 shows the results of application of the three methods to an EEG signal. Although the exact locations where the signal changes its characteristics are not known for the EEG signal, the boundaries indicated by the GLR method appear to be the most accurate. It may be desirable in real-life applications to err on the side of superfluous segmentation; a subsequent clustering step could merge adjacent segments with similar model parameters.

8.6 Use of Adaptive Filters for Segmentation

We saw in Sections 3.10.2 and 3.10.3 that the coefficient (tap-weight) vectors of the adaptive LMS and RLS filters are expressed as functions of time. The filters adapt to changes in the statistics of the primary and reference signals. Could we, therefore, use the tap-weight vector $\mathbf{w}(n)$ to detect nonstationarities in a signal?

Problem: *Investigate the potential use of the RLS adaptive filter for adaptive segmentation of nonstationary signals.*

Solution: When we have only one signal to work with — the signal that is to be segmented — the question arises as to how we may provide two inputs, namely, the primary and reference signals, to the adaptive filter. If we assume that the signal to be segmented (applied at the primary input) was generated by an AR system, then we may provide the same signal with a delay as the reference input to the adaptive filter. The delay is to be set such that the reference input at a given instant of

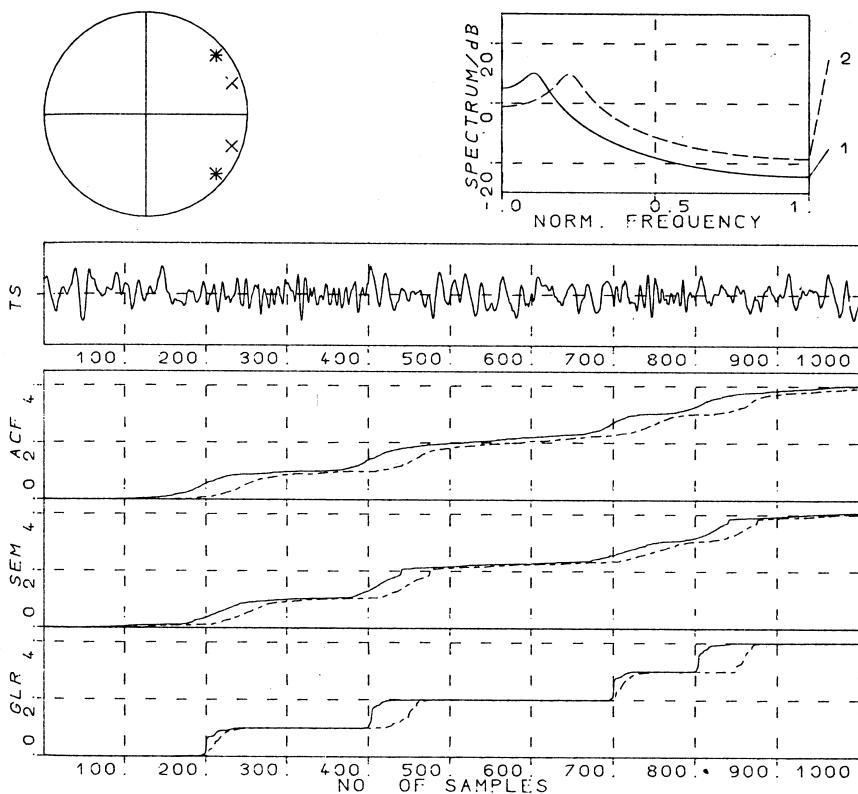


Figure 8.11 Comparative analysis of the ACF, SEM, and GLR methods for adaptive segmentation with the pole angle varied. Upper panel: pole positions and the related PSDs. Note: Norm. Frequency is normalized frequency such that the maximum frequency present in the sampled signal is unity. Middle panel: sample test signal; TS = time series. Lower panel: variability in the detected segment indices (dashed curve) and the estimated segment boundary positions (solid curves) for the three methods over 200 realizations of the test signals. See the text for more details. Reproduced with permission from U. Appel and A. v. Brandt, A comparative analysis of three sequential time series segmentation algorithms, *Signal Processing*, 6:45–60, 1984. ©Elsevier Science Publishers B.V. (North Holland).

time is uncorrelated with the primary input; the delay may also be set on the basis of the order of the filter. (It is also possible to apply white noise at the reference input.) In essence, the adaptive filter then acts the role of an adaptive AR model. The filter tap-weight vector is continually adapted to changes in the statistics (ACF) of the input signal. The output represents the prediction error. Significant changes in the tap-weight vector or the prediction error may be used to mark points of prominent nonstationarities in the signal. Figure 8.14 shows a signal-flow diagram of the adaptive filter as described above; the filter structure is only slightly different from that in Figure 3.94.

8.6.1 Monitoring the RLS filter

The RLS filter as in Figure 8.14 attempts to predict the current signal sample from the available knowledge of the previous samples stored in the filter's memory units. If a large change occurs in the signal, the prediction error exhibits a correspondingly large value. In response, the adaptive filter's tap-weight vector is modified by the RLS algorithm.

Moussavi et al. [57] applied the RLS filter for segmentation of VAG signals. The order of the filter was set to be 5 in order to be low enough to detect transient changes and also to provide fast

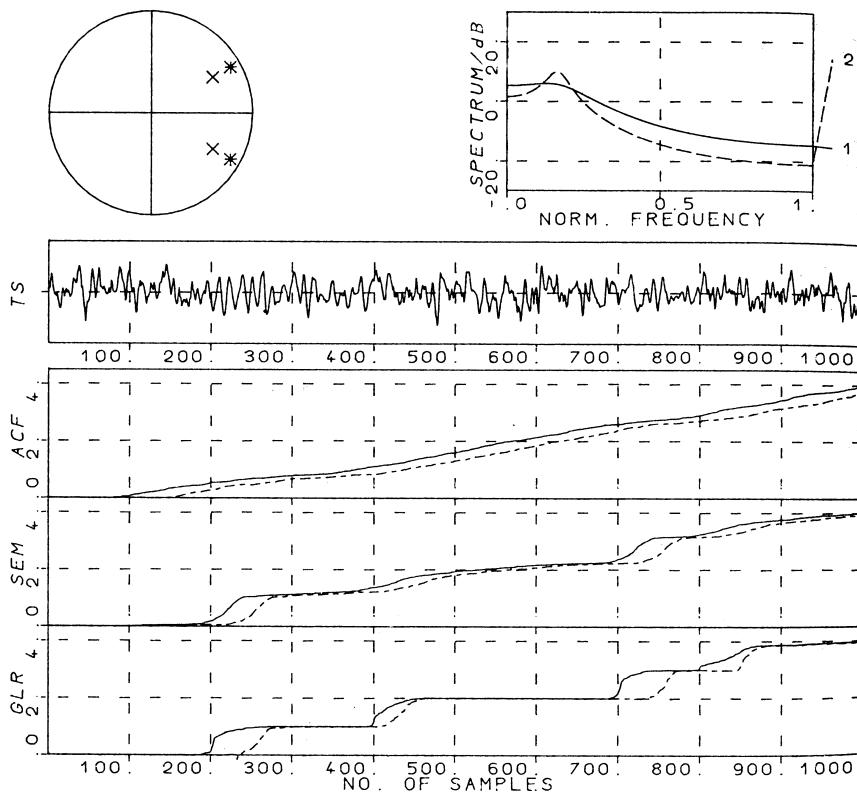


Figure 8.12 Comparative analysis of the ACF, SEM, and GLR methods for adaptive segmentation with the pole radius varied. Upper panel: pole positions and the related PSDs. Note: Norm. Frequency is normalized frequency such that the maximum frequency present in the sampled signal is unity. Middle panel: sample test signal; TS = time series. Lower panel: variability in the detected segment indices (dashed curve) and the estimated segment boundary positions (solid curves) for the three methods over 200 realizations of the test signals. See the text for more details. Reproduced with permission from U. Appel and A. v. Brandt, A comparative analysis of three sequential time series segmentation algorithms, *Signal Processing*, 6:45–60, 1984. ©Elsevier Science Publishers B.V. (North Holland).

convergence. The forgetting factor was defined as $\lambda = 0.98$ so that the filter may be assumed to operate in an almost-stationary situation. The delay between the input and the reference input was set to be 7 samples (which corresponds to 3.5 ms with $f_s = 2 \text{ kHz}$).

The adaptive segmentation algorithm of Moussavi et al. is as follows:

1. Initialize the RLS algorithm.
2. Find the squared Euclidean distance between the current tap-weight vector $\mathbf{w}(n)$ and the preceding vector $\mathbf{w}(n - 1)$ as

$$\Delta(n) = |\mathbf{w}(n) - \mathbf{w}(n - 1)|^2. \quad (8.32)$$

3. After computing $\Delta(n)$ for all samples of the signal available (in off-line processing), compute the *SD* of the $\Delta(n)$ values. Define a threshold as three times the *SD*.
4. Label all samples n for which $\Delta(n)$ exceeds the threshold as primary segment boundaries.

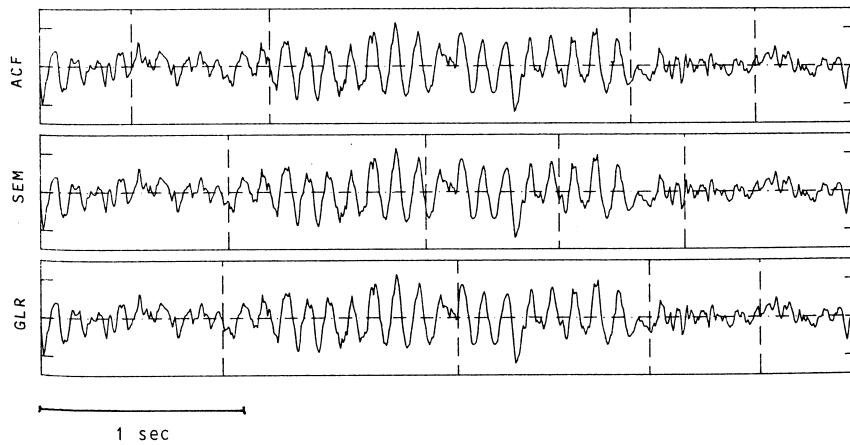


Figure 8.13 Comparative analysis of the ACF, SEM, and GLR methods for adaptive segmentation of an EEG signal. Reproduced with permission from U. Appel and A. v. Brandt, A comparative analysis of three sequential time series segmentation algorithms, *Signal Processing*, 6:45–60, 1984. ©Elsevier Science Publishers B.V. (North Holland).

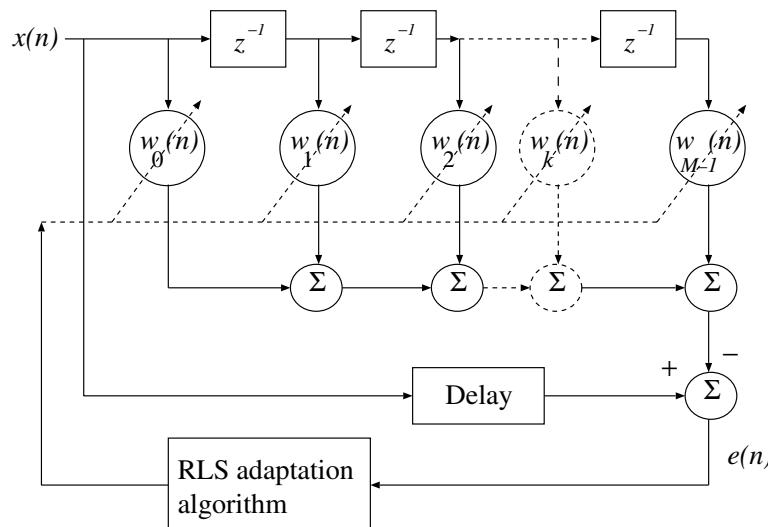


Figure 8.14 Adaptive RLS filter for segmentation of nonstationary signals [56]. See also Figure 3.94.

5. Compute the primary segment lengths (durations) as the differences between successive primary segment boundaries. Reject all primary segment boundaries that result in segment duration less than a preset minimum (defined in the work of Moussavi et al. [57] as 120 samples or 60 ms, corresponding to a knee-joint angle range of approximately 4°).
6. Mark the remaining boundary points as the final segment boundaries.

The main advantage of the RLS method is that there are no explicit reference and test windows as in the case of the ACF, SEM, and GLR methods. The RLS method computes a new filter tap-

weight vector at each sample of the incoming signal. The method was found to perform well in the detection of gradual changes as well as sudden variations in VAG signals.

Illustration of application: Figures 8.15 and 8.16 illustrate the segmentation of the VAG signals of a normal subject and a patient with arthroscopically confirmed cartilage pathology, respectively. The figures also illustrate the spectrograms of the two signals. While the segmentation of the abnormal signal in Figure 8.16 may appear to be superfluous at first sight, close inspection of the corresponding spectrogram indicates that the spectral characteristics of the signal do change within short intervals. It is evident that the RLS method has detected the different types of nonstationarity present in the signals. Moussavi et al. [57] tested the method with 46 VAG signals and observed that the segmentation boundaries agreed well with the nature of the joint sounds heard via auscultation with a stethoscope as well as with the spectral changes observed in the spectrograms of the signals. See Section 10.12 for further discussions on this application.

8.6.2 The RLS lattice filter

In order to apply the RLS method for adaptive segmentation in a nonstationary environment, it is necessary to solve the least-squares problem recursively and rapidly. The *recursive least-squares lattice* (RLSL) algorithm is well suited for such purposes. Since the RLRL method uses a lattice filter, and is based on forward and backward prediction and time-varying reflection coefficients, it is necessary to define some of the related procedures.

Forward and backward prediction: Let us rewrite Equation 7.17 related to LP or AR modeling as

$$\tilde{y}(n) = - \sum_{k=1}^M a_{M,k} y(n-k), \quad (8.33)$$

with the inclusion of the order of the model M as a subscript for the model coefficients a_k . In this procedure, M past samples of the signal, $y(n-1), y(n-2), \dots, y(n-M)$, are used in a linear combination to predict the current sample $y(n)$ in the *forward direction*. The *forward prediction error* is

$$e_{M,f}(n) = y(n) - \tilde{y}(n) = \sum_{k=0}^M a_{M,k} y(n-k), \quad (8.34)$$

with $a_{M,0} = 1$. This equation is a restatement of Equation 7.18 with the inclusion of the order of the model M as a subscript for the error e as well as the subscript f to indicate that the prediction is being performed in the forward direction.

The term *backward prediction* refers to the estimation of $y(n-M)$ from the samples $y(n), y(n-1), \dots, y(n-M+1)$ as

$$\tilde{y}(n-M) = - \sum_{k=0}^{M-1} a_{M,k}^\# y(n-k), \quad (8.35)$$

where $a_{M,k}^\#$ are the backward prediction coefficients. Application of the least-squares method described in Section 7.5 for a stationary signal leads to the result

$$a_{M,k}^\# = a_{M,M-k}, \quad k = 0, 1, 2, \dots, M, \quad (8.36)$$

that is, the backward prediction coefficients are the same as the forward prediction coefficients, but in reverse order [24]. The *backward prediction error* is, therefore, given by

$$\begin{aligned} e_{M,b}(n) &= y(n-M) - \tilde{y}(n-M) \\ &= \sum_{k=0}^M a_{M,k}^\# y(n-k) = \sum_{k=0}^M a_{M,M-k} y(n-k). \end{aligned} \quad (8.37)$$

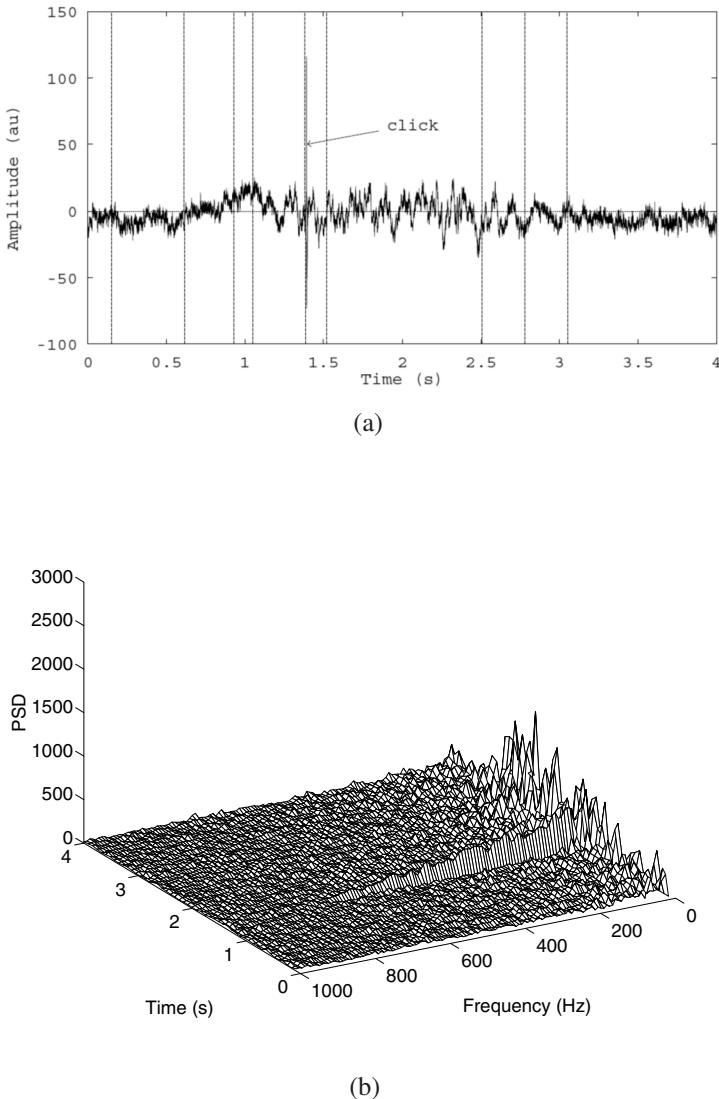


Figure 8.15 (a) Segmentation of the VAG signal of a normal subject using the RLS method. A click heard in auscultation of the knee joint is labeled. (b) Spectrogram (STFT) of the signal. Reproduced with permission from Z.M.K. Moussavi, R.M. Rangayyan, G.D. Bell, C.B. Frank, K.O. Ladly, and Y.T. Zhang, Screening of vibroarthrographic signals via adaptive segmentation and linear prediction modeling, *IEEE Transactions on Biomedical Engineering*, 43(1):15–23, 1996. ©IEEE.

The Burg-lattice method: The Burg-lattice method [24] is based on minimizing the sum of the squared forward and backward prediction errors. Assuming that the input $y(n)$ is ergodic, the *performance index* ξ_m is given by

$$\xi_m = \sum_{n=m+1}^N [e_{m,f}^2(n) + e_{m,b}^2(n)], \quad (8.38)$$

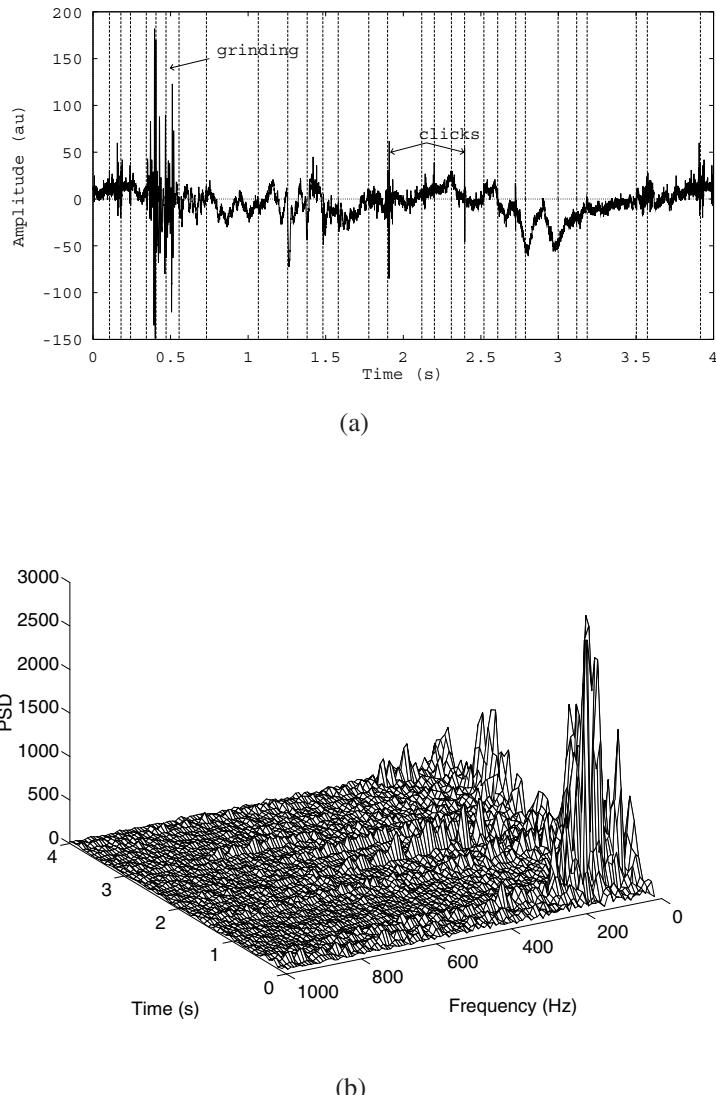


Figure 8.16 (a) Segmentation of the VAG signal of a subject with cartilage pathology using the RLS method. Clicking and grinding sounds heard during auscultation of the knee joint are labeled. (b) Spectrogram (STFT) of the signal. Reproduced with permission from Z.M.K. Moussavi, R.M. Rangayyan, G.D. Bell, C.B. Frank, K.O. Ladly, and Y.T. Zhang, Screening of vibroarthrographic signals via adaptive segmentation and linear prediction modeling, *IEEE Transactions on Biomedical Engineering*, 43(1):15–23, 1996. ©IEEE.

where $e_{m,f}(n)$ is the forward prediction error and $e_{m,b}(n)$ is the backward prediction error, with the model order m being recursively updated as $m = 1, 2, \dots, M$. The length of the available block of data is N samples.

If we use the Levinson–Durbin method to estimate the forward prediction coefficients, we get (see Section 7.5 and Equation 7.38)

$$a_{m,k} = a_{m-1,k} + \gamma_m a_{m-1,m-k}, \quad (8.39)$$

where γ_m is the reflection coefficient for order m . Similarly, for the case of backward prediction, we get

$$a_{m,m-k} = a_{m-1,m-k} + \gamma_m a_{m-1,m}, \quad (8.40)$$

including the substitution $a_{m,k}^\# = a_{m,m-k}$.

Combining the relationships in Equations 8.34, 8.37, 8.39, and 8.40 leads to the lattice structure for computation of the forward and backward prediction errors, where the two prediction error series are interrelated recursively as [24]

$$e_{m,f}(n) = e_{m-1,f}(n) + \gamma_m e_{m-1,b}(n-1) \quad (8.41)$$

and

$$e_{m,b}(n) = e_{m-1,b}(n-1) + \gamma_m e_{m-1,f}(n). \quad (8.42)$$

(All coefficients are assumed to be real-valued in this derivation; Haykin [24] allows for all coefficients to be complex-valued.) Figure 8.17 illustrates a basic unit of the lattice structure that performs the recursive operations in Equations 8.41 and 8.42. The reflection coefficient γ_m may be chosen so as to minimize the performance index given in Equation 8.38, that is, by setting

$$\frac{\partial \xi_m}{\partial \gamma_m} = 2 \sum_{n=m+1}^N \left[e_{m,f}(n) \frac{\partial e_{m,f}(n)}{\partial \gamma_m} + e_{m,b}(n) \frac{\partial e_{m,b}(n)}{\partial \gamma_m} \right] = 0. \quad (8.43)$$

Partial differentiation of Equations 8.41 and 8.42 with respect to γ_m yields

$$\frac{\partial e_{m,f}(n)}{\partial \gamma_m} = e_{m-1,b}(n-1) \quad (8.44)$$

and

$$\frac{\partial e_{m,b}(n)}{\partial \gamma_m} = e_{m-1,f}(n). \quad (8.45)$$

Substituting the results above in Equation 8.43, we get

$$\sum_{n=m+1}^N [e_{m,f}(n) e_{m-1,b}(n-1) + e_{m,b}(n) e_{m-1,f}(n)] = 0. \quad (8.46)$$

Substituting Equations 8.41 and 8.42 in Equation 8.46, we get

$$\begin{aligned} & \sum_{n=m+1}^N [\{e_{m-1,f}(n) + \gamma_m e_{m-1,b}(n-1)\} e_{m-1,b}(n-1) \\ & + \{e_{m-1,b}(n-1) + \gamma_m e_{m-1,f}(n)\} e_{m-1,f}(n)] = 0. \end{aligned} \quad (8.47)$$

The reflection coefficients γ_m can then be calculated as

$$\gamma_m = -2 \frac{\sum_{n=m+1}^N e_{m-1,f}(n) e_{m-1,b}(n-1)}{\sum_{n=m+1}^N [e_{m-1,f}^2(n) + e_{m-1,b}^2(n-1)]}. \quad (8.48)$$

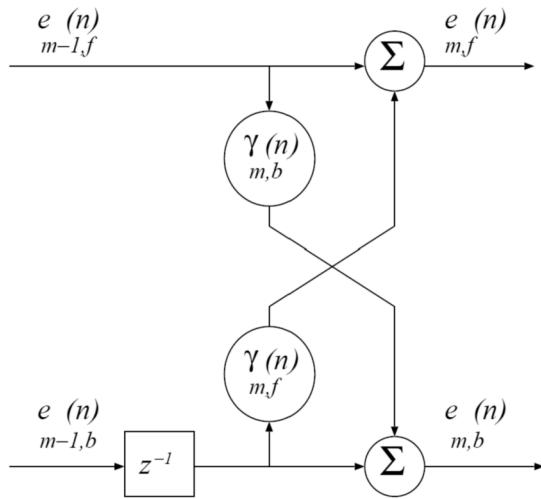


Figure 8.17 Basic unit of the lattice structure that performs the recursive operations in Equations 8.41 and 8.42 as well as the recursive operations in Equations 8.52 and 8.53. In the case of the former, due to the stationarity of the processes involved, the forward and backward reflection coefficients are the same and are independent of time (γ_m) [56].

The magnitudes of the reflection coefficients are less than unity. The Burg formula always yields a minimum-phase design for the lattice predictor.

The prediction coefficients or the AR-model parameters can be computed from the reflection coefficients by using the relationship in Equation 8.39. The order m is updated recursively as $m = 1, 2, \dots, M$, with $a_{m,0} = 1$, and $a_{m-1,k} = 0$ for $k > m - 1$. From Equation 8.39 and Figure 8.17, it can be observed that the AR coefficients can be computed for any model order by simply adding more lattice stages without affecting the earlier computations for lower orders. This is one of the main advantages of the Burg-lattice AR-modeling algorithm, especially in situations where the order of the system being modeled is not known in advance.

RLSL algorithm for adaptive segmentation: A general schematic representation of the RLSL filter structure is given in Figure 8.18. Two levels of updating are used in the RLSL algorithm:

1. *Order-update:* This involves updating the forward prediction error $e_{m,f}(n)$, the backward prediction error $e_{m,b}(n)$, the forward prediction error power $\varepsilon_{m,f}(n)$, and the backward prediction error power $\varepsilon_{m,b}(n)$. Here, m indicates the model order and n indicates the time instant.
2. *Time-update:* This involves time-updating of the parameters that ensure adaptation, including the forward reflection coefficients $\gamma_{m,f}(n)$ and backward reflection coefficients $\gamma_{m,b}(n)$. Note that, in the general nonstationary environment, $\gamma_{m,f}(n) \neq \gamma_{m,b}(n)$.

Order-updating and time-updating together enable the RLSL algorithm to achieve fast convergence and excellent tracking capability.

The RLSL algorithm can be expressed in three stages [24, 56, 58]:

1. *Initialization of the algorithm and lattice for filter order M :* The parameters of the algorithm are initialized at $n = 0$ and for each order $m = 1, 2, \dots, M$ by setting the forward prediction error power $\varepsilon_{m-1,f}(0)$ and the backward prediction error power $\varepsilon_{m-1,b}(0)$ equal to a small positive constant; the forward reflection coefficients $\gamma_{m,f}(0) = 0$; the backward reflection coefficients $\gamma_{m,b}(0) = 0$; the conversion factor $\gamma_{0,c}(0) = 1$; and an auxiliary variable $\Delta_{m-1}(0) = 0$.

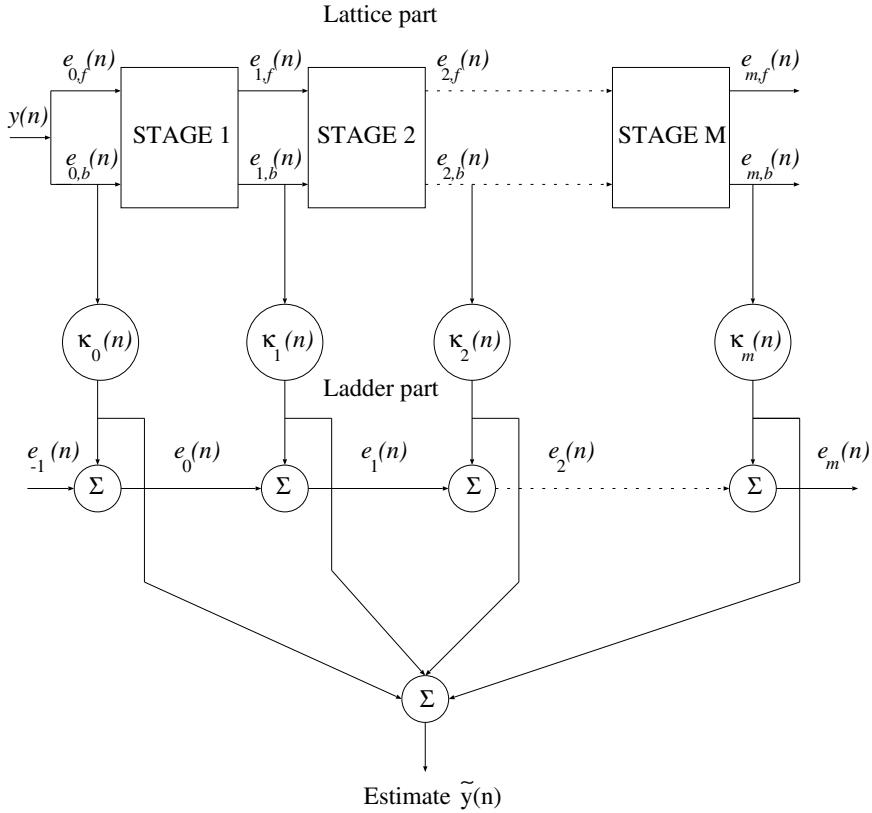


Figure 8.18 General schematic representation of the RLSL filter structure for adaptive segmentation of nonstationary signals [56].

For each time instant $n \geq 1$, the following zeroth-order variables are generated: the forward prediction error $e_{0,f}(n)$ equal to the data input $y(n)$; the backward prediction error $e_{0,b}(n) = y(n)$; $\varepsilon_{0,f}(n) = \varepsilon_{0,b}(n) = \lambda \varepsilon_{0,f}(n-1) + |y(n)|^2$, where λ is the forgetting factor; and $\gamma_{0,c}(n) = 1$.

The variables involved in joint process estimation, for each order $m = 0, 1, \dots, M$ at time $n = 0$, are initialized by setting the scalar $\rho_m(0) = 0$, and for each instant $n \geq 1$ the zeroth-order variable of *a priori* estimation error $e_0 = d(n)$, where $d(n)$ is the desired response of the system.

2. *Prediction part of the RLSL algorithm:* For $n = 1, 2, \dots, N_s$, where N_s is the number of signal samples available, the various order updates are computed in the sequence $m = 1, 2, \dots, M$, where M is the final order of the least squares predictor, as

$$\Delta_{m-1}(n) = \lambda \Delta_{m-1}(n-1) + \frac{e_{m-1,b}(n-1) e_{m-1,f}(n)}{\gamma_{m-1,c}(n-1)}, \quad (8.49)$$

where $\Delta_{m-1}(n)$ is the cross-correlation between the delayed backward prediction error $e_{m-1,b}(n-1)$ and the forward prediction error $e_{m-1,f}(n)$ of the lattice filter.

The forward reflection coefficient $\gamma_{m,f}(n)$ is then updated as

$$\gamma_{m,f}(n) = -\frac{\Delta_{m-1}(n)}{\varepsilon_{m-1,b}(n)}. \quad (8.50)$$

Similarly, the backward reflection coefficient is updated as

$$\gamma_{m,b}(n) = -\frac{\Delta_{m-1}(n)}{\varepsilon_{m-1,f}(n-1)}. \quad (8.51)$$

In general, $\varepsilon_{m-1,f}(n)$ and $\varepsilon_{m-1,b}(n-1)$ are unequal, so that in the RLSL algorithm, unlike in the Burg algorithm described earlier in this section, we have $\gamma_{m,f}(n) \neq \gamma_{m,b}(n)$.

From the lattice structure as described earlier in the context of Equations 8.41 and 8.42 and depicted in Figure 8.17, and noting that the reflection coefficients $\gamma_{m,f}(n)$ and $\gamma_{m,b}(n)$ are now different and time-variant parameters, we can write the order-update recursion of the forward prediction error as (see Figure 8.17)

$$e_{m,f}(n) = e_{m-1,f}(n) + \gamma_{m,f}(n) e_{m-1,b}(n-1), \quad (8.52)$$

and the order-update recursion of the backward prediction error as

$$e_{m,b}(n) = e_{m-1,b}(n-1) + \gamma_{m,b}(n) e_{m-1,f}(n). \quad (8.53)$$

The prediction error powers are updated as

$$\varepsilon_{m,f}(n) = \varepsilon_{m-1,f}(n) + \gamma_{m,f}(n) \Delta_{m-1}(n) \quad (8.54)$$

and

$$\varepsilon_{m,b}(n) = \varepsilon_{m-1,b}(n-1) + \gamma_{m,b}(n) \Delta_{m-1}(n). \quad (8.55)$$

The conversion factor $\gamma_{m,c}(n-1)$ is updated as

$$\gamma_{m,c}(n) = \gamma_{m-1,c}(n) - \frac{e_{m-1,b}^2(n)}{\varepsilon_{m-1,b}(n)}. \quad (8.56)$$

The equations in this step constitute the basic order-update recursions for the RLSL predictor. The recursions generate two sequences of prediction errors: the forward prediction error and the backward prediction error. The two error sequences play key roles in the recursive solution of the linear least-squares problem.

3. *Filtering part of the RLSL algorithm:* For $n = 1, 2, \dots, N_s$, the various order-updates are computed in the sequence $m = 0, 1, \dots, M$ as

$$\rho_m(n) = \lambda \rho_m(n-1) + \frac{e_{m,b}(n)}{\gamma_{m,c}(n)} e_{m-1}(n). \quad (8.57)$$

The regression coefficients $\kappa_m(n)$ of the joint process estimator are defined in terms of the scalar $\rho_m(n)$ as

$$\kappa_m(n) = \frac{\rho_m(n)}{\varepsilon_{m,b}(n)}. \quad (8.58)$$

The order-update recursion of the *a posteriori* estimation error $e_m(n)$ is then given as

$$e_m(n) = e_{m-1}(n) - \kappa_m(n) e_{m,b}(n). \quad (8.59)$$

The dynamics of the input signal, that is, the statistical changes occurring in the signal, are reflected in the lattice filter parameters. Parameters such as the reflection coefficients (γ_f and γ_b) and the MS value of the estimation error {that is, $E[e_m^2(n)]$ } may, therefore, be used to monitor the statistical changes.

The conversion factor γ_c that appears in the algorithm can be used as a good statistical detection measure of the “unexpectedness” of the recent data samples. As long as the data belong to the same distribution, the variable γ_c will be near unity. If the recent data samples belong to a different distribution, γ_c will tend to fall from unity. This will cause the factor $\frac{1}{\gamma_c}$ appearing in the time-update formula (Equation 8.49) to be large, which leads to abrupt changes in the lattice parameters. The quantities γ_c , $\frac{1}{\gamma_c}$, or $\frac{1}{1-\gamma_c}$ may be used for fast tracking of changes in the input data and to test for segment boundaries in a nonstationary environment.

Illustration of application: The advantage in using the RLSL filter for segmentation of VAG signals is that the statistical changes in the signals are well reflected in the filter parameters, and hence segment boundaries can be detected by monitoring any one of the filter parameters, such as the MSE, the conversion factor, or the reflection coefficients. Krishnan et al. [59] and Krishnan [56] used the conversion factor (γ_c) to monitor statistical changes in VAG signals. In a stationary environment, γ_c starts with a low initial value and remains small during the early part of the initialization period. After a few iterations, γ_c begins to increase rapidly toward the final value of unity. In the case of nonstationary signals, such as VAG, γ_c will fall from its steady-state value of unity whenever a change occurs in the statistics of the signal. This can be used in segmenting VAG signals into quasistationary components. The segmentation procedure proposed by Krishnan et al. [59] and Krishnan [56] is summarized as follows:

1. The VAG signal is passed twice through the segmentation filter: The first pass is used to allow the filter to converge, and the second pass is used to test the γ_c value at each sample against a threshold value for the detection of segment boundaries.
2. Whenever γ_c at a particular sample during the second pass is less than the threshold, a primary segment boundary is marked.
3. If the difference between two successive primary segment boundaries is less than the minimum desired segment length (120 samples in the work of Krishnan et al.), the later of the two boundaries is deleted.

Figures 8.19 and 8.20 show the results of application of the RLSL segmentation method to two VAG signals. Plots of $\gamma_c(n)$ are also included in the figures. It may be observed that the value of $\gamma_c(n)$ drops whenever there is a significant change in the characteristics of the signal. Whereas the direct application of a threshold on $\gamma_c(n)$ would result in superfluous segmentation, inclusion of the condition on the minimum segment length that is meaningful in the application is seen to provide practically useful segmentation. The number of segments was observed to be, on the average, eight segments per VAG signal. Signals of patients with cartilage pathology were observed to result in more segments than normal signals.

An advantage of the RLSL method of adaptive segmentation is that a fixed threshold may be used; Krishnan et al. [59] found a fixed threshold value of 0.9985 to give good segmentation results with VAG signals. The adaptive segmentation procedure was found to provide segments that agreed well with manual segmentation based on auscultation and/or arthroscopy. Adaptive analysis of VAG signals is described further in Sections 9.9 and 10.12.

8.7 The Kalman Filter

For individuals suffering from motor-related neurological disorders, assistive devices such as wearable prostheses can offer support via a brain-machine interface (BMI). In BMI, neural signals ob-

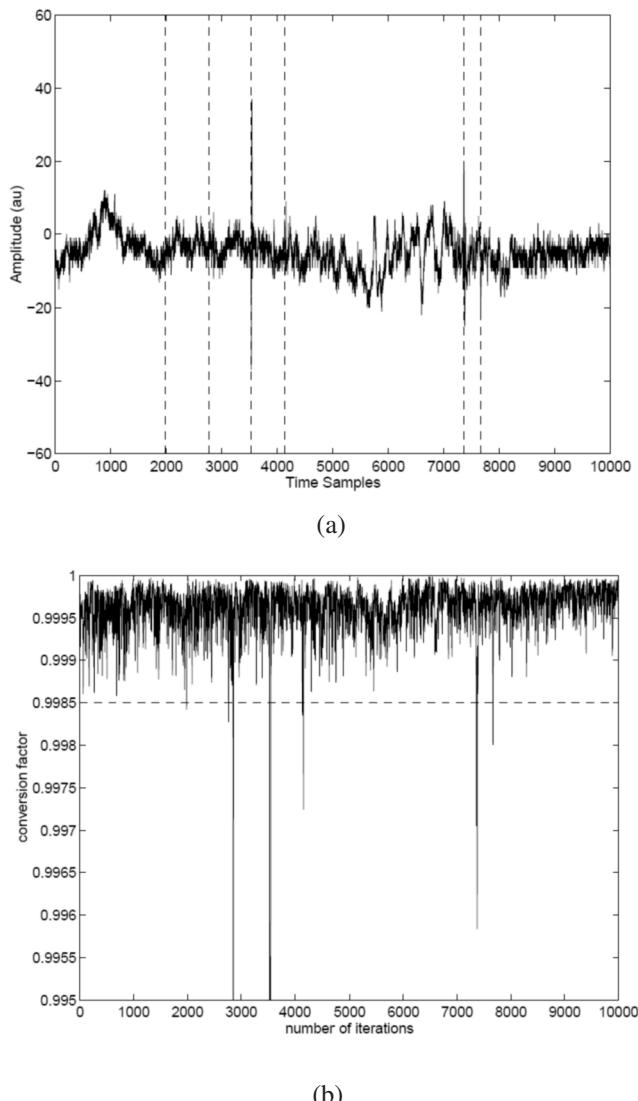


Figure 8.19 (a) VAG signal of a normal subject with the final segment boundaries given by the RLSL method shown by vertical dashed lines. (b) Plot of the conversion factor $\gamma_c(n)$; the horizontal dashed line represents the fixed threshold used to detect segment boundaries. The duration of the signal is 5 s, with $f_s = 2 \text{ kHz}$. Reproduced with permission from S. Krishnan, R.M. Rangayyan, G.D. Bell, C.B. Frank, and K.O. Ladly, Adaptive filtering, modelling, and classification of knee joint vibroarthrographic signals for non-invasive diagnosis of articular cartilage pathology, *Medical and Biological Engineering and Computing*, 35(6):677–684, 1997. ©IFMBE.

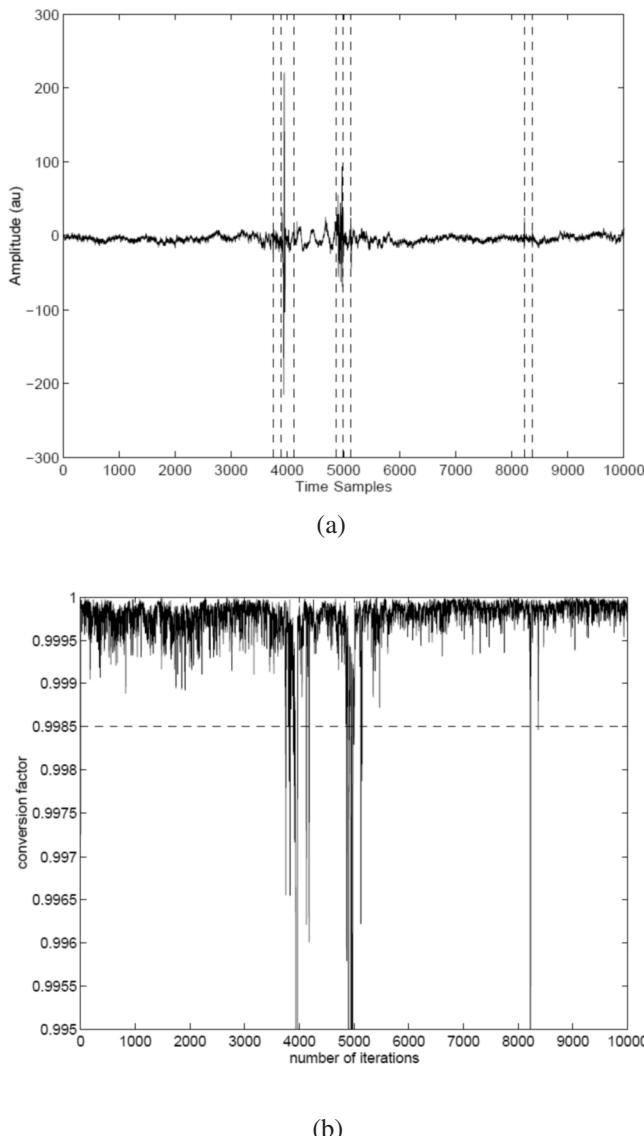


Figure 8.20 (a) VAG signal of a subject with cartilage pathology, with the final segment boundaries given by the RLSL method shown by vertical dashed lines. (b) Plot of the conversion factor $\gamma_c(n)$; the horizontal dashed line represents the fixed threshold used to detect segment boundaries. The duration of the signal is 5 s, with $f_s = 2 \text{ kHz}$. Reproduced with permission from S. Krishnan, R.M. Rangayyan, G.D. Bell, C.B. Frank, and K.O. Ladly, Adaptive filtering, modelling, and classification of knee joint vibroarthrographic signals for non-invasive diagnosis of articular cartilage pathology, *Medical and Biological Engineering and Computing*, 35(6):677–684, 1997. ©IFMBE.

tained noninvasively or semiinvasively from regions in the brain are converted into control signals to facilitate the guidance and operation of prosthetic devices. The control of a prosthetic device may be viewed as a classical system identification problem [60]. Among linear models, there are non-state-space models, such as the Wiener filter and the optimal linear estimator (see Section 3.9), and state-space models based on the Kalman filter. Nonlinear approaches include artificial neural networks (ANNs) and unscented Kalman filters [22, 24, 61]. Due to the ease of implementation, dynamic tracking capabilities, and the advantages of real-time operation, Kalman filters are typically preferred to RLS and RLSL filters for motion control in neurorehabilitation and assistive technology applications [60].

The impulse response of a Wiener filter defines its structure, which is stationary. The Kalman filter can be applied in stationary and nonstationary scenarios by integrating the concept of states and state transitions. The Kalman filter is a good choice for the design of dynamic systems, especially when the filter order is likely to be high. The Kalman filter is a recursive procedure that estimates the internal state of a linear dynamic system from a sequence of noisy observations. In the majority of applications, the internal state is characterized by many more attributes or parameters than the few “observable” characteristics that are monitored. The Kalman filter may estimate the internal state by combining a sequence of observations. By estimating a combined probability distribution across the variables for each period of observation, the Kalman filter generates estimates of unknown variables that are typically more accurate than those based on a single measurement alone. Specifically, Kalman filtering employs a dynamic model of the system, known control inputs to the system, and many sequential measurements (such as signals from sensors) to get a more accurate estimate of the system’s variable quantities (its state) than what would be possible with only one measurement.

The Kalman filter employs a two-phase operation as shown by the general block diagram in Figure 8.21. In the process system phase, the filter generates estimates of the current state variables and their associated uncertainty. When the outcome of the next measurement is observed, the estimates are updated using a weighted average, giving higher weight to more definite estimations. The algorithm has a recursive structure: It can work in real-time utilizing only the current input measurements and the previously determined state along with its uncertainty matrix; no further historical information is necessary. The mathematical formulation of the Kalman filtering procedure is described in the following paragraphs [22–24, 61, 62].

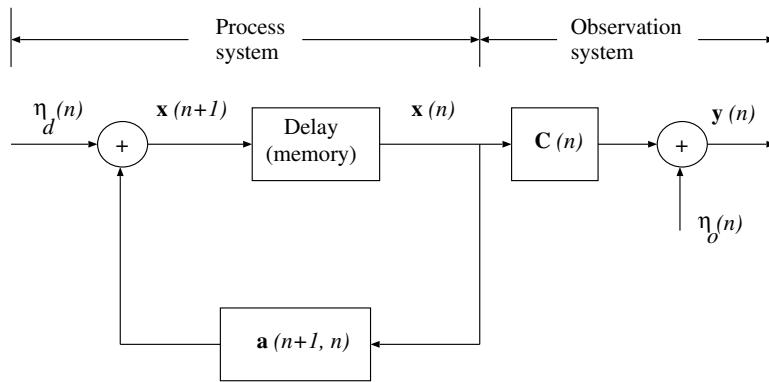


Figure 8.21 A general block diagram of the Kalman filter. Adapted from Rangayyan [23] and Haykin [24].

The signals are represented as a state vector $\mathbf{x}(n)$ and an observation vector $\mathbf{y}(n)$, where n could refer to the time instant of the state or observation vectors; see Figure 8.21. The noise source η_d generates the input process, and the observation noise source η_o affects the output. A state transition matrix represented as $\mathbf{a}(n + 1, n)$ indicates the state vector transition from one instant n to the next

instant $(n + 1)$, and is typically represented by an LP or AR model. The mapping from the state vector to the observation vector is represented by an observation matrix $\mathbf{C}(n)$.

The Kalman filter formulation [24] has two main equations: a process or message model and a measurement or observation model. The process model represents an updated value of the state vector \mathbf{x} recursively in terms of its new input and previous value as

$$\mathbf{x}(n + 1) = \mathbf{a}(n + 1, n) \mathbf{x}(n) + \boldsymbol{\eta}_d(n). \quad (8.60)$$

It could be assumed that the driving noise vector $\boldsymbol{\eta}_d$ is the source of excitation of the process system represented by the state vector \mathbf{x} and the state transition matrix \mathbf{a} . The noise process $\boldsymbol{\eta}_d$ is a zero-mean white-noise process which would be statistically independent of the stochastic characteristics of the state vector \mathbf{x} . The noise process $\boldsymbol{\eta}_d$ could be represented by an ACF ϕ_{η_d} given as

$$E [\boldsymbol{\eta}_d(n) \boldsymbol{\eta}_d^T(k)] = \begin{cases} \phi_{\eta_d}(n) & \text{if } n = k \\ \mathbf{0} & \text{otherwise.} \end{cases} \quad (8.61)$$

The state transition matrix $\mathbf{a}(n + 1, n)$ is characterized by the following properties [24]:

$$\begin{aligned} \mathbf{a}(l, m) \mathbf{a}(m, n) &= \mathbf{a}(l, n) && \text{product rule;} \\ \mathbf{a}^{-1}(m, n) &= \mathbf{a}(n, m) && \text{inverse rule;} \\ \mathbf{a}(m, m) &= \mathbf{I} && \text{identity.} \end{aligned} \quad (8.62)$$

For a stationary system, the state transition matrix would be a constant matrix \mathbf{a} that is stationary or independent of time.

The measurement or observation matrix $\mathbf{C}(n)$ transforms the state vector $\mathbf{x}(n)$ to the output vector $\mathbf{y}(n)$. The second main equation of the Kalman filter, the measurement or observation equation, is

$$\mathbf{y}(n) = \mathbf{C}(n) \mathbf{x}(n) + \boldsymbol{\eta}_o(n). \quad (8.63)$$

The measurement or observation noise $\boldsymbol{\eta}_o$ is assumed to be a zero-mean white-noise process that is statistically independent of the processes related to the state vector \mathbf{x} and the driving noise $\boldsymbol{\eta}_d$. The observation noise $\boldsymbol{\eta}_o$ is characterized by its ACF matrix ϕ_{η_o} . Due to the assumption of statistical independence of $\boldsymbol{\eta}_d$ and $\boldsymbol{\eta}_o$, we have

$$E [\boldsymbol{\eta}_d(n) \boldsymbol{\eta}_o^T(k)] = \mathbf{0} \quad \forall n, k. \quad (8.64)$$

Overall, the Kalman filtering problem may be stated as follows: given a series of the observation signal $\mathcal{Y}_n = \{\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(n)\}$, for each $n \geq 1$, provide the MMSE estimate of the state vector $\mathbf{x}(l)$. The derivation of the Kalman filter is provided in the following paragraphs; the material presented closely follows Haykin [24], Boulfelfel et al. [62], Sage and Melsa [63], and Rangayyan [23].

(Note: If \mathbf{x} and $\boldsymbol{\eta}_d$ are of size $M \times 1$, \mathbf{a} and ϕ_{η_d} are of size $M \times M$. In general, \mathbf{y} and $\boldsymbol{\eta}_o$ could be of a different size, $N \times 1$; then, ϕ_{η_o} would be of size $N \times N$ and \mathbf{C} would be of size $N \times M$. For the sake of simplicity, one could let $M = N$.)

The innovation process: The innovation process of obtaining a solution to the Kalman filtering problem involves a recursive estimation procedure using a one-step prediction process [24, 63]. Suppose that, based on the set of observations $\mathcal{Y}_{n-1} = \{\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(n-1)\}$, the MMSE estimate of $\mathbf{x}(n-1)$ has been obtained; let this estimate be denoted as $\tilde{\mathbf{x}}(n-1|\mathcal{Y}_{n-1})$. Given a new observation $\mathbf{y}(n)$, we could update the previous estimate and obtain a new state vector $\tilde{\mathbf{x}}(n|\mathcal{Y}_n)$. Because the state vector $\mathbf{x}(n)$ and the observation $\mathbf{y}(n)$ are related via the observation system, we may transfer the estimation procedure to the observation variable, and let $\tilde{\mathbf{y}}(n|\mathcal{Y}_{n-1})$ denote the MMSE estimate of $\mathbf{y}(n)$ given \mathcal{Y}_{n-1} . Then, the innovation process is defined as

$$\zeta(n) = \mathbf{y}(n) - \tilde{\mathbf{y}}(n|\mathcal{Y}_{n-1}), \quad n = 1, 2, \dots \quad (8.65)$$

The innovation process $\zeta(n)$ represents the new information contained in $\mathbf{y}(n)$ that cannot be estimated from \mathcal{Y}_{n-1} . Using the observation equation 8.63, we get

$$\begin{aligned}\tilde{\mathbf{y}}(n|\mathcal{Y}_{n-1}) &= \mathbf{C}(n) \tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1}) + \tilde{\boldsymbol{\eta}}_o(n|\mathcal{Y}_{n-1}) \\ &= \mathbf{C}(n) \tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1}),\end{aligned}\quad (8.66)$$

noting that $\tilde{\boldsymbol{\eta}}_o(n|\mathcal{Y}_{n-1}) = \mathbf{0}$ because the observation noise is orthogonal to the past observations. Combining Equations 8.65 and 8.66, we have

$$\zeta(n) = \mathbf{y}(n) - \mathbf{C}(n) \tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1}). \quad (8.67)$$

Using Equation 8.63, Equation 8.67 becomes

$$\zeta(n) = \mathbf{C}(n) \boldsymbol{\epsilon}_p(n, n-1) + \boldsymbol{\eta}_o(n), \quad (8.68)$$

where $\boldsymbol{\epsilon}_p(n, n-1)$ is the predicted state error vector at the instant n using the information available up to $(n-1)$, given by

$$\boldsymbol{\epsilon}_p(n, n-1) = \mathbf{x}(n) - \tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1}). \quad (8.69)$$

Some of the properties of the innovation process are [24]:

- $\zeta(n)$ is orthogonal to the past observations $\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(n-1)$, and hence

$$E [\zeta(n) \mathbf{y}^T(m)] = \mathbf{0}, \quad 1 \leq m \leq n-1. \quad (8.70)$$

- The innovation process is a series of vectors composed of random variables that are mutually orthogonal, and hence

$$E [\zeta(n) \zeta^T(m)] = \mathbf{0}, \quad 1 \leq m \leq n-1. \quad (8.71)$$

- A one-to-one correspondence exists between the observations $\{\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(n)\}$ and the vectors of the innovation process $\{\zeta(1), \zeta(2), \dots, \zeta(n)\}$.

The ACF matrix representation of the innovation process $\zeta(n)$ is given by

$$\begin{aligned}\phi_\zeta(n) &= E [\zeta(n) \zeta^T(n)] \\ &= \mathbf{C}(n) \phi_{\boldsymbol{\epsilon}_p}(n, n-1) \mathbf{C}^T(n) + \phi_{\boldsymbol{\eta}_o}(n),\end{aligned}\quad (8.72)$$

where

$$\phi_{\boldsymbol{\epsilon}_p}(n, n-1) = E [\boldsymbol{\epsilon}_p(n, n-1) \boldsymbol{\epsilon}_p^T(n, n-1)] \quad (8.73)$$

is the ACF matrix of the predicted state error, and the property that $\boldsymbol{\epsilon}_p(n, n-1)$ and $\boldsymbol{\eta}_o(n)$ are mutually orthogonal has been used. The ACF matrix of the predicted state error provides a statistical representation of the error in the predicted state vector $\tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1})$.

In order to obtain the MMSE estimate of the state vector $\mathbf{x}(l)$ using the innovation process, we may express the state vector as a linear transform of the innovation process as

$$\tilde{\mathbf{x}}(l|\mathcal{Y}_n) = \sum_{k=1}^n \mathbf{L}_l(k) \zeta(k), \quad (8.74)$$

where $\mathbf{L}_l(k)$, $l = 1, 2, \dots, n$, is a series of transformation matrices. Because the predicted state error vector is orthogonal to the innovation process, we have

$$\begin{aligned}E [\boldsymbol{\epsilon}_p(l, n) \zeta^T(m)] &= E [\{\mathbf{x}(l) - \tilde{\mathbf{x}}(l|\mathcal{Y}_n)\} \zeta^T(m)] \\ &= \mathbf{0}, \quad m = 1, 2, \dots, n.\end{aligned}\quad (8.75)$$

Using Equations 8.74, 8.75, and 8.71, we obtain

$$\begin{aligned} E[\mathbf{x}(l) \zeta^T(m)] &= \mathbf{L}_l(m) E[\zeta(m) \zeta^T(m)] \\ &= \mathbf{L}_l(m) \phi_\zeta(m). \end{aligned} \quad (8.76)$$

Furthermore, we get

$$\mathbf{L}_l(m) = E[\mathbf{x}(l) \zeta^T(m)] \phi_\zeta^{-1}(m). \quad (8.77)$$

Substituting the expression above for $\mathbf{L}_l(m)$ in Equation 8.74, we have

$$\tilde{\mathbf{x}}(l|\mathcal{Y}_n) = \sum_{k=1}^n E[\mathbf{x}(l) \zeta^T(k)] \phi_\zeta^{-1}(k) \zeta(k), \quad (8.78)$$

and the outcome can be expressed as

$$\begin{aligned} \tilde{\mathbf{x}}(l|\mathcal{Y}_n) &= \sum_{k=1}^{n-1} E[\mathbf{x}(l) \zeta^T(k)] \phi_\zeta^{-1}(k) \zeta(k) \\ &\quad + E[\mathbf{x}(l) \zeta^T(n)] \phi_\zeta^{-1}(n) \zeta(n). \end{aligned} \quad (8.79)$$

For $l = n + 1$, we obtain

$$\begin{aligned} \tilde{\mathbf{x}}(n+1|\mathcal{Y}_n) &= \sum_{k=1}^{n-1} E[\mathbf{x}(n+1) \zeta^T(k)] \phi_\zeta^{-1}(k) \zeta(k) \\ &\quad + E[\mathbf{x}(n+1) \zeta^T(n)] \phi_\zeta^{-1}(n) \zeta(n). \end{aligned} \quad (8.80)$$

Using the process equation 8.60, we have, for $0 \leq k \leq n$,

$$\begin{aligned} E[\mathbf{x}(n+1) \zeta^T(n)] &= E[\{\mathbf{a}(n+1, n) \mathbf{x}(n) + \boldsymbol{\eta}_d(n)\} \zeta^T(n)] \\ &= \mathbf{a}(n+1, n) E[\mathbf{x}(n) \zeta^T(n)]. \end{aligned} \quad (8.81)$$

The property that $\boldsymbol{\eta}_d(n)$ and $\zeta(k)$ are mutually orthogonal for $0 \leq k \leq n$ has been used to arrive at Equation 8.81. Using Equation 8.81 and Equation 8.78 with $l = n$, we get

$$\begin{aligned} \sum_{k=1}^{n-1} E[\mathbf{x}(n+1) \zeta^T(k)] \phi_\zeta^{-1}(k) \zeta(k) \\ &= \mathbf{a}(n+1, n) \sum_{k=1}^{n-1} E[\mathbf{x}(n) \zeta^T(k)] \phi_\zeta^{-1}(k) \zeta(k) \\ &= \mathbf{a}(n+1, n) \tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1}). \end{aligned} \quad (8.82)$$

The Kalman gain: Let

$$\mathbf{K}(n) = E[\mathbf{x}(n+1) \zeta^T(n)] \phi_\zeta^{-1}(n). \quad (8.83)$$

The expectation in the equation given above represents the CCF matrix between the state vector $\mathbf{x}(n+1)$ and the innovation process $\zeta(n)$. Using Equations 8.83 and 8.82, Equation 8.80 may be simplified to

$$\tilde{\mathbf{x}}(n+1|\mathcal{Y}_n) = \mathbf{a}(n+1, n) \tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1}) + \mathbf{K}(n) \zeta(n). \quad (8.84)$$

An important outcome of this equation is that one could obtain the MMSE estimate of the state vector $\tilde{\mathbf{x}}(n+1|\mathcal{Y}_n)$ by applying the state transition matrix $\mathbf{a}(n+1, n)$ to the previous estimate of

the state vector $\tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1})$ and adding a correction term. The correction term $\mathbf{K}(n) \zeta(n)$ includes the innovation process $\zeta(n)$ multiplied with the $M \times N$ matrix $\mathbf{K}(n)$, which is referred to as the Kalman gain.

Implementation of the Kalman gain matrix involves the following considerations. Equation 8.69 could be written as

$$\begin{aligned} E & [\epsilon_p(n, n-1) \epsilon_p^T(n, n-1)] \\ &= E [\mathbf{x}(n) \epsilon_p^T(n, n-1)] - E [\tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1}) \epsilon_p^T(n, n-1)] \\ &= E [\mathbf{x}(n) \epsilon_p^T(n, n-1)]. \end{aligned} \quad (8.85)$$

It should be noted that the estimated state vector $\tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1})$ and the predicted state error vector $\epsilon_p(n, n-1)$ are mutually orthogonal. Using Equations 8.81, 8.68, and 8.85, the following result is obtained:

$$\begin{aligned} E[\mathbf{x}(n+1) \zeta^T(n)] &= \mathbf{a}(n+1, n) E[\mathbf{x}(n) \zeta^T(n)] \\ &= \mathbf{a}(n+1, n) E[\mathbf{x}(n) \{ \mathbf{C}(n) \epsilon_p(n, n-1) + \boldsymbol{\eta}_o(n) \}^T] \\ &= \mathbf{a}(n+1, n) E[\mathbf{x}(n) \epsilon_p^T(n, n-1)] \mathbf{C}^T(n) \\ &= \mathbf{a}(n+1, n) E[\epsilon_p(n, n-1) \epsilon_p^T(n, n-1)] \mathbf{C}^T(n) \\ &= \mathbf{a}(n+1, n) \phi_{\epsilon_p}(n, n-1) \mathbf{C}^T(n). \end{aligned} \quad (8.86)$$

In arriving at the result above, Equation 8.73 has been used; the property that \mathbf{x} and $\boldsymbol{\eta}_o$ are independent processes has also been used. The Kalman gain matrix is obtained by using Equation 8.86 in Equation 8.83 as

$$\mathbf{K}(n) = \mathbf{a}(n+1, n) \phi_{\epsilon_p}(n, n-1) \mathbf{C}^T(n) \phi_{\zeta}^{-1}(n). \quad (8.87)$$

Using Equation 8.72 for $\phi_{\zeta}(n)$, we get

$$\mathbf{K}(n) = \mathbf{a}(n+1, n) \phi_{\epsilon_p}(n, n-1) \mathbf{C}^T(n) [\mathbf{C}(n) \phi_{\epsilon_p}(n, n-1) \mathbf{C}^T(n) + \phi_{\eta_o}(n)]^{-1}. \quad (8.88)$$

Practical computation of the Kalman gain matrix could use the following steps [24]. Extending Equation 8.69 one step further, we have

$$\epsilon_p(n+1, n) = \mathbf{x}(n+1) - \tilde{\mathbf{x}}(n+1|\mathcal{Y}_n). \quad (8.89)$$

By combining Equations 8.60, 8.67, 8.84, and 8.89, we have

$$\begin{aligned} \epsilon_p(n+1, n) &= \mathbf{a}(n+1, n) [\mathbf{x}(n) - \tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1})] \\ &\quad - \mathbf{K}(n) [\mathbf{y}(n) - \mathbf{C}(n) \tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1})] + \boldsymbol{\eta}_d(n). \end{aligned} \quad (8.90)$$

Using the measurement equation 8.63 and Equation 8.69, the last equation above could be written as

$$\begin{aligned} \epsilon_p(n+1, n) &= \mathbf{a}(n+1, n) \epsilon_p(n, n-1) \\ &\quad - \mathbf{K}(n) [\mathbf{C}(n) \mathbf{x}(n) + \boldsymbol{\eta}_o(n) - \mathbf{C}(n) \tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1})] + \boldsymbol{\eta}_d(n) \\ \\ &= \mathbf{a}(n+1, n) \epsilon_p(n, n-1) \\ &\quad - \mathbf{K}(n) \mathbf{C}(n) [\mathbf{x}(n) - \tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1})] + \boldsymbol{\eta}_d(n) - \mathbf{K}(n) \boldsymbol{\eta}_o(n) \\ \\ &= [\mathbf{a}(n+1, n) - \mathbf{K}(n) \mathbf{C}(n)] \epsilon_p(n, n-1) + \boldsymbol{\eta}_d(n) - \mathbf{K}(n) \boldsymbol{\eta}_o(n). \end{aligned} \quad (8.91)$$

Combining Equations 8.73 and 8.91, and noting the property that the processes ϵ_p , η_d , and η_o are mutually uncorrelated, the ACF matrix of the predicted state error $\epsilon_p(n+1, n)$ is

$$\begin{aligned}
\phi_{\epsilon_p}(n+1, n) &= E [\epsilon_p(n+1, n) \epsilon_p^T(n+1, n)] \\
&= [\mathbf{a}(n+1, n) - \mathbf{K}(n) \mathbf{C}(n)] \phi_{\epsilon_p}(n, n-1) \\
&\quad \times [\mathbf{a}(n+1, n) - \mathbf{K}(n) \mathbf{C}(n)]^T + \phi_{\eta_d}(n) + \mathbf{K}(n) \phi_{\eta_o}(n) \mathbf{K}^T(n) \\
&= \mathbf{a}(n+1, n) \phi_{\epsilon_p}(n, n-1) \mathbf{a}^T(n+1, n) \\
&\quad - \mathbf{K}(n) \mathbf{C}(n) \phi_{\epsilon_p}(n, n-1) \mathbf{a}^T(n+1, n) \\
&\quad - \mathbf{a}(n+1, n) \phi_{\epsilon_p}(n, n-1) \mathbf{C}^T(n) \mathbf{K}^T(n) \\
&\quad + \mathbf{K}(n) \mathbf{C}(n) \phi_{\epsilon_p}(n, n-1) \mathbf{C}^T(n) \mathbf{K}^T(n) \\
&\quad + \phi_{\eta_d}(n) + \mathbf{K}(n) \phi_{\eta_o}(n) \mathbf{K}^T(n) \\
&= \mathbf{a}(n+1, n) \phi_{\epsilon_p}(n, n-1) \mathbf{a}^T(n+1, n) \\
&\quad - \mathbf{K}(n) \mathbf{C}(n) \phi_{\epsilon_p}(n, n-1) \mathbf{a}^T(n+1, n) \\
&\quad - \mathbf{a}(n+1, n) \phi_{\epsilon_p}(n, n-1) \mathbf{C}^T(n) \mathbf{K}^T(n) \\
&\quad + \mathbf{K}(n) [\mathbf{C}(n) \phi_{\epsilon_p}(n, n-1) \mathbf{C}^T(n) + \phi_{\eta_o}(n)] \mathbf{K}^T(n) + \phi_{\eta_d}(n) \\
&= \mathbf{a}(n+1, n) \phi_{\epsilon_p}(n, n-1) \mathbf{a}^T(n+1, n) \\
&\quad - \mathbf{K}(n) \mathbf{C}(n) \phi_{\epsilon_p}(n, n-1) \mathbf{a}^T(n+1, n) \\
&\quad - \mathbf{a}(n+1, n) \phi_{\epsilon_p}(n, n-1) \mathbf{C}^T(n) \mathbf{K}^T(n) \\
&\quad + \mathbf{K}(n) \phi_{\zeta}(n) \mathbf{K}^T(n) + \phi_{\eta_d}(n), \tag{8.92}
\end{aligned}$$

which results in

$$\phi_{\epsilon_p}(n+1, n) = \mathbf{a}(n+1, n) \phi_{\epsilon_p}(n) \mathbf{a}^T(n+1, n) + \phi_{\eta_d}(n). \tag{8.93}$$

Equations 8.72 and 8.87 have been used in arriving at the result given above. A new $M \times M$ matrix $\phi_{\epsilon_p}(n)$ has been introduced, defined as

$$\phi_{\epsilon_p}(n) = \phi_{\epsilon_p}(n, n-1) - \mathbf{a}(n, n+1) \mathbf{K}(n) \mathbf{C}(n) \phi_{\epsilon_p}(n, n-1). \tag{8.94}$$

The property $\mathbf{a}^{-1}(n+1, n) = \mathbf{a}(n, n+1)$ has been used, which follows from the inverse rule in Equation 8.62. Equation 8.93 is commonly referred to as the Riccati equation, which helps in the recursive computation of the ACF matrix of the predicted state error.

Summary of the Kalman filter: The following steps summarize the main procedures in implementing the Kalman filter [24].

Data available: The observation vectors $\mathcal{Y}_n = \{\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(n)\}$.

System parameters assumed to be known:

- The state transition matrix $\mathbf{a}(n+1, n)$.
- The observation system matrix $\mathbf{C}(n)$.
- The ACF matrix of the driving noise $\phi_{\eta_d}(n)$.

- The ACF matrix of the observation noise $\phi_{\eta_o}(n)$.

Initial conditions:

- $\tilde{\mathbf{x}}(1|\mathcal{Y}_0) = E[\mathbf{x}(1)] = \mathbf{0}$;
- $\phi_{\epsilon_p}(1, 0) = \mathbf{D}_0$, a diagonal matrix with values of the order of 10^{-2} .

Iterative steps: For $n = 1, 2, 3, \dots$, do the following:

1. Using Equation 8.88, compute the Kalman gain matrix as

$$\begin{aligned} \mathbf{K}(n) &= \mathbf{a}(n+1, n) \phi_{\epsilon_p}(n, n-1) \mathbf{C}^T(n) \\ &\times [\mathbf{C}(n) \phi_{\epsilon_p}(n, n-1) \mathbf{C}^T(n) + \phi_{\eta_o}(n)]^{-1}. \end{aligned} \quad (8.95)$$

2. Compute the innovation process vector using Equation 8.67 as

$$\zeta(n) = \mathbf{y}(n) - \mathbf{C}(n) \tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1}). \quad (8.96)$$

3. Compute the estimate of the state vector using Equation 8.84 as

$$\tilde{\mathbf{x}}(n+1|\mathcal{Y}_n) = \mathbf{a}(n+1, n) \tilde{\mathbf{x}}(n|\mathcal{Y}_{n-1}) + \mathbf{K}(n) \zeta(n). \quad (8.97)$$

4. Using Equation 8.94, compute the ACF matrix of the filtered state error as

$$\phi_{\epsilon_p}(n) = \phi_{\epsilon_p}(n, n-1) - \mathbf{a}(n, n+1) \mathbf{K}(n) \mathbf{C}(n) \phi_{\epsilon_p}(n, n-1). \quad (8.98)$$

5. Update the ACF matrix of the predicted state error by using Equation 8.93 as

$$\phi_{\epsilon_p}(n+1, n) = \mathbf{a}(n+1, n) \phi_{\epsilon_p}(n) \mathbf{a}^T(n+1, n) + \phi_{\eta_d}(n). \quad (8.99)$$

The Kalman filter may be used for estimation, prediction, system identification, and filtering applications [24]. Kalman filtering in the feature space has been used to improve tremor detection associated with Parkinson's disease [64]. Extended Kalman filtering has shown improved fetal ECG extraction from single-channel recordings of the ECG obtained from the maternal abdomen [65]. Kalman filtering for grasping force estimation in myoelectric prosthesis was reported by Dutra et al. [66]. A 2D Kalman filter was developed by Boulfelfel et al. [62] for space-variant restoration of nuclear medicine images.

Illustration of application: Using the Kalman filter, Gowda et al. [67] developed a BMI system to maximize task-specific performance. In their investigation, Gowda et al. utilized two monkeys. The experimental setup and procedure are depicted in Figure 8.22. Microwire electrode arrays were implanted in the monkeys' brains to record neural signals. The monkeys were taught to undertake a self-paced delayed 2D center-out reaching exercise to eight uniformly spaced circular objects.

The monkeys learned to accomplish the BMI task after being instructed to complete it using arm motions. The monkeys were controlled through a Kalman filter decoder to move the cursor directly using brain activity rather than overt arm movements. The monkeys were given a sufficient amount of time to enter the center and begin a trial. The peripheral target emerged after entering the center. The monkeys were cued to begin the reach once the center-hold phase was completed (the center changed color); then, they had to move the mouse to the peripheral target within $3 - 7\text{s}$ and maintain it there for $250 - 400\text{ ms}$ to obtain a reward. Failure to retain the center, hold the peripheral target, or reach the peripheral target within the time restriction resulted in the trial being reset to the same target with no reward.

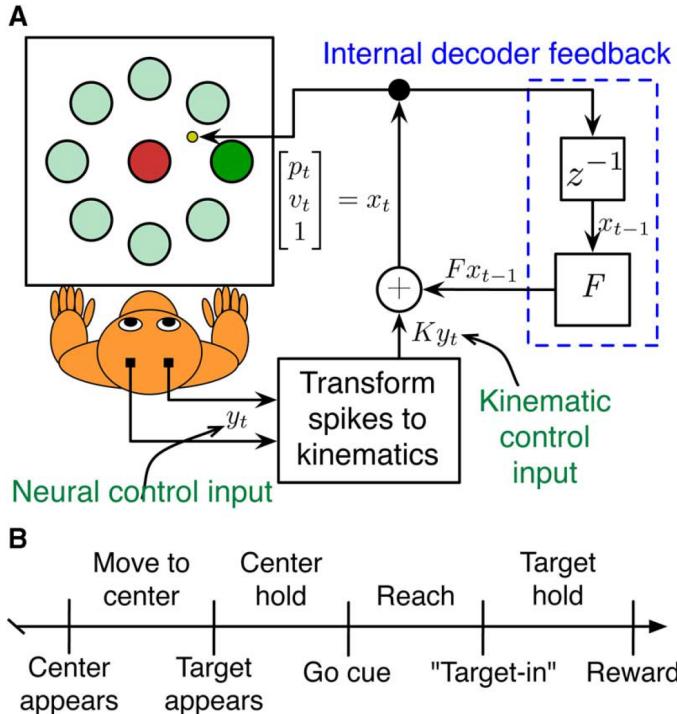


Figure 8.22 A. Experimental setup for BMI tasks performed by a monkey. B. Task events timeline [67]. See the text for details; the notation of the variables in the figure is different from the notation in the text. Reproduced with permission from S. Gowda, A.L. Orsborn, S.A. Overduin, H.G. Moorman, and J.M. Carmena. Designing dynamical properties of brain-machine interfaces to optimize task-specific performance. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(5):911–920, 2014. ©IEEE.

The BMI system model used by Gowda et al. [67] included the Kalman filter as a linear decoder, and could be mathematically represented as

$$x_n = (I - K_n C) Ax_{n-1} + K_n y_n = F_n x_{n-1} + K_n y_n, \quad (8.100)$$

where x_n denotes the intended kinematics of the cursor position and velocity using the previous estimate x_{n-1} , y_n is the current neural spike observation, C represents the neural firing parameter, F_n is the state transition matrix, and K_n is the control input matrix.

Frequent center-hold failures occurred at the beginning of the experiment because the penalty for the monkey was mild. Failure to maintain contact with the object occurred significantly more often than failure to reach the target within the given time limit. Targets were randomized to ensure that trials were distributed uniformly to each target in random order. Performance indicators such as cursor speed, hold error rate, and movement error were used to evaluate BMI performance [67]; for all of these performance indicators, lower numbers indicate better performance. Due to the variation in the duration of experimental blocks and the task was self-paced, the number of trials varied among sessions. The results showed that discrepancies between true BMI dynamics and the Kalman filter model used for control could create errors in BMI performance measures and the intended trajectory. Analysis of the influence of the dynamic features of the Kalman filter on performance could help to build future BMI systems that are suitable for complex neural control tasks.

Kalman filters with time-invariant parameters, such as those employed in BMI applications, converge to a steady-state form [60]. The capability to enforce the requirement that the motions of

a prosthetic device conform to physical rules is an advantage provided by dynamic modeling of the kinematic variables. A major drawback of Kalman filtering is that the state and observation noise processes are usually assumed to be Gaussian, which is a simplification. See Section 8.18 for discussions on control of prostheses using the Kalman filter.

8.8 Wavelet Analysis

Biomedical signals are often mixtures of multiple components with time-varying properties; to complicate matters, such components may have varying durations and overlap one another. For example, PCG signals could include not only S1 and S2 but also murmurs and opening snaps of valves. Furthermore, S1 and S2 are themselves made up of multiple overlapping components, murmurs are noise-like high-frequency sounds, and opening snaps are transients. We have seen that VAG signals could have multiple components related to various types and stages of deterioration of knee-joint surfaces causing sustained grinding sounds or transient clicks. When analyzing such signals, it would be desirable to obtain a joint distribution of their temporal and frequency-domain characteristics. The STFT and the various segmentation-based approaches described in the preceding sections of the present chapter are approaches that provide information of this nature, but impose several limitations.

A basic limitation of Fourier analysis is that the sinusoidal signals used as the basis functions are of infinite duration: The integral (see Equation 3.75) or summation (see Equation 3.80) used to compute the projection of the given signal on to each of the basis functions is defined over all time (or the entire duration of the given signal). Therefore, the properties of the given signal are averaged over its duration. Such global averaging provides good frequency resolution but no temporal resolution: The procedure obscures transients and components or events that exist over short durations of time. The STFT provides some improvement in temporal resolution by using short windows along the time axis but suffers from related limitations as described in Sections 8.4.1 and 8.4.2. Another limitation of Fourier analysis is that the basis function is limited to sinusoids; this limitation applies also to the STFT.

Problem: *How can we overcome the limitations of Fourier analysis and study the time-varying characteristics of nonstationary and multicomponent signals with improved accuracy and flexibility? How can we tailor the tool or procedure of analysis to adapt to the particular nature of a given signal?*

Solution: The answers to the questions stated above are provided by wavelets and methods for joint TF analysis that were developed over the past few decades and have gained substantial popularity in the analysis of biomedical signals [37,39,68–78]. TFDs may be used for representation and analysis of nonstationary signals; furthermore, TF and other decomposition methods may be used for parameterization of signals and feature extraction for classification. The following sections provide descriptions of wavelet analysis, TFDs, and signal decomposition methods with illustrations of their application. Discussions focused on signal analysis based on adaptive decomposition are presented in Chapter 9.

8.8.1 Approximation of a signal using wavelets

The analysis of a signal using a transform may be viewed as a decomposition of the given signal in the form of a weighted combination of the basis functions that define the transform (see Equation 3.80). For example, in Fourier analysis, the given signal is expressed as a weighted combination of mutually orthonormal sinusoids (see Equation 3.81). If only some of the basis functions are retained and others are rejected (or given selective weighting), we obtain an approximation of the original signal, commonly referred to as a filtered version of the original signal. We will now generalize this concept and extend its definition and application.

In linear approximation, the given signal is projected over M mutually orthogonal basis functions that are predetermined or chosen *a priori*. A linear approximation of a signal $x(t)$ may be written as

$$\tilde{x}(t) = \sum_{m=0}^{M-1} \langle x, \psi_m \rangle \psi_m(t), \quad (8.101)$$

where $\langle x, \psi_m \rangle$ denotes the projection, inner product, or dot product of $x(t)$ with one of the M orthogonal basis vectors $\psi_m(t)$; see also Equation 3.85. An optimal linear approximation is provided by the Karhunen–Loëve transform (KLT) [79]; see Section 9.7.1.

An adaptive approximation may be achieved by tailoring the M orthogonal basis functions to the properties of the given signal. The use of signal-adaptive basis functions leads to nonlinear decomposition, which may be expressed as

$$\tilde{x}(t) = \sum_{m \in B_M} \langle x, \psi_m \rangle \psi_m(t), \quad (8.102)$$

where B_M denotes a group of basis functions from a dictionary that provides the first M inner-product values $\langle x, \psi_m \rangle$ arranged in decreasing order. The functions in B_M are chosen as functions that correlate best with $x(t)$ and may be interpreted as the main features of $x(t)$. Instead of sinusoids as in the Fourier transform, one could define a set of wavelets that facilitate analysis of the characteristics of a certain type of signals.

Filtering may be performed using wavelets by approximating the given signal with a small number of nonzero wavelet coefficients. If $w_m = \langle x, \psi_m \rangle$ denotes the wavelet coefficients, hard thresholding is performed as

$$\tilde{w}_m = \begin{cases} w_m & \text{if } |w_m| \geq T, \\ 0 & \text{if } |w_m| < T, \end{cases} \quad (8.103)$$

where T denotes a threshold value. Soft thresholding is given by

$$\tilde{w}_m = \begin{cases} \operatorname{sgn}(w_m)(|w_m| - T) & \text{if } |w_m| \geq T, \\ 0 & \text{if } |w_m| < T. \end{cases} \quad (8.104)$$

The filtered signal (or the approximation) is given by

$$\tilde{x}(t) = \sum_{m \in B_M} \tilde{w}_m \psi_m(t). \quad (8.105)$$

The wavelet transform provides an approximation as above, where the basis functions or wavelets are obtained by dilating and translating a prototype function, $\psi(t)$, also known as the mother wavelet, written as

$$\psi_{\tau,s}(t) = \frac{1}{\sqrt{s}} \psi \left(\frac{t - \tau}{s} \right), \quad (8.106)$$

where s denotes the dilation or scale parameter and τ is the translation or shift parameter. The parameter τ represents the position on the time axis where the wavelet ψ is placed; the notion of

scale represented by the parameter s is related to the concept of frequency or a band of frequencies. Formally, the continuous wavelet transform (CWT) is defined as

$$X(\tau, s) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t - \tau}{s} \right) dt. \quad (8.107)$$

The integral in the CWT equation represents the projection of the given signal $x(t)$ on to the scaled wavelet with a shift so as to be positioned at $t = \tau$. The coefficient $X(\tau, s)$ resulting from the projection indicates the commonality between the two functions, or the extent or strength of the wavelet that is present in the signal being analyzed. The CWT equation also represents analysis of the signal using several wavelets so as to derive the corresponding projections or coefficients $X(\tau, s)$.

In practice, the CWT may be implemented with suitable discretization of the (τ, s) space. Typically, the time axis is discretized to the same sampling interval as that of the signal being processed. Due to substantial overlap between the shifted and scaled wavelets used, the CWT results in a redundant representation of the signal being analyzed. The (τ, s) space is referred to as the time-scale space and facilitates TF analysis if the relationships between the scaled wavelets and their spectral characteristics are taken into account.

Functions to be used as wavelets are required to possess certain properties. The property of finite energy is expressed as

$$\int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty. \quad (8.108)$$

The admissibility condition is stated as

$$C_{\psi} = \int_0^{\infty} \frac{|\Psi(f)|^2}{f} df < \infty, \quad (8.109)$$

where $\Psi(f)$ is the Fourier transform of $\psi(t)$ and C_{ψ} is known as the admissibility constant. If the wavelet is complex, its Fourier transform should be real and be zero for negative frequencies.

The inverse CWT is defined as

$$x(t) = \frac{1}{C_{\psi}} \int_{\tau=-\infty}^{\infty} \int_{s=0}^{\infty} X(\tau, s) \frac{1}{\sqrt{s}} \psi \left(\frac{t - \tau}{s} \right) d\tau \frac{ds}{s^2}. \quad (8.110)$$

The integral in the inverse CWT represents reconstruction or synthesis of the signal $x(t)$ using a weighted combination of scaled and shifted wavelets $\psi_{\tau,s}(\tau)$; each wavelet is weighted by the corresponding CWT coefficient $X(\tau, s)$.

The DWT is defined as

$$X(m, n) = \int_{-\infty}^{\infty} x(t) \psi_{m,n}^*(t) dt \quad (8.111)$$

with

$$\psi_{m,n}(t) = \frac{1}{\sqrt{s_0^m}} \psi \left(\frac{t - n \tau_0 s_0^m}{s_0^m} \right), \quad (8.112)$$

where the integers m and n control dilation and translation of the wavelet on a discrete grid; the parameters s_0 and τ_0 are the step parameters for dilation and translation, respectively. (In the equation given above, only the wavelet parameters have been shown with discrete variables; if the input signal has been sampled, the functions x and ψ need to be changed to discrete-time notation and the integral should be changed to summation.) The dyadic grid is obtained by setting $s_0 = 2$ and $\tau_0 = 1$; then, we have

$$\psi_{m,n}(t) = \frac{1}{\sqrt{2^m}} \psi(2^{-m}t - n). \quad (8.113)$$

The use of an appropriate grid for the DWT reduces the redundancy of representation as compared to the CWT. (The critically sampled DWT is a shift-variant transform; see Bradley [80] for related discussions.) The wavelet coefficients $X(m, n)$ obtained as above may be subjected to thresholding or selection procedures and an approximation of the signal may be reconstructed as

$$\tilde{x}(t) = \sum_m \sum_n \tilde{X}(m, n) \psi_{m,n}(t), \quad (8.114)$$

where \sim indicates the result of approximation or filtering and the range of summation depends upon the choices made and the application.

The wavelet approach permits simultaneous analysis of the local characteristics of a given nonstationary signal in both the temporal and spectral domains. The method facilitates the analysis of short-duration high-frequency signal components as well as long-duration low-frequency components. A multicomponent signal may be analyzed using multiple types of wavelets that match the expected characteristics of the various components so as to extract measures related to their properties in separate channels. Several types of wavelets and related procedures have been described in the literature [37, 39, 68–78]; the presentation in this chapter is limited to a few concepts that can be related to the remaining topics in the chapter. The use of appropriate wavelets can assist in the identification and analysis of transient, aperiodic, multicomponent, and nonstationary signals.

The approach of selecting the best basis among a dictionary of bases by minimizing a cost function is known as the method of wavelet packet [81]. The wavelet packet approach uses a family of orthogonal bases that include different types of localized TF functions (also known as TF atoms). The wavelet packet approach decomposes the given signal into TF atoms that are adapted to the TF structures expected or known to be present in the signal. The wavelet packet coefficients may then be used to obtain filtered or approximate versions of the signal; the coefficients may also be used in the form of a compact representation of the signal for further analysis or classification. See Chapter 9 for further details on related topics.

Illustration of application: Some of the commonly used wavelets are the Mexican hat (or sombrero), Morlet, Daubechies, and symmlet functions [37, 39, 68–78]. The Mexican hat function is defined as the negative of the second derivative of a Gaussian, and is given by

$$\psi(t) = (1 - t^2) \exp(-0.5 t^2). \quad (8.115)$$

Figure 8.23 shows the wavelet defined above for three scales of 1, 2, and 4 (top to bottom), scaled according to Equation 8.113. It is evident that, as the scale value increases, the duration or width of the wavelet is increased and its frequency bandwidth is reduced; thus, there exists an inverse relationship between scale and frequency.

The Morlet function is defined as

$$\psi(t) = \frac{1}{\pi^{1/4}} [\exp(j \omega_0 t) - \exp(-0.5 \omega_0^2)] \exp(-0.5 t^2), \quad (8.116)$$

where ω_0 is the central frequency of the frequency response of the mother wavelet. The term $\exp(-0.5 \omega_0^2)$ is negligible for $\omega_0 > 5$ and may be ignored. Figures 8.24 and 8.25 show Morlet wavelets obtained using the real part of Equation 8.116 for three scales of 1, 2, and 4 and $\omega_0 = 2$ and 6 radians/s, respectively; the wavelets have been scaled according to Equation 8.113. The Morlet wavelets in Figure 8.24 are similar in appearance to the Mexican hat wavelets in Figure 8.23.

Figure 8.26 shows a noisy ECG signal over three cardiac cycles, sampled at 200 Hz. The figure also shows the STFT of the ECG signal using a window width of 32 samples, which was moved in

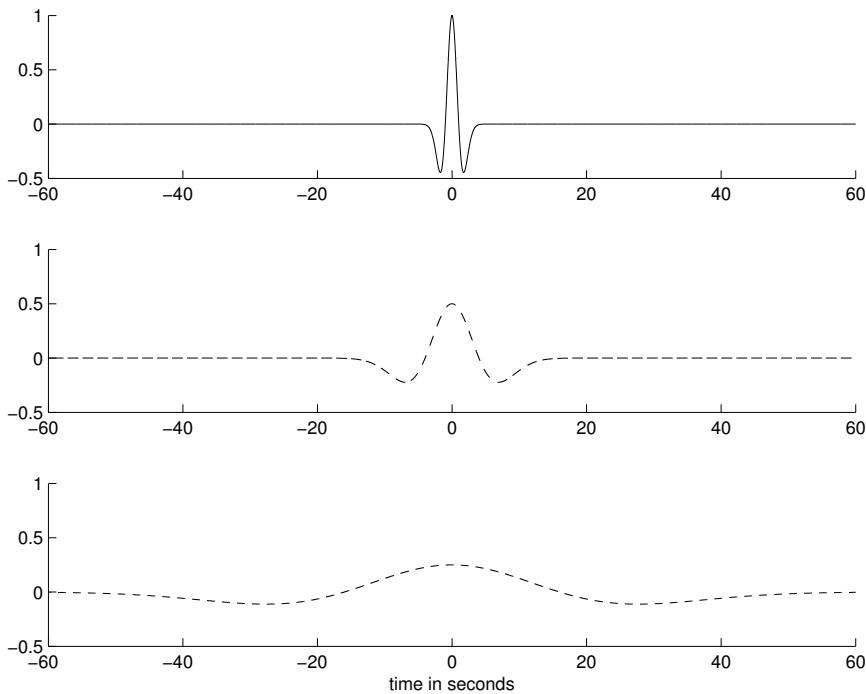


Figure 8.23 Top to bottom: The Mexican hat wavelet in Equation 8.115 for three scales of 1, 2, and 4. The wavelets have been scaled according to Equation 8.113.

steps of 16 samples. It is evident from the STFT that the spectrum of the signal changes substantially over its duration, with increased high-frequency content during the intervals corresponding to the QRS complexes and predominantly low-frequency composition in periods related to the P and T waves.

In order to analyze the ECG signal using wavelets, Morlet wavelets were prepared as per Equation 8.116 (using only the real part) with $\omega_0 = 200 \text{ radians/s}$ and by changing the Gaussian term to $\exp(-5000 t^2)$. The parameters were selected so as to create wavelets that are comparable to typical QRS waves in ECG signals. Figure 8.27 shows the wavelets for four scales of 1, 2, 4, and 8. Figure 8.28 shows the ECG signal (top) and the results of the CWT for the four scales used (labeled correspondingly as wt1, wt2, wt4, and wt8 in the figure). In each case, the wavelet was translated by one sample per step. It is evident that the wavelets of smaller scale result in higher coefficients corresponding to the sharp QRS complexes, whereas the wavelets of larger scale provide larger coefficients at the locations of the slower T waves. One may, if desired, compute the CWT for several values of the scale parameter and compose the result for display as an image, known as the scalogram (see Figure 8.29).

The lowest plot in Figure 8.28 (labeled as wr) shows an approximation of the ECG signal reconstructed using only the coefficients for scales of 2, 4, and 8; the result for the smallest scale of 1 was discarded. To obtain the reconstructed signal, the wavelet transform for each scale was convolved with the corresponding wavelet according to Equation 8.110; however, due to the redundancy of the CWT, a reconstruction may be obtained by simply adding the results of the wavelet transform for the selected scales [77, 78]. The result of approximation is seen to retain the P and QRS waves with most of the small-scale noise removed; however, the T waves have not been reproduced well. This is a simplified illustration of the CWT with only four scales used; in practice, one may have to use a large number of scales depending upon the nature of the signal and the desired result.

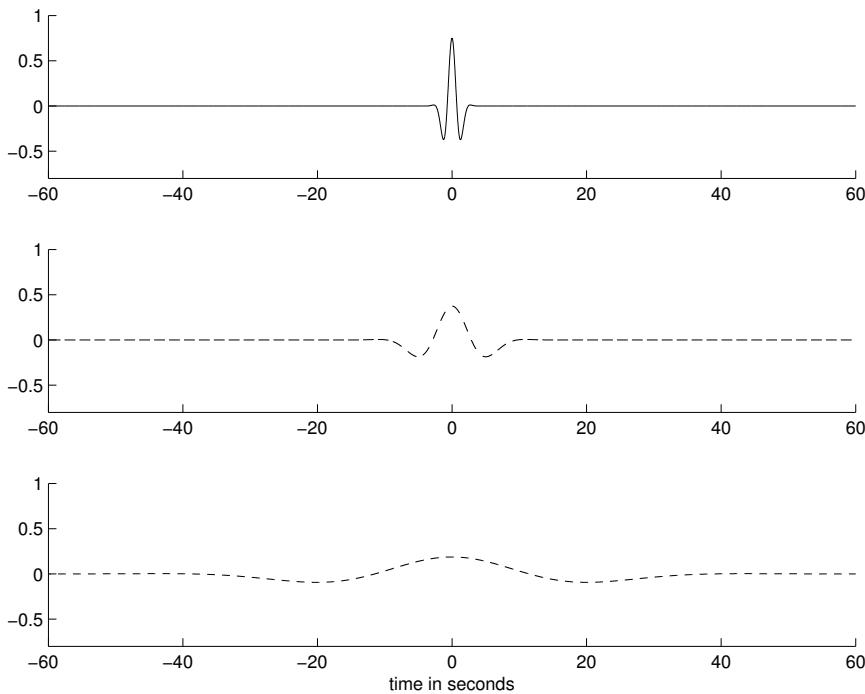


Figure 8.24 Top to bottom: The Morlet wavelet in Equation 8.116 for three scales of 1, 2, and 4, and $\omega_0 = 2 \text{ radians/s}$. Only the real part has been used and the wavelets have been scaled according to Equation 8.113.

Representation and analysis of a given signal, as shown above, may be viewed as an application of matched filters, correlation filters, and multiscale filter banks. The CWT given in Equation 8.107 represents a correlation operation between the given signal and a wavelet. When a symmetric wavelet is used, convolution is the same as correlation. The CWT operation is then comparable to matched filtering (see Section 4.6.1). Thus, the CWT may be interpreted as seeking a match between a collection of wavelets and the given signal: the illustration in Figure 8.28 demonstrates this aspect of wavelet analysis.

Further insight into the interpretation of wavelet analysis as an application of multiscale filter banks can be gained by examining the PSDs of the wavelets used. Figure 8.30 shows the normalized PSDs of the four wavelets in Figure 8.27; it is evident that the central frequency of the frequency response of the filter corresponding to each wavelet shifts towards lower values as the scale is increased. When several wavelets are used over a range of scales, a number of filtered versions of the given signal are obtained, as illustrated in Figure 8.28. The coefficients obtained using the CWT or DWT may also be used for further analysis of signals such as the ECG for various purposes, including detection of waves and classification [69, 70, 76]. See Section 8.17 for a discussion on the application of the DWT to analyze EEG signals.

8.9 Bilinear TFDs

Bilinear TFDs are a form of quadratic TFDs that provide joint TF representation by transforming the time-varying autocorrelation (a second-order function) of the given signal. If $x(t)$ is a signal and TFD(t, ω) is its TFD, the important criteria that the TFD is expected to satisfy are listed in the following paragraphs [37, 39].

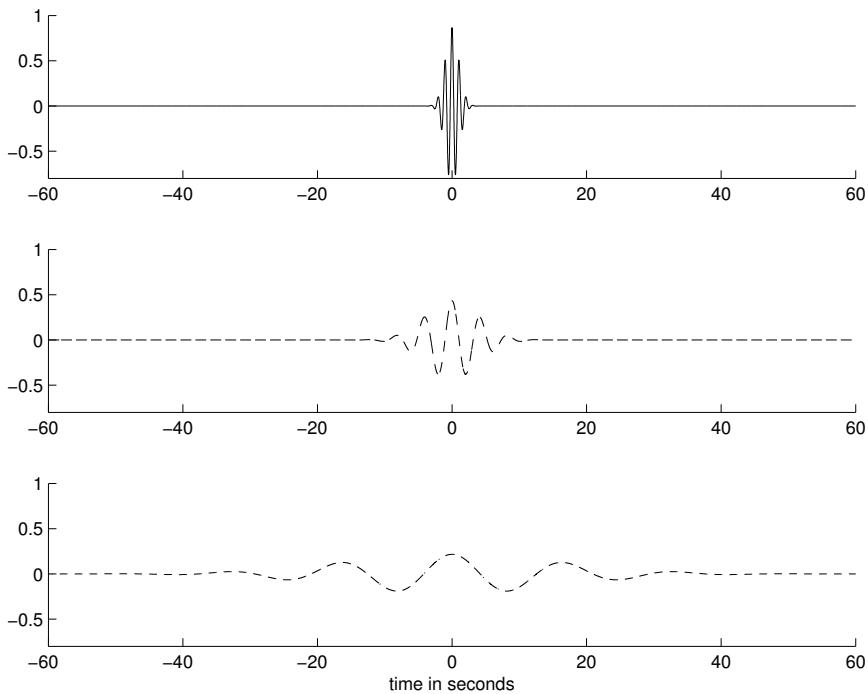


Figure 8.25 Top to bottom: The Morlet wavelet in Equation 8.116 for three scales of 1, 2, and 4, and $\omega_0 = 6 \text{ radians/s}$. Only the real part has been used and the wavelets have been scaled according to Equation 8.113.

Total energy:

$$\int \int \text{TFD}(t, \omega) dt d\omega = \int |x(t)|^2 dt = \int |X(\omega)|^2 d\omega, \quad (8.117)$$

where $X(\omega)$ is the Fourier transform of $x(t)$. This criterion indicates that, at particular values of t and ω , $\text{TFD}(t, \omega)$ gives the fractional energy of the signal; thus, $\text{TFD}(t, \omega)$ may be viewed as a joint TF density function.

Invariance: The TFD should be invariant to any linear shift in time and frequency; it is also expected that the TFD is scale-invariant.

Nonnegativity: $\text{TFD}(t, \omega) \geq 0$ for all t and ω . This criterion helps in treating the TFD as a density function.

Marginals: The instantaneous energy of the signal is given by

$$\int \text{TFD}(t, \omega) d\omega = |x(t)|^2. \quad (8.118)$$

Integration along the time axis gives the PSD of the signal as

$$\int \text{TFD}(t, \omega) dt = |X(\omega)|^2. \quad (8.119)$$

Expectation values:

$$E[g(t, \omega)] = \int \int g(t, \omega) \text{TFD}(t, \omega) dt d\omega. \quad (8.120)$$

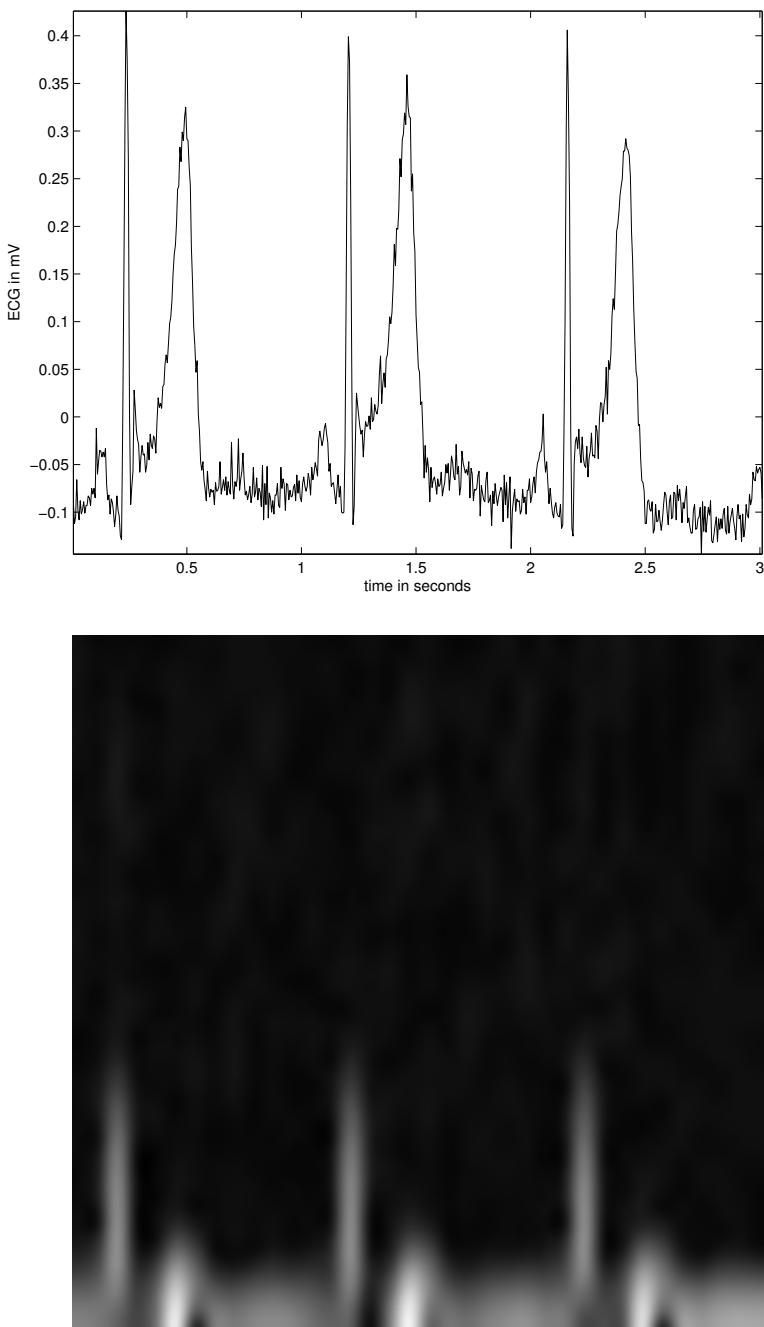


Figure 8.26 A noisy ECG signal over three cardiac cycles or beats (top), and its spectrogram (STFT) (bottom). The horizontal axis represents the time axis from 0 to 3 s (left to right) and the vertical axis represents frequency from 0 to 100 Hz (bottom to top).

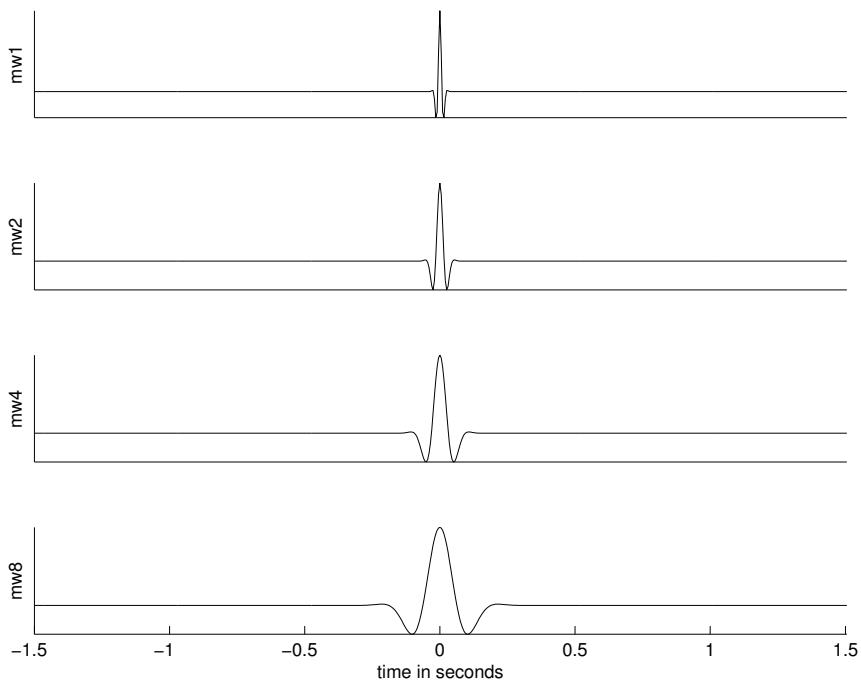


Figure 8.27 Morlet wavelets to analyze the ECG signal in Figure 8.26. The wavelets were prepared using Equation 8.116 with $\omega_0 = 200$ radians/s and by changing the Gaussian term to $\exp(-5000t^2)$, and are shown for four scales of 1, 2, 4, and 8. The amplitude details have been suppressed and all signals have been mapped to the full range of the ordinate available for improved display.

The equation given above expresses a generalized moment of a TFD. The function $g(t, \omega)$ needs to be chosen according to the desired moment. The instantaneous mean frequency (IMF) is given by the time-varying first moment of the TFD along the frequency axis as

$$E_t[\omega] = \frac{1}{|x(t)|^2} \int \omega \text{TFD}(t, \omega) d\omega. \quad (8.121)$$

The group delay is given by the frequency-varying first moment of the TFD along the time axis as

$$E_\omega[t] = \frac{1}{|X(\omega)|^2} \int t \text{TFD}(t, \omega) dt. \quad (8.122)$$

See Boashash [2] for discussions on analysis of the instantaneous frequency of nonstationary signals.

Finite support: If $x(t)$ is zero at $t = t_1$, then $\text{TFD}(t_1, \omega)$ should be zero. Similarly, if $X(\omega)$ is zero at $\omega = \omega_1$, then $\text{TFD}(t, \omega_1)$ should be zero.

The simplest TFD is the spectrogram, which is obtained by taking the squared magnitude of the STFT defined in Equation 8.9.

The Wigner–Ville distribution (WVD) of a signal $x(t)$ is given by [37]

$$\text{WVD}(t, \omega) = \int_{-\infty}^{\infty} x(t + \tau/2) x^*(t - \tau/2) \exp(-j\omega\tau) d\tau. \quad (8.123)$$

A drawback of the WVD is that, in the case of multicomponent signals, cross-terms (referred to as “ghost frequencies”) are generated in the TFD that could lead to misinterpretation.

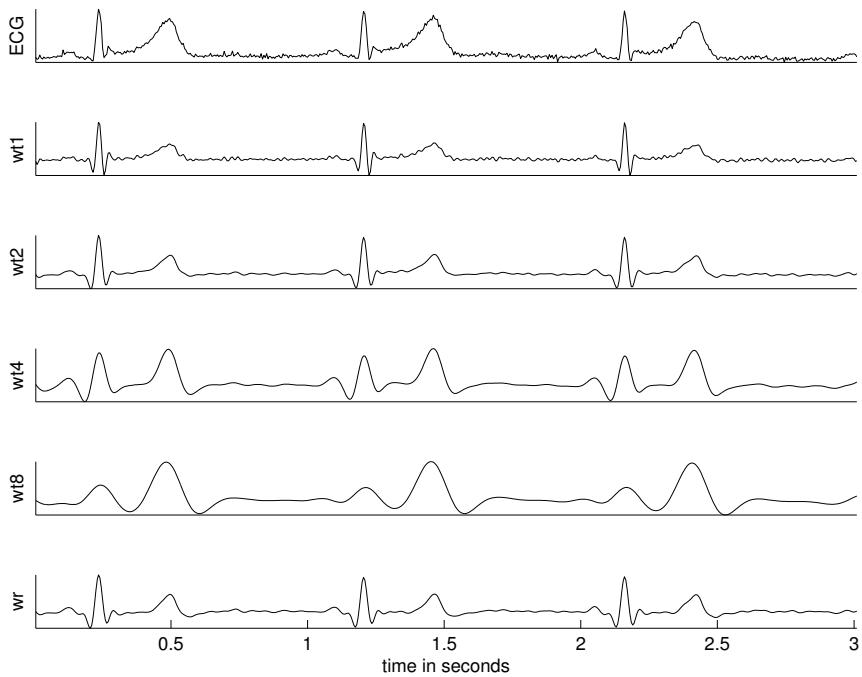


Figure 8.28 Results of analysis of the ECG signal in Figure 8.26 using the wavelets shown in Figure 8.27. Top to bottom: the original ECG signal; results of the CWT for the four scales used (wt1, wt2, wt4, and wt8); and approximation of the signal (wr) reconstructed using only wt2, wt4, and wt8. The amplitude details have been suppressed and all signals have been mapped to the full range of the ordinate available for improved display.

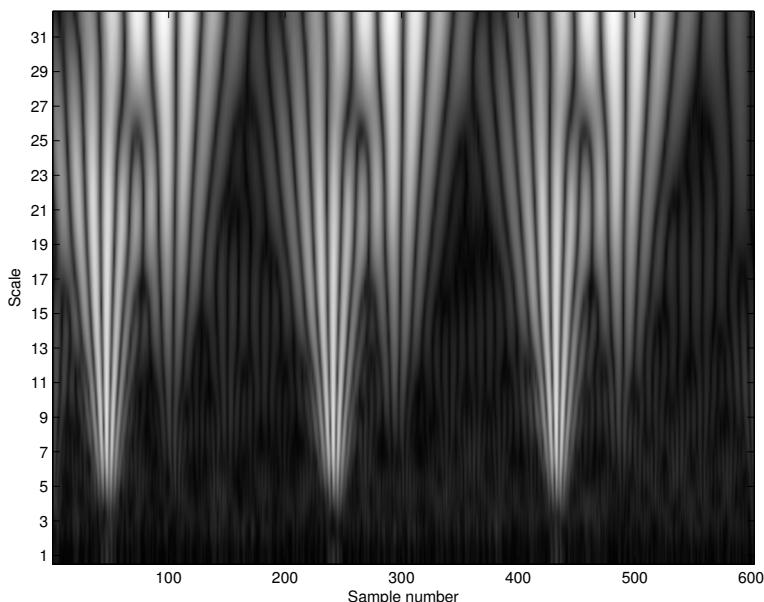


Figure 8.29 Scalogram resulting from the CWT of the ECG signal in Figure 8.26 using Morlet wavelets. The horizontal axis represents the time axis from 0 to 3 s sampled at 200 Hz and the vertical axis represents scale from 1 to 32.

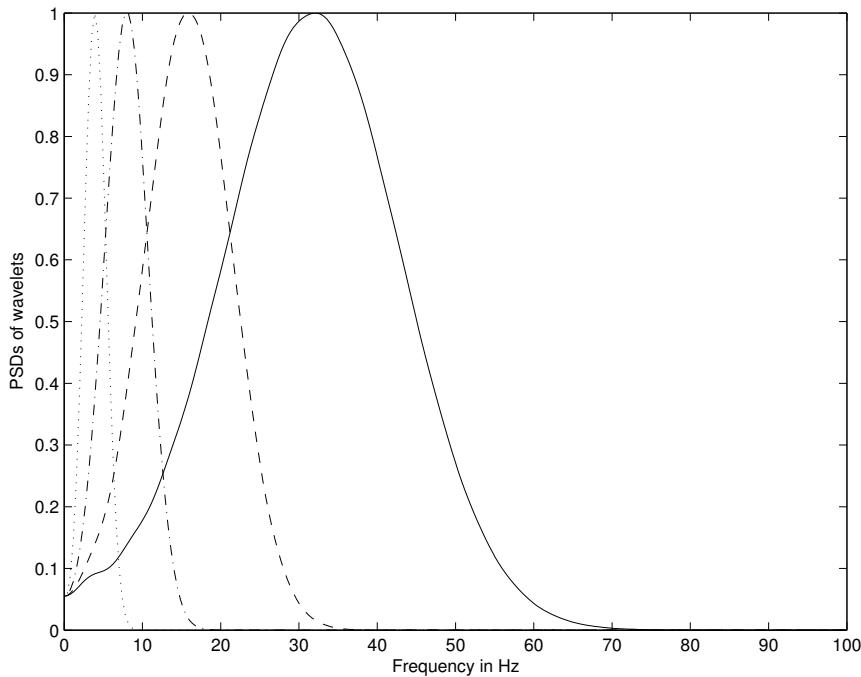


Figure 8.30 Normalized PSDs of the four Morlet wavelets in Figure 8.27. The shift towards lower values in the central frequency of the filter corresponding to each wavelet is evident as the scale is increased from 1 (solid line), to 2 (dashed line), 4 (dot-dash line), and 8 (dotted line).

The Cohen's class of generalized TFD (GTFD) is given by

$$\begin{aligned} \text{GTFD}(t, \omega) = & \frac{1}{4\pi^2} \int \int \int x(u + \tau/2) x^*(u - \tau/2) \Phi(\theta, \tau) \\ & \times \exp(-j\theta t - j\tau\omega + j\theta u) du d\tau d\theta, \end{aligned} \quad (8.124)$$

where $\Phi(\theta, \tau)$ is the TF transformation kernel. The kernel acts as a lowpass filter and minimizes cross-terms. The selection of the kernel, which may be signal-independent or signal-dependent, determines the properties of the resulting distribution [36, 37].

Considering a signal-independent kernel, cross-terms may be minimized by 2D filtering of the WVD; the smoothed version of the WVD (SWVD) can be expressed as

$$\text{SWVD}(t, \omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(u, \Omega) \text{WVD}(t-u, \omega-\Omega) du d\Omega, \quad (8.125)$$

where $\phi(u, \Omega)$ is a kernel function with lowpass-filter characteristics. The kernel function may be expressed as

$$\phi(u, \Omega) = g(u)H(\Omega), \quad (8.126)$$

where $g(u)$ and $H(\Omega)$ are smoothing functions or windows in the time and frequency domains, respectively. The smoothing windows are independent of the signal; Gaussian functions are commonly used as smoothing windows. The resulting TFD, known as the smoothed pseudo-WVD

(SPWVD) [82] is expressed as

$$\text{SPWVD}(t, \omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u) H(\Omega) \text{WVD}(t-u, \omega-\Omega) du d\Omega. \quad (8.127)$$

Smoothing windows suppress cross-terms but smear localized components, leading to less accurate TF localization of signal components as compared to the WVD. A reassignment method was proposed by Auger and Flandrin [83] to improve TF localization in smoothed TFDs, such as SPWVDs. In the reassignment method, the window is moved from a point (t, ω) to the center of energy, $(\bar{t}, \bar{\omega})$, of the TFD. The center of energy is given by the group delay and the mean frequency computed as the mean values of the WVD:

$$\bar{t}(t, \omega) = t - \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u g(u) H(\Omega) \text{WVD}(t-u, \omega-\Omega) du d\Omega}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u) H(\Omega) \text{WVD}(t-u, \omega-\Omega) du d\Omega}, \quad (8.128)$$

$$\bar{\omega}(t, \omega) = \omega - \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Omega g(u) H(\Omega) \text{WVD}(t-u, \omega-\Omega) du d\Omega}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u) H(\Omega) \text{WVD}(t-u, \omega-\Omega) du d\Omega}. \quad (8.129)$$

The modified TFD, known as the reassigned SPWVD (or RSPWVD), can be expressed as

$$\begin{aligned} \text{RSPWVD}(t', \omega') &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \text{SPWVD}(t, \omega) \\ &\times \delta[t' - \bar{t}(t, \omega)] \delta[\omega' - \bar{\omega}(t, \omega)] dt d\omega. \end{aligned} \quad (8.130)$$

Bilinear TFDs are termed bilinear because they provide both time and frequency shift-invariance. The invariance property helps in feature extraction applications that are useful in pattern classification and machine learning applications. Due to the presence of cross-terms and the possible loss of TF resolution due to smoothing operations in the time and frequency domains, bilinear TFDs are not readily suitable for pattern analysis applications. Alternative methods need to be developed to overcome the cross-term and TF resolution trade-off; this could be achieved by analyzing the interaction between individual signal components through the signal decomposition frameworks described in Section 9.3.

8.10 Application: Adaptive Segmentation of EEG Signals

Problem: Propose a method for parametric representation of nonstationary EEG signals.

Solution: Bodenstein and Praetorius [44] applied their adaptive segmentation procedure based upon the SEM (see Section 8.5.1) for representation and analysis of EEG signals with the following propositions.

1. An EEG signal consists of quasistationary segments upon which transients may be superimposed.
2. A segment is specified by its time of occurrence, duration, and PSD (represented by its AR-model coefficients). A transient is specified by its time of occurrence and a set of graph elements (or directly by its samples).

3. An EEG signal consists of a finite number of recurrent states.

It should be noted that, whereas the adaptive segments have variable length, each adaptive segment is represented by the same number of AR-model coefficients. The number of parameters is, therefore, independent of segment duration, which is convenient when pattern classification techniques are applied to the segments. Since the AR model is computed once at the beginning of each segment and a limited amount of prediction error is permitted in the moving analysis window, the initial AR model may not adequately represent the entire adaptive segment. A new model may be computed using the signal samples over the entire duration of each adaptive segment. Instead, Bodenstein and Praetorius maintained the initial AR model of order P of each adaptive segment, and an additional *corrective predictor* of order M was derived for each adaptive segment using the ACF of the prediction error which is computed and readily available in the segmentation procedure. Each adaptive segment was then represented by the $(P + M)$ AR-model coefficients, the associated prediction error RMS values, and the segment length. The PSD of the segment may be derived from the two sets of AR-model coefficients.

With the EEG signals bandpass filtered to the range $1 - 25\text{ Hz}$ and sampled at 50 Hz in the work of Bodenstein and Praetorius [44], the ACF window length was set to be 2 s with $2N + 1 = 101$ samples. Bodenstein and Praetorius used the rule of thumb that the AR-model order should be at least twice the number of expected resonances in the PSD of the signal. Short segments of EEG signals rarely demonstrate more than two spectral peaks, which suggests that an AR-model order of $P = 5$ should be adequate. Regardless, Bodenstein and Praetorius used $P = 8$, which met the Akaike information criterion as well (see Section 7.5.2). The order of the ACF of the prediction error and the associated corrective predictor was set to a low value of $M = 3$, allowing for one spectral peak (the error should ideally have a uniform PSD). The thresholds were defined as $Th_1 = 0.5$ (empirical) and $Th_2 = 2.5\sigma$, where σ is the RMS value of the prediction error (see Section 8.5.1). The range of $20\sigma^2$ to $40\sigma^2$ was recommended for Th_3 . A transitional delay of 25 samples was allowed between each segmentation boundary and the starting point of the following fixed window to prevent the inclusion of the spectral components of one segment into the following segment.

Figure 8.31 shows a few examples of adaptive segmentation of EEG signals. A clustering procedure was included to remove spurious boundaries, some examples of which may be seen in Figure 8.31 (d); neighboring segments with similar parameters were merged in a subsequent step. Visual inspection of the results indicates that most of the adaptive segments are stationary (that is, they have the same appearance) over their durations. It is worth noting that the longest segment in Figure 8.31 (d) of duration 16 s or 800 samples is represented by just 12 parameters.

Figure 8.32 shows examples of detection of transients in two contralateral channels of the EEG of a patient with epilepsy. The EEG signal between seizures (interictal periods) is expected to exhibit a large number of sharp waves. The length of the arrows shown in the figure was made proportional to the cumulated suprathreshold part of the squared prediction error in order to indicate how pronounced the event was regarded to be by the algorithm.

The method was further extended to parallel analysis of multichannel EEG signals by Bodenstein et al. [47] and Creutzfeldt et al. [48]. Procedures were proposed for computerized pattern classification and labeling of EEG signals, including clustering of similar segments and state diagrams indicating the sequence of the types of activity found in an EEG record. Figure 8.33 illustrates the record produced by the application of the procedure to two channels of an EEG signal. Typical EEG segments belonging to the four clusters detected in the signal are shown on the left and right sides of the upper portion of the figure. Each signal segment is labeled with the frequencies (FRQ, in Hz) and amplitudes (AMP, in μV) of the resonances detected using an eighth-order AR model. The central column of the upper portion of the figure illustrates the PSDs of the corresponding segments on the left (solid line) and right (dashed line). The middle portion of the figure provides the state diagram, indicating the transitions between the four states (represented by the four clusters of the EEG segments) detected in the two channels of the signal. The states represent: 1, background; 2,

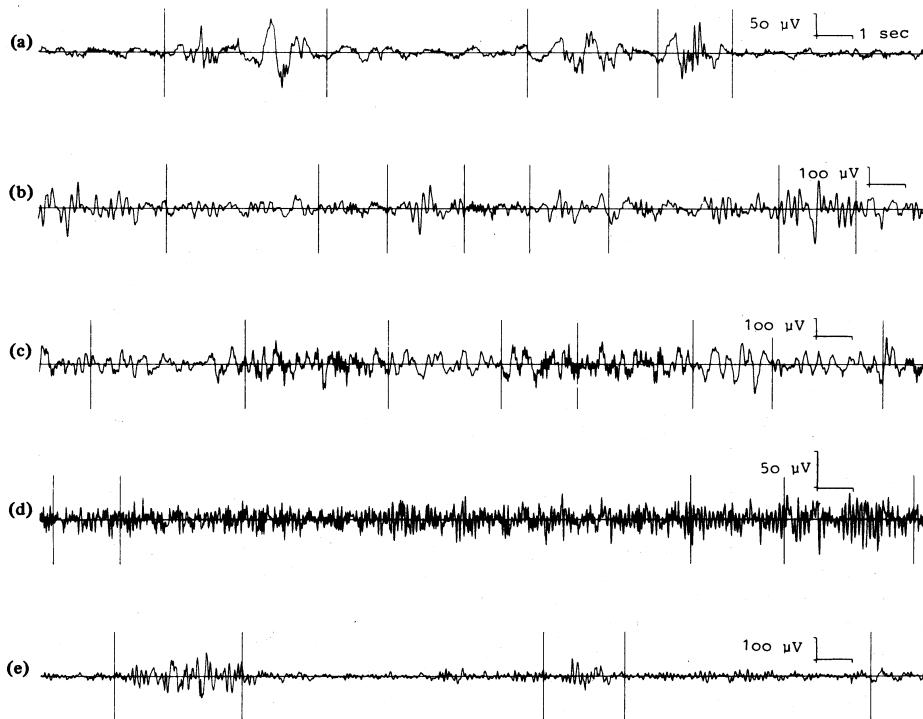


Figure 8.31 Examples of segmentation of EEG signals. (a) Newborn in non-REM sleep. REM = rapid eye movement. (b) Child of age 7 years in sleep stage I. (c) Child of age 8 years in sleep stage III. (d) Alpha rhythm of an adult. (e) EEG of an adult with paroxysms. The vertical lines represent the adaptive segmentation boundaries. Reproduced with permission from G. Bodenstein and H.M. Praetorius, Feature extraction from the electroencephalogram by adaptive segmentation, *Proceedings of the IEEE*, 65(5):642–652, 1977. ©IEEE.

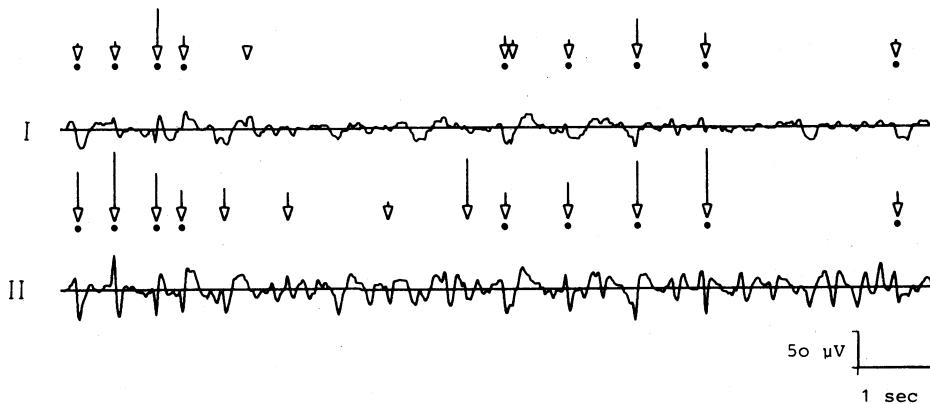


Figure 8.32 Example of detection of transients in the EEG signal of a patient with epilepsy. The signals shown are from contralateral channels between seizures (interictal period). The longer the arrow, the more pronounced is the transient detected at the corresponding time instant. Transients detected simultaneously in the two contralateral channels are marked with dots. Reproduced with permission from G. Bodenstein and H.M. Praetorius, Feature extraction from the electroencephalogram by adaptive segmentation, *Proceedings of the IEEE*, 65(5):642–652, 1977. ©IEEE.

eyes open; 3, paroxysm; and 4, epileptiform spike-and-wave complexes. The values on the right of the state diagram give the percentage of the total duration of the signal for which the EEG was in the corresponding states. The bottom portion of the figure illustrates singular events, that is, segments that could not be grouped with any of the four clusters. It was indicated that the segments of most EEG signals could be clustered into at most five states, and that the summarized record as illustrated in Figure 8.33 could assist clinicians in analyzing lengthy EEG records in an efficient manner.

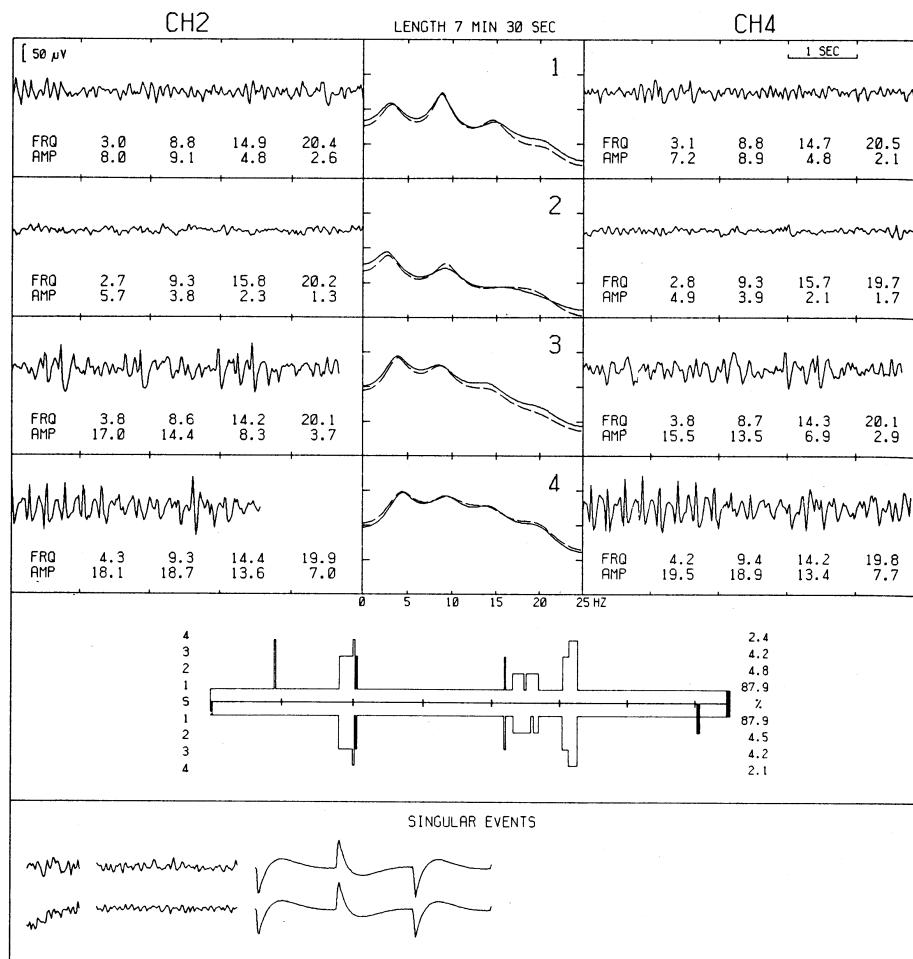


Figure 8.33 Example of application of segmentation and pattern analysis to the EEG signal of a patient with epileptiform activity. Refer to the text for details. Reproduced with permission from G. Bodenstein, W. Schneider, and C.V.D. Malsburg, Computerized EEG pattern classification by adaptive segmentation and probability-density-function classification. Description of the method, *Computers in Biology and Medicine*, 15(5):297–313, 1985. ©Elsevier Science.

Time-varying AR-modeling techniques have also been applied for the analysis of EEG signals by Gath et al. [28]. Time-varying ARMA modeling techniques were applied to analyze EEG signals by Chen et al. [29].

8.11 Application: Adaptive Segmentation of PCG Signals

We have noted several times that the PCG signal is nonstationary. Let us now assess the feasibility of adaptive segmentation of PCG signals using the RLSL method (see Section 8.6.2), with no other signal being used as a reference.

Figure 8.34 illustrates the results of segmentation of the PCG signal of a normal subject. The top trace shows the PCG signal over three cardiac cycles; the segment boundaries detected are indicated by the vertical dotted lines as well as by the triangular markers on the time axis. The second trace illustrates a plot of the conversion factor γ_c : The conversion factor drops from unity whenever there is a change in the signal's characteristics, in particular at the boundaries of S1 and S2. A threshold of 0.995 (indicated by the horizontal line overlaid on the second trace) applied to γ_c and a condition imposing a minimum segment length of 50 samples (50 ms) were used to obtain the segment boundaries. The third and fourth traces illustrate the ECG and carotid pulse signals of the subject acquired simultaneously with the PCG. The segment boundaries obtained by the RLSL method agree well with the readily noticeable S1 and S2 boundaries as well as the QRS and dicrotic notch positions. (See also Sections 1.2.9, 2.3, and 4.9 for related details.)

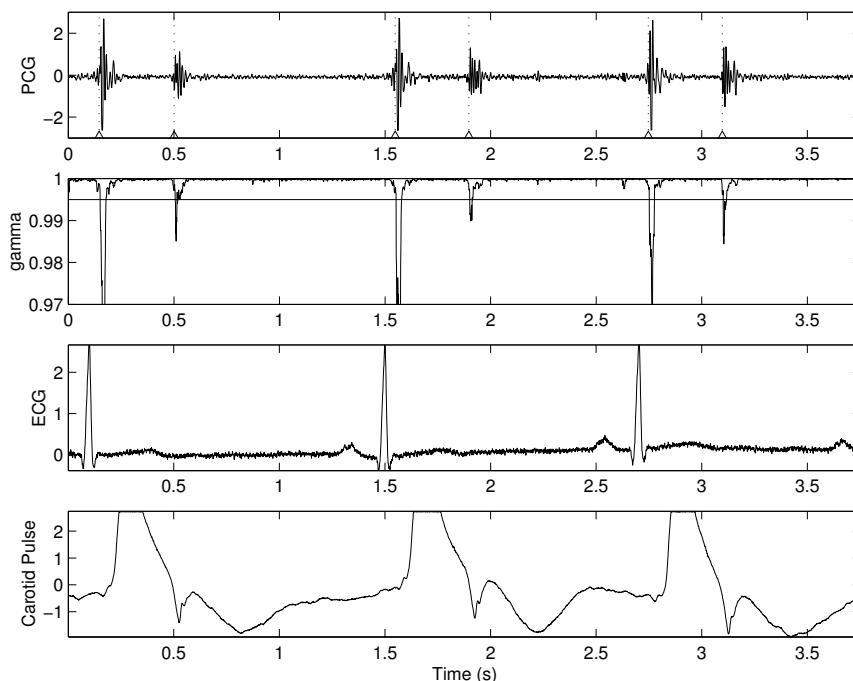


Figure 8.34 Adaptive segmentation of the PCG signal of a normal subject using the RLSL method. Top to bottom: PCG signal (the vertical dotted lines and triangular markers represent the segmentation boundaries); conversion factor γ_c (the horizontal line is the threshold used); ECG; carotid pulse (clipped due to saturation).

Figure 8.35 illustrates the results of adaptive segmentation of the PCG signal of a subject with systolic murmur due to aortic stenosis. The results in this case, however, are not as clear or as easy to interpret as in the preceding case. The method has identified the beginning of S1 and S2; furthermore, the split nature of S2 has been identified by an additional segment boundary within each S2. However, the method has not reliably identified the boundaries between the episodes of S1 and systolic murmur illustrated: The condition on the minimum segment length has affected the placement of the segment boundary after the beginning of S1. Use of other conditions on γ_c may provide better segmentation results.

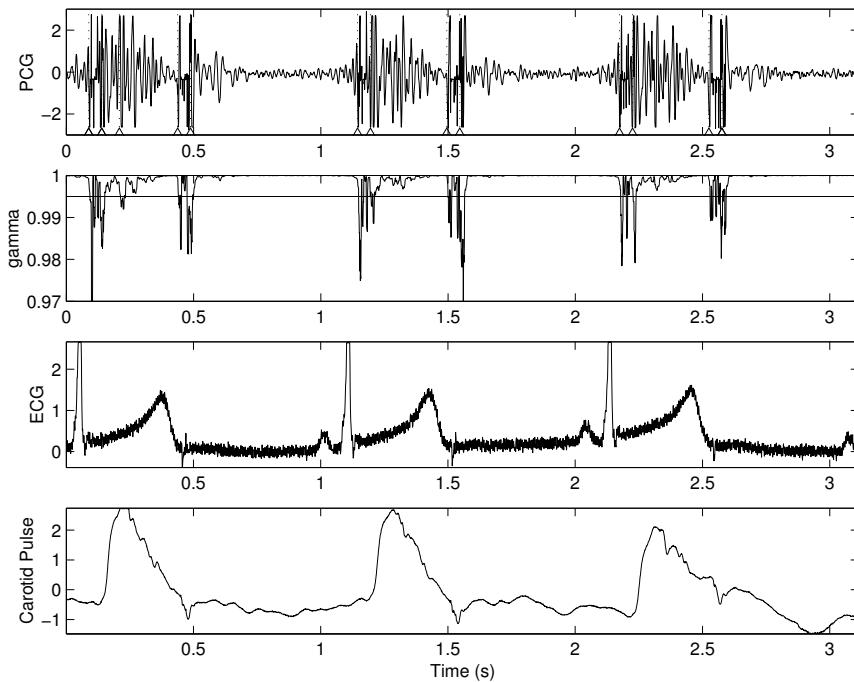


Figure 8.35 Adaptive segmentation of the PCG signal of a subject (female, 11 years) with systolic murmur due to aortic stenosis. Top to bottom: PCG signal (the vertical lines and triangular markers represent the segmentation boundaries); conversion factor γ_c (the horizontal line is the threshold used); ECG; carotid pulse.

8.12 Application: Time-varying Analysis of HRV

The heart rate is controlled by the ANS and the CNS: The vagal and sympathetic activities lead to a decrease or increase, respectively, in the heart rate (see Section 1.2.5). We saw in Sections 2.2.4 and 7.9 how respiration affects heart rate and how Fourier analysis may be extended to analyze HRV. When heart rate data such as beat-to-beat *RR* intervals are collected over long periods of time (several hours), the signal could be expected to be nonstationary.

Bianchi et al. [31] extended AR-modeling techniques for time-variant PSD analysis of HRV data in order to study transient episodes related to ischemic attacks. The prediction error was weighted with a forgetting factor, and a time-varying AR model was derived. The RLS algorithm was used to update the AR-model coefficients at every *RR* interval sample (every cardiac cycle). The AR coefficients were then used to compute a time-varying PSD. The following frequency bands were indicated to be of interest in the analysis of *RR* interval PSDs: very-low-frequency (VLF) band in the range $0 - 0.03 \text{ Hz}$ related to humoral and thermoregulatory factors; low-frequency (LF) band in the range $0.03 - 0.15 \text{ Hz}$ related to sympathetic activity; and high-frequency (HF) band in the range $0.18 - 0.4 \text{ Hz}$ related to respiration and vagal activity.

(*Note:* According to the standards of measurement, physiological interpretation, and clinical use of HRV published by the Task Force of the European Society of Cardiology and the North American Society of Pacing Electrophysiology [84], the frequency bands of HRV are slightly different from those given above, and are: VLF, $\leq 0.04 \text{ Hz}$; LF, $0.04 - 0.15 \text{ Hz}$; and HF, $0.15 - 0.4 \text{ Hz}$.)

Figure 8.36 shows an *RR* interval series including an ischemic episode (delineated by B for beginning and E for ending points, respectively). Figure 8.37 shows the time-varying PSD in the form of a spectrogram. Figure 8.38 shows a segment of *RR* interval data and a few measures derived from the data.

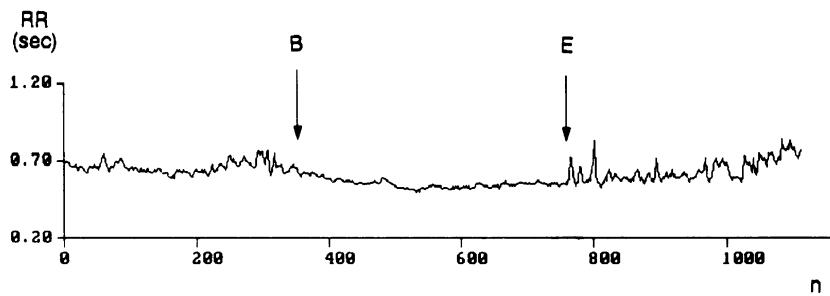


Figure 8.36 *RR* interval series including an ischemic episode. B denotes the beginning and E denotes the end of the episode. Reproduced with permission from A.M. Bianchi, L. Mainardi, E. Petrucci, M.G. Signorini, M. Mainardi, and S. Cerutti, Time-variant power spectrum analysis for the detection of transient episodes in HRV signal, *IEEE Transactions on Biomedical Engineering*, 40(2):136–144, 1993. ©IEEE.

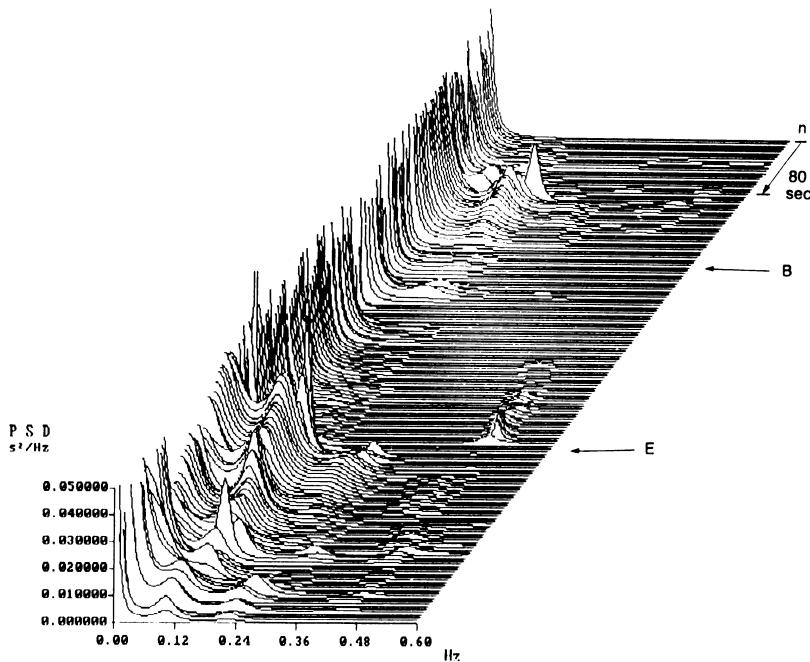


Figure 8.37 Spectrogram of the *RR* interval series in Figure 8.36. Time progresses from the top to the bottom. B denotes the beginning and E denotes the end of an ischemic episode. Reproduced with permission from A.M. Bianchi, L. Mainardi, E. Petrucci, M.G. Signorini, M. Mainardi, and S. Cerutti, Time-variant power spectrum analysis for the detection of transient episodes in HRV signal, *IEEE Transactions on Biomedical Engineering*, 40(2):136–144, 1993. ©IEEE.

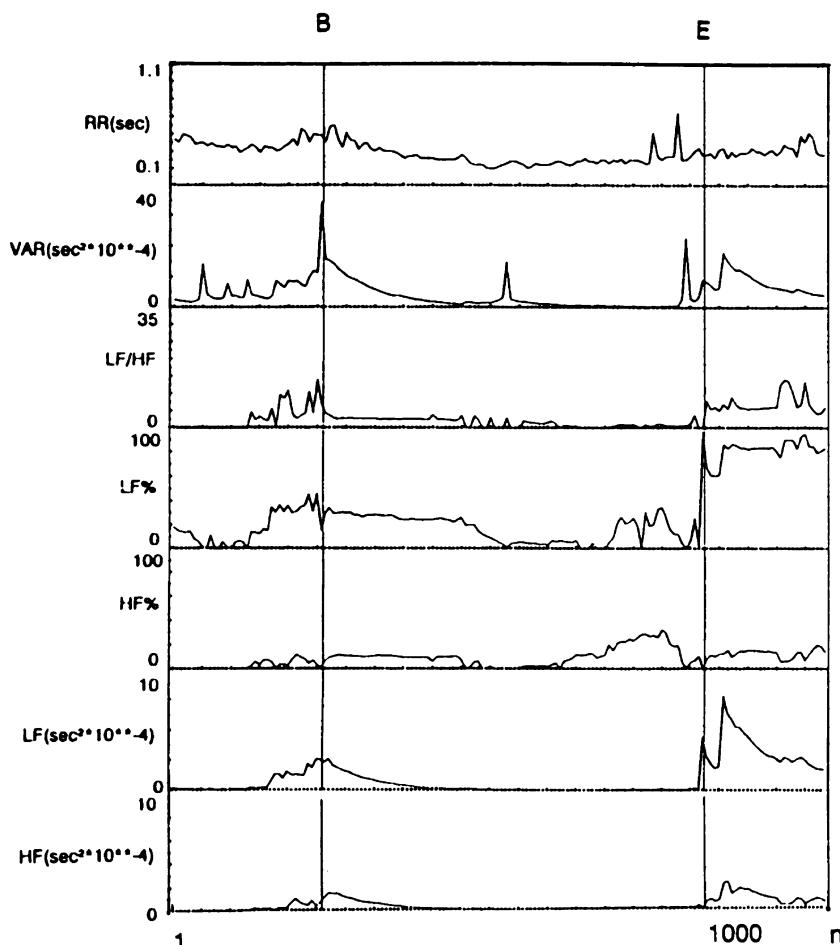


Figure 8.38 Top to bottom: RR interval series including an ischemic episode; variance; low-frequency (LF) to high-frequency (HF) power ratio; percentage of LF power; percentage of HF power; LF power; and HF power. B denotes the beginning and E denotes the end of the episode. Reproduced with permission from A.M. Bianchi, L. Mainardi, E. Petrucci, M.G. Signorini, M. Mainardi, and S. Cerutti, Time-variant power spectrum analysis for the detection of transient episodes in HRV signal, *IEEE Transactions on Biomedical Engineering*, 40(2):136–144, 1993. ©IEEE.

Some of the important observations made by Bianchi et al. (and illustrated by the spectrogram in Figure 8.37 and the parameters in Figure 8.38) are as follows:

- There is an increase in LF power about 1.5 – 2 minutes before an ischemic event.
- The RR variance decreases as an episode begins.
- There is a predominant rise in LF power at the end of an ischemic episode.
- A small HF component appears toward the end of an episode.
- Early activation of an LF component precedes tachycardia and ST displacement in the ECG that are generally indicative of the onset of an ischemic episode.
- The results suggest an arousal of the sympathetic system before an acute ischemic attack.

Bianchi et al. concluded that their methods could provide quantitative parameters that describe temporal modifications of the signal generation model.

8.13 Application: Analysis of Crying Sounds of Infants

Várallyay [85] developed techniques to analyze time-varying spectral patterns in crying sounds of infants. It was observed that during a crying episode, the volume, pitch, and tone of the signal vary over time. The varying pattern of the pitch (or fundamental frequency) was referred to as the *cry melody*. Just as the melody of normal speech expresses the mood and intention of a speaker, the melody of crying is expected to relate to the physical and physiological state of the infant. Cry melodies are expected to possess characteristic patterns for various states, such as pain, hunger, discomfort, and boredom.

The spectral structure of a cry segment is expected to be regular, containing a fundamental frequency and its harmonics. Dynamic variations of the fundamental frequency characterize the cry melody. In order to follow the variations of the fundamental frequency over time, Várallyay [85] divided cry segments into short-time windows of duration 40 ms. The fundamental frequencies of several infant cries that were analyzed were found to vary between 200 and 1,000 Hz, with the most common values being between 300 and 600 Hz. In order to follow variations of the fundamental frequency, three fundamental units of cry melody patterns were defined as falling (-1), flat (0), and rising (+1). Cry melodies were coded as sequences of these fundamental patterns.

Figure 8.39 shows the temporal and spectral variations in a cry signal. The topmost trace shows the time-varying amplitude of five cry segments. The second trace shows the spectrogram, in which the time-varying patterns of the fundamental frequency and its harmonics (that is, the cry melody) are clearly seen. The two traces at the bottom of the figure show the cry melody as a function of time; the lowest trace has five guidelines superimposed on the cry melody to facilitate its interpretation over predefined frequency ranges.

Várallyay [85] analyzed 2,762 crying sounds recorded from 316 infants. The mean value of the fundamental frequency of the melodies was higher for infants with hearing impairment (425.51 ± 78.10 Hz) than for the control group (408.74 ± 64.77 Hz). Hirschberg [86, 87] observed high-pitched cry melodies with fundamental frequencies in the range of 1,000 to 2,000 Hz in the case of infants with dysphonia, as compared to the range of 400 to 500 Hz for cries of healthy infants. It was noted that, whereas a person with normal hearing can control the sound or speech produced, a person with hearing impairment is unable to exercise such control due to the lack of auditory feedback.

8.14 Application: Wavelet Denoising of PPG Signals

The PPG signal is commonly used for the estimation of oxygen saturation and pulse rate, which in turn could be used for the estimation of heart rate, HRV, BP, and respiration rate. One of the common problems encountered with PPG is motion artifact. Given the temporal and spectral overlap of the signal of interest and the motion artifact signal, fixed filtering approaches will not be optimal. The adaptive filtering technique needs a reference channel consisting of a clean version of the signal, which would be difficult to obtain. A possible alternative would be to use a wavelet-based denoising technique. Signal denoising using wavelets has shown success with a wide variety of biomedical applications [88–90]. An example presented here is the work of Raghuram et al. [91], in which the wavelet transform was applied to denoise PPG signals. As shown in Figure 8.40, five different types of wavelets: Daubechies, biorthogonal, reverse biorthogonal, symlet, and Coiflet were applied for denoising, and it was found that the Daubechies wavelets provided the best denoising result. The performance of the denoising procedure was assessed via estimation of SpO_2 values from the

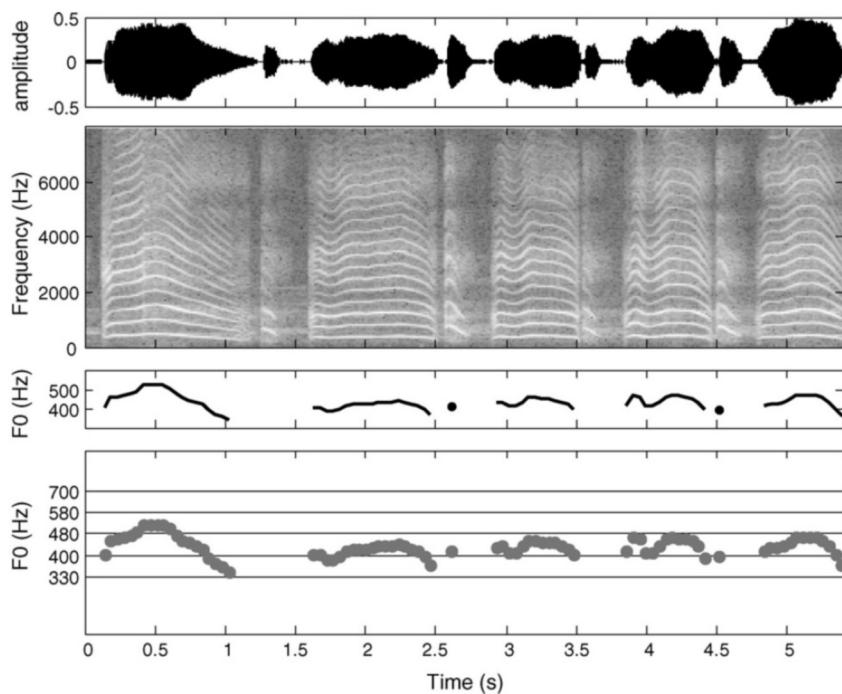


Figure 8.39 Top to bottom: Time-varying amplitude of five cry segments; spectrogram; cry melody as a function of time; the previous item with five guidelines superimposed. Reprinted with permission from G. Várallyay Jr., The melody of crying, *International Journal of Pediatric Otorhinolaryngology*, 71(11):1699–1708, 2007. ©Elsevier.

PPG signals after removal of the motion artifact; all of the wavelets were found to perform well in this regard. However, it was observed that, whereas the respiratory activity signal in the 0.2 – 0.35 Hz band gets modified or removed by most denoising methods, filtering with the Daubechies wavelet preserved the respiratory signal component. Although wavelet techniques can provide good denoising results, care should be exercised in selecting the appropriate type of wavelet for a given application.

8.15 Application: Wavelet Analysis for CPR Studies

During cardiac arrest, the heart cannot pump blood to the lungs, brain, and the rest of the body. Without timely intervention, death could occur in minutes. Although defibrillation with an electric shock could lead to return of spontaneous circulation (ROSC), its success is conditional on a variety of factors, including the time delay between the onset of ventricular fibrillation and defibrillation. Cardiopulmonary resuscitation (CPR) is a life-saving procedure used to restore and sustain circulation and respiration in a person who has gone into cardiac arrest. CPR simulates the heart's pumping action with chest compression. Compression of the chest aids in the circulation of blood throughout the body. Timely CPR can double or triple the chances of survival following cardiac arrest. Additionally, studies indicate that performing CPR prior to giving a defibrillating shock improves survival rates, especially when ventricular fibrillation is left untreated for more than five minutes.

Due to the short time available to treat ventricular fibrillation and improve survival prospects, it is critical to select the appropriate therapeutic procedure, which includes shock parameters, CPR first or shock first, and medication delivery. To assist in making this decision, it would be advantageous

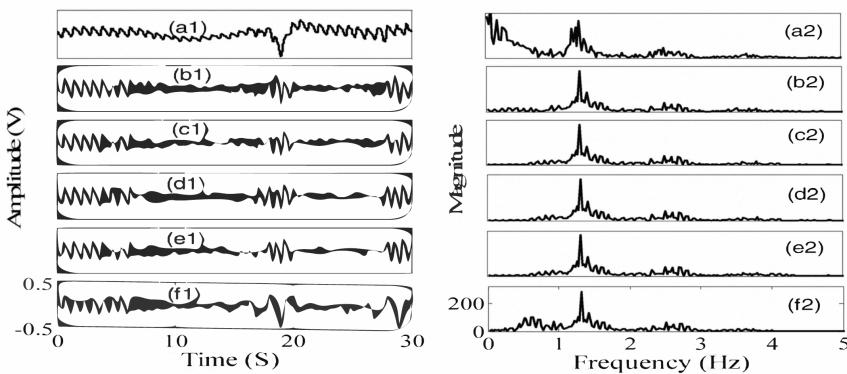


Figure 8.40 (al) Plot of a PPG signal with motion artifact. Plots of wavelet denoising of the PPG signal in (a1) using: (bl) Daubechies wavelet, (cl) biorthogonal wavelet, (dl) reverse biorthogonal wavelet, (el) symlet wavelet, and (fl) Coiflet wavelet. The corresponding spectral plots are shown in (a2)-(f2). Reproduced with permission from M. Raghuram, K.V. Madhav, E.H. Krishna, and K.A. Reddy, Evaluation of wavelets for reduction of motion artifacts in photoplethysmographic signals. In Proceedings of the 10th International Conference on Information Science, Signal Processing and their Applications (pp 460-463), 2010. ©IEEE.

for emergency medical staff if objective measures on the characteristics of the ventricular fibrillation waveform could be informed in near-real time to guide the shock treatment and analyze its outcome. While the majority of available prediction techniques are based on temporal or spectral features, an optimal strategy to study the nonstationary characteristics of the ventricular fibrillation waveform could benefit from TF and time-scale analysis. Umapathy et al. [92] used time-scale analysis due to its better characterization of morphological characteristics that could help in understanding the ventricular fibrillation phenomenon. The wavelet transform approach is computationally less expensive than the Fourier transform, which makes it appropriate for near-real-time feature extraction.

Pig (porcine) models are frequently employed in ventricular fibrillation investigations due to their anatomical and physiological (heart size, coronary structure, and inflammatory response) as well as electrical parallels to humans. The dataset used in the work of Umapathy et al. [92] comprised 16 healthy pigs weighing between 25 and 35 kg. Throughout the trial, each pig was sedated. The surface ECG was obtained by lead II, and the coronary perfusion pressure was determined by taking the difference between the pressures in the aorta and the right atrium, which were both monitored using Millar catheters. The surface ECG and pressures were recorded at 1 kHz and 500 Hz; all signals were downsampled to 250 Hz for analysis to reduce computational complexity.

Burst pacing was utilized to induce ventricular fibrillation. A 10 V, 60 Hz pulse was delivered to the heart for two seconds. The pig was kept untreated for five minutes in ventricular fibrillation, simulating the ischemic phase. At the completion of this period, chest compression at the rate of 100 compressions per minute commenced using a pneumatic device, in conjunction with manual ventilation at the rate of 6 breaths per minute utilizing 5–6 l/min of 100% O₂ with an artificial breathing unit.

CPR was administered at the rate of 30:2 for three minutes (compressions to respirations). Following CPR, the first attempt at defibrillation with a 150 J shock was made. If the animal did not respond, CPR was maintained for an additional two minutes and then 200 J defibrillation was performed. In the case of another failure, the combination of CPR and defibrillation was repeated, but the energy of the defibrillation shock was gradually increased to the maximum of 360 J. Throughout the experimental methodology, the ECG as well as the aortic, right atrial, and airway pressures were continuously measured following the beginning of ventricular fibrillation. Return to sinus rhythm and its persistence for 10 minutes after shock was classified as ROSC; otherwise, the case was classified as non-ROSC.

Wavelet feature analysis: As Umapathy et al. [92] have demonstrated, wavelet analysis has several advantages in processing ventricular fibrillation data and delivering near-real-time feedback: effectiveness in analysis of nonstationary ventricular fibrillation data, suitability for extraction of information from signal morphology, flexibility in choosing alternative wavelet bases to model diverse signal features, low processing complexity, and ease of hardware implementation.

The approach described here employs the CWT; see Equation 8.107. Specifically, $x(t)$ in this context indicates the ischemic component of the ventricular fibrillation waveform (preshock and pre-CPR). Umapathy et al. [92] assessed several wavelet functions, including Daubechies, Gaussian, Morlet, and Shannon wavelets, and determined that the Morlet wavelet provided the best fit to the ventricular fibrillation waveforms, particularly when the ventricular fibrillation waveforms exhibited organization. Prior to using wavelet analysis to deconstruct the ventricular fibrillation waveforms, a bandpass filter ($3 - 21 \text{ Hz}$) was implemented to remove low- and high-frequency artifacts. Filtered ventricular fibrillation waveforms captured during the ischemic phase were segmented into 5 s segments (1,250 samples at 250 Hz) to replicate real-time acquisition with a 5 s buffer. Each 5 s data segment was decomposed across a range of wavelet scales whose associated frequencies encompassed the ventricular fibrillation bandwidth.

Umapathy et al. [92] studied the wavelet decomposition coefficients derived by decomposing 5 s ventricular fibrillation segments and observed that the energy distribution throughout the range of scales differed depending on the ventricular fibrillation waveform's signal composition. To be more specific, the analyzing wavelet acquired varying amounts of signal energy at each scale, depending on the signal's characteristics. If E_x is the signal's total energy and s_1 to s_N are the wavelet scales that completely characterize the signal, the signal energy may be written as $E_x = E_{s_1} + E_{s_2} + E_{s_3} + \dots + E_{s_N}$.

Umapathy et al. [92] proposed a novel wavelet-based feature in which the normalized energy distribution was computed at all scales and the width of the distribution was used as the feature, denoted as the scale distribution width (SDW). The number of scales that contribute significantly to the total signal energy varies with the signal's waveform or structural content, as indicated by the distribution's width. The SDW of a single-component ventricular fibrillation waveform is expected to be a tall and sharp peak, indicating that only a few scales were required to characterize most of the signal's energy. In the case of a multicomponent segment of a ventricular fibrillation waveform with a larger distribution width, the opposite is expected. This feature could be used to indirectly track the signal's composition over time as a proxy for the morphological changes that occur during ventricular fibrillation. As an illustration, in Figures 8.41 and 8.42, two 5 s ventricular fibrillation segments taken at distinct points during a 5-minute period of ischemia in a pig are plotted. In comparison to the 5 s section in the top panel of Figure 8.42, the 5 s segment in the top panel of Figure 8.41 is more orderly and requires fewer scales to model. The bottom panels in Figures 8.41 and 8.42 correspond to the wavelet scale-energy distribution and emphasize the inverse relationship between the SDW and the wavelet scale-energy distribution. The examples demonstrate that SDW may be used as a feature to represent the varying characteristics of ventricular fibrillation during its temporal progression. Umapathy et al. [92] opted to measure the width of the distribution at half the height of the apex. This choice was made to ensure that the width measurement remains unaffected by minor or insignificant signal fluctuations while being suitably sensitive.

The top panel of Figure 8.43 depicts the SDW feature for all 11 successful ROSC cases over about 4.1 minutes (that is, over 50 points of 5 s each). The evolution of the SDW feature over time has a definite rising tendency. Given that smaller widths imply more ordered components and larger widths indicate disorganized components, it is possible to deduce from the upper plot that the ventricular fibrillation waveforms of successfully resuscitated pigs tend to become disorganized with time after ischemia. In the plot, one of the successful examples is denoted by a dashed line to indicate that it deviates from the other curves and noticeable changes in morphology were observed in the waveforms, as reflected by SDW.

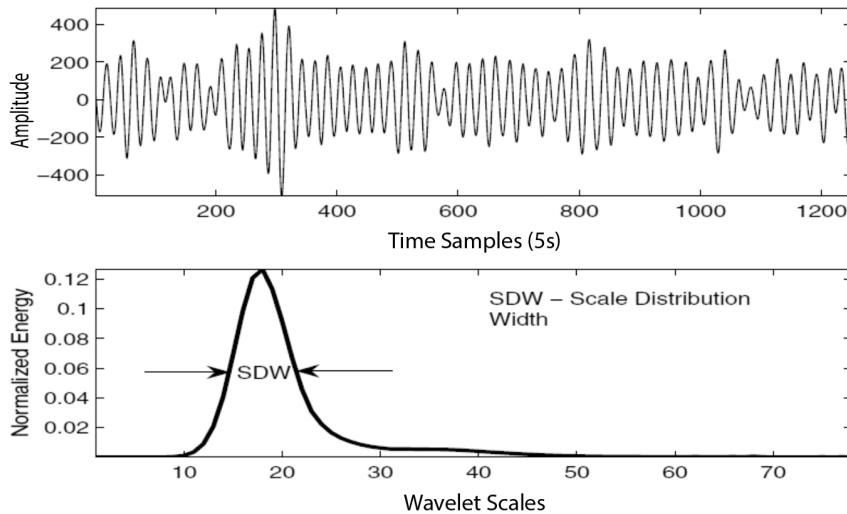


Figure 8.41 Illustration of SDW extracted from an organized portion of ventricular fibrillation. Reproduced with permission from K. Umapathy, S. Krishnan, S. Massé, X. Hu, P. Dorian, and K. Nanthakumar. Optimizing cardiac resuscitation outcomes using wavelet analysis. In Proceedings of the 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 6761–6764, Sept. 2009. ©IEEE.

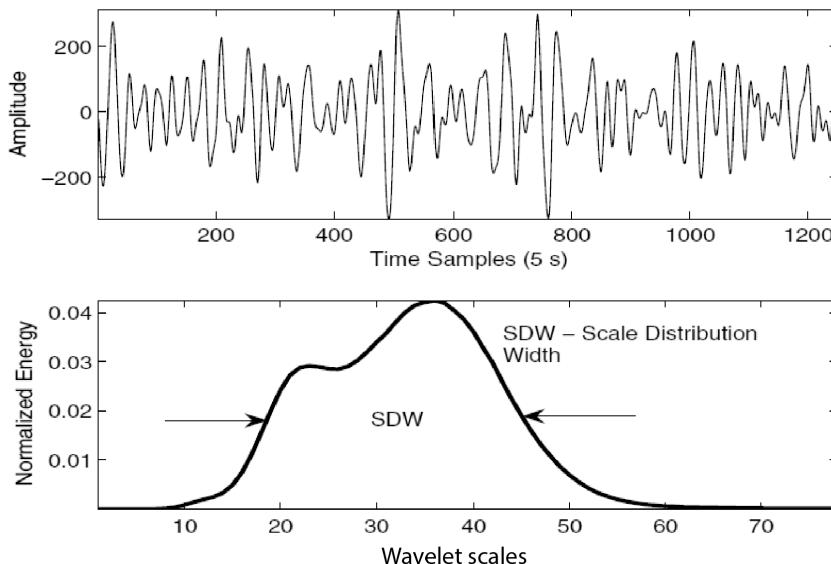


Figure 8.42 Illustration of SDW extracted from a disorganized portion of ventricular fibrillation. Reproduced with permission from K. Umapathy, S. Krishnan, S. Massé, X. Hu, P. Dorian, and K. Nanthakumar. Optimizing cardiac resuscitation outcomes using wavelet analysis. In Proceedings of the 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 6761–6764, Sept. 2009. ©IEEE.

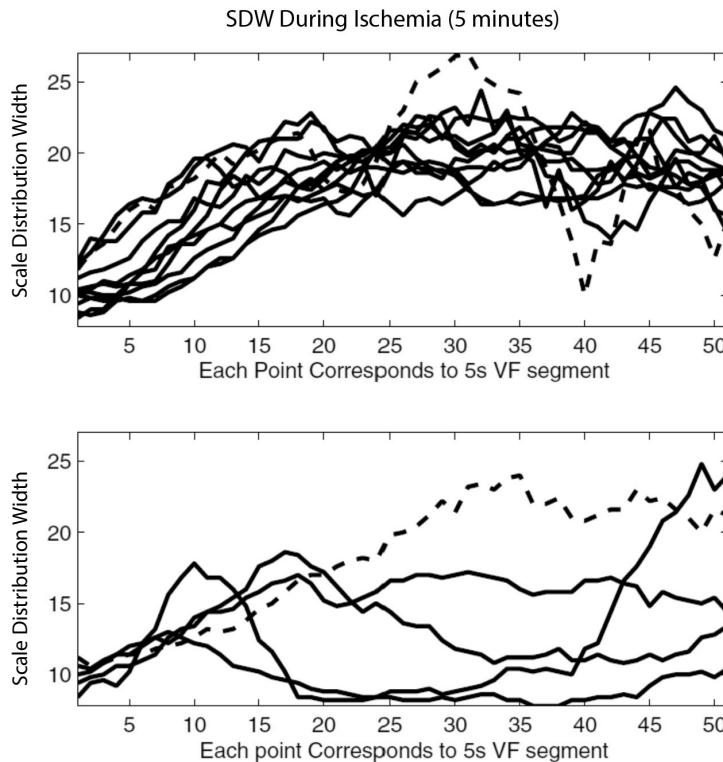


Figure 8.43 Temporal evolution of the SDW feature for all successful (top panel) and unsuccessful (bottom panel) cases of defibrillation. VF - ventricular fibrillation. Reproduced with permission from K. Umapathy, S. Krishnan, S. Massé, X. Hu, P. Dorian, and K. Nanthakumar. Optimizing cardiac resuscitation outcomes using wavelet analysis. In Proceedings of the 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 6761-6764, Sept. 2009. ©IEEE.

When contrasted to the five unsuccessful examples in the bottom panel of Figure 8.43, it is clear that the trend in most cases is in the opposite direction of the successful cases. Apart from one of the unsuccessful examples, the mean values of the curves' plateaus (that is, the region of the curve following the initial segment) are notably different. One example in the unsuccessful set that deviates from the others is shown in a dashed line to indicate that it fits better with the successful cases.

The SDW feature can be considered to be a time-dependent bandwidth measure that can also be derived from short-time data. However, the TF resolution as well as the ability to choose the analyzing wavelet and window size make the SDW feature useful for the application described. The SDW feature closely follows the methodology given by Rosso et al. [93] to compute the scale-energy distribution, except that it measures the distribution's width rather than computing the wavelet entropy from the distribution values. Entropy is an information-theoretic measure of uncertainty that quantifies the notion of randomness or a broadly distributed PDF without assigning a physical meaning. Because the SDW characteristic is always recorded near the peak of the scale-energy distribution, that is, near the dominant scale, or, alternatively, a proportional dominant frequency, it can be considered to be an objective and relevant bandwidth measure around the dominant signal frequency. Additionally, the SDW feature is quite different from the wavelet entropy marker described by

Watson et al. [94], which computes entropy in the temporal direction by utilizing the scalogram's maximum modulus at a specified scale. SDW, in other ways, is calculated from the scale-energy distribution, and the distribution width is measured along the vertical axis of the scalogram. Some other related works on wavelet analysis of CPR include those by Kwok et al. [95] and Box et al. [96].

8.16 Application: Detection of Ventricular Fibrillation in ECG Signals

According to Wiggers [97], Ludwig and Hoffa were the first to characterize ventricular fibrillation as a chaotic asynchronous action of the heart muscle. During ventricular fibrillation, fast impulse production from a single focus or entry can initiate the process, resulting in uncoordinated contraction that generates no functional heart beats despite the myocardium's high metabolic rate. As a result, arterial blood pressure rapidly decreases to dangerously low levels, and anemia of the brain and spinal cord can result in death in as quickly as eight minutes. For over a century, researchers have been investigating ventricular fibrillation in an attempt to determine what causes it and, as a result, offer therapeutic procedures; see Section 8.15.

Ventricular fibrillation is a nonstationary phenomenon that alters the temporal waveform shape, phase, and frequency dynamics of cardiac-surface electrograms [98]. Dominant frequency analysis and phase maps are two commonly used complementary techniques to analyze the development of the ventricular fibrillation process. Despite the fact that ventricular fibrillation is a nonstationary process, the majority of current research restricts frequency analysis to segmented, time-averaged spectrum analysis, omitting critical information on the temporal evolution of spectral features. To circumvent this constraint, one potential solution is a combined TF technique that is well-suited for analysis of ventricular fibrillation and demonstrates a process to understand ventricular fibrillation episodes using the IMF. Human ventricular fibrillation sources are seldom anatomically fixed and frequently migrate over the ventricles. The term "rotors," also known as "spiral waves," is used to denote the drivers of the ventricular fibrillation source; dominant frequency techniques have been unable to account for such migratory behavior. On the other hand, the IMF is more suited than dominant frequency techniques to cope with migratory sources and conduction blockages [98].

Umapathy et al. [98] obtained their ventricular fibrillation dataset from the Toronto General Hospital in Toronto, Ontario, Canada, which has an active heart transplant program. Isolated human hearts from individuals undergoing heart transplantation were obtained for scientific purposes. The University Health Network Ethics Committee authorized the protocol, and informed consent was obtained from each patient. Thirteen isolated hearts were employed in the investigation, and the ventricular fibrillation data were acquired using the following experimental methodology.

A Langendorff apparatus was used to study the isolated human hearts [99]. To summarize, shortly after the transplant recipient's heart was removed, the heart was immersed in cold Tyrode's solution and taken to an adjacent room in less than five minutes, where it was extensively cleansed to eliminate blood particles. The heart was cannulated selectively in the right and left coronary arteries and placed in the Langendorff configuration. Perfusion of the heart with Tyrode's solution ($95\% O_2 + 5\% CO_2$) at the flow rate of $0.9\text{--}1.1\text{ ml/g}\cdot\text{min}$ was then performed. Adjustment of the perfusion pressure was made to maintain a pressure range of $60\text{--}70\text{ mmHg}$. The temperature was maintained at $37^\circ C$ and was checked continuously. The epicardium could encounter temperature variations with respect to the endocardium during isolated heart tests. Therefore, the temperature was monitored to ensure that the difference between the epi- and endocardium was never more than $0.25^\circ C$.

As illustrated in Figure 8.44, the mapping system is capable of simultaneously acquiring electrograms from the epicardium, left-ventricular (LV) endocardium, and right-ventricular (RV) endocardium electrode arrays. Each of the three electrode arrays used contains 112 electrode pairs. Each pair is comprised of two silver beads (2 mm in diameter) spaced apart by 2.1 mm center-to-center. The system facilitates the acquisition of 112 unipolar or 112 bipolar electrograms.

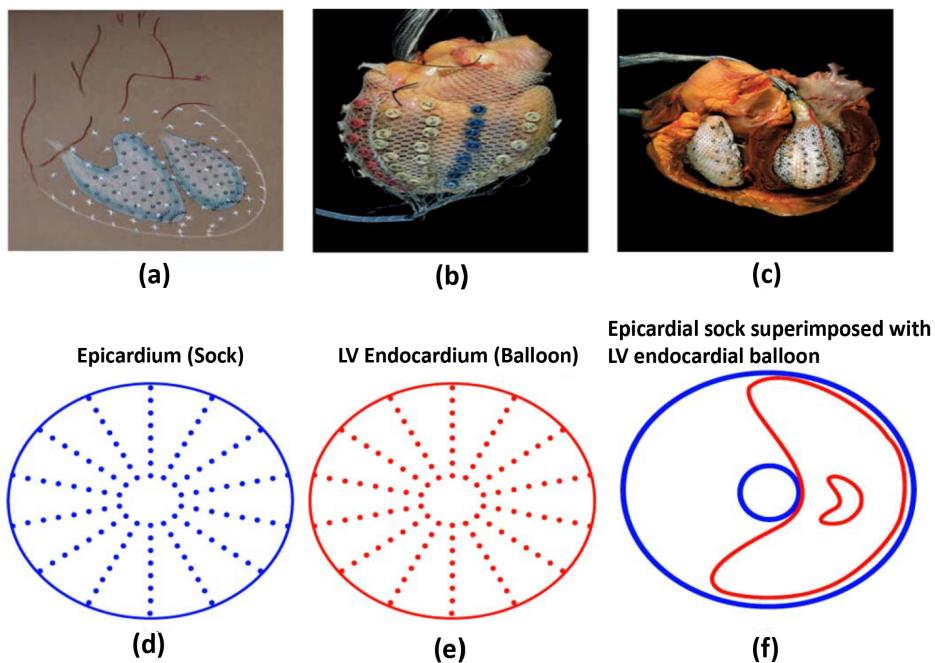


Figure 8.44 Electrode array placement on the heart. (a) Sock array exterior (epicardium) electrodes are indicated by white cross points, LV/RV balloon array interior (endocardium) electrodes are indicated by circular dots. (b) Actual placement of sock array. (c) Actual placement of LV and RV balloon arrays. (d) and (e) Electrode geometry of sock and LV balloon arrays. (f) Sock and LV balloon array superimposed to show the approximate geometrical relation between them. Reproduced with permission from K. Umapathy, S. Massé, E. Sevaptsidis, J. Asta, S. Krishnan, and K. Nanthakumar, Spatiotemporal frequency analysis of ventricular fibrillation in explanted human hearts. *IEEE Transactions on Biomedical Engineering*, 56(2), 328–335, 2008. ©IEEE.

Using the electrode arrays as described above, Umapathy et al. collected 112 simultaneous channels of electrograms from the entire epicardium and endocardium. The sock electrode array positioned over an actual heart is shown in Figure 8.44(b), and a view of the endocardium is shown in Figure 8.44(c); both the LV and RV balloon electrode arrays are visible. The data from the epicardium (sock array) and the LV endocardium (LV balloon array) were used in the study of Umapathy et al. [98]. Figures 8.44(d) and (e) depict the sock and LV balloon arrays, each of which contains 112 electrodes dispersed across 14 radial arms, each with eight pairs of electrodes. The LV balloon is located within the LV endocardium, which spans a large region comparable to the epicardium's inner wall; Figure 8.44(f) illustrates the approximate geometric alignment of the two arrays on either side of the myocardium.

After fitting the electrode arrays and connecting them to the data acquisition system, the heart was paced with an interval of 600 ms. Ventricular fibrillation was induced by briefly contacting the heart in the same location with the two poles of a 9 V battery. Ventricular fibrillation was permitted to continue for 30 s prior to defibrillation. The recordings were made 5 s after ventricular fibrillation stabilized into a chaotic rhythm with no myocardial contraction, and ventricular fibrillation data were acquired for 20 s. Each heart was provoked into five episodes of ventricular fibrillation, on average. Between ventricular fibrillation episodes, a minimum of 7 min was allowed. Each 20 s episode was segmented into 4 s epochs. Simultaneous unipolar recordings were made using the mapping technique described by Sevaptsidis et al. [100]. For unipolar signals, the filter parameters

were $0.5 - 200 \text{ Hz}$, and the sampling rate was set to 1 kHz . Umapathy et al. [98] obtained 204 ventricular fibrillation electrogram data segments of 4 s each from 13 hearts.

TF feature analysis: Umapathy et al. [98] investigated the spatiotemporal evolution of frequencies by extracting the IMF of electrograms during ventricular fibrillation using bilinear TFDs [101, 102]. There are numerous techniques to derive the IMF from a signal [103]. One method to obtain the IMF is to convert the 1D signal to a 2D TFD, and then compute the IMF as the distribution's first moment in the frequency direction; see Section 9.9.

Le and Krishnan [103] evaluated the suitability of various TFDs to extract the IMF from mono- and multicomponent signals; they concluded that the SPWVD is the most suitable TFD with the most favorable trade-off between cross-term suppression, TF resolution, and computational complexity. Umapathy et al. [98] analyzed the ventricular fibrillation electrograms and extracted the IMF using the SPWVD. The SPWVD energy distribution is a well-known variation of the traditional WVD [37]; see Section 8.9. If $x(n)$ is an electrogram, its SPWVD is given by [104]

$$\begin{aligned} SPWVD_x(n, k) = & \frac{N_h}{2} \sum_{m=-N_h+1}^{N_h-1} |h(m)|^2 \sum_{l=-N_g+1}^{N_g-1} g(l) \\ & \times x(n+l+m) x^*(n+l-m) \exp\left(-j2\pi \frac{km}{N_h}\right), \end{aligned} \quad (8.131)$$

where h and g are window functions for frequency and time smoothing with N_h and N_g samples, respectively. In contrast to windowed approaches such as the STFT, where both time and frequency resolution are dependent on a single parameter, the window length, the smoothing functions h and g above operate independently on the time and frequency axes, respectively.

In general, multiple surface electrograms, phase maps, and dominant frequency maps are processed to analyze the ventricular fibrillation process [99, 105]. Dominant frequency maps depict the spatial distribution of an electrogram's dominant frequency, and are typically generated across a time segment. The dominant frequency maps do not contain time-varying information about the spectrum, but rather time-averaged information. To compare the information provided by the IMF during ventricular fibrillation to that provided by the dominant frequency, dominant frequency maps were created for all 204 ventricular fibrillation segments using a technique similar to that described by Massé et al. [99]. Each ventricular fibrillation segment was downsampled by a factor of four (to 250 Hz) and filtered using a bandpass filter with 1.5 and 15 Hz low- and high-frequency cutoffs, respectively. The PSDs of the ventricular fibrillation segments were calculated using the Welch method. PSDs were computed using a 256-sample window length. The dominant frequency was determined using the weighted mean frequency in each of the ventricular fibrillation segments.

While the primary objective of Umapathy et al. [98] was to develop a versatile method for time-varying frequency analysis in lieu of static dominant frequency analysis, the correlation between the IMF and phase maps was used exclusively to find rotors and correlate the spatiotemporal frequency variation with the rotor mechanism. The Hilbert transform was used to construct phase maps. As with the dominant frequency maps, the ventricular fibrillation segments were downsampled and preprocessed to eliminate the mean of each signal segment. After applying the Hilbert transform to the signals, they were filtered and converted to the corresponding analytic signal. The phase at each moment of time was then calculated by determining the analytic signal's angle. The phase map for each instant of time was generated using the same interpolation approach used for the dominant frequency maps. Phase information is valuable in the analysis of ventricular fibrillation because it gives an amplitude-independent tool to locate the reentrant wave's tip or, in simpler terms, the core of a rotor where phase singularities occur. Phase singularities are described as sites on a phase map where the phase is indeterminable and the phase cycles around such sites.

All 204 ventricular fibrillation segments were analyzed using the dominant frequency, IMF, and phase maps. After evaluating all of the measures, 26 ventricular fibrillation segments were selected

that clearly demonstrated the rotor mechanism, and used to demonstrate the relationship between the rotor mechanism and IMF.

To understand the spatiotemporal changes during ventricular fibrillation, the IMF and dominant frequency maps were constructed for each of the ventricular fibrillation segments: one dominant frequency map (since a dominant frequency map only provides the time-averaged spatial distribution of dominant frequency) and “N” IMF maps (spatial distribution of the IMF for each time instant of the electrogram). Rather than doing “N” comparisons, the dominant frequency map was associated with IMF maps at 1, 2, and 3 s of the ventricular fibrillation segments for computational convenience. It was found that 81% of the ventricular fibrillation segments had a low correlation value (0.7), showing that frequency does vary significantly across a ventricular fibrillation segment and that the IMF technique is capable of elucidating such changes. Only 3% of the remaining ventricular fibrillation segments had a higher correlation value (>0.9), specifying that the spatial frequency distribution was consistent over time and that there was no substantial advantage to using IMF maps in such segments. The findings imply that, in the majority of situations, the processes that sustain ventricular fibrillation are not anatomically or temporally consistent, and thus using time averages to track the progression of ventricular fibrillation is inadequate.

Figure 8.45 illustrates the evolution of the IMF map through time for a ventricular fibrillation segment. The leftmost image in Figure 8.45 depicts the dominant frequency map (time-averaged frequency over 4 s); the remaining three images depict the IMF map at 1, 2, and 3 s, respectively. It is evident that the IMF maps contain information about spatiotemporal frequency evolution during ventricular fibrillation. For the entire database used, the mean and SD of the average maximum dominant frequency, IMF (at 1, 2, and 3 s), and IMF over all-time instances were 6.243 ± 0.9028 Hz, 8.086 ± 1.36 Hz, and 6.61 ± 0.947 Hz, respectively. The differences in IMF values at specific time instants compared to those of over all time instances and the average maximum dominant frequency indicate the sensitivity of the IMF parameter in tracking spatiotemporal frequency evolution.

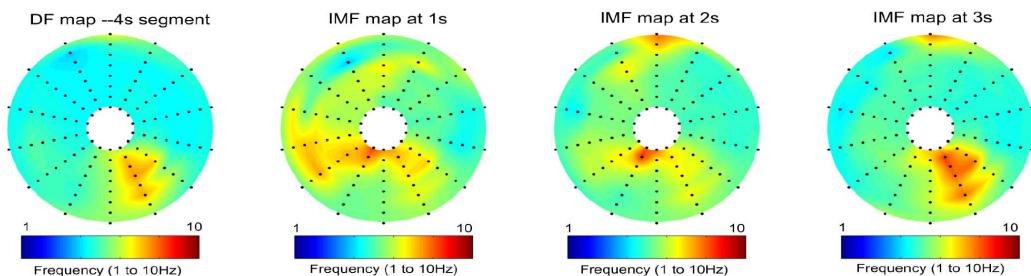


Figure 8.45 The leftmost plot shows the dominant frequency (DF) map computed over a 4 s ventricular fibrillation segment; the remaining three plots show the IMF maps at 1, 2, and 3 s, respectively. The black dots are the electrode locations. By analyzing the maps, the temporal evolution of the frequency phenomenon during a ventricular fibrillation segment can be studied. Reproduced with permission from K. Umapathy, S. Massé, E. Sevaptsidis, J. Asta, S. Krishnan, and K. Nanthakumar, Spatiotemporal frequency analysis of ventricular fibrillation in explanted human hearts. *IEEE Transactions on Biomedical Engineering*, 56(2), 328–335, 2008. ©IEEE.

To explain rotor activity during ventricular fibrillation, the methods reported in the existing literature relate the presence of a rotor to the presence of a steady high-frequency source on the epi- or endocardial surface [106,107]. The IMF maps illustrate the location and temporal evolution of high-frequency zones that may be associated with the rotor core [108]. The immediate updating of the spatial frequency distribution over time may be associated with the migration pattern of a rotor. To illustrate this point, Figure 8.46 shows a sequence of three dominant frequency, IMF, and matching

phase maps from an LV endocardial ventricular fibrillation segment. The first column displays the time-averaged (4 s) dominant frequency maps; the second and third columns display the IMF and phase maps for three points in time. By comparing the dominant frequency, IMF, and phase maps in each row, we can see that the IMF corresponds better with the location of phase singularities in the phase maps than the dominant frequency. For instance, in the second row, the spatial frequency pattern in the dominant frequency map is unrelated to the location of the rotor in the phase map, but an edge of the high-frequency area in the IMF map coincides with the location of the rotor's core in the phase map. This is critical information that cannot be gathered via time-averaged dominant frequency maps. The preceding examples demonstrate that structured and migratory ventricular fibrillation activities can be more accurately analyzed with IMF maps than with dominant frequency maps. This capability of IMF maps could lead to the development of focused therapeutics aimed at terminating ventricular fibrillation [109].

Umapathy et al. [98] utilized unique data and highlighted the advantages of IMF mapping over traditional dominant frequency mapping techniques. IMF holds promise for the development of novel focused therapeutics for ventricular fibrillation, suggests the prospect of tracking the emergence and migration of rotors in a focused manner, and enables the understanding of the genesis and progression of ventricular fibrillation. By determining the precise location and evolution of a rotor and its migration, it is feasible to identify myocardial lesions that isolate or control the rotor in specific locations. Further studies utilizing mapping methods based on the IMF would be required to evaluate the usefulness of such information in treating ventricular fibrillation. Electrophysiologists and cardiologists may be able to develop new catheter-based focal therapeutic procedures to treat ventricular fibrillation using the IMF approach.

8.17 Application: Detection of Epileptic Seizures in EEG Signals

Zhou et al. [110] proposed measures of lacunarity and fluctuation intensity (FI) of DWT coefficients of EEG signals for the detection of epileptic seizures. EEG signals were subjected to DWT-based decomposition with the Daubechies 4 wavelet and five scales, and wavelet coefficients at scales 3, 4, and 5 were selected for further processing. The selection of the scales was based on the observation that seizure signals in intracranial EEG signals usually occur in the range of 3 to 29 Hz.

Noting that ictal EEG segments typically display larger fluctuations than interictal segments, the feature FI was defined to measure the intensity of the fluctuations in an EEG signal as

$$FI(s) = \frac{1}{N} \sum_{i=1}^{N-1} |d(i+1) - d(i)|, \quad (8.132)$$

where N is the number of DWT coefficients $d(i)$ for scale s . It was demonstrated that the values of FI during seizures were typically greater than those during other periods.

Lacunarity is a scale-dependent measure of heterogeneity of a signal or image, and quantifies the presence of gaps [111, 112]. If $p(m, l)$ represents the probability of data samples whose amplitude is equal to m , l represents the length of an epoch, and $[A, B]$ is the range of values of the data samples, it follows that

$$\sum_{m=A}^B p(m, l) = 1. \quad (8.133)$$

With

$$M_1(l) = \sum_{m=A}^B m p(m, l) \quad (8.134)$$

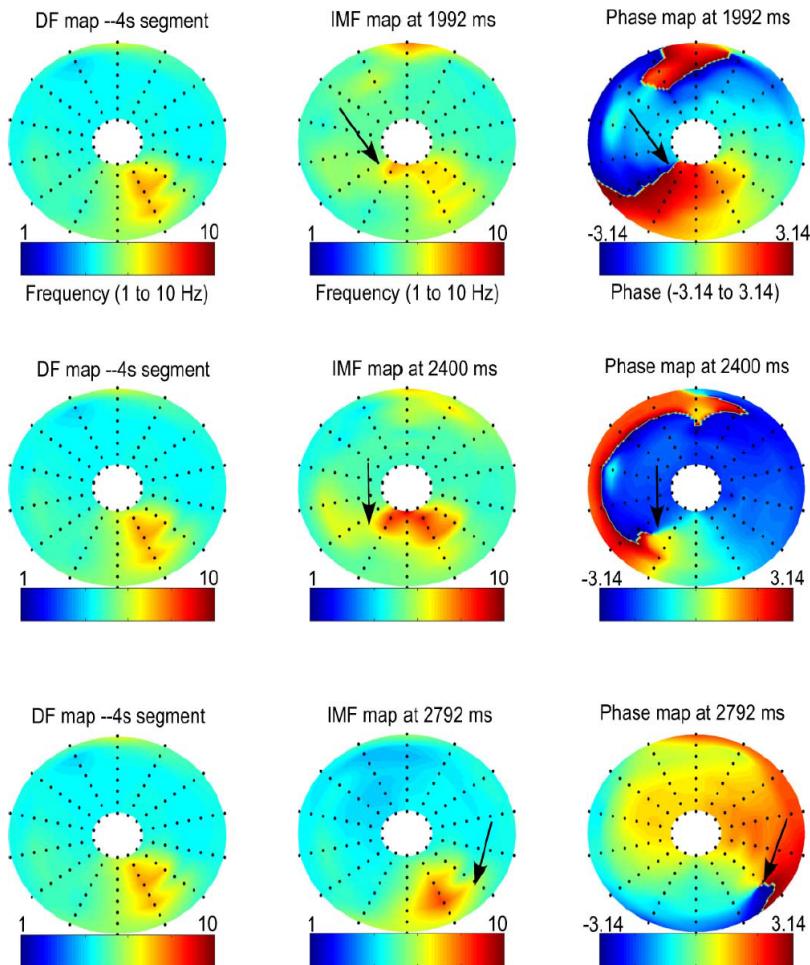


Figure 8.46 The first column of the figure shows the dominant frequency (DF) maps, the second column shows three time instances of IMF maps, and the third column shows three time instances of phase maps. (The DF maps are the same as they are averaged over the entire segment.) Observing the maps row-wise, it is seen that the IMF maps better correspond with the locations of phase singularities (indicated by the arrows) or rotating phase patterns than the dominant frequency maps. For example, in the second row, there is no correspondence between the high-frequency activity in the dominant frequency map and the phase singularity or the rotating phase pattern in the phase map. However, the boundary of the high-frequency activity observed in the IMF map shows a correspondence with the location of the phase singularity or rotating phase pattern in the phase map. Reproduced with permission from K. Umapathy, S. Massé, E. Sevaptsidis, J. Asta, S. Krishnan, and K. Nanthakumar, Spatiotemporal frequency analysis of ventricular fibrillation in explanted human hearts. *IEEE Transactions on Biomedical Engineering*, 56(2), 328–335, 2008. ©IEEE.

and

$$M_2(l) = \sum_{m=A}^B m^2 p(m, l), \quad (8.135)$$

the measure of lacunarity was obtained as

$$\lambda(l) = \frac{M_2(l) - [M_1(l)]^2}{[M_1(l)]^2}. \quad (8.136)$$

Zhou et al. [110] computed lacunarity values of the EEG wavelet coefficients at scales 3, 4, and 5. It was demonstrated that the values of lacunarity during seizures were typically lower than those during other periods.

Including further postprocessing and classification steps, Zhou et al. [110] analyzed a dataset of 289.14 h of intracranial EEG data recorded from 21 patients, and obtained a sensitivity of 96.25% in the detection of epileptic seizures with a false detection rate of 0.13/h and a mean delay time of 13.8 s. See Zhou et al. [110] for a comparative analysis of their results with those obtained in other works on the detection of epileptic seizures in EEG signals.

8.18 Application: Neural Decoding for Control of Prostheses

The amputation of a limb has a significant impact on daily activities. From physical damage to infection to illnesses, there are many causes of limb loss. Whatever the reason, losing a limb has a profound impact on a person's life and makes many everyday chores extremely challenging. In the last two decades, electrical signals from indirect muscles have been incorporated into prosthetic limbs for user control; this is referred to as conventional prosthetic control [113]. Furthermore, the growing discipline of brain prosthetics interprets the user's cerebral activity to enable natural control of prosthetic devices [113]. Raw voltage waveforms from multielectrode recordings (or other procedures) are used to derive a control signal including kinematic parameters to command a robotic arm or other devices. There are two main processes in such a signal flow: first, the raw voltage must be divided into spike trains from one or more neural units, a process known as "spike sorting," and second, the spike trains should be processed by a decoding algorithm to generate behavioral control signals. Brain decoding is the process of deriving direct motion from neural activity.

High-performance BMI systems require decoding algorithms, which convert recorded neuronal activity into control signals for prostheses. Decoder designs have historically drawn inspiration from linear estimates, statistical inference, neural network theory, and neuroscientific theories of the motor cortex. Algorithms used in BMI decoding are trained by simultaneously observing the kinematics of a genuine arm or prosthesis and the activity of its related neural population. A movement estimate is produced by continuous decoders that varies over the possible movement range. Numerous neural prosthesis applications have shown continuous decoders that use ECoG signals [114–116].

In a research study [113] conducted at the University of Utah, four amputee participants were implanted with two 16-channel ECoG electrode arrays. The neural decoding for prosthesis control was performed using the Kalman filter representation explained in Section 8.7. There is no knowledge about hand kinematics in the setting of motor impairment, and the neural decoder makes an attempt to estimate this information. As shown in Equations 8.60 and 8.63, given the observation vector $\mathbf{y}(n)$, the matrices \mathbf{a} and \mathbf{C} , and the noise covariance matrices $\boldsymbol{\eta}_d(n)$ and $\boldsymbol{\eta}_o(n)$, the neural decoding required for prosthesis control would be that of predicting the state vector $\mathbf{x}(n)$ using Equation 8.60.

Figure 8.47 depicts an illustration of a real-time Kalman filter decoder's performance. For each movement of the thumb, index, and middle fingers, the top three traces in the figure display the movement requested in solid black lines and the movement estimated in dashed colored lines. The

decoder estimated the movements with modest errors, as it did during offline analysis [113]. The vertical bars in the figure's bottom panel display times of spike events in six separate electrodes. It is seen that each finger's motion produced a different pattern of activity.

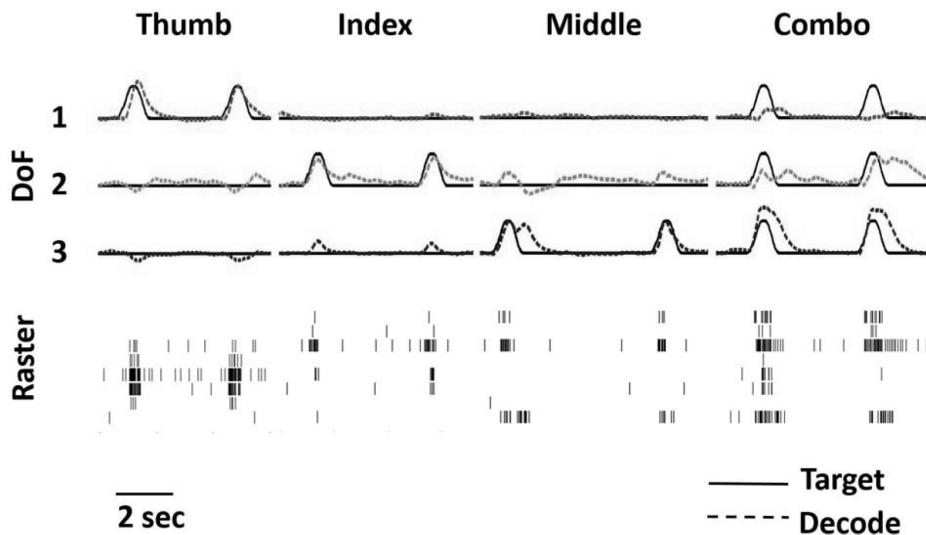


Figure 8.47 Performance of a Kalman filter decoder for the movement of thumb, index, and middle fingers. DoF indicated on the y-axis implies degrees of freedom. The target location (solid black line) and subject-controlled virtual finger location (dashed colored lines) for each movement of the thumb, index, and middle fingers are shown in the top three traces. A raster representation of the timings of spike events from six chosen electrodes during this task is plotted in the lower panel. Reproduced with permission from D.J. Warren, S. Kellis, J.G. Nieven, S.M. Wendelken, H. Dantas, T.S. Davis, D.T. Hutchinson, R.A. Normann, G.A. Clark, and J.V. Mathews, Recording and decoding for neural prostheses. *Proceedings of the IEEE*, 104(2):374–391, 2016. ©IEEE.

From mathematical models and animal testing, prostheses design and control research has advanced to include human test participants. Assuming further advancement in signal processing and control systems techniques, it is fair to anticipate that low-cost prosthetic limbs for practical use may soon be a reality [113].

8.19 Remarks

We have now reached the stage where we have extended the application of a number of signal processing, modeling, and analysis techniques to nonstationary and multicomponent biomedical signals. Fixed or adaptive segmentation of the signals into quasistationary segments is one approach to facilitate the analysis of such signals using traditional techniques, and we studied several approaches for segmentation. Adaptive segmentation facilitates not only the identification of distinct and separate events at previously unknown time instants in the given signal, but also the characterization of events of variable duration using the same number of parameters. We have also studied a few advanced techniques, such as wavelets, bilinear TFDs, and the Kalman filter that facilitate the analysis of nonstationary signals without segmentation. Chapter 9 presents further advanced techniques for the analysis of nonstationary and multicomponent signals via adaptive decomposition. The results obtained using the techniques studied in the present chapter provide advantages in pat-

decoder estimated the movements with modest errors, as it did during offline analysis [113]. The vertical bars in the figure's bottom panel display times of spike events in six separate electrodes. It is seen that each finger's motion produced a different pattern of activity.

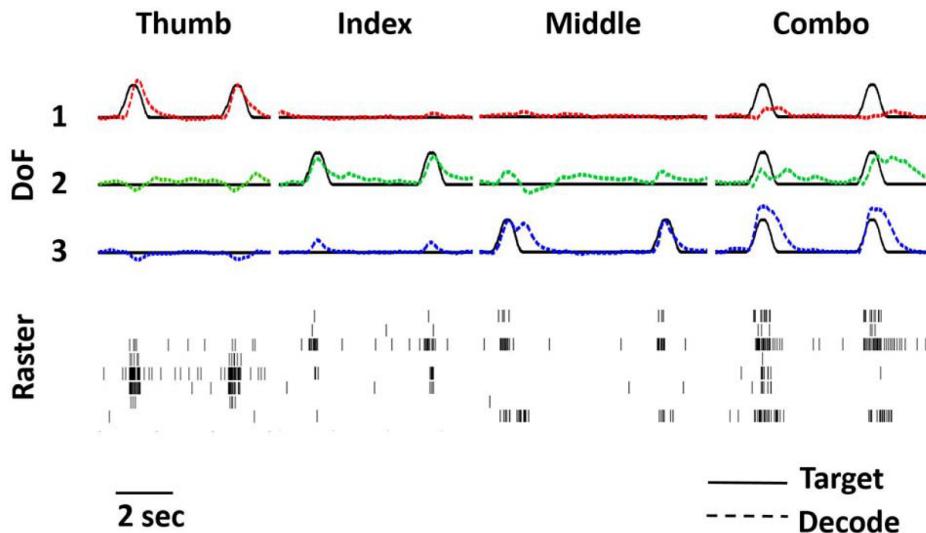


Figure 8.47 Performance of a Kalman filter decoder for the movement of thumb, index, and middle fingers. DoF indicated on the y-axis implies degrees of freedom. The target location (solid black line) and subject-controlled virtual finger location (dashed colored lines) for each movement of the thumb, index, and middle fingers are shown in the top three traces. A raster representation of the timings of spike events from six chosen electrodes during this task is plotted in the lower panel. Reproduced with permission from D.J. Warren, S. Kellis, J.G. Nieven, S.M. Wendelken, H. Dantas, T.S. Davis, D.T. Hutchinson, R.A. Normann, G.A. Clark, and J.V. Mathews, Recording and decoding for neural prostheses. *Proceedings of the IEEE*, 104(2):374–391, 2016. ©IEEE.

From mathematical models and animal testing, prostheses design and control research has advanced to include human test participants. Assuming further advancement in signal processing and control systems techniques, it is fair to anticipate that low-cost prosthetic limbs for practical use may soon be a reality [113].

8.19 Remarks

We have now reached the stage where we have extended the application of a number of signal processing, modeling, and analysis techniques to nonstationary and multicomponent biomedical signals. Fixed or adaptive segmentation of the signals into quasistationary segments is one approach to facilitate the analysis of such signals using traditional techniques, and we studied several approaches for segmentation. Adaptive segmentation facilitates not only the identification of distinct and separate events at previously unknown time instants in the given signal, but also the characterization of events of variable duration using the same number of parameters. We have also studied a few advanced techniques, such as wavelets, bilinear TFDs, and the Kalman filter that facilitate the analysis of nonstationary signals without segmentation. Chapter 9 presents further advanced techniques for the analysis of nonstationary and multicomponent signals via adaptive decomposition. The results obtained using the techniques studied in the present chapter provide advantages in pat-

tern classification (see Chapter 10) as well as efficient representation and analysis of nonstationary signals.

8.20 Study Questions and Problems

1. Describe the characteristics of PCG signals that would make them nonstationary. Propose signal processing strategies to break a PCG signal into quasistationary segments.
2. Discuss features of the EEG that make the signal nonstationary. Propose signal processing strategies to detect each type of nonstationarity and to break an EEG signal into quasistationary segments.
3. Investigate features of the EMG that make the signal nonstationary. Propose signal processing strategies to track the time-varying characteristics of the signal. Under what conditions can the signal be partitioned into quasistationary segments? What are the physiological features that you would be able to derive from each segment?
4. Propose an algorithm to perform the segmentation of PCG signals into four parts per cardiac cycle as: (a) the first heart sound; (b) systolic murmur, if present; (c) the second heart sound; and (d) diastolic murmur, if present. If your proposal includes the use of other signals, explain the need and rationale for the use of such signals. Explain the relationship between events in your reference signals and the events of interest in the PCG signal. Provide sketches of typical signals and the expected results of your methods to illustrate your procedures. Document your procedures using a flowchart or an algorithmic listing. Give at least three nontrivial equations representing important steps in your procedures.
5. Describe a few pathological conditions that could be detected from speech signal analysis. Propose adaptive segmentation approaches to segment pathological speech signals into quasistationary segments.
6. Provide the mathematical expression of a Gaussian signal, and comment on its time and frequency resolution. Prove that the lower bound of the uncertainty principle or time-bandwidth product is satisfied in the case of Gaussian signals.
7. Prove that the Mexican hat function $\psi(t) = (1 - t^2) \exp(-0.5 t^2)$ satisfies the two desired properties to qualify as a wavelet: finite energy and admissibility condition.
8. What are the basis functions of the Fourier expansion? Comment on their suitability to qualify as wavelets.
9. Provide a schematic plot of the time–frequency tiling of the STFT and the wavelet transform in a 2D time–frequency plane. Comment on the uncertainty principle or time-bandwidth product of the two tiling configurations.
10. Give representative plots of the TFDs of a sinusoid, a transient, and a linearly frequency-modulated signal. Comment on the time and frequency localization of the three signals.
11. For a linearly frequency-modulated signal, also known as a chirp signal, given by $x(t) = \exp(j2\pi f t^2)$, provide the mathematical equation of the WVD of the signal. Plot the 2D WVD with the axes clearly labeled with normalized values.

8.21 Laboratory Exercises and Projects

Note: Data files related to the exercises are available at the site

<https://github.com/srikrishnan1972/Biomedical-Signal-Analysis>

1. The speech signal of the word “safety” is given in the file safety.wav. You may use the program safety.m to read the data. Explore the use of short-time statistics such as *ZCR* and *RMS* values for segmentation of the signal. Study the effect of the duration of the short-time analysis window on the trends in the parameters computed and on segmentation.
2. The files pec1.dat, pec22.dat, pec33.dat, and pec52.dat give the PCG, ECG, and carotid pulse signals of two normal subjects and two patients with systolic murmur. You may use the program plotpec.m to read

the data. Explore the use of short-time *ZCR*, *RMS*, and AR-model coefficients for segmentation of the signals. Evaluate the segment boundaries obtained in relation to the events in the PCG signals as well as the corresponding events in the ECG and carotid pulse channels.

3. Using a synthetic nonstationary signal, compare the adaptive segmentation outcomes obtained using the SEM, ACF, and GLR.
4. Implement an STFT algorithm using fixed signal segments and adaptive signal segments. Plot and compare the spectrograms of a few PCG signals.
5. Implement the following three adaptive segmentation (filtering) algorithms: LMS, RLS, and RLSL. Apply them to a nonstationary signal, such as a PCG, and comment on the convergence speed and estimates of the mean-squared error values. Study the effect of the filter order on the convergence speed and error values.
6. Compare the spectrogram and scalogram plots of a pathological speech signal. Comment on the time and frequency localization provided by the two representations. Perform additional simulations by varying the segment length in the case of the spectrogram and by changing the wavelet function for the scalogram, and provide a qualitative assessment of the results.
7. Plot the WVD of monocomponent signals such as a sinusoid, a transient, and a chirp signal separately, and comment on the time and frequency resolution. Compare the WVD representations to the spectrogram and the scalogram, and discuss the time and frequency localization of the various representations.
8. Plot the WVD of a multicomponent signal that is a mixture of a sinusoid, a transient, and a chirp signal. Comment on the time and frequency resolution of the components. Compare the WVD representation to the spectrogram and the scalogram, and discuss the time and frequency localization of the various representations.
9. Implement a program to compute two cross-term reducing TFDs (for example, SPWVD and RSPWVD). Apply the methods to a synthetic multicomponent signal that has a mixture of a sinusoid, a transient, and a chirp signal. Comment on the cross-term reduction properties of the TFDs when compared to the WVD.
10. Plot both the time and frequency marginal distributions of the WVDs of a few PCG signals (see *pec*.dat*). Comment on the energy distribution of the PCG signals in the time domain alone and the frequency domain alone.

References

- [1] Cohen L. What is a multicomponent signal? In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5:113–116, San Francisco, CA, March 1992.
- [2] Boashash B. Estimating and interpreting the instantaneous frequency of a signal — Part 1: Fundamentals. *Proceedings of the IEEE*, 80:520–538, 1992.
- [3] Iwata A, Suzumura N, and Ikegaya K. Pattern classification of the phonocardiogram using linear prediction analysis. *Medical and Biological Engineering and Computing*, 15:407–412, 1977.
- [4] Goodfellow J, Hungerford DS, and Woods C. Patellofemoral joint mechanics and pathology. *Journal of Bone and Joint Surgery*, 58B:921, 1976.
- [5] Krishnan S. *Adaptive Signal Processing Techniques for Analysis of Knee Joint Vibroarthrographic Signals*. PhD thesis, Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB, Canada, June 1999.
- [6] Woo SLY and Buckwalter JA, editors. *Injury and Repair of the Musculoskeletal Soft Tissues*. American Academy of Orthopaedic Surgeons, Park Ridge, IL, 1987.
- [7] Ellison AE. *Athletic Training and Sports Medicine*. American Academy of Orthopaedic Surgeons, Chicago, IL, 1984.
- [8] Frankel VH and Nordin M, editors. *Basic Biomechanics of the Skeletal System*. Lea and Febiger, Philadelphia, PA, 1980.

- [9] Hwang WS, Li B, and Jin LH. Collagen fibril structure of normal, aging, and osteoarthritic cartilage. *Journal of Pathology*, 167:425–433, 1992.
- [10] Fulkerson JP and Hungerford DS, editors. *Disorders of the Patello-femoral Joint*. Williams/Wilkins, Baltimore, MD, 1990.
- [11] Noyes FR and Stabler CL. A system for grading articular cartilage lesions at arthroscopy. *American Journal of Sports Medicine*, 17(4):505–513, 1989.
- [12] Kulund DN, editor. *The Injured Athlete*. Lippincott, Philadelphia, PA, 2nd edition, 1988.
- [13] Meisel AD and Bullough PG. Osteoarthritis of the knee. In Krieger A, editor, *Atlas of Osteoarthritis*, pages 5.1–5.19. Gower Medical Publishing, New York, NY, 1984.
- [14] Smillie IS. *Injuries of the Knee Joint*. Churchill Livingstone, Edinburgh, Scotland, 5th edition, 1978.
- [15] Mankin HJ. The articular cartilages, cartilage healing, and osteoarthritis. In Cruess RL and Rennie WRJ, editors, *Adult Orthopaedics*, pages 163–270. Churchill Livingstone, New York, NY, 1984.
- [16] Frank CB, Rangayyan RM, and Bell GD. Analysis of knee sound signals for non-invasive diagnosis of cartilage pathology. *IEEE Engineering in Medicine and Biology Magazine*, pages 65–68, March 1990.
- [17] McCoy GF, McCrea JD, Beverland DE, Kernohan WG, and Mollan RAB. Vibration arthrography as a diagnostic aid in disease of the knee. *Journal of Bone and Joint Surgery*, 69-B(2):288–293, 1987.
- [18] Kernohan WG, Beverland DE, McCoy GF, Hamilton A, Watson P, and Mollan RAB. Vibration arthrometry. *Acta Orthopædica Scandinavica*, 61(1):70–79, 1990.
- [19] Appel U and v. Brandt A. Adaptive sequential segmentation of piecewise stationary time series. *Information Sciences*, 29:27–56, 1983.
- [20] Appel U and v. Brandt A. A comparative analysis of three sequential time series segmentation algorithms. *Signal Processing*, 6:45–60, 1984.
- [21] Kalman RE. Design of a self-optimizing control system. *Transactions of the ASME*, 80:468–478, 1958.
- [22] Kalman RE and Bucy RS. New results in linear filtering and prediction theory. *Transactions of the American Society of Mechanical Engineers: Journal of Basic Engineering*, 83:95–108, 1961.
- [23] Rangayyan RM. *Biomedical Image Analysis*. CRC Press, Boca Raton, FL, 2005.
- [24] Haykin S. *Adaptive Filter Theory*. Prentice-Hall, Upper Saddle River, NJ, 3rd edition, 1996.
- [25] Arnold M, Witte H, Leger P, Boccalon H, Bertuglia S, and Colantuoni A. Time-variant spectral analysis of LDF signals on the basis of multivariate autoregressive modelling. *Technology and Health Care*, 7:103–112, 1999.
- [26] Arnold M, Miltner WHR, Witte H, Bauer R, and Braun C. Adaptive AR modeling of nonstationary time series by means of Kalman filtering. *IEEE Transactions on Biomedical Engineering*, 45(5):553–562, 1998.
- [27] Bohlin T. Analysis of EEG signals with changing spectra using a short-word Kalman estimator. *Mathematical Biosciences*, 35:221–259, 1977.
- [28] Gath I, Feuerstein C, Pham DT, and Rondouin G. On the tracking of rapid dynamic changes in seizure EEG. *IEEE Transactions on Biomedical Engineering*, 39(9):952–958, 1992.
- [29] Chen JDZ, Stewart Jr. WR, and McCallum RW. Spectral analysis of episodic rhythmic variations in the cutaneous electrogastrogram. *IEEE Transactions on Biomedical Engineering*, 40(2):128–135, 1993.
- [30] Avendaño-Valencia LD, Godino-Llorente JI, Blanco-Velasco M, and Castellanos-Dominguez G. Feature extraction from parametric time-frequency representations for heart murmur detection. *Annals of Biomedical Engineering*, 38(8):2716–2732, 2010.
- [31] Bianchi AM, Mainardi L, Petrucci E, Signorini MG, Mainardi M, and Cerutti S. Time-variant power spectrum analysis for the detection of transient episodes in HRV signal. *IEEE Transactions on Biomedical Engineering*, 40(2):136–144, 1993.
- [32] Oppenheim AV and Lim JS. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.

- [33] Hayes MH, Lim JS, and Oppenheim AV. Signal reconstruction from phase or magnitude. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(6):672–680, 1980.
- [34] Nikias CL and Mendel JM. Signal processing with higher-order spectra. In Ackenhusen JG, editor, *Signal Processing Technology and Applications*, pages 7–34. IEEE Technology Update Series, New York, NY, 1995.
- [35] Nikias CL and Raghubeer MR. Bispectrum estimation — A digital signal processing framework. *Proceedings of the IEEE*, 75:869–891, 1987.
- [36] Hlawatsch F and Boudreux-Bartels GF. Linear and quadratic time-frequency signal representations. *IEEE Signal Processing Magazine*, pages 21–67, April 1992.
- [37] Cohen L. Time-frequency distributions — A review. *Proceedings of the IEEE*, 77:941–981, 1989.
- [38] Boashash B, editor. *Time-Frequency Signal Analysis*. Wiley, New York, NY, 1992.
- [39] Akay M, editor. *Time Frequency and Wavelets in Biomedical Signal Processing*. IEEE, New York, NY, 1998.
- [40] Allen JB and Rabiner LR. A unified approach to short-time Fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558–1564, 1977.
- [41] Portnoff MR. Time-frequency representation of digital signals and systems based on short-time Fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(1):55–69, 1980.
- [42] Rabiner LR and Schafer RW. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, 1978.
- [43] Iwata A, Ishii N, Suzumara N, and Ikegaya K. Algorithm for detecting the first and the second heart sounds by spectral tracking. *Medical and Biological Engineering and Computing*, 18:19–26, 1980.
- [44] Bodenstein G and Praetorius HM. Feature extraction from the electroencephalogram by adaptive segmentation. *Proceedings of the IEEE*, 65(5):642–652, 1977.
- [45] Praetorius HM, Bodenstein G, and Creutzfeldt OD. Adaptive segmentation of EEG records: A new approach to automatic EEG analysis. *Electroencephalography and Clinical Neurophysiology*, 42:84–94, 1977.
- [46] Ferber G. Treatment of some nonstationarities in the EEG. *Neuropsychobiology*, 17:100–104, 1987.
- [47] Bodenstein G, Schneider W, and Malsburg CVD. Computerized EEG pattern classification by adaptive segmentation and probability-density-function classification. Description of the method. *Computers in Biology and Medicine*, 15(5):297–313, 1985.
- [48] Creutzfeldt OD, Bodenstein G, and Barlow JS. Computerized EEG pattern classification by adaptive segmentation and probability density function classification. Clinical evaluation. *Electroencephalography and Clinical Neurophysiology*, 60:373–393, 1985.
- [49] Tavathia S, Rangayyan RM, Frank CB, Bell GD, Ladly KO, and Zhang YT. Analysis of knee vibration signals using linear prediction. *IEEE Transactions on Biomedical Engineering*, 39(9):959–970, 1992.
- [50] Michael D and Houchin J. Automatic EEG analysis: A segmentation procedure based on the autocorrelation function. *Electroencephalography and Clinical Neurophysiology*, 46:232–235, 1979.
- [51] Barlow JS, Creutzfeldt OD, Michael D, Houchin J, and Epelbaum H. Automatic adaptive segmentation of clinical EEGs. *Electroencephalography and Clinical Neurophysiology*, 51:512–525, 1981.
- [52] Duda RO and Hart PE. *Pattern Classification and Scene Analysis*. Wiley, New York, NY, 1973.
- [53] Cohen A. *Biomedical Signal Processing*. CRC Press, Boca Raton, FL, 1986.
- [54] Willsky AS and Jones HL. A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems. *IEEE Transactions on Automatic Control*, 21:108–112, February 1976.
- [55] Basseville M and Benveniste A. Sequential segmentation of nonstationary digital signals using spectral analysis. *Information Sciences*, 29:57–73, 1983.
- [56] Krishnan S. Adaptive filtering, modeling, and classification of knee joint vibroarthrographic signals. Master's thesis, Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB, Canada, April 1996.

- [57] Moussavi ZMK, Rangayyan RM, Bell GD, Frank CB, Ladly KO, and Zhang YT. Screening of vibroarthrographic signals via adaptive segmentation and linear prediction modeling. *IEEE Transactions on Biomedical Engineering*, 43(1):15–23, 1996.
- [58] Sesay AB. *ENEL 671: Adaptive Signal Processing*. Unpublished lecture notes, Department of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, Canada, 1995.
- [59] Krishnan S, Rangayyan RM, Bell GD, Frank CB, and Ladly KO. Adaptive filtering, modelling, and classification of knee joint vibroarthrographic signals for non-invasive diagnosis of articular cartilage pathology. *Medical and Biological Engineering and Computing*, 35(6):677–684, 1997.
- [60] Kao JC, Stavisky SD, Sussillo D, Nuyujukian P, and Shenoy KV. Information systems opportunities in brain-machine interface decoders. *Proceedings of the IEEE*, 102(5):666–682, 2014.
- [61] Kalman RE. A new approach to linear filtering and prediction problems. *Transactions of the American Society of Mechanical Engineers: Journal of Basic Engineering*, 82:35–45, 1960.
- [62] Boulfelfel D, Rangayyan RM, Hahn LJ, Kloiber R, and Kuduvalli GR. Restoration of single photon emission computed tomography images by the Kalman filter. *IEEE Transactions on Medical Imaging*, 13(1):102–109, 1994.
- [63] Sage AP and Melsa JL. *Estimation Theory with Applications to Communications and Control*. McGraw-Hill, New York, NY, 1971.
- [64] Yao L, Brown P, and Shoaran M. Resting tremor detection in Parkinson’s disease with machine learning and Kalman filtering. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4. IEEE, 2018.
- [65] Niknazar M, Rivet B, and Jutten C. Fetal ECG extraction by extended state Kalman filtering based on single-channel recordings. *IEEE Transactions on Biomedical Engineering*, 60(5):1345–1352, 2012.
- [66] Dutra B, Silveira A, and Pereira A. Grasping force estimation using state-space model and Kalman filter. *Biomedical Signal Processing and Control*, 70:103036, 2021.
- [67] Gowda S, Orsborn AL, Overduin SA, Moorman HG, and Carmena JM. Designing dynamical properties of brain-machine interfaces to optimize task-specific performance. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(5):911–920, 2014.
- [68] Cerutti S and Marchesi C, editors. *Advanced Methods of Biomedical Signal Processing*. IEEE and Wiley, New York, NY, 2011.
- [69] Addison PS, Walker J, and Guido RC. Time-frequency analysis of biosignals. *IEEE Engineering in Medicine and Biology Magazine*, 28(5):14–29, 2009.
- [70] Addison PS. Wavelet transforms and the ECG: A review. *Physiological Measurement*, 26:R155–R199, 2005.
- [71] Unser M and Aldroubi A. A review of wavelets in biomedical applications. *Proceedings of the IEEE*, 84(4):626–638, 1996.
- [72] Mallat SG. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.
- [73] Mallat SG and Zhang Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.
- [74] Daubechies I. The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005, 1990.
- [75] Weiss LG. Wavelets and wideband correlation processing. *IEEE Signal Processing Magazine*, 11:13–32, January 1994.
- [76] Li C, Zheng C, and Tai C. Detection of ECG characteristic points using wavelet transforms. *IEEE Transactions on Biomedical Engineering*, 42(1):21–28, 1995.
- [77] Torrence C and Compo GP. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, 1998.
- [78] Farge M. Wavelet transforms and their application to turbulence. *Annual Review of Fluid Mechanics*, 24:395–457, 1992.

- [79] Mallat S. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, CA, 1998.
- [80] Bradley AP. Shift-invariance in the discrete wavelet transform. In Sun C, Talbot H, Ourselin S, and Adriaansen T, editors, *Proceedings of VIIth Digital Image Computing: Techniques and Applications*, pages 29–38, Sydney, Australia, December 2003.
- [81] Wickerhauser MV. *Adapted Wavelet Analysis from Theory to Software*. IEEE Press, Piscataway, NJ, 1994.
- [82] Flandrin P and Martin W. A general class of estimators for the Wigner–Ville spectrum of nonstationary processes. In Bensoussan A and Lions JL, editors, *Systems Analysis and Optimization of Systems, Lecture Notes in Control and Information Sciences*, pages 15–23. Springer-Verlag, Berlin, Germany, 1984.
- [83] Auger F and Flandrin P. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, 43(5):1068–1089, May 1995.
- [84] Task Force of the European Society of Cardiology and the North American Society of Pacing Electrophysiology. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065, 1996.
- [85] Várrallyay Jr. G. The melody of crying. *International Journal of Pediatric Otorhinolaryngology*, 71(11):1699–1708, 2007.
- [86] Hirschberg J. Acoustic analysis of pathological cries, stridor and coughing sounds in infancy. *International Journal of Pediatric Otorhinolaryngology*, 2(4):287–300, 1980.
- [87] Hirschberg J. Dysphonia in infants. *International Journal of Pediatric Otorhinolaryngology*, 49(Supplement 1):S293–S296, 1999.
- [88] Unser M and Aldroubi A. A review of wavelets in biomedical applications. *Proceedings of the IEEE*, 84(4):626–638, 1996.
- [89] Seena V and Yomas J. A review on feature extraction and denoising of ECG signal using wavelet transform. In *2nd International Conference on Devices, Circuits and Systems (ICDCS)*, pages 1–6. IEEE, 2014.
- [90] Vázquez RR, Velez-Perez H, Ranta R, Dorr VL, Maquin D, and Maillard L. Blind source separation, wavelet denoising and discriminant analysis for EEG artefacts and noise cancelling. *Biomedical Signal Processing and Control*, 7(4):389–400, 2012.
- [91] Raghuram M, Madhav KV, Krishna EH, and Reddy KA. Evaluation of wavelets for reduction of motion artifacts in photoplethysmographic signals. In *10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010)*, pages 460–463. IEEE, 2010.
- [92] Umapathy K, Krishnan S, Massé S, Hu X, Dorian P, and Nanthakumar K. Optimizing cardiac resuscitation outcomes using wavelet analysis. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6761–6764. IEEE, 2009.
- [93] Rosso OA, Blanco S, Yordanova J, Kolev V, Figliola A, Schürmann M, and Başar E. Wavelet entropy: A new tool for analysis of short duration brain electrical signals. *Journal of Neuroscience Methods*, 105(1):65–75, 2001.
- [94] Watson JN, Addison PS, Clegg GR, Steen PA, and Robertson CE. Wavelet transform-based prediction of the likelihood of successful defibrillation for patients exhibiting ventricular fibrillation. *Measurement Science and Technology*, 16(10):L1, 2005.
- [95] Kwok H, Coulter J, Blackwood J, Sotoodehnia N, Kudenchuk P, and Rea T. A method for continuous rhythm classification and early detection of ventricular fibrillation during CPR. *Resuscitation*, 176:90–97, 2022.
- [96] Box MS, Watson JN, Addison PS, Clegg GR, and Robertson CE. Shock outcome prediction before and after CPR: a comparative study of manual and automated active compression-decompression CPR. *Resuscitation*, 78(3):265–274, 2008.
- [97] Wiggers CJ. The mechanism and nature of ventricular fibrillation. *American Heart Journal*, 20(4):399–412, 1940.

- [98] Umapathy K, Massé S, Sevaptsidis E, Asta J, Krishnan S, and Nanthakumar K. Spatiotemporal frequency analysis of ventricular fibrillation in explanted human hearts. *IEEE Transactions on Biomedical Engineering*, 56(2):328–335, 2008.
- [99] Massé S, Downar E, Chauhan V, Sevaptsidis E, and Nanthakumar K. Ventricular fibrillation in myopathic human hearts: Mechanistic insights from in vivo global endocardial and epicardial mapping. *American Journal of Physiology: Heart and Circulatory Physiology*, 292(6):H2589–H2597, 2007.
- [100] Sevaptsidis E, Massé S, Parson ID, Downar E, and Kimber S. Simultaneous unipolar and bipolar recording of cardiac electrical activity. *Pacing and Clinical Electrophysiology*, 15(1):45–51, 1992.
- [101] Moghe SA, Qu F, Leonelli FM, and Patwardhan AR. Time-frequency representation of epicardial electrograms during ventricular fibrillation. *Biomedical Sciences Instrumentation*, 36:45–50, 2000.
- [102] Patwardhan A, Moghe S, Wang KE, and Leonelli F. Frequency modulation within electrocardiograms during ventricular fibrillation. *American Journal of Physiology: Heart and Circulatory Physiology*, 279(2):H825–H835, 2000.
- [103] Le L and Krishnan S. Time-frequency signal synthesis and its application in multimedia watermark detection. *EURASIP Journal on Advances in Signal Processing*, 2006:1–14, 2006.
- [104] Ebden MJ, Tarassenko L, Payne SJ, Darowski A, and Price JD. Time-frequency analysis of the ECG in the diagnosis of vasovagal syndrome in older people. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 1, pages 290–293. IEEE, 2004.
- [105] Nash MP, Mourad A, Clayton RH, Sutton PM, Bradley CP, Hayward M, Paterson DJ, and Taggart P. Evidence for multiple mechanisms in human ventricular fibrillation. *Circulation*, 114(6):536–542, 2006.
- [106] Samie FH, Berenfeld O, Anumonwo J, Mironov SF, Udassi S, Beaumont J, Taffet S, Pertsov AM, and Jalife J. Rectification of the background potassium current: A determinant of rotor dynamics in ventricular fibrillation. *Circulation Research*, 89(12):1216–1223, 2001.
- [107] Chen J, Mandapati R, Berenfeld O, Skanes AC, and Jalife J. High-frequency periodic sources underlie ventricular fibrillation in the isolated rabbit heart. *Circulation Research*, 86(1):86–93, 2000.
- [108] Liu YB, Peter A, Lamp ST, Weiss JN, Chen PS, and Lin SF. Spatiotemporal correlation between phase singularities and wavebreaks during ventricular fibrillation. *Journal of Cardiovascular Electrophysiology*, 14(10):1103–1109, 2003.
- [109] Biktashev VN and Holden AV. Reentrant waves and their elimination in a model of mammalian ventricular tissue. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 8(1):48–56, 1998.
- [110] Zhou W, Liu Y, Yuan Q, and Li X. Epileptic seizure detection using lacunarity and Bayesian linear discriminant analysis in intracranial EEG. *IEEE Transactions on Biomedical Engineering*, 60(12):3375–3381, 2013.
- [111] Mandelbrot BB. *Fractal Geometry of Nature*. WH Freeman, San Francisco, CA, 1983.
- [112] Guo Q, Shao J, and Ruiz VF. Characterization and classification of tumor lesions using computerized fractal-based texture analysis and support vector machines in digital mammograms. *International Journal of Computer Assisted Radiology and Surgery*, 4(1):11–25, January 2009.
- [113] Warren DJ, Kellis S, Nieveen JG, Wendelken SM, Dantas H, Davis TS, Hutchinson DT, Normann RA, Clark GA, and Mathews VJ. Recording and decoding for neural prostheses. *Proceedings of the IEEE*, 104(2):374–391, 2016.
- [114] Hermiz J, Rogers N, Kaestner E, Ganji M, Cleary DR, Carter BS, Barba D, Dayeh SA, Halgren E, and Gilja V. Sub-millimeter ECoG pitch in human enables higher fidelity cognitive neural state estimation. *NeuroImage*, 176:454–464, 2018.
- [115] Gilja V, Chestek CA, Diester I, Henderson JM, Deisseroth K, and Shenoy KV. Challenges and opportunities for next-generation intracortically based neural prostheses. *IEEE Transactions on Biomedical Engineering*, 58(7):1891–1899, 2011.
- [116] Jiang T, Jiang T, Wang T, Mei S, Liu Q, Li Y, Wang X, Prabhu S, Sha Z, and Ince NF. Characterization and decoding the spatial patterns of hand extension/flexion using high-density ECoG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(4):370–379, 2017.

CHAPTER 9

SIGNAL ANALYSIS VIA ADAPTIVE DECOMPOSITION

We shall now take up a study of specialized techniques and procedures to analyze a signal by “decomposing” its constituent parts in an adaptive manner. Before delving into the methods of analysis, it is worthwhile examining, in detail, the terms *multichannel*, *multicomponent*, and *multisource* signals.

To begin with, one could take the position that the multiplicity of channels in a recording depends simply on the number of channels of instrumentation available or affordable in an application. EEG signals, for example, could be acquired using a single channel with two electrodes on the scalp, or a couple of hundred channels with as many or more electrodes sewn into a cap worn by the subject. See Chapter 2 for several examples of multichannel signals.

The term *multicomponent* could have several meanings depending upon usage and the context. For example, a cycle of an ECG signal could be said to have the P, QRS, and T waves as its components (see Section 1.2.5), and the S2 part of a PCG could be viewed as a combination of its A2 and P2 components (see Section 1.2.9). If one were to break down a given signal into parts using transforms, filters, and decomposition techniques, the results could be referred to as components of the original signal. Thus, the components of a biomedical signal could be related to the physiological aspects of its genesis or to its decomposition using signal processing techniques. It should be noted that technical decomposition of a given signal into components may not facilitate comprehension of potentially related physiological or pathological processes.

The third term, *multisource*, deserves more careful and detailed consideration than the other two. Considering the specifics at the level of cells, fibers, or tissue, most biomedical signals could be said to be generated by multiple sources: the EMG (see Section 1.2.4) and ECG (see Section 1.2.5) signals include superimposed action potentials of numerous muscle fibers or myocytes. At a higher level, the ECG could be seen to arise from the four anatomical chambers of the heart: the left and right atria and ventricles. Under normal conditions, the P waves from the atria and the QRS

waves from the ventricles appear in an orderly time sequence and do not overlap; however, they could overlap in the case of atrial flutter, AV dissociation, and ectopic beats (PVCs). In the case of an expectant mother, the ECGs of the mother and the fetus are clearly from different independent sources that overlap along the temporal and spectral axes. Separation of the fetal ECG from the maternal ECG is an important application in monitoring the well-being of the fetus.

The A2 and P2 components of S2 are related to closure of the aortic and pulmonary valves; hence, they could be considered to arise from different sources, although Rushmer's theory views vibration of the entire cardiohemic system as the net source of all heart sounds (see Section 1.2.9). It is worth recollecting that A2 and P2 may overlap or be well separated, and may even appear in reverse order, depending upon physiological and pathological conditions. The relative appearance and characteristics of A2 and P2 also depend upon the location where the microphone or transducer is placed on the chest. These observations indicate the need for multichannel recording of the PCG if separate analysis of its components is desired.

The EEG is almost always a multisource signal as the brain simultaneously carries out a multitude of functions (see Section 1.2.6). Due to the numerous control processes, mental activities, and sensory inputs being analyzed, several parts of the brain are concurrently active. The combined activities of numerous neurons in various parts or regions of the brain get projected on to the recording electrodes used in various ways. The identification and spatial localization of the various active sources underlying an EEG signal, especially of epileptic foci, constitute an important application.

It should be evident that analysis and separation of multisource signal components require multichannel recording of the relevant signals, with the number of channels determined by the expected number of sources and the complexity of the application. Regardless, in some applications, it may be desirable to minimize the number of channels of signals, to reduce data transmission and processing costs. Whereas several signal processing techniques may break down or decompose such signals into numerous components, the recognition of the source underlying each component could be a substantial additional challenge. In this chapter, we shall investigate techniques for adaptive decomposition of signals, and explore their usefulness in various biomedical applications.

Signal decomposition techniques that could adaptively provide the components of a given signal represent a flexible framework for analysis and interpretation of a signal on a parts-basis. The key to successful representation of signals lies in the selection of the signal decomposition algorithm. The components obtained from a decomposition algorithm depend largely on the type of basis functions (dictionary) used. For example, the basis function of the Fourier transform decomposes the signal into tonal (sinusoidal) components, and the basis function of the wavelet transform decomposes the signal into components with good time and scale properties. Finding the most sparse representation of a signal using a redundant dictionary is an NP-hard problem, that is, the solution of the problem cannot be obtained in polynomial time. For suboptimal implementation, either the search algorithm or the dictionary can be modified. In traditional signal representation methods, such as the DCT or various wavelet transforms, the dictionary is simply a basis: the number of functions (atoms) in the dictionary is equal to the dimensionality of the signal space and the representation is unique. In the case of an overcomplete dictionary, the number of atoms is greater than the dimensionality of the signal space and the representation is redundant and no longer unique. This enables flexibility in representation with certain specified or desired properties, but requires a criterion to select the final result from various possible representations.

One could develop a criterion for basis selection either from a predefined dictionary or from the signal itself that is intrinsically well adapted to represent a class of signals. This could be done in two ways: pursuit-based approaches or empirical approaches. Sparse modeling of data assumes that the data could be represented with a few atoms from a predefined dictionary. The choice of the dictionary is an important aspect in deriving a sparse representation of the data. Dictionary design has evolved in the last few decades from orthogonal or biorthogonal dictionaries to overcomplete dictionaries. As orthogonal bases are not crucial in the subsequent processing of signal coefficients, the freedom of choice could be enlarged by approximating a given signal with nonorthogonal bases.

An iterative approach for signal decomposition is the matching pursuit algorithm, the details of which are described in Section 9.3. Instead of projecting the signal on to a predefined dictionary as in the pursuit-based approach, empirical mode decomposition (EMD) finds the bases in an adaptive way. EMD decomposes a given signal into components with different time scales called intrinsic mode functions (IMFs) and a residue. The details of EMD for signal decomposition are elaborated upon in Section 9.4. Further specialized methods for signal decomposition using dictionary learning, adaptive TFDs, and factorization techniques as well as their applications are presented in the subsequent sections.

9.1 Problem Statement

Design advanced signal processing techniques for physiological signal analysis and interpretation under the circumstances of the presence of multiple signal components, either acquired from multiple physiological sites or as the result of a mixture of signals recorded at a single site.

Decomposition of a single-channel biomedical signal or multichannel analysis of several biomedical signals could provide complementary information, and in some cases, help in understanding overlapping sources or components. In order to facilitate analysis of multicomponent and multichannel signals, instead of the 1D vectorial format used for a single signal, the multichannel data are represented as a matrix or a tensor. Techniques that are commonly used for matrix and image analysis [1] are extended to multichannel and mixed-signal analysis to achieve the desired outcomes in applications for adaptive filtering of interference, source separation, pattern analysis, and information fusion. The techniques and examples of application described in the remaining sections of the present chapter provide details and illustrations of the notions mentioned here.

9.2 Illustration of the Problem with Case Studies

9.2.1 Separation of the fetal ECG from a single-channel abdominal ECG

Fetal health monitoring can be performed via the analysis of the ECG of the fetus for early detection of congenital heart defects. The fetal ECG allows for monitoring of the fetal heart rate, HRV, and analysis of the electrical activity of the heart of the fetus. Continuous remote monitoring is also possible with methods based on the internet of things (IoT).

Typically, monitoring and analyzing fetal ECG signals can be done in one of two ways. The first, which is an invasive method, is an approach in which an electrode is placed on the fetus' scalp, requiring rupture of the amniotic membranes. The invasive method can lead to infection and risks to the fetus' safety, and is not applicable for long-term continuous monitoring. Due to these drawbacks, the invasive method is not typically used in prenatal clinics, but could be used during labor and the process of birth.

The second method, which is more common as it is noninvasive, is extracting the fetal ECG from the expectant mother's abdominal ECG using well-known adaptive filtering techniques. In the noninvasive method, sensors are placed over the abdomen close to the fetus' position such that the fetal signal is stronger than the maternal ECG signal. There are drawbacks with noninvasive monitoring as well: the ECG is contaminated with noise from various sources such as the maternal ECG signal, muscle contraction, EGG, and fetal brain activity. Due to the overlap with artifacts, interference, and noise, obtaining the desired signal, the fetal ECG, becomes a challenging task.

Existing adaptive filtering methods for extraction of the fetal ECG (see Sections 3.3.5 and 9.7.2) require multilead abdominal and chest-lead ECG signals, which consume substantial power and bandwidth. Furthermore, a multisensor system is not well-suited for the development of IoT health-

care systems. It would be desirable to extract the fetal ECG from a single-channel recording of the expectant mother's ECG.

Section 9.7.2 presents multichannel signal analysis methods to extract the fetal ECG from maternal ECG signals. Section 9.11 presents an approach in which a single signal is represented as a matrix, and decomposition of the matrix helps in extracting the fetal ECG from a single-channel ECG recording obtained from the maternal abdomen.

9.2.2 Patient-specific EEG channel selection for BCI applications

BCI systems are used to convert human intention of action into physical movement of devices such as neural prostheses, wheelchairs, and household equipment [2]. The EEG is typically nonstationary and has a low SNR with temporal resolution in the millisecond range. EEG signals can be used to drive an intention–action relationship between the human brain and computer-controlled devices [2].

BCI systems also provide a unique approach to understand hemiparesis, which causes partial loss of motor function. Hemiparesis can be provoked by stroke; stroke is a neurological event or illness that can cause disability in all age groups, and has been associated with gait disturbance. Robotic BCI devices can be used for the rehabilitation of gait disorders [2]. A patient-specific BCI system allows for adaptive customization of the system to the patient's stage and progress in therapy. Using such devices, stroke survivors have been able to improve their walking and pedaling speeds; improve muscle strength, activity, and balance; and achieve functional independence [2].

Current patient-specific BCI systems use a predetermined fixed configuration of EEG channels. However, it is becoming increasingly apparent that the input optimization problem is challenging in multichannel environments. Patient-specific selection of a small number of EEG channels that are suitable for a particular application is desirable. A related application scenario is explained in Section 9.12.

9.2.3 Detection of microvolt T-wave alternans in long-term ECG recordings

Sudden cardiac death (SCD) is a phenomenon that affects 400,000 North Americans annually [3]. Predictive measures to identify individuals who may be at risk of SCD remains to be a challenge [4]. A typical ECG waveform consists of P, QRS, and T wave components. ECG signals analyzed from long-term recordings could be further investigated to identify markers that could predict SCD. Specifically, information related to T-wave alternans (TWA), has been seen as a potential pointer for noninvasive prediction of SCD [4]. TWA is a heart-rate-dependent anomaly observed in surface ECG in which the amplitude and/or shape of the T wave changes every second heart beat [3].

It is important to note that the TWA signal amplitude is fairly small, typically in the microvolt range. Hence, the algorithms used for TWA detection in the multicomponent ECG must be accurate, robust, and capable of handling noise and data nonstationarity. TWA detection techniques rely on its magnitude. Essentially, if the TWA magnitude is relatively large, it can be taken as a warning signal for fatal arrhythmia of the ventricles and subsequent SCD.

Although TWA analysis holds promise as an important tool to diagnose SCD for individuals at risk, there are limitations that need to be addressed. In particular, a spectral method (SM) [5] and the modified moving average (MMA) method [6] used for this purpose do not handle signal nonstationarity. In Section 9.10, methods to detect TWA using well-established time-domain and frequency-domain techniques as well as more advanced joint TF analysis approaches are presented.

9.3 Matching Pursuit

Matching pursuit is a generic iterative signal decomposition framework, whose properties depend on the type of the dictionary of TF atoms used. In the matching pursuit algorithm [7], the given signal

$x(t)$ is projected on to a dictionary of TF atoms obtained by scaling, translating, and modulating a window function $g(t)$ as

$$x(t) = \sum_{n=0}^{\infty} a_n g_{\gamma_n}(t), \quad (9.1)$$

where

$$g_{\gamma_n}(t) = \frac{1}{\sqrt{s_n}} g\left(\frac{t - \tau_n}{s_n}\right) \exp[j(2\pi f_n t + \phi_n)]. \quad (9.2)$$

Here, a_n is an expansion coefficient, the scale factor s_n is used to control the width of the window function, the parameter τ_n controls temporal placement, and the parameters f_n and ϕ_n are the frequency and phase of the exponential function, respectively. The subscript or index γ_n represents the set of parameters $(s_n, \tau_n, f_n, \phi_n)$. If the window is a Gaussian function, that is,

$$g(t) = 2^{\frac{1}{4}} \exp(-\pi t^2), \quad (9.3)$$

the TF atoms are known as Gabor atoms.

In the matching pursuit algorithm, the given signal is iteratively projected on to a dictionary of TF atoms. The first projection decomposes the signal into two parts:

$$x(t) = \langle x, g_{\gamma_0} \rangle g_{\gamma_0}(t) + R^1 x(t), \quad (9.4)$$

where $\langle x, g_{\gamma_0} \rangle$ denotes the inner product or projection of $x(t)$ on to the first TF atom $g_{\gamma_0}(t)$; the term $R^1 x(t)$ represents the residue after approximating $x(t)$ using $g_{\gamma_0}(t)$. The process of decomposition is continued by projecting the residue on to the subsequent functions in the dictionary, and after M iterations, we have

$$x(t) = \sum_{n=0}^{M-1} \langle R^n x, g_{\gamma_n} \rangle g_{\gamma_n}(t) + R^M x(t), \quad (9.5)$$

with $R^0 x(t) = x(t)$.

The iterative decomposition process may be stopped after using a prespecified number M of the TF atoms or continued until the energy of the residue $R^M x(t)$ decreases below a certain value. A decay parameter, defined as [7]

$$\lambda(m) = \sqrt{1 - \frac{\|R^m x\|^2}{\|R^{m-1} x\|^2}}, \quad (9.6)$$

may also be used; the process is continued until the decay parameter does not reduce substantially any further. The selected atoms or components may be considered to represent the coherent structures present in the signal with respect to the dictionary used. The residue may be assumed to be due to random noise. The approximate signal reconstructed using the M selected (coherent) structures is given by

$$\tilde{x}(t) = \sum_{n=0}^{M-1} \langle R^n x, g_{\gamma_n} \rangle g_{\gamma_n}(t), \quad (9.7)$$

which may be taken to be a filtered version of the original signal. As in the case of wavelets, the coefficients $a_n = \langle R^n x, g_{\gamma_n} \rangle$, $n = 0, 1, 2, \dots, M - 1$, may be used, along with the set of the corresponding parameters, as a compact representation of the signal for further analysis or classification.

By adapting the TF atoms to suit the properties of the given signal, more representative matching-pursuit analysis could be performed. Further discussions on related methods and applications are presented in Sections 9.5, 9.6, and 9.9. See Kovács et al. [8, 9] for illustrations of matching-pursuit-based analysis of fetal PCG signals.

9.4 Empirical Mode Decomposition

EMD is a decomposition procedure [10–14] with which a multicomponent and nonstationary signal can be decomposed into a small number of simpler signals referred to as IMFs. IMFs may be analyzed using the Hilbert transform and other methods to derive TFDs [10]. The decomposition procedure is based on the local characteristics of the given signal; it is a data-driven algorithm that does not use predetermined basis functions.

An IMF is required to have the following properties [10, 12]:

- Over the duration of an IMF, the number of extrema and the number of zero-crossings must be equal or differ at most by one.
- At any location, the mean of the envelopes defined by the local maxima and the local minima is zero.

An IMF is not restricted to be a narrowband signal; it could be modulated in both amplitude and frequency and also be nonstationary [10].

IMFs are extracted level by level [10, 11]. First, high-frequency local oscillations riding on the corresponding low-frequency parts of the given signal is extracted. Then, the next level of high-frequency local oscillations in the residual of the signal are extracted. This iterative procedure is continued until no complete oscillation can be identified in the residual, which is considered to be the low-frequency trend in the original signal; then, the final residual is a monotonic function.

In practice, EMD is implemented through a sifting process that detects and uses the local extrema. The standard EMD algorithm works as follows [10–14]:

1. Find the locations of all of the extrema in the given signal $x(n)$.
2. Interpolate between the minima to obtain the lower signal envelope, $x_{\min}(n)$; do the same with the maxima to obtain $x_{\max}(n)$. The cubic spline function is recommended for interpolation.
3. Compute the mean $x_m(n) = [x_{\min}(n) + x_{\max}(n)]/2$.
4. Subtract the mean obtained in the previous step from the signal to obtain the oscillatory mode signal $s(n) = x(n) - x_m(n)$.
5. If $s(n)$ meets the criteria for an IMF, define $c(n) = s(n)$ as an IMF; otherwise, set $x(n) = s(n)$ and repeat the process from Step 1.
6. Initially, the procedure given above results in the first IMF. To get the remaining IMFs, define the residual as $x(n) - c(n)$ and repeat the procedure.

If $c(n)$ is an IMF obtained as above, let

$$c_a(n) = c(n) + j c_H(n), \quad (9.8)$$

where $c_H(n)$ is the Hilbert transform of $c(n)$ and $c_a(n)$ is the related analytic signal (see Section 5.5.3). Let

$$c_a(n) = a(n) \exp[j\theta(n)], \quad (9.9)$$

with

$$a(n) = \sqrt{c^2(n) + c_H^2(n)} \quad (9.10)$$

and

$$\theta(n) = \arctan \left[\frac{c_H(n)}{c(n)} \right]. \quad (9.11)$$

Then, according to one of the definitions used, the instantaneous frequency $\omega(n)$ is given by the derivative of $\theta(n)$ with respect to time. Analysis as above can be performed on all IMFs and the results combined to perform TFD analysis of the original signal. See Sections 5.5.1 and 5.5.3, and Huang et al. [10] for related discussions.

The original signal may be expressed using the IMFs as

$$x(n) = \text{real} \left\{ \sum_i a_i(n) \exp[j\theta_i(n)] \right\}, \quad (9.12)$$

where the index i indicates the i^{th} IMF and the summation is over all IMFs. The expression provided above gives both the amplitude and the frequency of each component as functions of time; the residual component is neglected. With EMD and expansion of the given signal into its IMFs, the AM and FM components in the original signal are separated: this facilitates TFD analysis.

Illustration of application: Figure 9.1 shows an example of the application of EMD to a VAG signal. The interpretation of IMFs is not straightforward because of the data-dependent nature of the components. In the example, the first three IMFs may be considered to be high-frequency or small-scale noise and not used in further analysis. The IMFs numbered 4 to 7 may be considered to contain most of the useful information. The remaining components may represent low-frequency or large-scale variations and trends that are not of interest.

Wu et al. [15] applied a modified method known as ensemble EMD or EEMD (see Section 9.4.1) and detrended fluctuation analysis for removal of artifacts in knee-joint VAG signals. They found that EMD can improve the quality of VAG signals and that the results obtained were superior to those achieved via decomposition by the matching pursuit algorithm.

Echeverria et al. [14] applied EMD to analyze HRV data. The results were said to facilitate the isolation of at least four components with overlapping and dynamic frequency excursions localized in the currently recognized spectral bands of autonomic modulation. In addition, EMD was observed to permit the separation of the VLF content as well.

Kaleem et al. [16] used EMD to analyze pathological speech signals. Instead of using sustained vowels, EMD was used to decompose randomly chosen portions of speech signals into IMFs, which were then analyzed to extract temporal and spectral features. Instantaneous features were used to capture discriminative information in signals that were suspected to be hidden at local time scales. Using a database consisting of continuous speech from 51 normal speakers and 161 speakers with speech pathology, a classification accuracy of 95.7% was obtained.

9.4.1 Variants of empirical mode decomposition

The original EMD method can provide good time and frequency characterization of signals that possess nonstationary and nonlinear characteristics. The technique is free of basis functions and provides a flexible and adaptive representation of the characteristics of signals. EMD has been

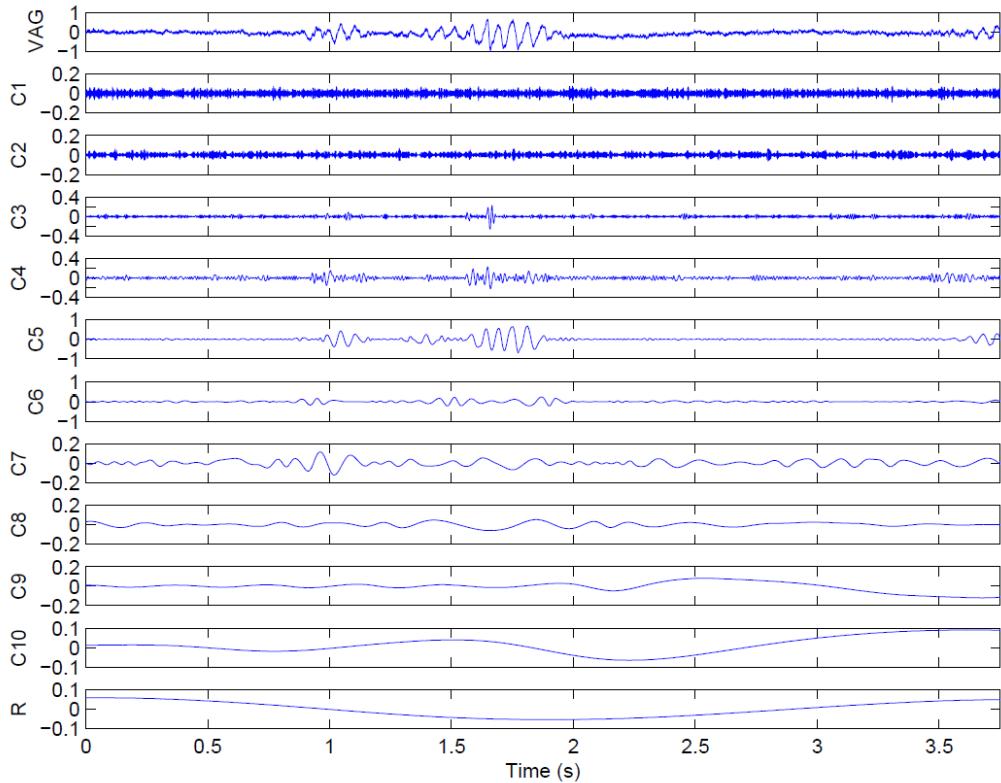


Figure 9.1 A VAG signal and 10 IMFs (C1 to C10) obtained via EMD. The last plot labeled “R” is the residual component. The number of IMFs was fixed at 10; the residual signal is not monotonic and the EMD procedure could be continued further. Figure courtesy of Yunfeng Wu, Xiamen University, China.

improved by various modifications to overcome issues associated with the original decomposition framework, including the mode-mixing problem [17]. In mode mixing, different frequency components get captured in the same IMF or a single frequency component gets decomposed into multiple IMFs, thereby making the interpretation of the IMFs vague.

Ensemble EMD: To overcome the mode-mixing problem associated with EMD, Wu and Huang [17] proposed the EEMD method. The main idea behind EEMD is the addition of white noise at different scales to the signal. Iteratively, through superposition and averaging, the added white noise averages out, and the resulting decomposition provides components free of mode mixing. In practice, EEMD is implemented as follows:

1. The signal $x(n)$ gets added to white noise $w_k(n)$, and the k^{th} version of the signal is expressed as
$$x_k(n) = x(n) + w_k(n). \quad (9.13)$$
2. The standard EMD method is applied to $x_k(n)$ to decompose it into M IMFs.
3. The preceding steps are repeated M times, with $k = 1, 2, 3, \dots, M$, by adding white noise of different sequences but with the same variance. As M increases, white noise averages to lower levels.
4. Once the different IMFs are obtained for each of the iterative steps, they are ensemble averaged to obtain the final IMFs. The resulting IMFs are free of mode mixing.

Several biomedical applications of EEMD have been reported; most notably, it has been applied for human activity recognition [18], heart rate and breathing rate estimation [19], and ECG denoising [20].

Multivariate EMD: The standard EMD method can be applied only to single-channel signals; an extension of EMD to multichannel signals is achieved through multivariate EMD (MEMD) [13]. In MEMD, multidimensional signal envelopes are generated by projecting the signals along different axes or directions, and the local mean of the signals is obtained by computing the average of the envelopes. The residual signal, $R(n) = x(n) - e(n)$, is computed, where $x(n)$ is the given signal and $e(n)$ is the average of the envelopes. If the stopping criteria specified are not met, then the procedure is applied to $R(n)$ and the iterative steps given above are continued.

MEMD has been used in a variety of biomedical signal analysis applications, including removal of muscle artifact from EEG recordings [21], estimation of motor imagery movement [22], and detection of seizure in EEG signals [23].

9.5 Dictionary Learning

The dictionary learning strategy incorporates the advantages of constructing a dictionary from a collection of functions, as in matching pursuit, and identifying dictionary functions via the EMD process. In describing and characterizing biomedical signals, this technique combines the advantages of data-driven decomposition and greedy approximations. (*Note:* A greedy approximation is an approach to determine the best available options at a given time instant without considering the overall or final outcomes.)

The decomposition behavior of EMD can be modeled as a dyadic filter bank [24], in which the spectral content of the given signal is partitioned into distinct IMFs (that is, the faster oscillatory components in a signal are represented in the low-index IMFs, whereas the slower oscillations in a signal are represented in the high-index IMFs). Separating the spectral content of a signal using a data-adaptive filter bank has been effectively applied in a variety of areas (for example, see Lee et al. [25]).

Kaleem et al. [26] developed an EMD-based dictionary technique in which training signals of different classes are decomposed into IMFs using EMD. The IMFs obtained define the raw dictionary's functions or atoms. The raw dictionary is trained using the same training signals with greedy approximation techniques as in matching pursuit. This stage produces a trained dictionary with a modest number of atoms.

As described by Kaleem et al. [26], the pursuit-based technique is used to learn a dictionary with attributes comparable to those of traditional dictionaries (for example, simple invertibility, $\mathbf{DD}^T = \mathbf{I}$, where \mathbf{D} is the dictionary matrix [27, 28]), rather than sparse coding as is the case with standard dictionary learning algorithms. The trained dictionary may subsequently be used to discover patterns of interest in other signals.

The use of the EMD process for the dictionary construction approach is critical because it provides the trained dictionary with discriminant power. EMD is used to decompose the variations in the signals into IMFs. The dictionary learning algorithm then picks IMFs with properties unique to each class [29], resulting in not just a small dictionary, but also a dictionary with atoms having properties unique to the signals of the class being investigated. The technique based on EMD does not correlate with the shift-invariant or convolutional model in this case, as EMD, much like wavelets, is not a shift-invariant decomposition process. As a result, the trained dictionary is not shift invariant [26].

The initial point of the dictionary learning framework is a training matrix $\mathbf{X}_{\text{Train}}^c \in \mathbb{R}^{N \times k^c}$, where the columns of the matrix consist of k^c training signals $x^c \in \mathbb{R}^N$ associated with class c ; $c \in \mathbf{C}$, $\mathbf{C} = \{c_1, c_2, \dots, c_K\}$, where K is the number of classes; and N is the number of samples in x^c . The first step of the methodology is to create a raw dictionary $\mathbf{D}_{\text{raw}}^{\mathbf{C}} = \{\psi^m\}_{m=1}^M$, where ψ

denotes the dictionary atoms. The dictionary atoms ψ are comprised of IMFs which are generated by applying EMD to the training signals x^c . The main objective of the dictionary learning approach is to create a trained dictionary $\mathbf{D}_{\text{Train}}^C = \{\psi^p\}_{p=1}^P$, with $P \ll M$.

The steps involved in forming and learning the dictionary are as follows [26]:

1. The IMFs a_j are obtained by decomposing the signals $x^c \in \mathbf{X}_{\text{Train}}^c$ by EMD, so that $x^c = \sum_{j=1}^J a_j^c$. The number of IMFs J depends on the implementation of the EMD algorithm, and may also differ for different signals x^c . In general, $J \leq \log_2(N)$ [30].
2. The IMFs obtained in Step 1 form the atoms of class-specific raw dictionaries $\mathbf{D}_{\text{raw}}^c = [\mathbf{d}_1 | \mathbf{d}_2, \dots, |\mathbf{d}_L]$, $L = k^c \times J$. The atoms of the class-specific raw dictionaries are constrained to have L_2 -norm ≤ 1 ; therefore, the atoms consist of normalized IMFs, such that $\mathbf{d}_l = \hat{\mathbf{a}}_j^c, l \in \{1, 2, \dots, L\}, j \in \{1, 2, \dots, J\}$, and $\hat{\mathbf{a}}_j^c = \mathbf{a}_j^c / \|\mathbf{a}_j^c\|_2$.
3. A combined raw dictionary $\mathbf{D}_{\text{raw}}^C = [\mathbf{D}_{\text{raw}}^{c_1} | \mathbf{D}_{\text{raw}}^{c_2}, \dots, |\mathbf{D}_{\text{raw}}^{c_K}]$ is obtained by combining the class-specific raw dictionaries.

Kaleem et al. [26] used a matching-pursuit-like algorithm for the EMD-based dictionary learning algorithm to obtain the trained dictionary $\mathbf{D}_{\text{Train}}^C$ from $\mathbf{D}_{\text{raw}}^C$. The dictionary learning algorithm is expressed in Algorithm 9.1, and the procedure to terminate the iterative algorithm is described next.

Algorithm 9.1: EMD-based Dictionary Learning Algorithm [26]

- 1: Initialize $\mathbf{X}_{\text{Train}}^c$ {all training matrices}, $\mathbf{D}_{\text{raw}}^C = \{\psi^m\}_{m=1}^M$ {merged raw dictionary}, $\mathbf{D}_{\text{Train}}^C = []$ {empty matrix}.
 - 2: **repeat** for each signal x^c :
 - 3: Find the projection coefficient: $\alpha_m = \langle x^c, \psi^m \rangle$;
 - 4: Select the dictionary atom $\widetilde{\psi^m}$ with the highest value of $|\alpha_m|$;
 - 5: **if** $\widetilde{\psi^m}$ not in $\mathbf{D}_{\text{Train}}^C$ **then** append $\widetilde{\psi^m}$ to $\mathbf{D}_{\text{Train}}^C$: $\mathbf{D}_{\text{Train}}^C = [\mathbf{D}_{\text{Train}}^C | \widetilde{\psi^m}]$;
 - 6: **end if**
 - 7: Calculate the residue $r^x = x^c - \langle x^c, \widetilde{\psi^m} \rangle \widetilde{\psi^m}$;
 - 8: Set the atom $\widetilde{\psi^m} = 0$ in $\mathbf{D}_{\text{raw}}^C$;
 - 9: Set $x^c = r^x$;
 - 10: **until** Termination
 - 11: Merge the class-specific trained dictionaries to form a combined trained dictionary $\mathbf{D}_{\text{Train}}^C = [\mathbf{D}_{\text{Train}}^{c_1} | \mathbf{D}_{\text{Train}}^{c_2}, \dots, |\mathbf{D}_{\text{Train}}^{c_K}]$.
-

The dictionary learning algorithm based on EMDs may be stopped in one of two ways [26]:

1. Algorithm 9.1 may be stopped after I iterations, where I is a predetermined number. The size of the dictionary is determined by the number of iterations, as more iterations result in the addition of more atoms to the dictionary. Furthermore, the “depth” of the residue (Step 7 of Algorithm 9.1) is dependent on the number of iterations, as more iterations result in a “deeper” residue, resulting in the addition of more IMFs with higher indices to the dictionary. Here, the term “depth” is used to indicate the number of iterations that are needed in the decomposition process. This is because Step 7 of Algorithm 9.1 and the EMD algorithm are identical in that an IMF is removed from the residue r^x .
2. Additionally, a validation mechanism may be used to determine the number of iterations necessary to end the dictionary learning algorithm. A possible validation scheme for a two-class classification scenario is to maximize the distance between two vectors representing the projection coefficients obtained using the validation signals \hat{x}^c belonging to the two classes, where

the validation signals are distinct from those used for dictionary creation and training. The projection coefficient $\alpha_m = \langle \hat{x}^c, \psi^m \rangle$ for each validation signal \hat{x}^c in each of the two classes is first calculated, and then appended to the projection coefficient vector Γ_c appropriate for the class c . If I iterations are used, the projection coefficient vector Γ_c^I is obtained. The stopping criterion that determines the number of iterations for the termination of the algorithm is then given by

$$\max_I \|\Gamma_{c_i}^I - \Gamma_{c_j}^I\|_2^2, \quad (9.14)$$

where c_i and c_j represent the two classes. It is worth noting that additional validation schemes may be applied. A technique to maximize the accuracy of classification using a certain classification method could be an example of such an additional scheme.

Another significant aspect of the method proposed by Kaleem et al. [26] is the reduction in dictionary size as a result of dictionary learning. The raw dictionary $\mathbf{D}_{\text{raw}}^C$ is composed of IMFs obtained through the application of EMD to the training signals. Although the number of IMFs generated is dependent on the execution of the EMD method, the number of IMFs J from each signal is constrained by $J \leq \log_2(N)$ [30]. Then, $M \leq K \times \log_2(N)$, where $M < N$, in general. After training the dictionary with Algorithm 9.1, the trained dictionary $\mathbf{D}_{\text{Train}}^C$ contains $P < M$ atoms.

As indicated in Step 4 of Algorithm 9.1, each iteration of the algorithm could result in the addition of up to k^c atoms to the trained dictionary, assuming that dictionary training is performed using the same K training signals as for dictionary construction. Thus, switching from an untrained to a trained dictionary changes the number of dictionary atoms from M to P . It should be noted that $P \leq K$, as an atom selected during dictionary learning is not added to the dictionary again if it already exists in the trained dictionary, as specified in Step 5 of Algorithm 9.1.

Additionally, if the dictionary is taught using fewer iterations of Algorithm 9.1, the size difference between the raw and trained dictionaries could be significant, such that $P \ll M$. Thus, starting with the raw dictionary and employing a two-stage dictionary generation and learning process could result in a considerable reduction in dictionary size. This enables the test signals to be projected against the trained dictionary in a computationally efficient manner.

Illustration of application: In the study of Kaleem et al. [26], the method outlined in the present section was applied in the following manner to long-term EEG segments of 23 patients experiencing seizures. In the first phase, 50% of the EEG segments were used to train the process. This was to ensure that sufficient signals are available to test the proposed algorithm across all patients. With the remaining segments, 35% of the segments were used for dictionary creation and 15% for dictionary validation. These values were chosen after evaluating three distinct approaches to divide the data segments between the dictionary construction and validation tasks. It was found that having a large dictionary is more advantageous than having more data for dictionary validation or having the same data size for dictionary creation and validation, which were the other two choices studied. The same number of segments were chosen from nonseizure segments, as an equal number of seizure and nonseizure segments were utilized for dictionary development, training, and validation. Additionally, the segmentation process was randomized to eliminate the possibility of over-fitting during the classification stage. See Section 9.8 for additional details on the results obtained and related discussions.

9.6 Decomposition-based Adaptive TFD

The bilinear TFDs described in Section 8.9 can provide improved joint TF resolution at the expense of increased cross-terms of the different signal components. Krishnan et al. [31] presented

the design of an adaptive TFD based on signal decomposition for the analysis of VAG signals; a schematic representation of their procedure is shown in Figure 9.2. In this method, the given signal is first decomposed into components of a specified mathematical representation; the matching pursuit algorithm described in Section 9.3 was used for this purpose. By knowing the components of the signal, the interaction between them can be established and used to remove or prevent cross-terms.

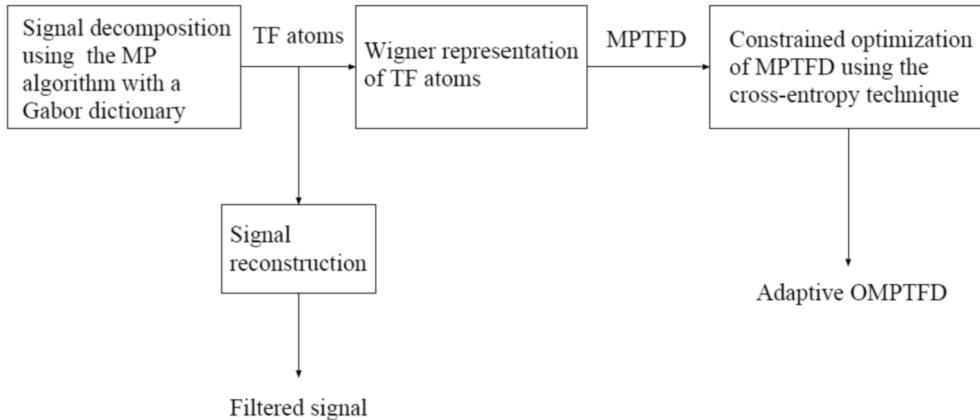


Figure 9.2 Schematic representation of the procedure to construct a decomposition-based adaptive TFD. MP, matching pursuit. Adapted with permission from S. Krishnan, R.M. Rangayyan, G.D. Bell, and C.B. Frank, Adaptive time-frequency analysis of knee joint vibroarthrographic signals for noninvasive screening of articular cartilage pathology, *IEEE Transactions on Biomedical Engineering*, 47(6):773–783, June 2000. ©IEEE.

The TFD based on signal decomposition using the matching pursuit algorithm (MPTFD) is obtained by taking the WVD of the TF atoms used to represent the signal as in Equation 9.7, and is given by [7]

$$\begin{aligned} W(t, \omega) &= \sum_{n=0}^{M-1} |\langle R^n x, g_{\gamma_n} \rangle|^2 W g_{\gamma_n}(t, \omega) \\ &+ \sum_{n=0}^{M-1} \sum_{\substack{m=0 \\ m \neq n}}^{M-1} \langle R^n x, g_{\gamma_n} \rangle \langle R^m x, g_{\gamma_m} \rangle^* W_{[g_{\gamma_n}, g_{\gamma_m}]}(t, \omega), \end{aligned} \quad (9.15)$$

where $W g_{\gamma_n}(t, \omega)$ is the WVD of the Gaussian window function. The double sum corresponds to the cross-terms of the WVD indicated by $W_{[g_{\gamma_n}, g_{\gamma_m}]}(t, \omega)$ and should be rejected in order to obtain a cross-term-free distribution of $x(t)$ in the TF plane. Thus, only the first term is retained, and the resulting TFD is given by

$$W'(t, \omega) = \sum_{n=0}^{M-1} |\langle R^n x, g_{\gamma_n} \rangle|^2 W g_{\gamma_n}(t, \omega). \quad (9.16)$$

Because the MPTFD is constructed using the WVD of TF atoms, most of the properties of the WVD are applicable to the MPTFD, except the marginals.

The MPTFD may be modified to satisfy the marginal requirements while preserving its other important characteristics. One way to optimize the MPTFD is by applying the cross-entropy minimization method [32–34]. Cross-entropy minimization is a general method of inference about an

unknown PDF when there exists a prior estimate of the PDF and new information is available in the form of constraints on its expected values. The optimized MPTFD (or OMPTFD), denoted by $M(t, \omega)$, should satisfy the condition on the marginals, stated as

$$\int M(t, \omega) d\omega = |x(t)|^2 = m(t), \quad (9.17)$$

$$\int M(t, \omega) dt = |X(\omega)|^2 = m(\omega). \quad (9.18)$$

The two conditions given above may be treated as constraints for optimization. The desired $M(t, \omega)$ may be obtained from $W'(t, \omega)$, which is taken to be a prior estimate of the density, by minimizing the cross-entropy between them, given by

$$H(M, W') = \int \int M(t, \omega) \log \left(\frac{M(t, \omega)}{W'(t, \omega)} \right) dt d\omega. \quad (9.19)$$

Considering the marginals, the OMPTFD may be written as [33]

$$M(t, \omega) = W'(t, \omega) \exp\{-[\alpha_0(t) + \beta_0(\omega)]\}, \quad (9.20)$$

where α and β are Lagrange multipliers that may be determined using the constraints in an iterative algorithm as proposed by Loughlin et al. [34].

At the first iteration, let

$$M_1(t, \omega) = W'(t, \omega) \exp[-\alpha_0(t)]. \quad (9.21)$$

Imposing the constraint related to the time marginal given by Equation 9.17 on Equation 9.21, we obtain

$$\alpha_0(t) = \ln \left(\frac{m'(t)}{m(t)} \right), \quad (9.22)$$

where $m(t)$ is the desired time marginal and $m'(t)$ is the time marginal estimated from $W'(t, \omega)$. Then,

$$M_1(t, \omega) = W'(t, \omega) \frac{m(t)}{m'(t)}. \quad (9.23)$$

To have the desired frequency marginal $m(\omega)$, let

$$M_2(t, \omega) = M_1(t, \omega) \exp[-\beta_0(\omega)]. \quad (9.24)$$

By imposing the constraint on the frequency marginal given by Equation 9.18 on Equation 9.24, we obtain

$$\beta_0(\omega) = \ln \left(\frac{m'(\omega)}{m(\omega)} \right), \quad (9.25)$$

where $m(\omega)$ is the desired frequency marginal and $m'(\omega)$ is the frequency marginal estimated from $W'(t, \omega)$. Now, we have

$$M_2(t, \omega) = M_1(t, \omega) \frac{m(\omega)}{m'(\omega)}. \quad (9.26)$$

At this point, $M_2(t, \omega)$ may be altered and need not necessarily give the desired time marginal. Successive iteration could overcome this problem and modify the TFD to get closer to $M(t, \omega)$ [34]. This follows from the fact that the cross-entropy between the desired TFD and the estimated TFD decreases with the number of iterations [33, 34]. Because the iterative procedure is started with a positive distribution, $W'(t, \omega)$, the TFD at the n^{th} iteration, $M_n(t, \omega)$, is guaranteed to be a positive distribution.

Illustrations of application: In the work of Krishnan and Rangayyan [35], several methods were applied to remove noise in multicomponent and nonstationary synthetic signals as well as VAG signals. Gaussian random noise was added to the synthetic signals such that the resulting signals had SNR of 10 dB and 0 dB. The symmlet 4 wavelet [36] was used for wavelet-based filtering. A soft threshold ($T = 0.5$) was applied to the wavelet coefficients. In the case of the wavelet packet method, the best basis was selected by using the Schur concavity cost function [37], and the filtered version was obtained by applying the soft thresholding method ($T = 0.5$) to the wavelet packet coefficients. Gaussian functions were used for the matching pursuit method. The stopping criterion was based on the decay parameter as given by Equation 9.6.

Figure 9.3 (a) shows a synthetic signal and part (b) of the same figure shows the signal with noise added such that $SNR = 0$ dB. The filtered versions of the signal using the wavelet method, the wavelet packet method, and the matching pursuit method are shown in Figures 9.4 and 9.5. Visual comparison indicates that the matching pursuit result has preserved most of the important characteristics of the signal, in particular, the transient component.

The wavelet packet method may lead to better results if the threshold is selected in an optimal manner. The matching pursuit method locally optimizes the choice of the decomposition function for the signal residue at each stage. In the case of a multicomponent signal where different types of structures are located at different times but in the same frequency band, there may be no wavelet packet basis that is well adapted to all of them. Matching pursuit is a translation-invariant method if a translation-invariant dictionary such as a Gabor dictionary is used. See Krishnan and Rangayyan [35] for more details and quantitative evaluation of the results of the procedures discussed in this section.

The VAG signal of a subject with cartilage pathology is shown in Figure 9.6 (a), its filtered version using matching pursuit is shown in part (b) of the same figure, and the difference between the original signal and the filtered result is shown in Figure 9.7. From Figure 9.7, it is evident that a substantial amount of random noise has been removed from the original signal.

The TFD of the original VAG signal in Figure 9.6 computed using the STFT is shown in Figure 9.8 (a); the TFD of the matching-pursuit-filtered version using STFT is shown in part (b) of the same figure. The two TFDs were computed using the same STFT parameters. Tonal and FM components are seen more clearly in the TFD of the filtered signal than that of the original signal. For demonstrations of the use of MPTFDs for feature identification and classification of VAG signals, see Krishnan et al. [31] and Rangayyan and Krishnan [38]; see also Section 9.9.

In the work of Krishnan [39], the synthetic signal shown in Figure 9.9 was prepared by combining overlapping chirp, impulse, and sinusoidal FM components; part (b) of the same figure shows the ideal TFD of the signal (known by design). Figure 9.10 shows a noisy version of the signal with $SNR = 10$ dB and its RSPWVD. Even though the basic nature of the ideal TFD is visible in the RSPWVD, a large amount of interference is present. Figure 9.11 shows the MPTFD of the noisy signal, which clearly demonstrates the various components present in the signal and their TF characteristics without degrading interference.

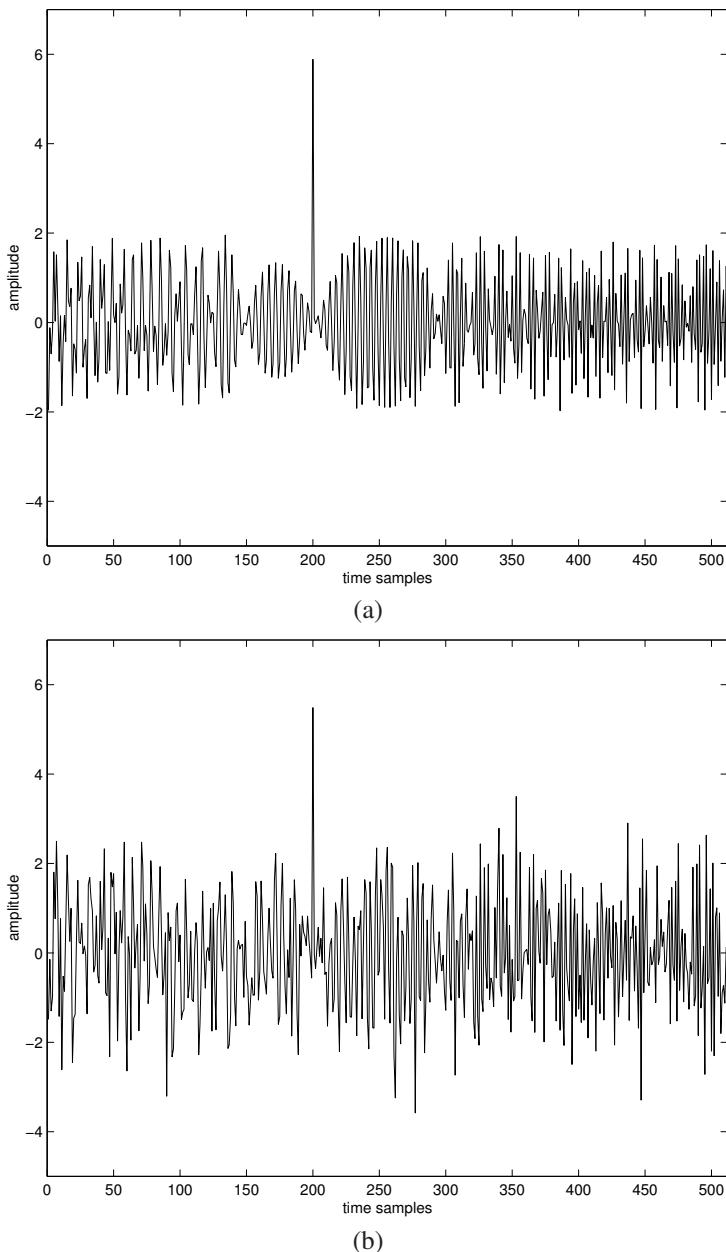


Figure 9.3 (a) Multicomponent and nonstationary synthetic signal with a linear FM, a nonlinear FM, and a transient component. (b) The signal with noise ($SNR = 0 \text{ dB}$). Reproduced from S. Krishnan and R. M. Rangayyan, Automatic de-noising of knee-joint vibration signals using adaptive time-frequency representations, *Medical and Biological Engineering and Computing*, 38(1):2–8, 2000, with kind permission from Springer Science+Business Media B.V. ©IFMBE and Springer.

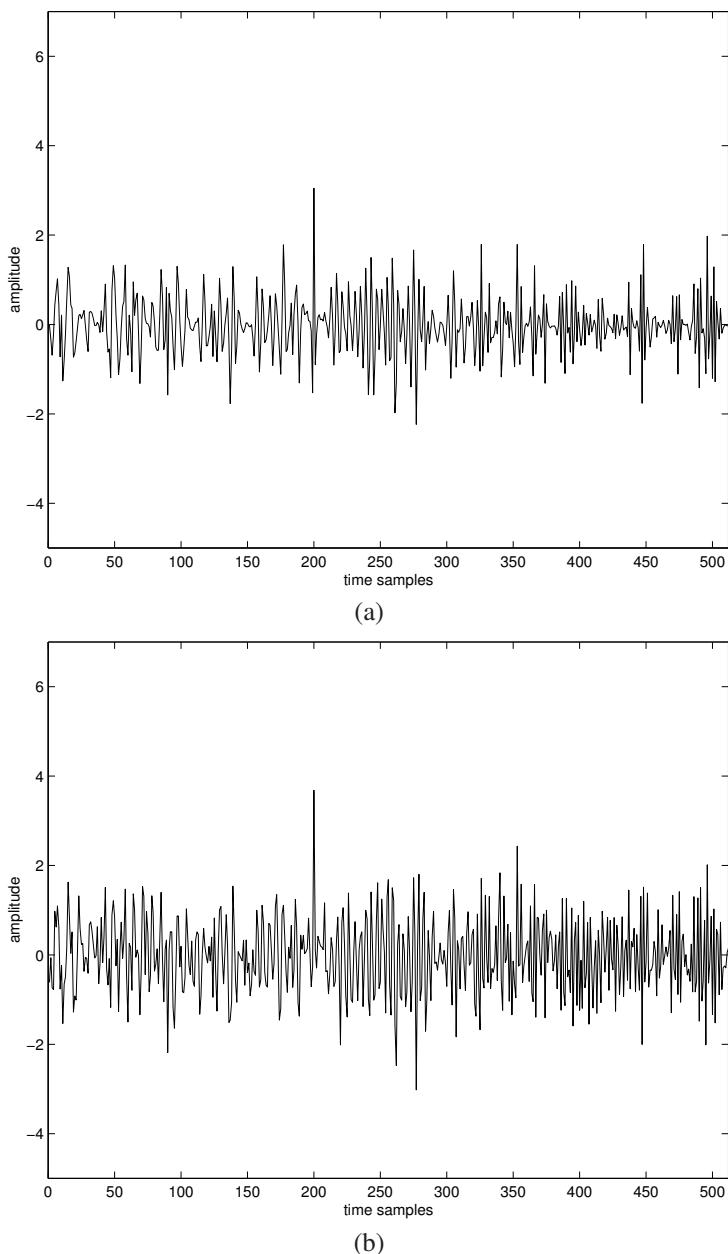


Figure 9.4 Filtered version of the noisy signal in Figure 9.3 (b) using (a) wavelets (symmlet 4) and (b) the wavelet packet method. Reproduced from S. Krishnan and R. M. Rangayyan, Automatic de-noising of knee-joint vibration signals using adaptive time-frequency representations, *Medical and Biological Engineering and Computing*, 38(1):2–8, 2000, with kind permission from Springer Science+Business Media B.V. ©IFMBE and Springer.

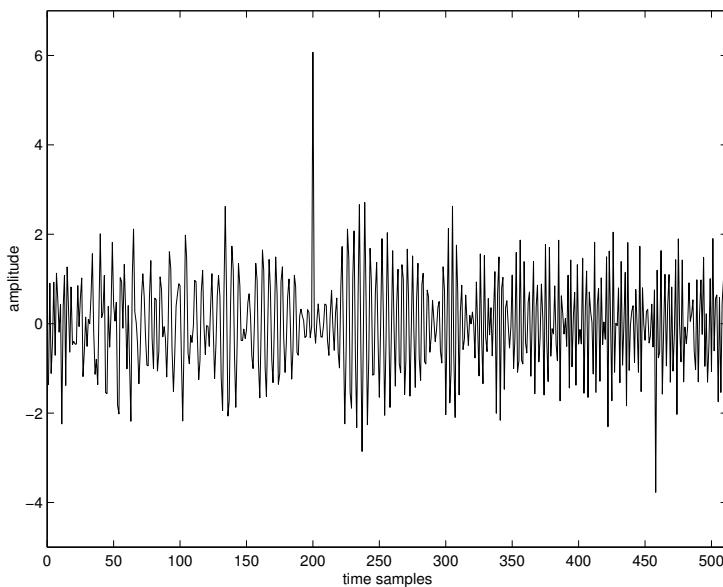


Figure 9.5 Filtered version of noisy signal in Figure 9.3 (b) using the matching pursuit method. Reproduced from S. Krishnan and R. M. Rangayyan, Automatic de-noising of knee-joint vibration signals using adaptive time-frequency representations, *Medical and Biological Engineering and Computing*, 38(1):2–8, 2000, with kind permission from Springer Science+Business Media B.V. ©IFMBE and Springer.

See Section 9.9 for further illustrations and discussion of the application of TFD analysis to VAG signals.

9.7 Separation of Mixtures of Signals

In studies of biomedical signals, it is common to acquire multiple channels of signals from several sources, some of which could provide mutually independent information while others could be correlated and cause redundancy. As shown in Chapter 2, studies of the cardiac, nervous, and neuromuscular systems are often conducted with multiple channels of signals, some of which could be inherently electrical while others could be basically of mechanical, acoustic, and other types, although all are acquired as electrical signals via suitable instrumentation. Regardless of the nature of the signals, cross-talk and cross-contamination can occur at their sources, along their paths of propagation, and at the sites of their acquisition. When analyzing such collections of multichannel signals, questions arise regarding their interrelationships, the potential for existence of correlation, the presence of redundancy in the data, and the quality of the information extracted. It would be desirable to remove redundant and correlated information for the sake of efficiency and to extract information related to each source without contamination from others that are active at the same time. These considerations lead to the need to separate mixtures of signals into their components. Depending upon the assumptions made and the techniques used, one can extract uncorrelated components via principal component analysis (PCA), statistically independent components via methods of independent component analysis (ICA), or local and global components using nonnegative matrix factorization (NMF), as described in the following sections.

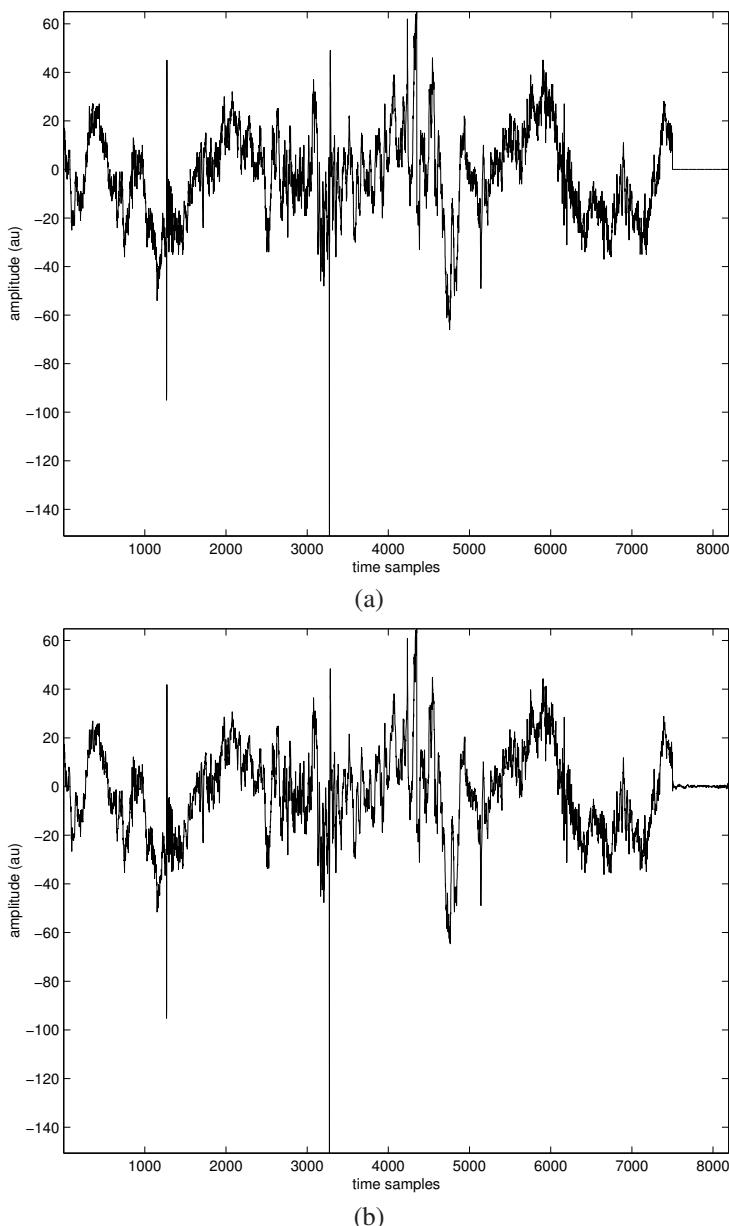


Figure 9.6 (a) VAG signal of a subject with cartilage pathology and (b) filtered version of the signal using matching pursuit. Reproduced from S. Krishnan and R. M. Rangayyan, Automatic de-noising of knee-joint vibration signals using adaptive time-frequency representations, *Medical and Biological Engineering and Computing*, 38(1):2–8, 2000, with kind permission from Springer Science+Business Media B.V. ©IFMBE and Springer.

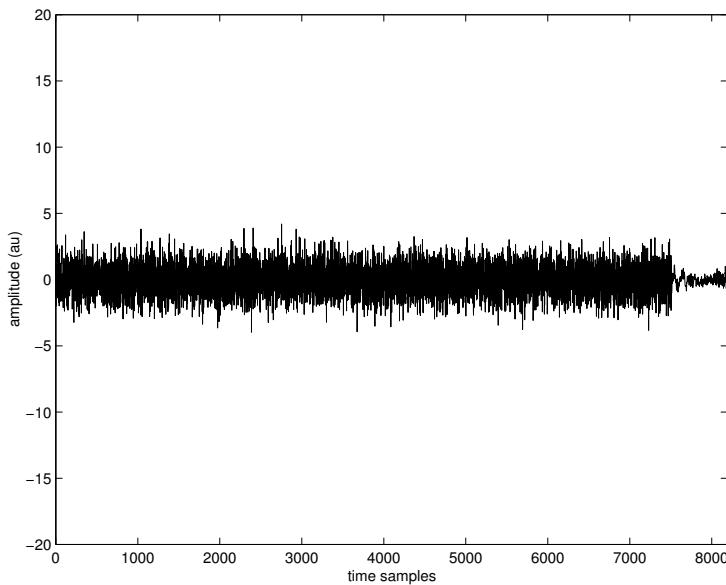


Figure 9.7 Difference between the original VAG signal in Figure 9.6 (a) and the filtered version in Figure 9.6 (b). Reproduced from S. Krishnan and R. M. Rangayyan, Automatic de-noising of knee-joint vibration signals using adaptive time-frequency representations, *Medical and Biological Engineering and Computing*, 38(1):2–8, 2000, with kind permission from Springer Science+Business Media B.V. ©IFMBE and Springer.

9.7.1 Principal component analysis

In order to facilitate efficient representation of collections of signals, measurements, and features for pattern analysis and classification, it is desirable to prepare a compact set of the data with no redundancy or mutual correlation, and to keep the dimension of the result as small as possible. An approach to reduce the dimension of a data vector is PCA, which facilitates the selection of a compact set of uncorrelated components. The PCA method, which is related to the KLT [1, 40, 41], is described briefly in the following paragraphs.

Consider a set of signal samples, measurements, or features represented as a vector \mathbf{y} , of size $K \times 1$. The vector \mathbf{y} may be represented without error by a deterministic linear transformation of the form

$$\mathbf{y} = \mathbf{W} \mathbf{x} = \sum_{k=1}^K x_k \mathbf{W}_k, \quad (9.27)$$

$$\mathbf{W} = [\mathbf{W}_1 \ \mathbf{W}_2 \cdots \mathbf{W}_K], \quad (9.28)$$

where $|\mathbf{W}| \neq 0$, and \mathbf{W}_k are $K \times 1$ column vectors that make up the $K \times K$ matrix \mathbf{W} . The matrix \mathbf{W} should be formulated such that the vector \mathbf{x} , also of size $K \times 1$, facilitates a reduced and efficient representation of the original vector \mathbf{y} .

The matrix \mathbf{W} may be viewed as being composed of K linearly independent column vectors in the K -dimensional space containing \mathbf{y} . Let \mathbf{W} be orthonormal, that is,

$$\mathbf{W}_k^T \mathbf{W}_l = \begin{cases} 1, & k = l, \\ 0, & k \neq l. \end{cases} \quad (9.29)$$

Then,

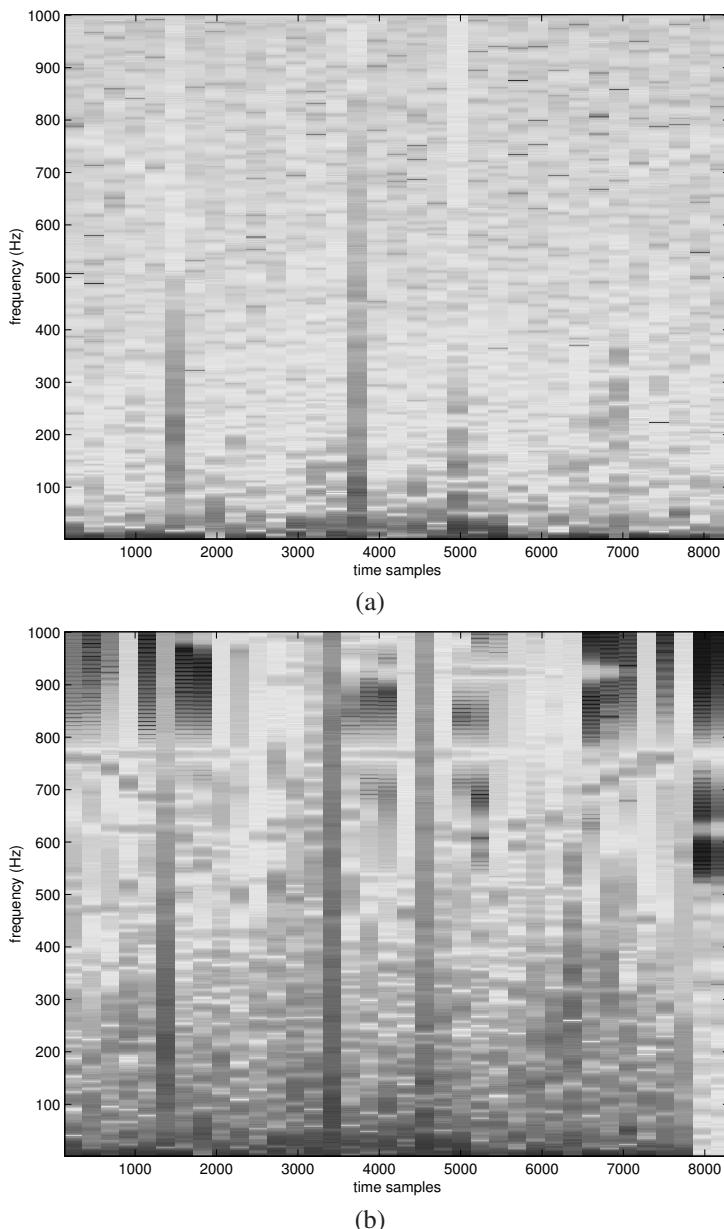


Figure 9.8 (a) TFD of the VAG signal in Figure 9.6 (a) and (b) the TFD of the matching-pursuit-filtered VAG signal in Figure 9.6 (b) computed using the STFT. $f_s = 2,000 \text{ Hz}$. Reproduced from S. Krishnan and R. M. Rangayyan, Automatic de-noising of knee-joint vibration signals using adaptive time-frequency representations, *Medical and Biological Engineering and Computing*, 38(1):2–8, 2000, with permission from Springer Science+Business Media B.V. ©IFMBE and Springer.

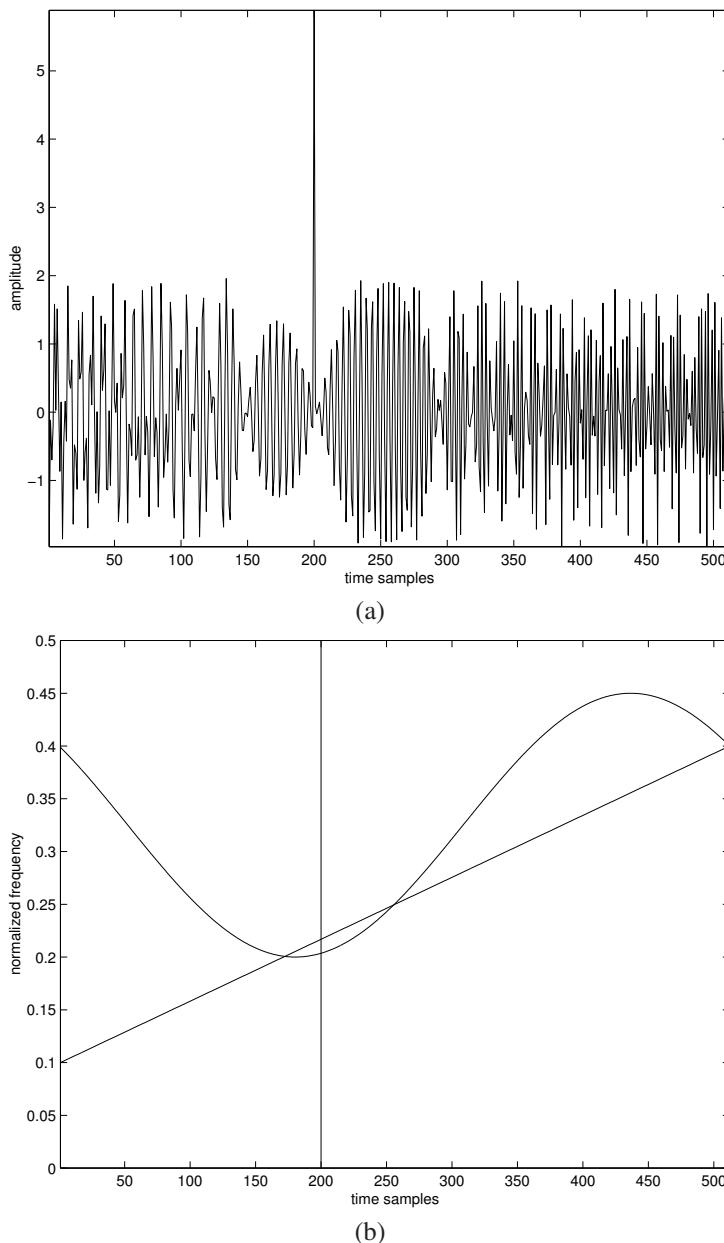


Figure 9.9 (a) Multicomponent and nonstationary synthetic signal consisting of overlapping chirp, impulse, and sinusoidal FM components and (b) its ideal TFD [39].

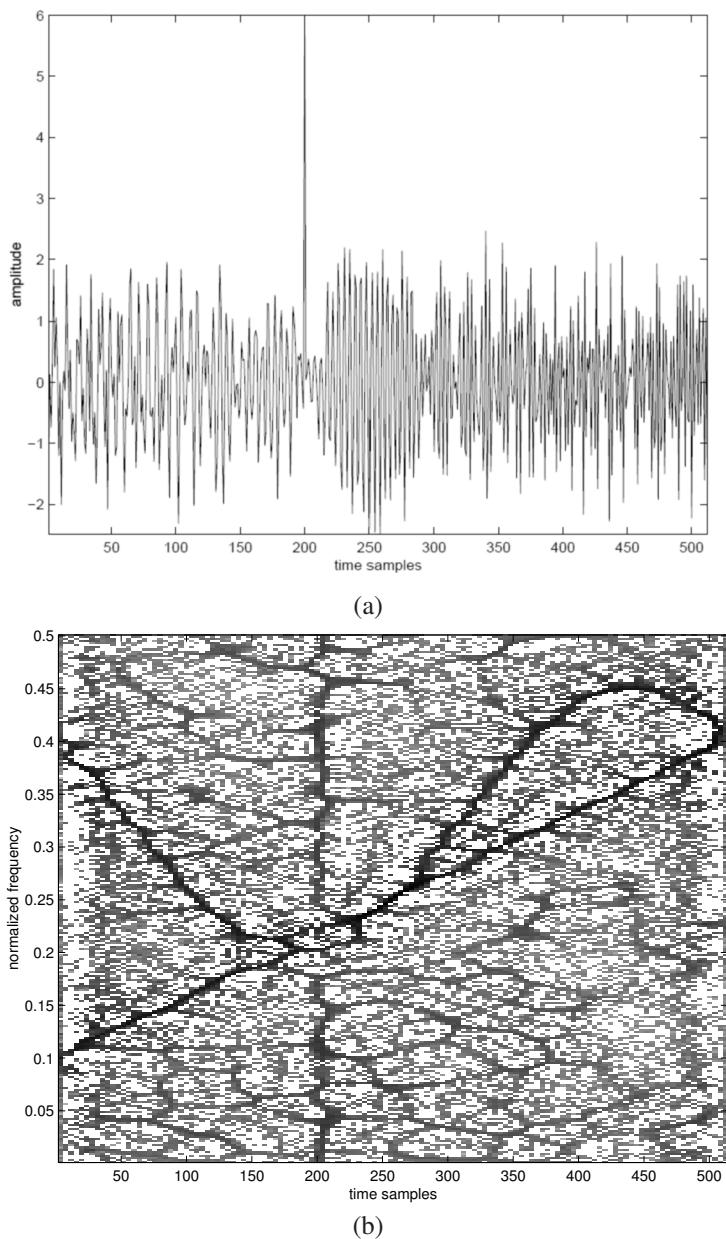


Figure 9.10 (a) The signal in Figure 9.9 (a) with random noise added such that $SNR = 10 dB$. (b) RSPWVD of the noisy signal [39].

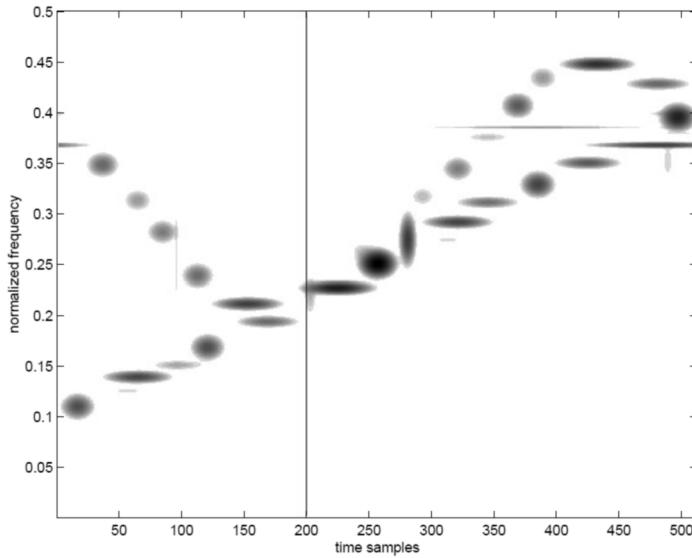


Figure 9.11 MPTFD of the noisy signal in Figure 9.10 (a) [39].

$$\mathbf{W}^T \mathbf{W} = \mathbf{I} \text{ or } \mathbf{W}^{-1} = \mathbf{W}^T. \quad (9.30)$$

The column vectors of \mathbf{W} form a set of orthonormal basis vectors of a linear transformation. The inverse relationship is given by

$$\mathbf{x} = \mathbf{W}^{-1} \mathbf{y} = \sum_{k=1}^K y_k \mathbf{W}_k^{-1}. \quad (9.31)$$

In the representation formulated above, each component of \mathbf{x} contributes to the representation of \mathbf{y} and \mathbf{W}_k^{-1} is the k^{th} $K \times 1$ column vector of the $K \times K$ matrix \mathbf{W}^{-1} . With the design of \mathbf{W} as a linear and reversible transformation, \mathbf{x} provides a complete or lossless representation of \mathbf{y} if all of its K elements are used.

To achieve compact and efficient representation of the original vector \mathbf{y} by extracting the most significant information contained, we may choose to use only $L < K$ components of \mathbf{x} . The omitted components of \mathbf{x} may be replaced with other values \tilde{x}_k , $k = L + 1, L + 2, \dots, K$. We now have an approximation of \mathbf{y} , represented as

$$\tilde{\mathbf{y}} = \sum_{k=1}^L x_k \mathbf{W}_k + \sum_{k=L+1}^K \tilde{x}_k \mathbf{W}_k. \quad (9.32)$$

The error of approximation is

$$\boldsymbol{\varepsilon} = \mathbf{y} - \tilde{\mathbf{y}} = \sum_{k=L+1}^K (x_k - \tilde{x}_k) \mathbf{W}_k. \quad (9.33)$$

The MSE is

$$\begin{aligned}
\overline{\varepsilon^2} &= E[\varepsilon^T \varepsilon] \\
&= E\left[\sum_{k=L+1}^K \sum_{l=L+1}^K (x_k - \tilde{x}_k)(x_l - \tilde{x}_l) \mathbf{W}_k^T \mathbf{W}_l\right] \\
&= \sum_{k=L+1}^K E[(x_k - \tilde{x}_k)^2]. \tag{9.34}
\end{aligned}$$

The last step above is due to the orthonormality of \mathbf{W} .

Taking the derivative of the MSE with respect to \tilde{x}_k and setting the result to zero, we get

$$\frac{\partial \overline{\varepsilon^2}}{\partial \tilde{x}_k} = -2 E[(x_k - \tilde{x}_k)] = 0. \tag{9.35}$$

The optimal MMSE choice for \tilde{x}_k is, therefore,

$$\tilde{x}_k = E[x_k] = \bar{x}_k = \mathbf{W}_k^{-1} E[y_k], \quad k = L+1, L+2, \dots, K. \tag{9.36}$$

This result indicates that the omitted components are replaced by their means estimated for the given population of the vectors \mathbf{y} . The MMSE is given by

$$\begin{aligned}
\overline{\varepsilon^2}_{\min} &= \sum_{k=L+1}^K E[(x_k - \bar{x}_k)^2] \\
&= \sum_{k=L+1}^K E[(\mathbf{W}_k^{-1} y_k - \mathbf{W}_k^{-1} \bar{y}_k)(\mathbf{W}_k^{-1} y_k - \mathbf{W}_k^{-1} \bar{y}_k)^T] \\
&= \sum_{k=L+1}^K E[\mathbf{W}_k^{-1} (y_k - \bar{y}_k)(y_k - \bar{y}_k)^T \{\mathbf{W}_k^{-1}\}^T] \\
&= \sum_{k=L+1}^K E[\mathbf{W}_k^{-1} (y_k - \bar{y}_k)(y_k - \bar{y}_k)^T \mathbf{W}_k] \\
&= \sum_{k=L+1}^K \mathbf{W}_k^T \boldsymbol{\sigma}_y \mathbf{W}_k, \tag{9.37}
\end{aligned}$$

where $\boldsymbol{\sigma}_y = E[(\mathbf{y} - \bar{\mathbf{y}})(\mathbf{y} - \bar{\mathbf{y}})^T]$ is the $K \times K$ covariance matrix of \mathbf{y} .

If the basis vectors \mathbf{W}_k are selected as the eigenvectors of $\boldsymbol{\sigma}_y$, that is,

$$\boldsymbol{\sigma}_y \mathbf{W}_k = \lambda_k \mathbf{W}_k, \tag{9.38}$$

where

$$\lambda_k = \mathbf{W}_k^T \boldsymbol{\sigma}_y \mathbf{W}_k \tag{9.39}$$

are the corresponding eigenvalues, then we have

$$\overline{\varepsilon^2}_{\min} = \sum_{k=L+1}^K \lambda_k. \tag{9.40}$$

Therefore, the MSE may be minimized by ordering the eigenvectors such that the corresponding eigenvalues are arranged in decreasing order, that is, $\lambda_1 > \lambda_2 > \dots > \lambda_K$. Then, if a component x_k of \mathbf{x} is replaced by $\tilde{x}_k = \bar{x}_k$, the MSE increases by λ_k . By replacing the components of \mathbf{x} corresponding to the eigenvalues at the lower end of the list by their respective mean values, the MSE is kept at its minimum for a chosen number of components L .

From the procedure and the properties described above, it is evident that the components of \mathbf{x} are mutually uncorrelated:

$$\boldsymbol{\sigma}_x = \mathbf{W}^T \boldsymbol{\sigma}_y \mathbf{W} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_K \end{bmatrix} = \boldsymbol{\Lambda}, \quad (9.41)$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues λ_k placed along its diagonal. Because the eigenvalues λ_k are equal to the variances of x_k , a selection of the larger eigenvalues implies the selection of the transformed components with the higher variance or information content across the ensemble of the data vectors considered.

Illustration of application: In a study reported by Mesin et al. [42], surface EMG signals were recorded from the biceps brachii muscle. The electrodes were placed 5 mm apart and two channels of EMG were acquired from approximately the same group of muscle fibers. Therefore, the two EMG signals, shown in Figure 9.12, parts (a) and (b), are expected to be highly correlated. The correlation between the two EMG signals is demonstrated in part (c) of the figure, which also shows the directions of the two principal components (eigenvectors). The two principal components of the EMG signals are shown in parts (d) and (e) of the figure; whereas the first component is close in appearance to the two EMG signals, the second component could be considered to be random noise. By design, the two principal components are uncorrelated. Due to the high degree of correlation between the two input signals, there is considerable redundancy present in them. PCA has extracted most of the power in the signals within the first principal component, which may be considered to be an adequate and redundancy-free representation of both input channels. In this manner, PCA can achieve data reduction and efficient representation of information extracted from multiple measurements. Although the two signals in the present example represent activity of the same muscle or source, the method may be applied to analyze signals from multiple sources.

PCA is commonly used to achieve reduction in the dimension of feature vectors to facilitate efficient pattern classification. Variations of PCA include sparse PCA [43], nonlinear PCA [44], and robust PCA (RPCA) [45]. PCA also has strong relation to other methods, such as K-means clustering [46], factor analysis [47], and correspondence analysis [48].

9.7.2 Independent component analysis

Let us consider the situation where we have an observation with K channels of signals, labeled as $y_1(n), y_2(n), \dots, y_K(n)$. The observation at one instant of time may be expressed as the $K \times 1$ vector $\mathbf{y}(n) = [y_1(n), y_2(n), \dots, y_K(n)]^T$. Suppose that the observed signals are being produced by L sources $x_1(n), x_2(n), \dots, x_L(n)$ that are mutually statistically independent. Let us also assume that we are able to gather more than the required number of observations to determine the sources, that is, the system is overdetermined, with $K \geq L$. Then, letting the $L \times 1$ vector $\mathbf{x}(n) = [x_1(n), x_2(n), \dots, x_L(n)]^T$ represent the values of the sources at the instant of time n , we have

$$\mathbf{y}(n) = \mathbf{M} \mathbf{x}(n) + \eta(n), \quad (9.42)$$

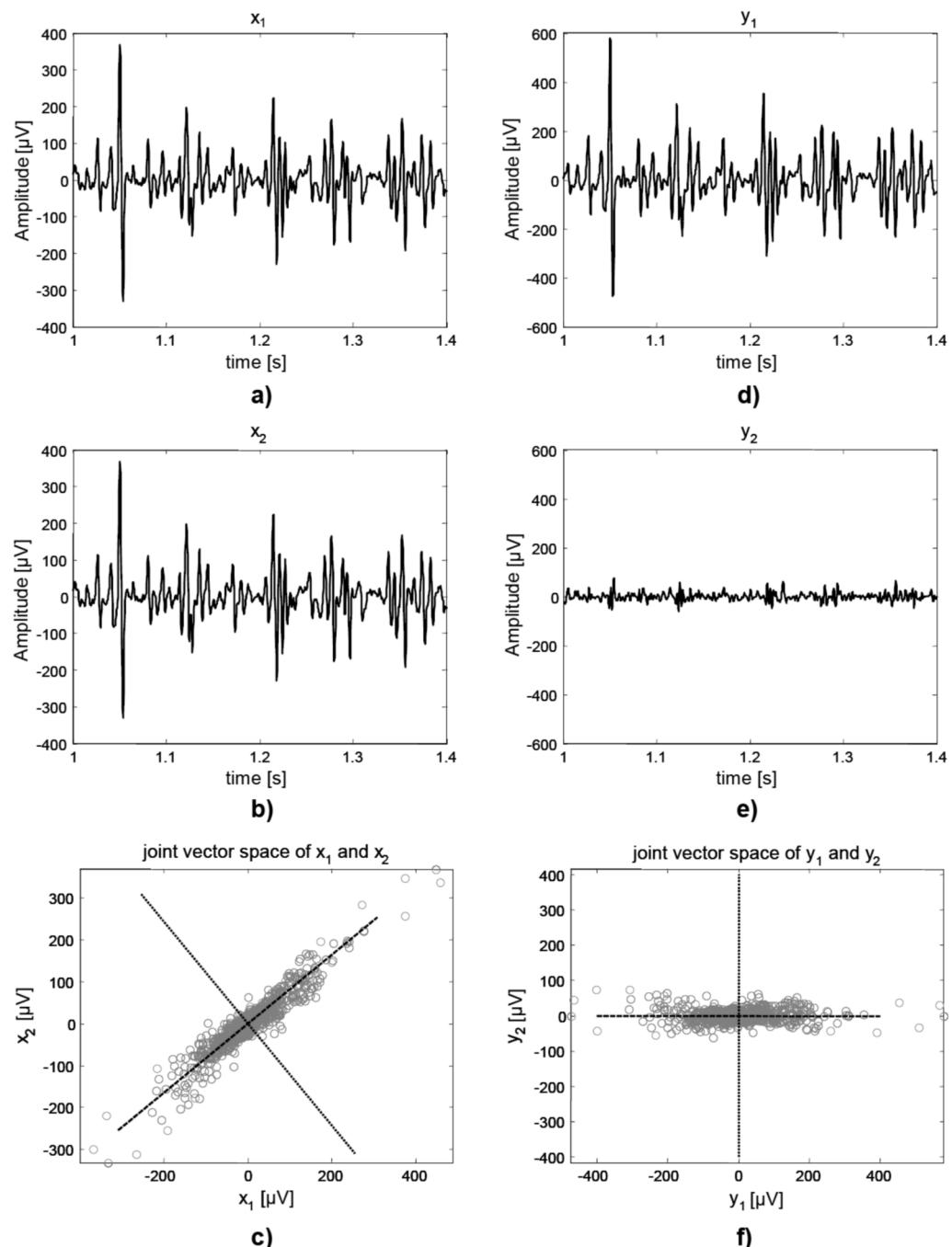


Figure 9.12 Illustration of PCA of correlated EMG signals. (a) and (b) Two channels of EMG signals from the biceps brachii muscle. (c) Directions of the principal components and demonstration of the correlation between the two EMG channels. (d) and (e) Principal components extracted via PCA. (f) Joint vector space of the principal components. Reproduced with permission from L. Mesin, A. Holobar, and R. Merletti, Blind source separation: Application to biomedical signals, Chapter 15 in *Advanced Methods of Biomedical Signal Processing*, IEEE and Wiley, 2011, Edited by S. Cerutti and C. Marchesi, pp 379–409. ©IEEE.

where η is random noise that is assumed to be independent of \mathbf{y} and \mathbf{x} . This model relates the K observed values in $\mathbf{y}(n)$ at the instant of time n to the corresponding values of the L sources in $\mathbf{x}(n)$ through the $K \times L$ mixing matrix \mathbf{M} . The problem in ICA and BSS is to derive the values of \mathbf{x} from the given observations \mathbf{y} without knowing \mathbf{M} (hence the term “blind” in BSS). We may assume that there is no noise in the observed data or let η be one of the sources contained in \mathbf{x} , and simplify Equation 9.42 to

$$\mathbf{y}(n) = \mathbf{M} \mathbf{x}(n). \quad (9.43)$$

PCA, as described in Section 9.7.1, facilitates the decomposition of a data vector into uncorrelated components. If the sources of the data or signal components have Gaussian PDFs, the property of being uncorrelated implies that the components are statistically independent; this, however, is not always true. PCA also does not assume any mixing matrix, and characterizes the redundancy in the input using its covariance matrix. Thus, PCA is based on second-order statistics.

ICA achieves separation of the sources by assuming them to be statistically independent and not only being uncorrelated as in PCA. Various measures of statistical independence can be used in optimization procedures to estimate a separation or “unmixing” matrix \mathbf{W} (of size $L \times K$) such that estimates of the original source signals can be obtained as

$$\tilde{\mathbf{x}}(n) = \mathbf{W} \mathbf{y}(n). \quad (9.44)$$

The central limit theorem [49] states that the PDF of a linear combination of several statistically independent random variables tends towards a Gaussian PDF regardless of the PDFs of the individual variables. Therefore, the mixed observed signal \mathbf{y} in Equation 9.43 is likely to have a PDF that is close to a Gaussian regardless of the PDFs of the sources of the input signals \mathbf{x} . One approach to achieve BSS is to employ means to minimize the Gaussian characteristics of the reconstructed or estimated signals $\tilde{\mathbf{x}}$. The kurtosis excess (see Section 3.2.1) of a Gaussian PDF is zero: This property may be used to construct a function $F(\tilde{\mathbf{x}})$ that can be used in an iterative optimization procedure to estimate the separation matrix as [42]

$$\tilde{\mathbf{W}}_{(n+1)} = \tilde{\mathbf{W}}_n - \mu_n \nabla F(\tilde{\mathbf{x}}), \quad (9.45)$$

where n is the iteration number, μ is the learning rate, and ∇ is a gradient measure. This procedure is a gradient-descent approach for optimization (see Section 3.10.2). Several other algorithms exist for ICA and BSS [42, 50–64]; some BSS methods use PCA as a preprocessing step.

Illustration of application: Zarzoso and Nandi [65] performed a comparative analysis of BSS with adaptive filtering (see Section 3.10.1) for the purpose of extraction of the fetal ECG from maternal ECG recordings. Figure 9.13 shows the input ECG signals used in their work, including eight channels of the maternal ECG obtained using abdominal and thoracic surface electrodes. The result of BSS is shown in Figure 9.14. Although BSS separates the mixed input signals and provides independent components, the method, by itself, does not identify the various sources that correspond to the components. However, knowing that the fetal heart rate is typically higher than the maternal heart rate, one can determine, by visual inspection, that the fifth and seventh traces from the top in Figure 9.14 represent the extracted fetal ECG with 24 beats over the interval of the recordings. Similarly, the first, second, and fourth traces can be taken to be representations of the maternal ECG (with 14 beats over the same interval) without interference from the fetal ECG. (The first trace in Figure 9.13 clearly shows the fetal and maternal QRS complexes, some of which are distinct and some are overlapped.) The third, sixth, and eighth traces in Figure 9.14 do not show any QRS complexes with a regular rhythm and, hence, could be considered to be noise. If the derivation of the fetal heart rate is the goal of the recording, it can be easily achieved by using one of the extracted fetal ECG channels. The maternal heart rate may also be estimated from one of the extracted maternal ECG channels without interference from the fetal ECG.

It should be observed that BSS could lead to multiple representations of a single source in the resulting output channels. In the present example, the electrical activity of the maternal heart is shown in three output channels and that of the fetal heart is shown in two channels; the user is required to analyze the results and select appropriate channels for further analysis. (Indeed, all of the eight input channels demonstrate various views of the maternal ECG, with some of them including the fetal ECG as well.) Zarzoso and Nandi [65] showed that BSS performs better than adaptive filtering in the extraction of fetal ECG from maternal ECG signals. See Callaerts et al. [66], Vanderschoot et al. [67], and De Lathauwer et al. [54] for other methods for the same application.

Fast ICA (FICA) [68] is a computationally efficient implementation of ICA in which the algorithm could be implemented using parallel processors with lower memory requirements compared to the standard ICA algorithm. See Section 9.11 for a method to extract the fetal ECG from a single-channel recording of the maternal ECG.

9.7.3 Nonnegative matrix factorization

NMF is a form of matrix decomposition technique used for 2D image analysis [69]. It is possible to represent a 1D signal as a 2D covariance matrix or as a TF matrix by time-domain operations and TF transformations; then, NMF may be applied for signal analysis.

Standard NMF [69, 70] is a technique to find a low-rank approximation of a given nonnegative matrix \mathbf{V} . Assume that \mathbf{V} is a matrix with size $M \times N$. The NMF problem entails determining two nonnegative matrices whose product is close to the nonnegative matrix \mathbf{V} . Let \mathbf{V} be approximated by the product of the matrices \mathbf{W} and \mathbf{H} of size $M \times r$ and $r \times N$, respectively, with $r < M$ and $r < N$:

$$\mathbf{V} = \mathbf{W} \mathbf{H}. \quad (9.46)$$

The decomposition provides two nonnegative matrices, \mathbf{W} and \mathbf{H} . The typical method of computing the product of two matrices as above is to obtain the $(i, j)^{\text{th}}$ element of \mathbf{V} as the dot product of the i^{th} row of \mathbf{W} (of size $1 \times r$) with the j^{th} column of \mathbf{H} (of size $r \times 1$), and do the same for $i = 1, 2, \dots, M$ and $j = 1, 2, \dots, N$:

```

for i = 1, M
    for j = 1, N
        for k = 1, r
            v(i, j) = v(i, j) + w(i, k) h(k, j)
        end
    end
end

```

(9.47)

By changing the order of accessing the elements of the matrices, that is, by changing the order of the three for loops given above, different implementations and their interpretations may be obtained [71, 72]. One such variation of interest in the present topic is

$$\mathbf{V} = \sum_{i=1}^r \mathbf{W}_i \mathbf{H}_i = \sum_{i=1}^r \mathbf{V}(i), \quad (9.48)$$

where \mathbf{W}_i is the i^{th} column of \mathbf{W} (of size $M \times 1$) and \mathbf{H}_i is the i^{th} row of \mathbf{H} (of size $1 \times N$). The product $\mathbf{W}_i \mathbf{H}_i$, referred to as an outer product (an inner product or dot product is not feasible in this case, unless $M = N$), provides an $M \times N$ submatrix, expressed as $\mathbf{V}(i)$ in Equation 9.48. Thus, the matrix \mathbf{V} is decomposed into r submatrices $\mathbf{V}(i)$, $i = 1, 2, \dots, r$. Depending upon the application and the nature and composition of the original matrix \mathbf{V} , various interpretations may be made of the

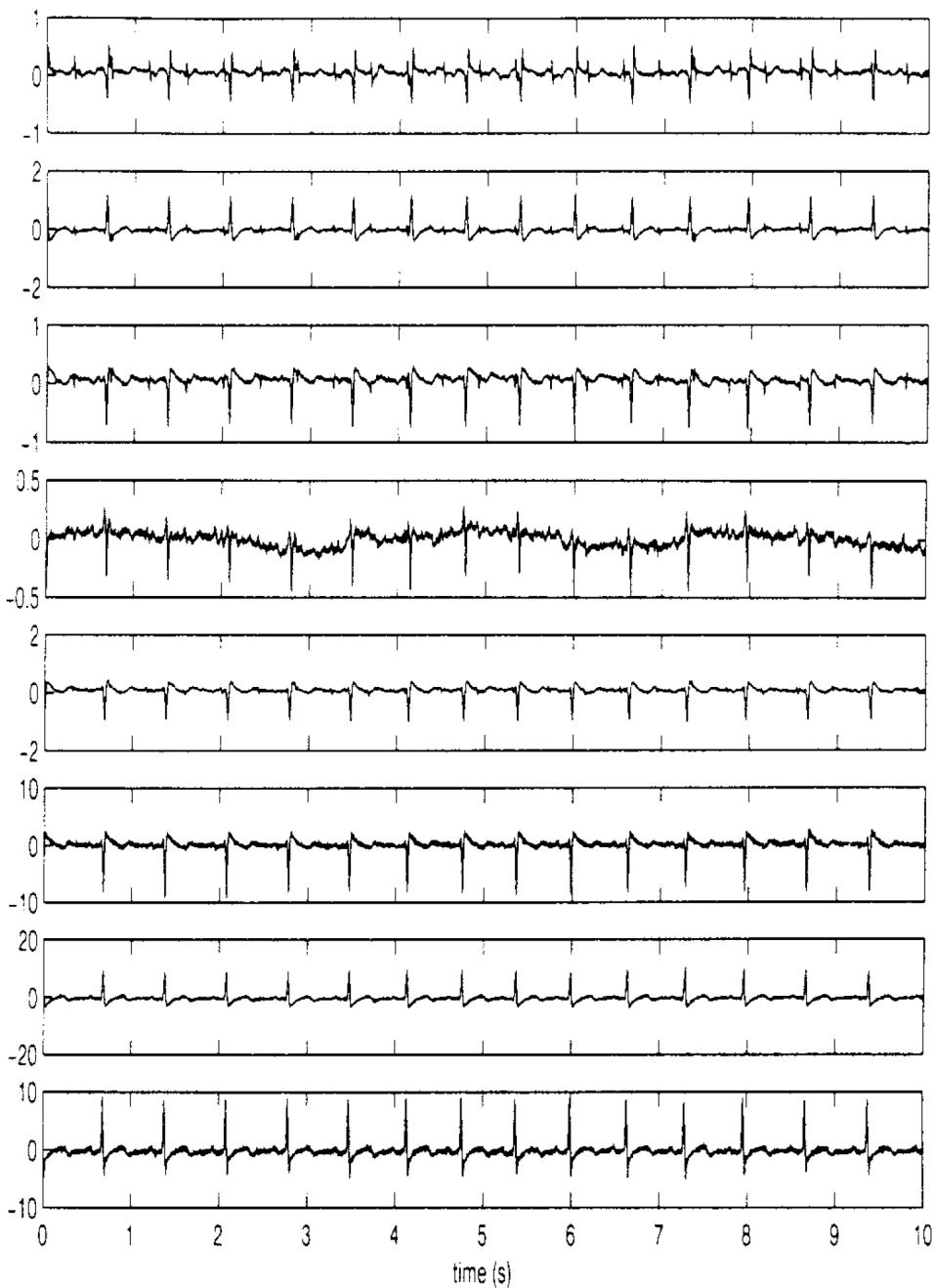


Figure 9.13 Eight channels of maternal ECG signals using abdominal and thoracic surface electrodes. Reproduced with permission from V. Zarzoso and A.K. Nandi, Noninvasive fetal electrocardiogram extraction: blind separation versus adaptive noise cancellation, *IEEE Transactions on Biomedical Engineering*, 48(1):12–18, 2001. ©IEEE.

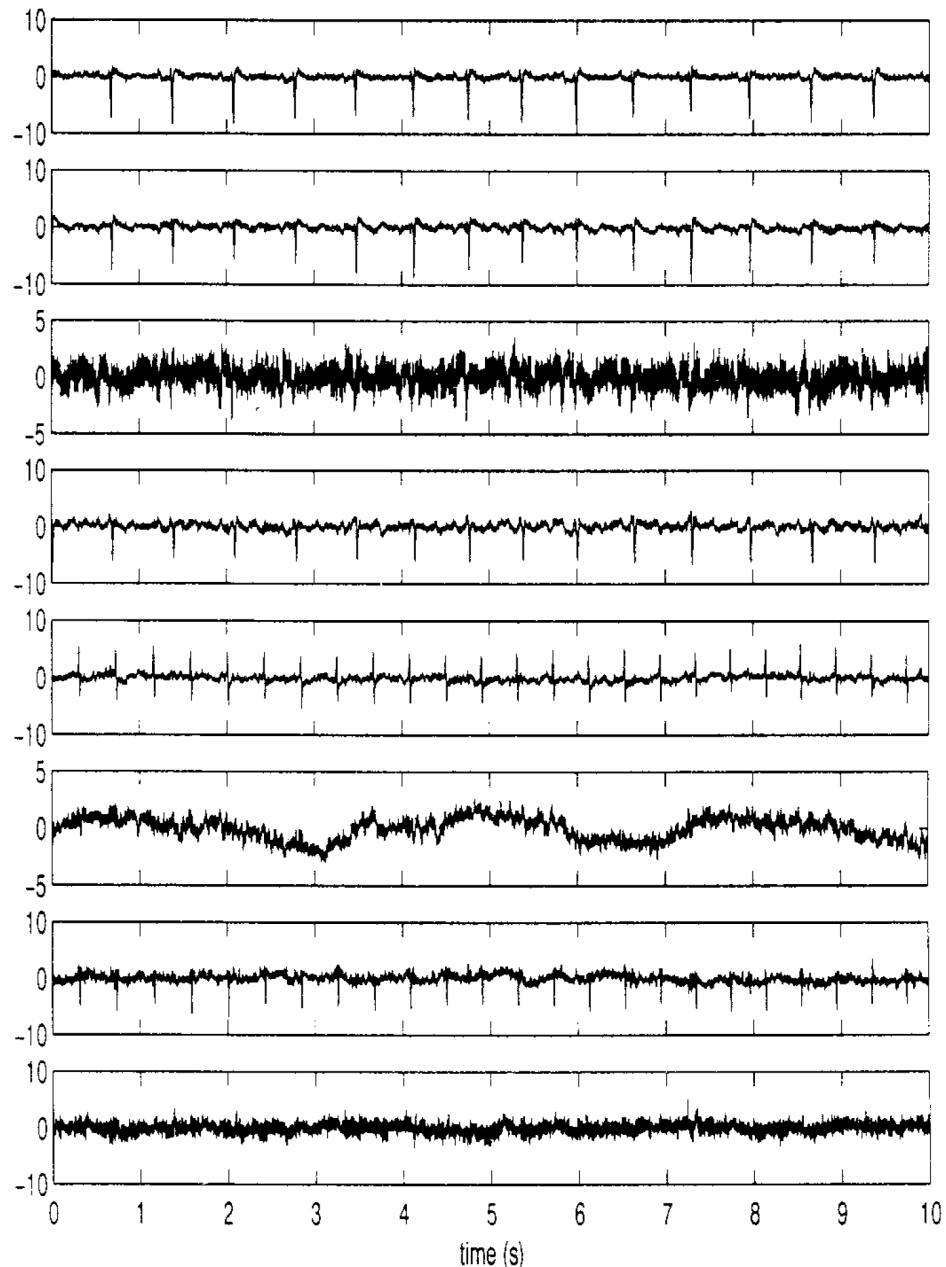


Figure 9.14 Results of BSS showing eight independent components. The fifth and seventh traces from the top show the extracted fetal ECG with 24 beats over the 10-s interval of the recordings. The first, second, and fourth traces show the maternal ECG with 14 beats over the same interval. The remaining traces represent noise. Reproduced with permission from V. Zarzoso and A.K. Nandi, Noninvasive fetal electrocardiogram extraction: blind separation versus adaptive noise cancellation, *IEEE Transactions on Biomedical Engineering*, 48(1):12–18, 2001. ©IEEE.

decomposition process and the components of the matrices \mathbf{W} and \mathbf{H} used in the process. In NMF and its applications, the columns of \mathbf{W} are typically considered to be the basis vectors and the rows of \mathbf{H} are considered to be the corresponding weights or activations. The submatrices $\mathbf{V}(i)$ could also be interpreted as suited for the application; specific submatrices may be selected for particular representations or interpretations. (*Note:* If the i^{th} row of \mathbf{H} is expressed as an $N \times 1$ vector, the product in Equation 9.48 would have to be expressed as $\mathbf{W}_i \mathbf{H}_i^T$.)

A variety of cost functions may be used to assess the success of the reconstruction process. Lee and Seung [69] investigated the squared error and KLD functions for NMF; each of these results in a unique NMF algorithm. By applying gradient descent to optimize the cost function and selecting a suitable step size, \mathbf{W} and \mathbf{H} may be estimated using iterative updating rules. For the squared-error version of NMF, $\varepsilon = \sum_{i,j} (V_{ij} - \hat{V}_{ij})^2$, where \hat{V} is an estimate of V . Then, the multiplicative update, for \mathbf{W} and \mathbf{H} may be given as

$$\mathbf{H} \leftarrow \mathbf{H} \odot ((\mathbf{W}^T \mathbf{V}) \oslash (\mathbf{W}^T \mathbf{W} \mathbf{H})), \quad (9.49)$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot ((\mathbf{V} \mathbf{H}^T) \oslash (\mathbf{W} \mathbf{H} \mathbf{H}^T)). \quad (9.50)$$

The mathematical operations \odot and \oslash in Equations 9.49 and 9.50 represent element-wise multiplication and element-wise division, respectively. Given matrices \mathbf{A} , \mathbf{B} , and \mathbf{C} , all of size $M \times N$, for $\mathbf{C} = \mathbf{A} \odot \mathbf{B}$, we have

$$C_{ij} = A_{ij} \times B_{ij}, \quad (9.51)$$

for $i = 1, 2, \dots, M$ and $j = 1, 2, \dots, N$. Similarly, for $\mathbf{C} = \mathbf{A} \oslash \mathbf{B}$, we have

$$C_{ij} = A_{ij} / B_{ij}. \quad (9.52)$$

For the KLD version of NMF, denoted as

$$D_I(\mathbf{V} \| \hat{\mathbf{V}}) = \sum_{ij} \left(V_{ij} \log \frac{V_{ij}}{\hat{V}_{ij}} - V_{ij} + \hat{V}_{ij} \right), \quad (9.53)$$

the submatrices \mathbf{W} and \mathbf{H} can be multiplicatively updated as [73, 74]

$$\mathbf{W} \leftarrow \mathbf{W} \odot \left(\{[\mathbf{V} \oslash (\mathbf{W} \mathbf{H})] \mathbf{H}^T\} \oslash (\mathbf{J} \mathbf{H}^T) \right), \quad (9.54)$$

where \mathbf{J} is an $M \times N$ matrix will all elements equal to unity [73].

$$\mathbf{H} \leftarrow \mathbf{H} \odot (\{\mathbf{W}^T [\mathbf{V} \oslash (\mathbf{W} \mathbf{H})]\} \oslash (\mathbf{W}^T \mathbf{J})). \quad (9.55)$$

The least-squares error approach is a classic bound-constrained optimization method. The KLD formula [75], on the other hand, is not a bound-constrained problem, which means that the objective function does not have to be well-defined at any point within the bounded region. As the log function is not well-defined if any element in the matrix \mathbf{V} or \mathbf{WH} is zero, KLD should be used with caution.

Various alternative minimization strategies have been proposed for NMF [76]. For its ease of implementation, the projected-gradient-bound-constrained optimization method proposed by Lin [77] is discussed here. The optimization method is performed on the function $f = \mathbf{V} - \mathbf{WH}$ and is composed of three steps:

(1) *Updating the matrix \mathbf{W} :* During this stage, the optimization problem of $f_{\mathbf{H}}(\mathbf{W})$ is solved with respect to \mathbf{W} , where $f_{\mathbf{H}}(\mathbf{W})$ is the function $f = \mathbf{V} - \mathbf{WH}$, and the matrix \mathbf{H} is assumed to be constant. During every iteration, \mathbf{W} is updated as

$$\mathbf{W}^{n+1} = \max \{[\mathbf{W}^n - \mu^n \nabla f_{\mathbf{H}}(\mathbf{W}^n)], 0\}, \quad (9.56)$$

where n is the iteration order, μ^n is the step size to update the matrix \mathbf{W} , $\nabla f_{\mathbf{H}}(\mathbf{W})$ is the projected gradient of the function f , and \mathbf{H} is constant. The step size is defined as $\mu^n = \beta^{K_n}$, where β^1, β^2, \dots , are the possible step sizes, and K_n is the first nonnegative integer for which

$$f(\mathbf{W}^{n+1}) - f(\mathbf{W}^n) \leq \sigma \langle \nabla f_{\mathbf{H}}(\mathbf{W}^n), \mathbf{W}^{n+1} - \mathbf{W}^n \rangle, \quad (9.57)$$

where the operator $\langle ., . \rangle$ is the inner product between two matrices:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_i \sum_j A_{ij} B_{ij}. \quad (9.58)$$

In the work of Lin [77], the values of σ and β are suggested to be 0.01 and 0.1, respectively. Once the suitable step size, μ^n , is found, the condition on stationarity of $f_{\mathbf{H}}(\mathbf{W})$ with the updated matrix is checked as

$$\| \nabla^P f_{\mathbf{H}}(\mathbf{W}^{n+1}) \| \leq \epsilon \| \nabla f_{\mathbf{H}}(\mathbf{W}^1) \|, \quad (9.59)$$

where $\| \nabla f_{\mathbf{H}}(\mathbf{W}^1) \|$ is the the projected gradient of the function $f_{\mathbf{H}}(\mathbf{W})$ at the first iteration ($n = 1$), ϵ is a small tolerance value, and $\nabla^P f_{\mathbf{H}}(\mathbf{W})$ is the projected gradient, defined as

$$\nabla^P f_{\mathbf{H}}(\mathbf{W}) = \begin{cases} \nabla f_{\mathbf{H}}(\mathbf{W}), & W_{mr} > 0, \\ \min(0, \nabla f_{\mathbf{H}}(\mathbf{W})), & W_{mr} = 0, \end{cases} \quad (9.60)$$

where W_{mr} is the $(m, r)^{\text{th}}$ element of \mathbf{W} . If the stationarity condition is met, the procedure stops; if not, the optimization process is repeated until the point \mathbf{W}^{n+1} becomes a stationary point of $f_{\mathbf{H}}$.

(2) *Updating the matrix \mathbf{H} :* This stage solves the optimization problem with respect to \mathbf{H} while assuming \mathbf{W} is constant. A procedure similar to that in Step 1 is repeated for this purpose.

(3) *The convergence test:* Once the suboptimal problems described above are solved, the stationarity of the \mathbf{W} and \mathbf{H} solutions together is checked as

$$\| \nabla f_{\mathbf{H}}(\mathbf{W}^n) \| + \| \nabla f_{\mathbf{W}}(\mathbf{H}^n) \| \leq \epsilon (\| \nabla f_{\mathbf{H}}(\mathbf{W}^1) \| + \| \nabla f_{\mathbf{W}}(\mathbf{H}^1) \|). \quad (9.61)$$

The optimization is complete if the global convergence rule given above is satisfied; otherwise, Steps 1 and 2 are repeated until the optimization is complete.

The gradient-based NMF is computationally competitive and offers better convergence properties than the standard approach. Variants of the standard NMF include: sparse NMF [78] (discussed in Section 9.11), orthogonal NMF [79], convolutional NMF [80], and kernel NMF [81].

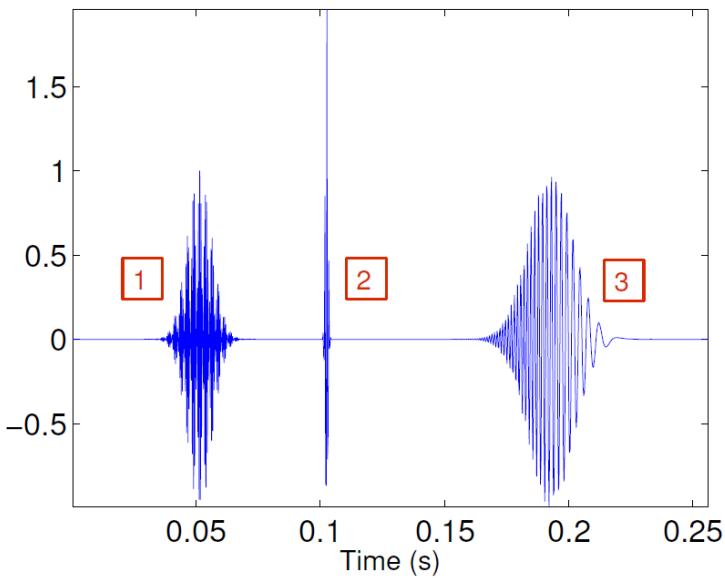
9.7.4 Comparison of PCA, ICA, and NMF

In this section, the most appropriate matrix decomposition technique for signal analysis is investigated. The performance of the three well-known matrix decomposition techniques, as explained in the preceding subsections, is analyzed. In the first experiment, the method that provides the most accurate decomposition is investigated. In the second experiment, the TF features obtained using different matrix decomposition methods are compared as related to their TF localization performance. The experiments provide guidelines for the selection of the appropriate matrix decomposition technique for source separation and signal analysis applications.

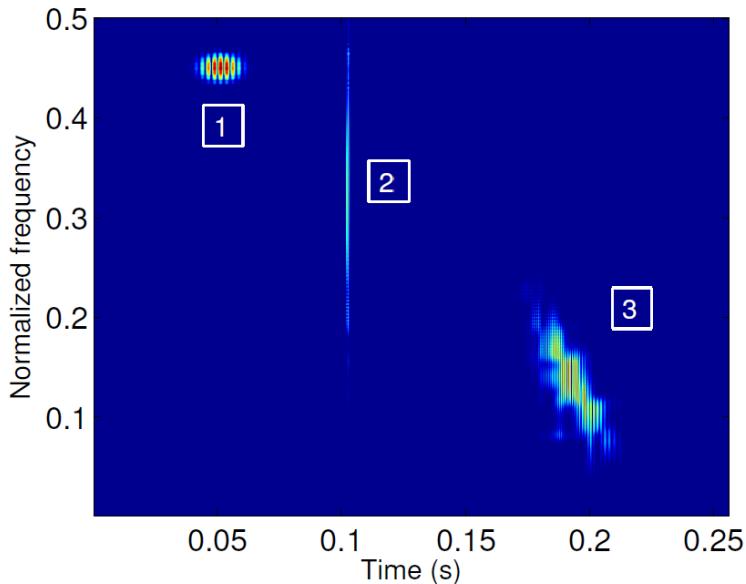
A synthetic signal was generated as

$$x(t) = \sum_{j=1}^3 x_j(t) = \sum_{j=1}^3 \alpha_j g(\sigma_j, \mu_j) \sin(2\pi f_j t), \quad (9.62)$$

where $g(\sigma_j, \mu_j)$ is a Gaussian function with mean μ and standard deviation σ , α_j represents the weight of the Gaussian function, and f_j is the frequency of a sinusoidal wave. Figure 9.15 depicts



(a) Synthetic signal in time domain.



(b) TFD of the synthetic signal.

Figure 9.15 (a) A synthetic signal with three components; from left to right, the components are frequency-localized (FL), transient, and FM. (b) The TF matrix constructed using MPTFD. The FM component has five subcomponents. Reproduced with permission from B. Ghoraani [82].

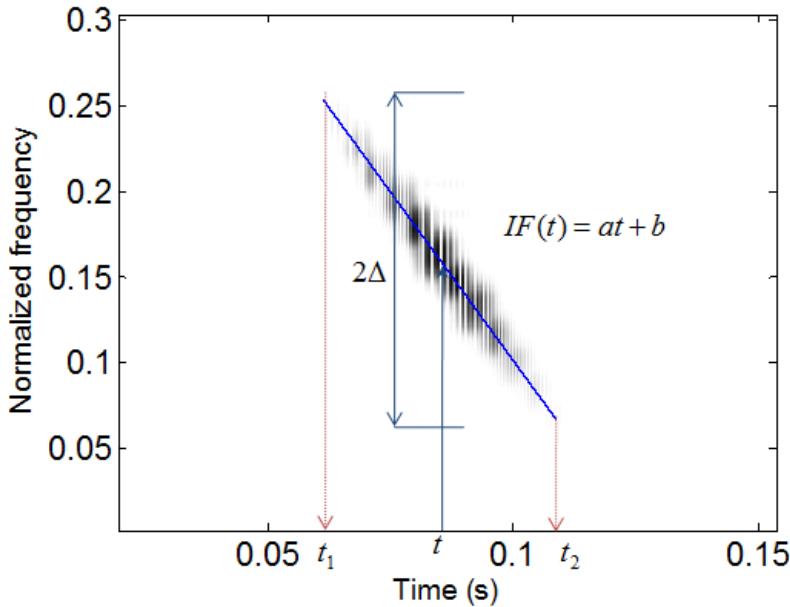


Figure 9.16 Each component in the synthetic signal is defined with four parameters: start time t_1 , ending time t_2 , and the parameters of the instantaneous frequency $IF(t)$ function a and b . Reproduced with permission from B. Ghoraani [82].

the synthetic signal and its MPFTD. The MPFTD shows the frequency-localized (FL), transient, and FM components of the synthetic signal. Note that the FM component, which is a chirp signal, is composed of five subcomponents in the MPFTD representation.

Next, the extracted basis and the coefficient matrices were used to reconstruct the TF matrix, denoted by \mathbf{V}_r . Then, using the reconstructed matrix and the previously known instantaneous frequency of the signal, a figure of merit was computed to measure the decomposition performance of each of the matrix decomposition techniques. A moment-based figure of merit was used as it calculates the reconstruction accuracy on the basis of both the structure and the energy of the samples.

The synthetic signal generated using Equation 9.62 is considered to be made up of components characterized by four parameters: beginning time, t_{j1} ; ending time, t_{j2} ; and the line parameters, a_j and b_j , which represent the instantaneous frequency, IF , of each component as

$$IF_j(t) = a_j t + b_j. \quad (9.63)$$

Figure 9.16 illustrates the instantaneous frequency parameters for a linearly frequency-modulated component.

For each component, the first-order moment of the reconstructed TF matrix around its instantaneous frequency was computed as the localization of the component, given by

$$Lcz_j = \sum_{n=t_{j1}}^{t_{j2}} \sum_{m=IF_j(n)-\Delta}^{IF_j(n)+\Delta} (|V_{mn}| \times |m - IF_j(n)|), \quad (9.64)$$

where 2Δ (shown in Figure 9.16) is the frequency interval around the instantaneous frequency over which the localization parameters are calculated. The percentage localization of component j was

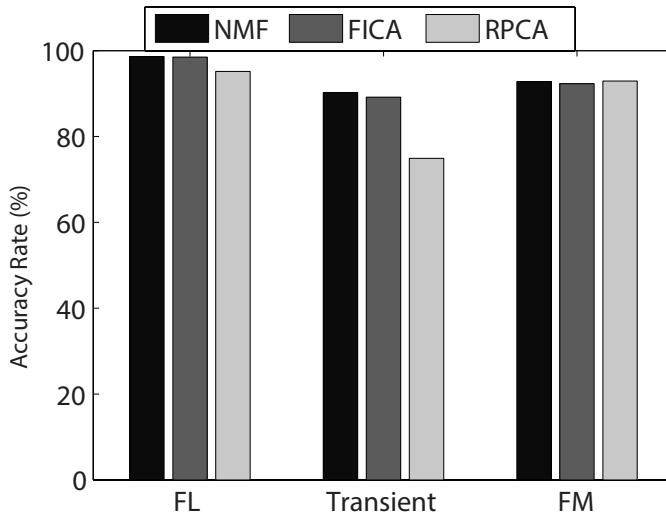


Figure 9.17 Localization performance of NMF, fast ICA (FICA), and robust PCA (RPCA) compared for FL, transient, and FM components. Reproduced with permission from B. Ghoraani [82].

calculated as

$$\text{Localization}_j(\%) = 100 - \left(\frac{|\text{Lcz}_j - \text{Lcz}_{O-j}|}{\text{Lcz}_{O-j}} \times 100 \right), \quad (9.65)$$

where Lcz_j and Lcz_{O-j} are the localization values calculated using Equation 9.64 from the reconstructed and the original TF component j , respectively, and Localization_j is the quantified localization of the same component.

When the matrix decomposition procedure provides an accurate decomposition of the TF matrix, $|\text{Lcz}_j - \text{Lcz}_{O-j}|$ will be small, and as a result, Localization_j in Equation 9.65 will be close to 100%. However, if matrix decomposition does not provide an accurate estimate, Localization_j will have a smaller value. Hence, the localization value is suitable as a figure of merit to evaluate decomposition performance, which was performed as described below:

1. Ten synthetic signals with varying characteristics were generated using Equation 9.62.
2. The adaptive TFD of each signal via MPTFD was constructed as the TF matrix representation.
3. PCA, ICA, and NMF were applied to the TF matrix \mathbf{V} to decompose it into submatrices \mathbf{W} and \mathbf{H} .
4. \mathbf{W} and \mathbf{H} were used to reconstruct the TF matrix: $\mathbf{V}_r = \mathbf{WH}$.
5. The localization of each component was calculated using Equation 9.64.

The average localization percentage of each technique is plotted in Figure 9.17. In this figure, the localization percentage of each component type is shown separately.

In the second experiment, the TF matrix features of the test signals were extracted as explained in the following steps:

1. The TF matrix, $\mathbf{V}_{M \times N}$, was constructed using an adaptive TFD technique.
2. TF matrix decomposition was performed as $\mathbf{V}_{M \times N} = \mathbf{W}_{M \times r} \mathbf{H}_{r \times N}$.
3. The first and second orders of spectral and temporal moments were extracted as in the following equations:

$$\begin{aligned} F_j &= (\hat{e}_j, \bar{t}_j, \bar{f}_j, \hat{t}_j, \hat{f}_j) \\ &= (10 \log(e_j/e_{\max}), \langle t \rangle_j, \langle f \rangle_j, \sqrt{\langle t^2 \rangle_j - \bar{t}_j^2}, \sqrt{\langle f^2 \rangle_j - \bar{f}_j^2}), \end{aligned} \quad (9.66)$$

where

$$\begin{aligned} \langle t^{(p)} \rangle_j &= \sum_{n=0}^N n^{(p)} H_{jn}, \\ \langle f^{(q)} \rangle_j &= \sum_{m=0}^M m^{(q)} W_{jm}, \end{aligned} \quad (9.67)$$

e_j is the average energy of the signal in a rectangular TF region for feature j , given by

$$e_j = \text{mean} \left[\sum_{n=\bar{t}_j-\hat{t}_j}^{\bar{t}_j+\hat{t}_j} \sum_{m=\bar{f}_j-\hat{f}_j}^{\bar{f}_j+\hat{f}_j} V_{mn} \right], \quad (9.68)$$

and e_{\max} is the feature vector with the highest energy.

Figure 9.18 provides a comprehensible demonstration of the extracted feature vectors as per Equation 9.66 for the three techniques. Each feature vector is associated with a rectangular TF region centered at \bar{t} and \bar{f} , and width of \hat{t} and \hat{f} in time and frequency, respectively.

The feature set consists of 35 values only (five features of F_j for each of the seven components in the MPTFD image), which is a substantial reduction in data size while capturing the important details of a TFD; yet, it retains considerable detail covering the TF energy distribution.

Some of the important observations from the second experiment are summarized as follows:

- NMF features are highly localized in the TF plane, whereas the PCA- and ICA-based features are spread out. The reason for this is that the bases and coefficients of ICA and PCA are not necessarily nonnegative. The negative elements of the \mathbf{W} and \mathbf{H} matrices cause artifacts in the reconstructed TF matrix, and therefore, absolute values of the elements of the matrices \mathbf{W} and \mathbf{H} were used.
- PCA decomposes a given TFD into orthogonal bases, and successfully reconstructs the TF matrix; however, because of the presence of negative elements in the decomposed matrices, the moment-based features are not able to localize signal components, especially the first component as shown in Figures 9.15 and 9.18.
- NMF provides the most efficient data reduction and the best TF feature representation as compared to ICA and PCA.

Table 9.1 summarizes the performance of the three matrix decomposition techniques considered as related to characterization of a TFD.

From the studies and the results presented above, it is evident that NMF would be the preferred choice for the extraction of valuable information from matrix representations of signals. However, a common shortcoming of all of the available NMF optimization approaches exists in the initialization of the \mathbf{W} and \mathbf{H} matrices. The NMF decomposition algorithms start with random initialization for

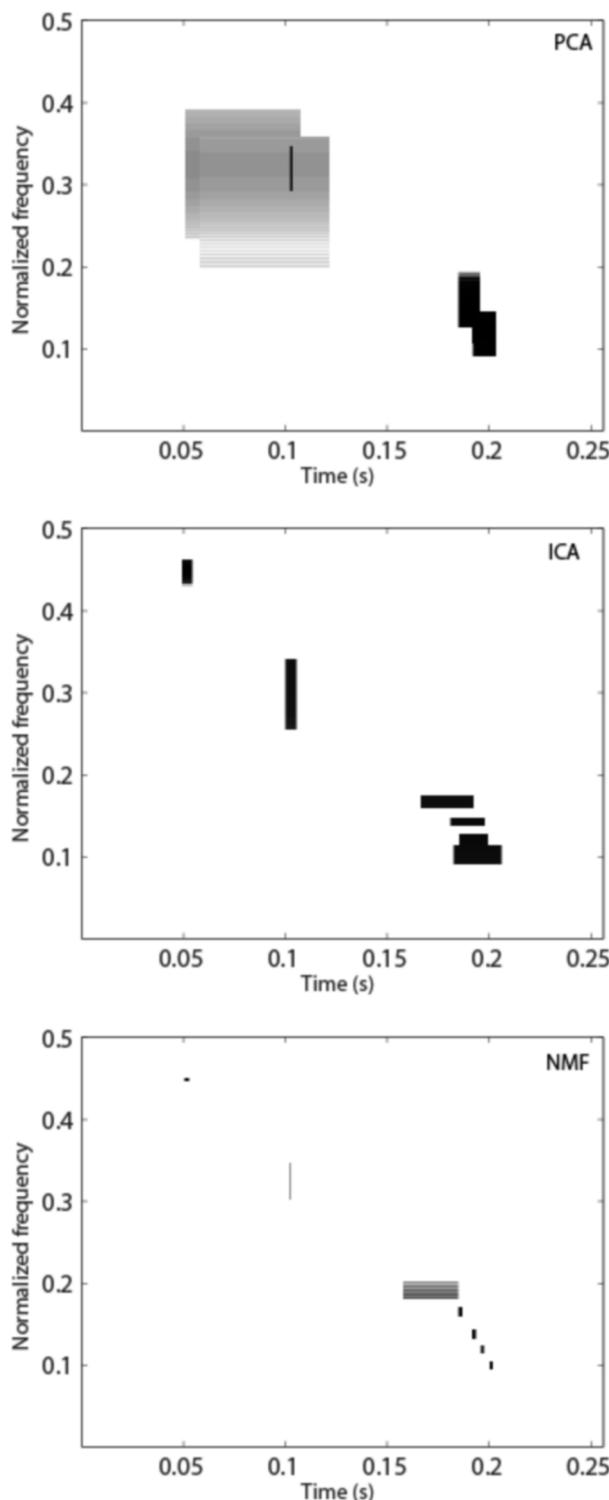


Figure 9.18 Top to bottom: TF features computed using PCA, ICA, and NMF for the signal in Figure 9.15. Each rectangle represents one feature vector. Reproduced with permission from B. Ghoraani [82].

Table 9.1 Matrix decomposition performance for TF quantification.

Method	Localized Reconstruction	TF Feature Localization	Transient Characterization	Efficient Representation
PCA	*	*	*	*
ICA	**	**	***	**
NMF	***	***	***	***

Note: The higher the number of the stars for a property, the better the performance of the corresponding method for the specific property. Reproduced with permission from B. Ghoraani [82].

W and **H** and modify these two matrices iteratively until a cost function is minimized. However, due to the nonconvexity of the cost function in both **W** and **H**, depending on the initial matrices, at each iteration of optimization, different local minima of the cost function may be achieved. As a result, each time the algorithm is applied, NMF might result in a different decomposition output. It was shown theoretically [83] that, under certain conditions, the NMF decomposition will be unique, but also that the same conditions are not generally satisfied in the case of real-world data. One solution to overcome this shortcoming is appropriate seeding of the **W** and **H** matrices. Various efforts have been directed to find seeding approaches in order to influence the convergence of NMF to the desired solution. Spherical k-means clustering was employed by Wild et al. [84], and SVD was used to seed the NMF matrices by Boutsidis and Gallopoulos [85]. Ghoraani and Krishnan [86] proposed a novel seeding method as explained in the following paragraphs.

The goal of the NMF seeding approach is to leverage the knowledge of a signal's TF structure to determine appropriate initialization values for the **W** and **H** matrices. The MPTFD matrix of a signal is obtained by adding the WVDs of the first I Gabor functions, and is given as

$$\mathbf{V}(t, f) = \sum_{i=1}^I |a_{\gamma_i}|^2 \mathbf{WVG}_{\gamma_i}(t, f). \quad (9.69)$$

The WVD of a Gabor function G_{γ_i} can be expressed as

$$\mathbf{WVG}_{\gamma_i}(t, f) = \mathbf{WVg}\left(\frac{t - p_i}{s_i}, s_i(f - f_i)\right), \quad (9.70)$$

where $\mathbf{WVg}(t, f)$ is the WVD of the Gaussian function $g(t)$, p_i is the temporal position of the function, s_i is the scale parameter, and f_i is the frequency parameter of the Gaussian function. It should be noted that the WVD of a Gaussian atom is a Gaussian, which can be written as [7]:

$$\mathbf{WVg}(t, f) = \hat{g}^2(f) g^2(t). \quad (9.71)$$

Here, $\hat{g}(f)$ is the Fourier transform of $g(t)$. Using Equation 9.71, the matrix \mathbf{WVg} can be written as

$$\mathbf{WVg} = \begin{bmatrix} \hat{g}^2(1) \\ \hat{g}^2(2) \\ \vdots \\ \hat{g}^2(M) \end{bmatrix} [g^2(1) \ g^2(2) \ \cdots \ g^2(N)]. \quad (9.72)$$

\mathbf{WVg} is an $M \times N$ matrix, where N and M are the number of time samples in $g(t)$ and the number of frequency samples in $\hat{g}(f)$, respectively.

Equations 9.69, 9.70, and 9.72 can be merged so that the MPTFD in Equation 9.69 can be represented in matrix format as

$$\mathbf{V}(t, f) = \sum_{i=1}^I |a_{\gamma_i}|^2 \begin{bmatrix} \hat{g}^2(s_i(1 - f_i)) \\ \hat{g}^2(s_i(2 - f_i)) \\ \vdots \\ \hat{g}^2(s_i(M - f_i)) \end{bmatrix} \times \left[g^2\left(\frac{1 - p_i}{s_i}\right) g^2\left(\frac{2 - p_i}{s_i}\right) \cdots g^2\left(\frac{N - p_i}{s_i}\right) \right]. \quad (9.73)$$

In the matching pursuit algorithm, some of the signal energy is modeled with optimal TF resolution in the TF plane at every iteration. Thus, the first few iterations of the Gabor dictionary have the best-correlated TF functions. Given that the first few atoms in the MPTFD decomposition provide a substantial amount of information about the TF structure of the signal, the first few Gabor atoms obtained through the matching-pursuit decomposition are utilized to initiate the NMF algorithm, leading to

$$\mathbf{W}_i^{\text{Init}} = |a_{\gamma_i}| \begin{bmatrix} \hat{g}^2(s_i(1 - f_i)) \\ \hat{g}^2(s_i(2 - f_i)) \\ \vdots \\ \hat{g}^2(s_i(M - f_i)) \end{bmatrix}, \text{ for } i = 1, 2, \dots, I. \quad (9.74)$$

$$\mathbf{H}_i^{\text{Init}} = |a_{\gamma_i}| \left[g^2\left(\frac{1 - p_i}{s_i}\right) g^2\left(\frac{2 - p_i}{s_i}\right) \cdots g^2\left(\frac{N - p_i}{s_i}\right) \right], \text{ for } i = 1, 2, \dots, I. \quad (9.75)$$

Because the seeding strategy is based on the signal's TF structure, it is projected to result in faster convergence. Furthermore, unlike random initialization, which produces a unique local minimum of the cost function each time the procedure is repeated, the MPTFD initialization approach discussed above produces a unique decomposition output.

To assess the performance of the seeding approach, a speech signal segment of 80 ms was decomposed using the MPTFD and random initialization methods. Figure 9.19 illustrates the MSE difference between the original and reconstructed TF matrices utilizing the two strategies. The MPTFD has a higher initial MSE than random seeding but converges faster. After eight iterations, NMF with MPTFD initialization achieved an MSE of 0.01, but the random initialization approach needed 18 iterations to achieve the same MSE. Therefore, rather than a random initialization of NMF matrices, a guided initialization process such as the one described here would lead to faster convergence.

9.8 Application: Detection of Epileptic Seizures Using Dictionary Learning Methods

Using the DWT and EMD, Kaleem et al. [87] presented methods for the detection of epilepsy, and compared their performance in automatic seizure detection. The various effects of seizures include physical harm [88]. Long-term analysis of the EEG is a potential approach for automatic seizure identification and early warning [89]. The nonstationary nature of EEG signals makes the development of algorithms for seizure detection difficult. Dictionary learning approaches [90–94] provide an adaptive strategy for EEG signal analysis and are a good choice for nonstationary EEG signal analysis. The general formulation of the seizure detection and signal classification problem is to find a sparse representation of the signal using dictionary functions [95].

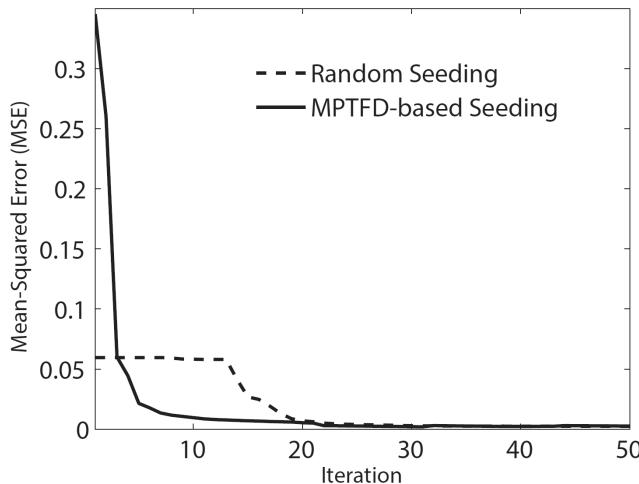


Figure 9.19 Comparison of NMF convergence with MPTFD seeding and seeding with a random number. Reproduced with permission from B. Ghoraani [82].

The matching-pursuit technique has been employed in conjunction with over-complete dictionaries of Gabor atoms to extract features for seizure detection [96]. The extracted features, such as the Gabor atom density, normalized Gabor entropy, and regularity statistics, have been employed in conjunction with a classifier to detect seizures. The rate of convergence of the iterative atomic decomposition process has been applied as a feature for seizure detection.

As illustrated in Section 9.4, EMD is a data-driven decomposition technique that does not require a basis function. It shares numerous characteristics with the DWT, including decomposition behavior similar to that of a dyadic filter bank [24]. Kaleem et al. [87] applied the following steps in their work on seizure detection in EEG signals: A DWT-based empirical dictionary technique was used, in which the dictionary's atoms are made of components derived via DWT decomposition. The projection coefficients, coefficient vector, and reconstruction error were employed as characteristics for automatic seizure detection. All of these characteristics are present in dictionary learning strategies, indicating the approach's adaptability. The dictionary building and learning performance of EMD and DWT-based dictionary techniques, as well as the seizure detection performance of both processes across all features, were compared. The seizure detection approach was structured in a realistic manner, with classifiers trained on small samples of the specific patient's past seizure and nonseizure data, and then tested on subsequent data. The results of seizure detection were confirmed using k-fold cross-validation to minimize or rule out any classifier bias. The seizure detection performance was evaluated with five widely used machine learning classification methods, allowing for comparisons between different classifiers.

EEG dataset for seizure detection: Kaleem et al. [87] used the CHB-MIT scalp EEG dataset [97]. Multiple long-term EEG recordings from 23 pediatric patients with uncontrollable seizures are included in this dataset. Each of these recordings is comprised of 23 EEG channels: FP1-F7, F7-T7, T7-P7, P7-O1, FP1-F3, F3-C3, C3-P3, P3-O1, FP2-F4, F4-C4, C4-P4, P4-O2, FP2-F8, F8-T8, T8-P8, P8-O2, FZ-CZ, CZ-PZ, P7-T7, T7-FT9, FT9-FT10, FT10-T8, and T8-P8 (see Figure 1.39). Recordings containing at least one seizure event were included. Seizure and nonseizure EEG recordings were obtained for a total of 2.9 h and 171 h, respectively. To facilitate processing, the long-term EEG recordings were divided into 4 s segments sampled at 256 Hz. Various segment durations have been used in prior research, including 1, 2, 3, 4, and even 10 s [98]. One reason for choosing shorter segment lengths, such as 1 s, is to assure stationarity, but this is not relevant if

the method can accommodate nonstationary signals. Another factor to consider when determining segment length is resolution of the lower frequencies in the EEG signals, which are typically filtered between 0.5 and 60 Hz in preprocessing steps.

Joint EMD and DWT analysis: The EMD algorithm acts at the oscillatory component level and is adaptive to the signal's local frequency content. The IMFs exhibit dyadic filter-bank behavior [24]; however, minor departures from this behavior are possible due to the data-adaptive characteristics of EMD. The IMFs reflect a hierarchical separation of the spectral content of the given signal, with lower index IMFs holding the signal's higher frequency components. The number of IMFs J is unknown in advance, but in general, $J \leq \log 2N$, where N is the number of samples in the signal [30].

The DWT can be used to split a given signal into its detail and approximation components via a predefined dyadic subband filtering step based on the mother wavelet basis function. These components can be acquired by using the DWT to decompose a signal $x(n)$ as follows:

$$x(n) = \sum_{k=1}^N a_{J,k} 2^{-J/2} \phi(2^{-J}n - k) + \sum_{j,k} d_{j,k} 2^{-j/2} \varphi(2^{-j}n - k), \quad (9.76)$$

where $\varphi(n)$ is the orthonormal basis function referred to as the mother wavelet, $\phi(n)$ is the scaling function orthogonal to $\varphi(n)$, k is the translation parameter, and $d_{j,k}$ and $a_{J,k}$ are the detail and approximation components, respectively. The number of DWT components depends on the decomposition levels specified, such that there are J detail components and one approximation component for a J -level decomposition.

Whereas EMD's filter-bank structure is similar to that of the DWT in terms of self-similarity, quasidecorrelation, and variance progression, wavelet decomposition at each level occurs according to a predetermined frequency division. The use of LTI filters for wavelet decomposition does not allow for adaptation to local variations in the signal's frequency content [99]. On the other hand, because EMD is data-adaptive, it is well-suited for nonlinear and nonstationary EEG signal processing.

Signal-derived dictionary algorithm: Kaleem et al. employed the Daubechies db4 mother wavelet for DWT decomposition, which is a commonly used mother wavelet in seizure detection research [100], and a seven-level decomposition was selected for the reasons detailed later in this section. The dictionary approach begins with a training matrix $\mathbf{X}_{\text{Train}}^c$. The columns of the matrix contain k^c training signals x^c (EEG segments with $N = 1,024$ samples) associated with class c . In this application, $c \in \mathbf{C}$, $\mathbf{C} = \{c_1, c_2\}$, where c_1 represents the seizure class and c_2 represents the nonseizure class. The dictionary approach begins with the creation of a raw dictionary, $\mathbf{D}_{\text{raw}}^c = \{\psi^m\}_{m=1}^M$ with M atoms ψ , where, in general, $M < N$.

The dictionary atoms ψ in the DWT-based dictionary approach are composed of approximation and detail components obtained by decomposing the signals $x^c \in \mathbf{X}_{\text{Train}}^c$ using the DWT, whereas the dictionary atoms ψ in the EMD-based dictionary approach are composed of the IMFs. For DWT, a seven-level decomposition was utilized, yielding seven detail components and one approximation component. The average number of IMFs derived by decomposing all segments, averaged across all patients, was found to be eight using the EMD-based dictionary technique. As a result, a seven-level DWT decomposition was utilized to compare the number of components acquired from EMD and DWT decomposition. The notation $a_q^c(n)$, $q \in \{1, 2, \dots, Q\}$ is used here to refer to both the DWT components and IMFs. The raw dictionary $\mathbf{D}_{\text{raw}}^c$ was constructed in the following manner [87]:

- For the DWT and EMD-based dictionaries, respectively, the DWT components and IMFs serve as the building blocks of class-specific raw dictionaries $\mathbf{D}_{\text{raw}}^c = [\mathbf{d}_1 | \mathbf{d}_2, \dots, |\mathbf{d}_L]$, $L = K^c \times Q$. The atoms of class-specific raw dictionaries are bound to have L_2 -norm ≤ 1 , so that $\mathbf{d}_l = \hat{\mathbf{a}}_q^c$, $l \in \{1, 2, \dots, L\}$, $q \in \{1, 2, \dots, Q\}$, and $\hat{\mathbf{a}}_q^c = \mathbf{a}_q^c / \|\mathbf{a}_q^c\|_2$. For a seven-level DWT decomposition, $J = 7$, hence $Q = J + 1 = 8$. However, Q is not known in advance in the case of EMD, and hence $Q = 8$ was selected.

2. $\mathbf{D}_{\text{raw}}^{\mathbf{C}} = [\mathbf{D}_{\text{raw}}^{c_1} | \mathbf{D}_{\text{raw}}^{c_2}]$ was formed by merging the class-specific raw dictionaries.

Once the combined raw dictionary was prepared, a trained dictionary $\mathbf{D}_{\text{Train}}^{\mathbf{C}} = \{\psi^p\}_{p=1}^P$ was learned from the raw dictionary $\mathbf{D}_{\text{raw}}^{\mathbf{C}} = \{\psi^m\}_{m=1}^M$. Here, $P \ll M$, which suggests that the dictionary learning stage results in a significant reduction in the size of the dictionary. The trained dictionary was created using a dictionary learning method and the training signals x^c , which are the same signals as in the training matrices $\mathbf{X}_{\text{Train}}^{c_1}$ and $\mathbf{X}_{\text{Train}}^{c_2}$ that belong to the seizure and nonseizure classes; see Section 9.5. After I iterations, the dictionary learning procedure was terminated, with I determined by using a validation scheme. The validation scheme made use of validation signals \hat{x}^c pertaining to the seizure and nonseizure classes, which were distinct from the dictionary creation signals x^c . The number of validation signals was fixed and identical for the seizure and nonseizure classes. The trained dictionary $\mathbf{D}_{\text{Train}}^{\mathbf{C}}$ is the outcome of the dictionary learning procedure being completed. The dictionary learning procedure as described by Kaleem et al. [87] is provided in Algorithm 9.2.

Algorithm 9.2: Signal-derived Dictionary Learning Algorithm [87]

- 1: Create a new trained dictionary $\mathbf{D}_{\text{Train}}^{\mathbf{C}} = []$.
 - 2: Initialize a raw dictionary $\mathbf{D}_{\text{raw}}^{\mathbf{C}}(x)$ for each training signal x .
 - 3: **repeat** $I = 1$ to 7:
 - 4: For each training signal $x \in \mathbf{X}_{\text{Train}}^{c_1} \cup \mathbf{X}_{\text{Train}}^{c_2}$:
 - 5: Determine the projection coefficient α_m of x against each atom ψ^m in the raw dictionary $\mathbf{D}_{\text{raw}}^{\mathbf{C}}(x) : \alpha_m = \langle x, \psi^m \rangle$;
 - 6: Select the atom ψ^m whose projection coefficient has the largest absolute value $|\alpha_m|$;
 - 7: If ψ^m is not in $\mathbf{D}_{\text{Train}}^{\mathbf{C}}$, then add it to the trained dictionary: $\mathbf{D}_{\text{Train}}^{\mathbf{C}} \leftarrow \mathbf{D}_{\text{Train}}^{\mathbf{C}} \cup \{\psi^m\}$;
 - 8: Remove ψ^m from the raw dictionary: $\mathbf{D}_{\text{raw}}^{\mathbf{C}}(x) \leftarrow \mathbf{D}_{\text{raw}}^{\mathbf{C}}(x) \setminus \{\psi^m\}$;
 - 9: Replace x by the residue after projecting on $\psi^m : x \leftarrow x - \langle x, \psi^m \rangle \widetilde{\psi^m}$;
 - 10: Save the trained dictionary $\mathbf{D}_{\text{Train}}^{\mathbf{C}}$ for the current value of I .
 - 11: Apply the following validation scheme:
 - 12: **repeat** for each class $c \in \{c_1, c_2\}$:
 - 13: Initialize an empty projection coefficient vector Γ_c ;
 - 14: **repeat** for each validation signal \hat{x}^c of class c :
 - 15: Compute the projection coefficient α_m of \hat{x}^c against each atom ψ^m in the current trained dictionary $\mathbf{D}_{\text{Train}}^{\mathbf{C}} : \alpha_m = \langle \hat{x}^c, \psi^m \rangle$;
 - 16: Select the projection coefficient with the largest absolute value $|\alpha_m|$ and append this value to Γ_c .
 - 17: **until** end of validation signals
 - 18: **until** end of class
 - 19: Calculate the distance between the projection coefficient vectors of both classes: $d = \|\Gamma_{c_1} - \Gamma_{c_2}\|_2^2$.
 - 20: If d is the largest distance encountered so far, retain the current trained dictionary as the best trained dictionary $\mathbf{D}_{\text{Train}}^{\mathbf{C}}$.
 - 21: **until** Termination
 - 22: Return the best trained dictionary $\mathbf{D}_{\text{Train}}^{\mathbf{C}}$.
-

The raw dictionary $\mathbf{D}_{\text{raw}}^{\mathbf{C}} = \{\psi^m\}_{m=1}^M$ has M atoms, where $M = 2L = 2KQ$, and K represents the number of training signals in each training matrix $\mathbf{X}_{\text{Train}}^c$, whereas the trained dictionary $\mathbf{D}_{\text{Train}}^{\mathbf{C}} = \{\psi^p\}_{p=1}^P$ has P atoms. When the same K training signals are used to train the dictionary, a maximum of $2K$ atoms can be added to the trained dictionary in a single iteration of the dictionary learning process. As a result, the transition from the raw to the trained dictionary results in a decrease in dictionary size from M to P . Additionally, if the number of iterations of dictionary learning I is minimal, the size difference between the raw and trained dictionaries is considerable,

such that $P \ll M$. A smaller vocabulary is predicted to result in faster feature extraction in terms of computation. The validation signals \hat{x}^c are organized in the order in which they occur in the EEG recordings. During the validation stage, however, the pairing utilized to calculate the distance between the projection coefficient vectors is arbitrary.

The signal-derived dictionary technique was used to create and learn dictionaries using segmented EEG recordings from the CHB-MIT database. To begin, 15% of the seizure segments from each patient's 23 channels were used for dictionary building and learning, and 5% were used for validation. The same number of segments were obtained from nonseizure segments. Additionally, 30% of seizure segments were retained for classifier training, and the same percentage of nonseizure segments were retained. These portions were not chosen at random; rather, their chronological order was maintained. This means that the first 20% of the EEG data were retained for dictionary building, learning, and validation, while the following 30% were retained for classifier training. The remaining 50% of seizure segments and all nonseizure segments were then retained for the purpose of evaluation of the performance of seizure detection. This division of EEG records was intended to replicate a real-world situation in which the system would be trained on existing data and then utilized for automatic seizure identification on subsequently acquired data [101].

Partitioning of the available data as described above enabled Kaleem et al. subsequently to retain 50% of the EEG records for seizure detection testing. To demonstrate the usefulness of the strategy utilizing smaller dictionaries and to have classifier training data equivalent to other approaches, such as that of Zabihi et al. [101], 20% of the available data were utilized for dictionary building, learning, and validation, and 30% for classifier training. To create and validate the dictionary, the 4s segments extracted from the data of all 23 channels were merged. This enabled the multidimensional nature of the data to be incorporated into the dictionary generation and learning process. With the segments of all 23 channels combined, the raw dictionary and trained dictionary were larger than the case when dictionaries were constructed and trained separately for each channel. After the dictionary learning and validation stages were complete, the trained dictionary was expected to contain the relevant atoms collected from all channels. A raw dictionary D_{raw}^C was built for each patient using the allotted segments from all channels. After I iterations, each raw dictionary was trained using the dictionary learning steps to yield D_{Train}^C , where I was selected using the validation step and the segments allotted for validation (having been combined from all channels).

Examples of EEG recordings corresponding to seizure and nonseizure cases are shown in Figures 9.20 (a) and (b), respectively. Figures 9.21 (a) and (b) present examples of trained dictionary atoms for both EMD and DWT-based dictionaries. As previously stated, the atoms are composed of IMFs or DWT detail and approximation components. It is worthwhile to emphasize here that, depending on the application, physical meaning can be provided to EMD and DWT components in order to justify the various contributions to a specific event revealed by signal decomposition [102]. However, attributing physical meaning to the components is not relevant in the present study, as the components constitute atoms of a dictionary that has been trained using a dictionary learning method, resulting in the selection of atoms that are most similar to the two classes (seizure and nonseizure). This also eliminates the need for any hand-crafted method for IMF selection, in which the most important or relevant IMFs are chosen for a specific application, such as seizure detection [103].

Preprocessing EEG recordings for artifact removal is a typical aspect of the majority of seizure detection investigations [98, 101]. The dictionary learning algorithm chooses for inclusion in the learned dictionary those atoms (DWT components or IMFs) that are most comparable to the seizure and nonseizure classes, thereby avoiding any components that may contain artifacts. As a result, no preprocessing is required in the methodology described here, which is another advantage of the approach proposed by Kaleem et al.

Kaleem et al. [87] extracted three features from the trained dictionary, which are listed below. To assess the performance of the three features in detecting seizures, all features were retrieved using EMD and DWT-based dictionaries.

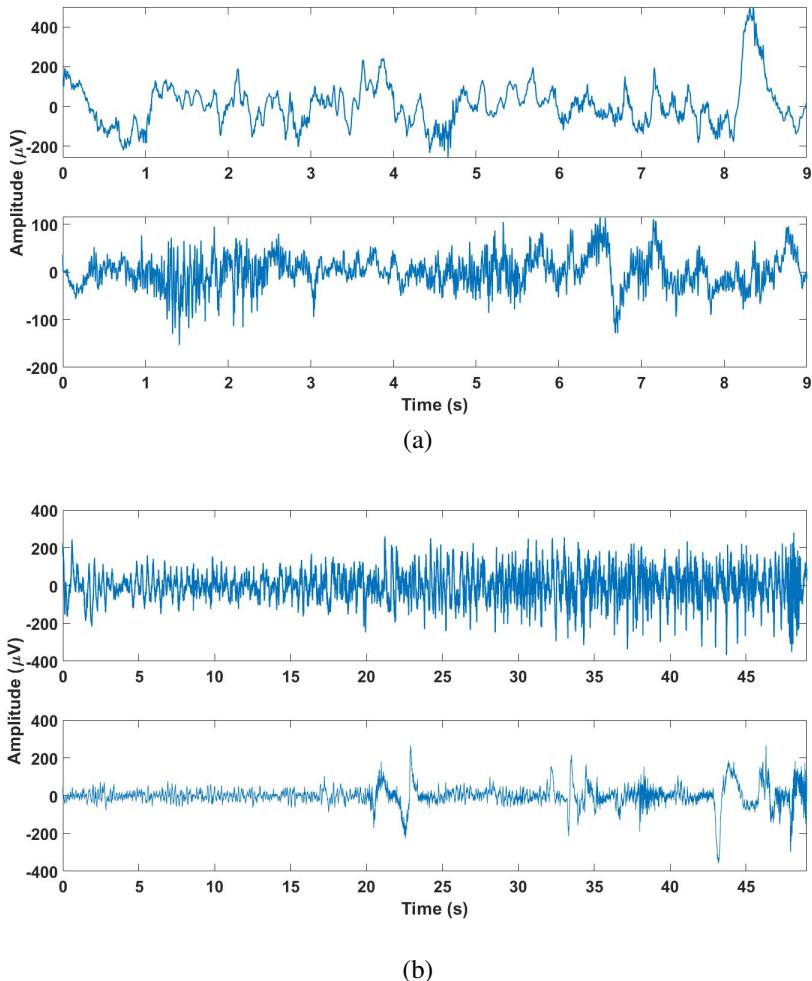


Figure 9.20 Examples of seizure and nonseizure recordings: (a) a 9 s long seizure recording (top plot) of channel 21 of patient 2, and the same length of nonseizure signal (bottom plot) from the same recording. (b) a 49 s long seizure recording (top plot) of channel 21 of patient 4, and the same length of nonseizure signal (bottom plot) from the same recording. Reproduced with permission from M.F. Kaleem, A. Guergachi, and S. Krishnan. Comparison of empirical mode decomposition, wavelets, and different machine learning approaches for patient-specific seizure detection using signal-derived empirical dictionary approach. *Frontiers in Digital Health*, 3, 738996, 2021.

Projection Coefficients (F1): The projection coefficients were determined during the dictionary learning procedure. The projection coefficient $\alpha_m = \langle x^c, \psi^m \rangle$ with the greatest magnitude was chosen as the feature coefficient, where x^c represents a seizure or nonseizure classifier training or testing segment and ψ^m represents an atom in the trained dictionary.

Coefficient Vector (F2): The feature vector \mathbf{a} was constructed using the learned dictionary $\mathbf{D}_{\text{Train}}^C$ and a signal \mathbf{x} as

$$\mathbf{a} = \mathbf{D}^\dagger \mathbf{x}, \quad (9.77)$$

where \mathbf{x} represents a classifier training or testing signal, and \mathbf{D}^\dagger is the left-pseudoinverse of the trained dictionary $\mathbf{D}_{\text{Train}}^C$.

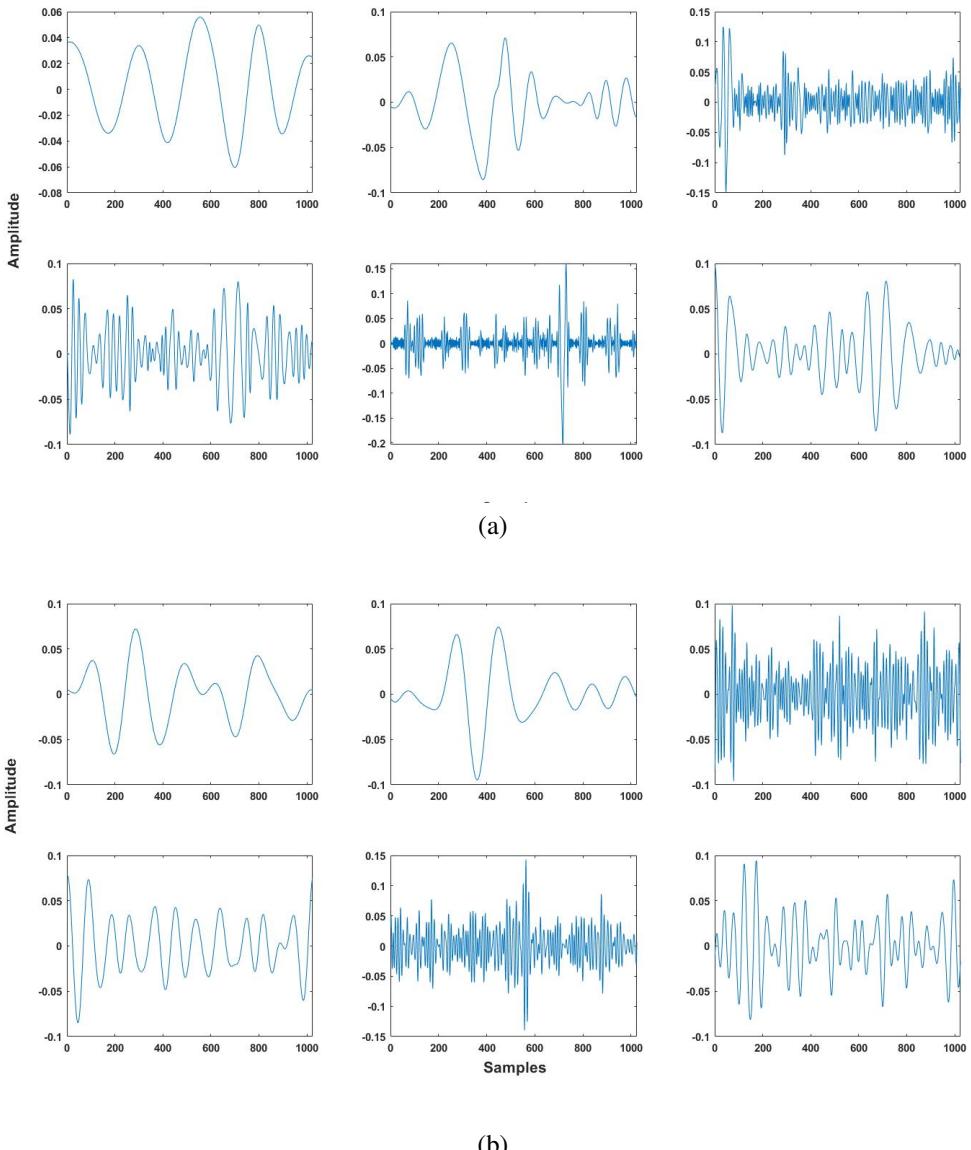


Figure 9.21 Examples of trained dictionary atoms from (a) EMD and (b) DWT-based dictionaries of patient 3. The atoms are of length 4 s or 1,024 samples each. The atoms consist of IMFs or detail and approximation components of both seizure and nonseizure segments, and have been normalized. Reproduced with permission from M.F. Kaleem, A. Guergachi, and S. Krishnan. Comparison of empirical mode decomposition, wavelets, and different machine learning approaches for patient-specific seizure detection using signal-derived empirical dictionary approach. *Frontiers in Digital Health*, 3, 738996, 2021.

Reconstruction Error (F3): The feature corresponding to reconstruction error ϵ for a classifier training or testing signal \mathbf{x} is given by $\|\mathbf{x} - \mathbf{D}\mathbf{a}\|_2$, where \mathbf{a} is the coefficient vector and \mathbf{D} is the trained dictionary. While the learned dictionary \mathbf{D} is not intended to be a reconstructive dictionary, it is feasible to create a coarse reconstruction of the training or testing signal represented by $\hat{\mathbf{x}}$ by applying the relation

$$\hat{\mathbf{x}} = \mathbf{D}\mathbf{D}^\dagger\mathbf{x}. \quad (9.78)$$

The reconstruction error feature is then obtained as $\epsilon = \|\mathbf{x} - \hat{\mathbf{x}}\|_2$.

In addition to evaluating the effectiveness of seizure detection using the three features described above individually, combinations of the three features were also used as $\{F1, F2, F3\}$, $\{F1, F2\}$, $\{F1, F3\}$, and $\{F2, F3\}$.

Performance of the EMD- and DWT-based dictionary approaches: The average number of iterations required to terminate dictionary learning was 3.26 ± 1.63 for EMD and 3.09 ± 1.56 for the DWT-based dictionary techniques. The average size of the DWT-based trained dictionaries of the 23 patients was smaller than that of EMD-based dictionaries. The average reduction in size from raw to trained dictionary was similar for EMD- (6,471 to 2,346) and DWT-based dictionaries, at 68% and 66%, respectively, with similar dispersion around the average. The number of atoms in the EMD-based raw dictionaries was typically higher than that in the DWT-based raw dictionaries.

The reduction in dictionary size for EMD- and DWT-based dictionaries was comparable at the individual patient level. Individual dictionaries trained with fewer iterations had the greatest reduction in size when compared to the corresponding raw dictionaries.

The EMD-based dictionary approach provided the highest values for accuracy, sensitivity, and specificity in seizure detection. $F3$ (reconstruction error) performed the best across all classifiers for both EMD- and DWT-based techniques. The best averaged sensitivity value (89.9%) for $F3$ was obtained using an SVM classifier and an EMD-based dictionary. Even with specificity (86.4%) taken into account, $F3$ with the SVM classifier provided the best seizure detection performance. To rule out any bias in the seizure detection results, five-fold cross-validation was used with the same testing data for each patient to validate the results. The obtained values were 90.3%, 93.5%, and 87.12% for accuracy, sensitivity, and specificity, respectively. Although the EEG channels identified in this case differed from those obtained in earlier experiments, they all originated from the same region of the brain. This demonstrates the robustness of the seizure detection approach of Kaleem et al. [87].

9.9 Application: Adaptive Time–Frequency Analysis of VAG Signals

Krishnan et al. [31] proposed several methods to derive time-varying parameters from TFDs for the analysis and classification of VAG signals. Specifically, four TF features were derived from OMPTFDs (see Section 9.6) of VAG signals: the energy parameter (EP), the energy spread parameter (ESP), the frequency parameter (FP), and the frequency spread parameter (FSP). EP was computed as the mean of the OMPTFD $M(t, \omega)$ along each time slice, which gives a representation of energy variation with time as

$$EP(t) = \frac{1}{\omega_m} \sum_{\omega=0}^{\omega_m} M(t, \omega), \quad (9.79)$$

where ω_m is the maximum frequency present in the signal. Signals generated by pathological knees are expected to be highly time-variant (nonstationary) because of differences in cartilage roughness and nonuniformity. Thus, $EP(t)$ of an abnormal VAG signal is expected to show large variations over time. Although $EP(t)$ may be computed directly from the signal, its derivation as above from the OMPTFD includes the benefits of adaptive filtering because only the coherent structures in the signal are retained (see Section 9.3).

The parameter ESP, computed as the *SD* of $M(t, \omega)$ along each time slice, measures the spread of energy over frequency for each time slice as

$$\text{ESP}(t) = \left[\frac{1}{\omega_m} \sum_{\omega=0}^{\omega_m} [M(t, \omega) - \text{EP}(t)]^2 \right]^{\frac{1}{2}}. \quad (9.80)$$

ESP could serve as a feature suitable for the analysis of multicomponent signals. Abnormal VAG signals generated as a result of friction between rough cartilage surfaces may have more components because of the nonuniformity of the surfaces, and hence a higher ESP than normal VAG signals.

The parameter FP, also known as IMF, is given by the first moment of $M(t, \omega)$ along each time slice, as

$$\text{FP}(t) = \frac{\sum_{\omega=0}^{\omega_m} \omega M(t, \omega)}{\sum_{\omega=0}^{\omega_m} M(t, \omega)}. \quad (9.81)$$

FP characterizes the frequency dynamics of the signal. Movement of the knee during VAG signal acquisition may cause linear or nonlinear FM components, with the modulation index depending on the state of lubrication, stiffness, and roughness of the cartilage surfaces [38]. Pathological knees have less lubricated and rougher cartilage surfaces than normal knees, and hence the variations in FP for pathological knees are expected to be different from those for normal knees.

The parameter FSP is given by the second central moment of $M(t, \omega)$ along each time slice as

$$\text{FSP}(t) = \left[\frac{\sum_{\omega=0}^{\omega_m} [\omega - \text{FP}(t)]^2 M(t, \omega)}{\sum_{\omega=0}^{\omega_m} M(t, \omega)} \right]^{\frac{1}{2}}. \quad (9.82)$$

FSP gives the spread of frequency about the mean frequency for each time instant. The spread of frequency at an instant of time could arise as a result of AM components. AM effects may occur in VAG signals and may be dependent on the quality and intensity of the sound produced due to joint vibration. FSP could serve as a feature to identify noisy knees.

In order to facilitate the derivation of a global decision on each VAG signal, the mean and *SD* of the arrays of each of the four parameters, as defined above, were calculated. The mean and *SD* characterize the central tendency and dispersion of the nonstationary parameters.

The VAG signal of a normal subject is shown in Figure 9.22. The signal demonstrates a normal noisy knee with a click and grinding sounds heard during auscultation. The OMPTFD of the signal is shown in part (b) of the same figure. The TFD was constructed using only the coherent structures of the signal and the number of TF atoms used was 588. Optimization of the marginals was achieved in two iterations. The click and grinding sounds are shown as high-frequency activities in the TFD and have been represented with good TF localization.

Figure 9.23 shows the VAG signal of a subject with chondromalacia patella. Grinding sound was heard during auscultation. The OMPTFD shown in part (b) of the same figure was constructed using 803 atoms, and the marginals were optimized in two iterations. The OMPTFD indicates two linear FM components (chirps) in the high-frequency range of 500 Hz to 800 Hz. Several tonal components below 400 Hz can also be identified in the TFD.

Figures 9.24 and 9.25 show the EP, ESP, FP, and FSP waveforms derived from the OMPTFD of the normal VAG signal shown in Figure 9.22. Figures 9.26 and 9.27 show the EP, ESP, FP, and FSP waveforms derived from the OMPTFD of the abnormal VAG signal shown in Figure 9.23. It may be readily seen that the parameters of the abnormal VAG signal have larger variations over time as compared to the corresponding parameters of the normal VAG signal. Whereas visual comparative analysis of the TFDs of the normal and abnormal signals could be difficult, several differences are readily observable in the plots of the related parameters.

The EP and ESP features are dependent on the data acquisition protocol and related parameters, in particular, the gain of the amplifier and the power of the VAG signal. To overcome this problem,

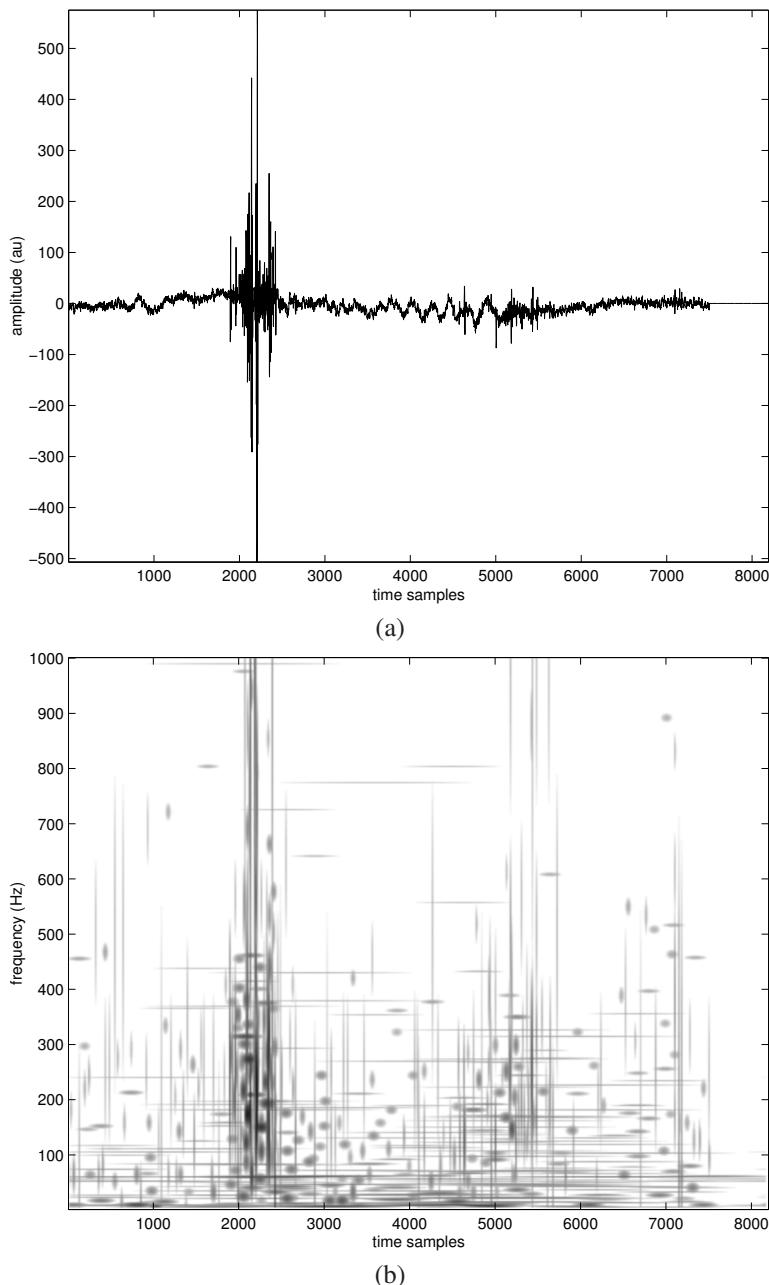


Figure 9.22 (a) VAG signal of a normal subject. A click and grinding sounds were heard during auscultation of the knee. au: arbitrary acceleration units. Sampling rate = 2 kHz. (b) OMPTFD of the signal. Reproduced with permission from S. Krishnan, R.M. Rangayyan, G.D. Bell, and C.B. Frank, Adaptive time-frequency analysis of knee joint vibroarthrographic signals for noninvasive screening of articular cartilage pathology, *IEEE Transactions on Biomedical Engineering*, 47(6):773–783, June 2000. ©IEEE.

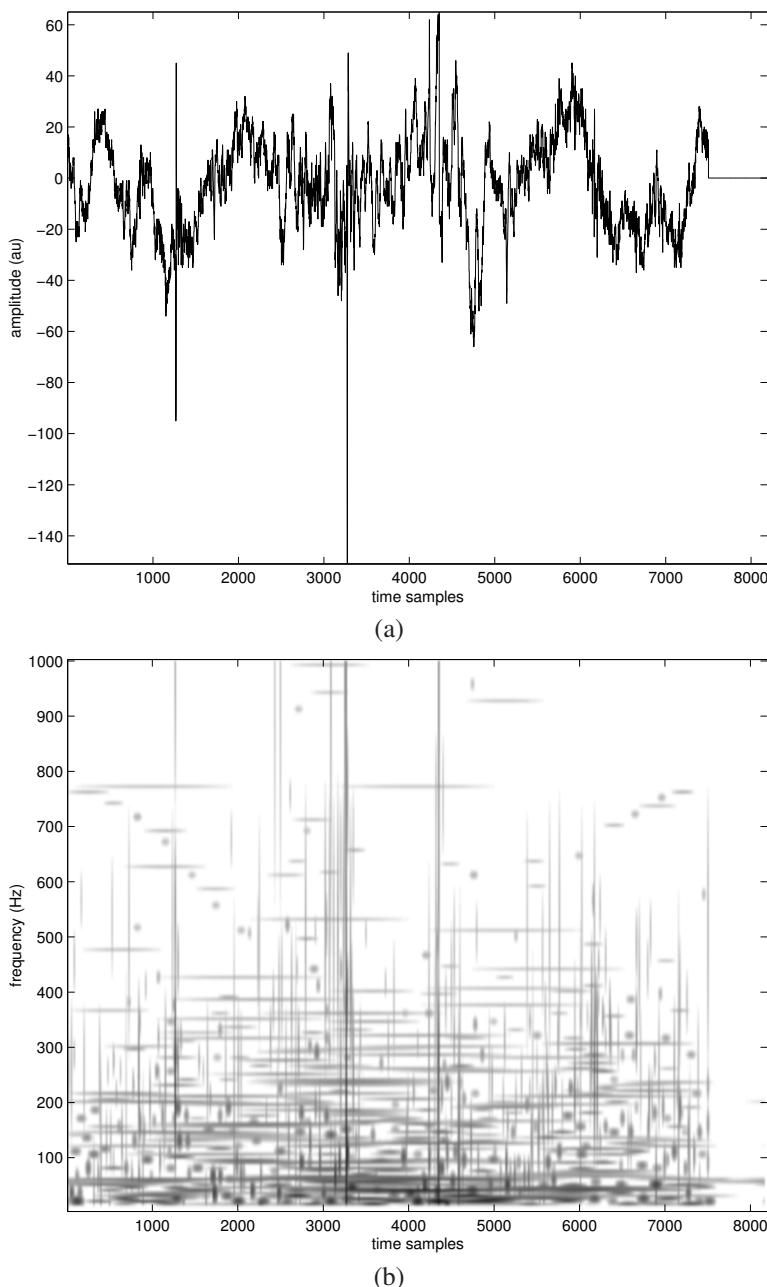


Figure 9.23 (a) VAG signal of a subject with knee-joint pathology. Grinding sound was heard during auscultation of the knee. au: arbitrary acceleration units. Sampling rate = 2 kHz. (b) OMPTFD of the signal. Reproduced with permission from S. Krishnan, R.M. Rangayyan, G.D. Bell, and C.B. Frank, Adaptive time-frequency analysis of knee joint vibroarthrographic signals for noninvasive screening of articular cartilage pathology, *IEEE Transactions on Biomedical Engineering*, 47(6):773–783, June 2000. ©IEEE.

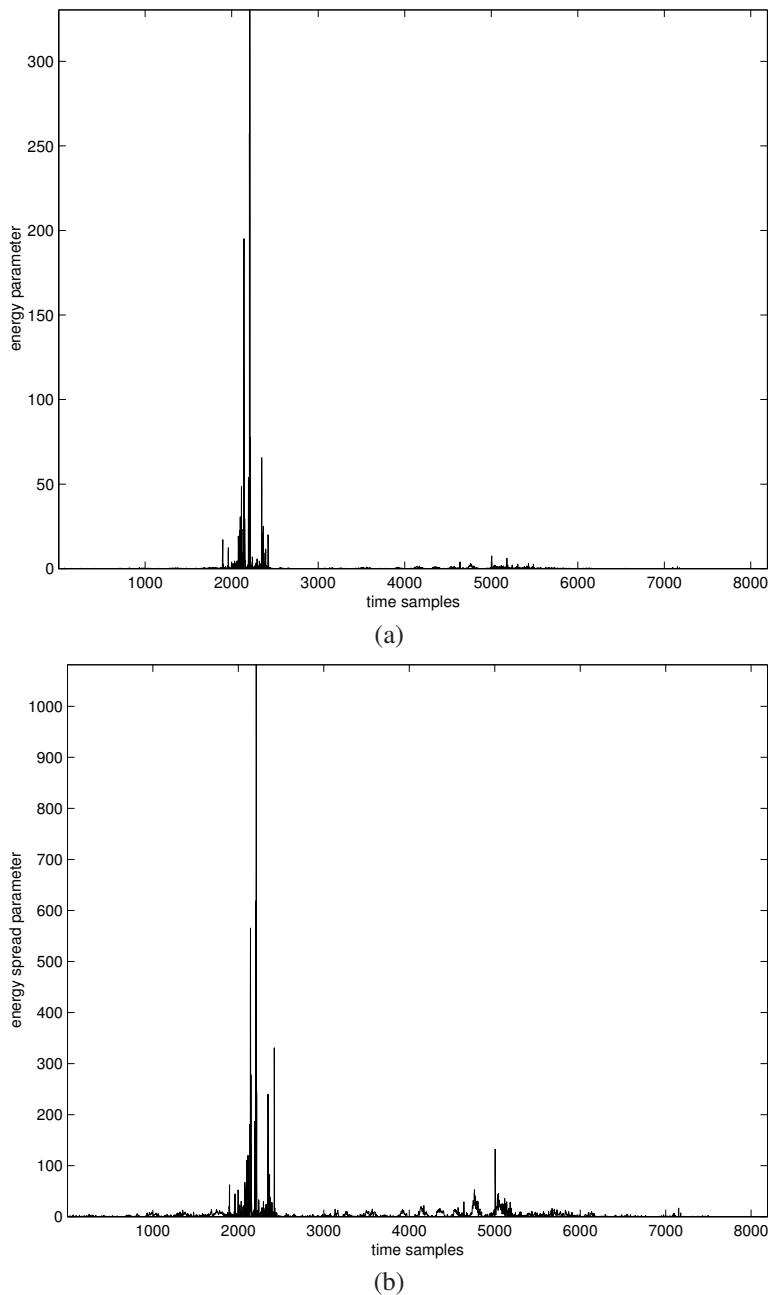


Figure 9.24 (a) EP and (b) ESP obtained from the OMPTFD of the normal VAG signal in Figure 9.22. Reproduced with permission from S. Krishnan, R.M. Rangayyan, G.D. Bell, and C.B. Frank, Adaptive time-frequency analysis of knee joint vibroarthrographic signals for noninvasive screening of articular cartilage pathology, *IEEE Transactions on Biomedical Engineering*, 47(6):773–783, June 2000. ©IEEE.

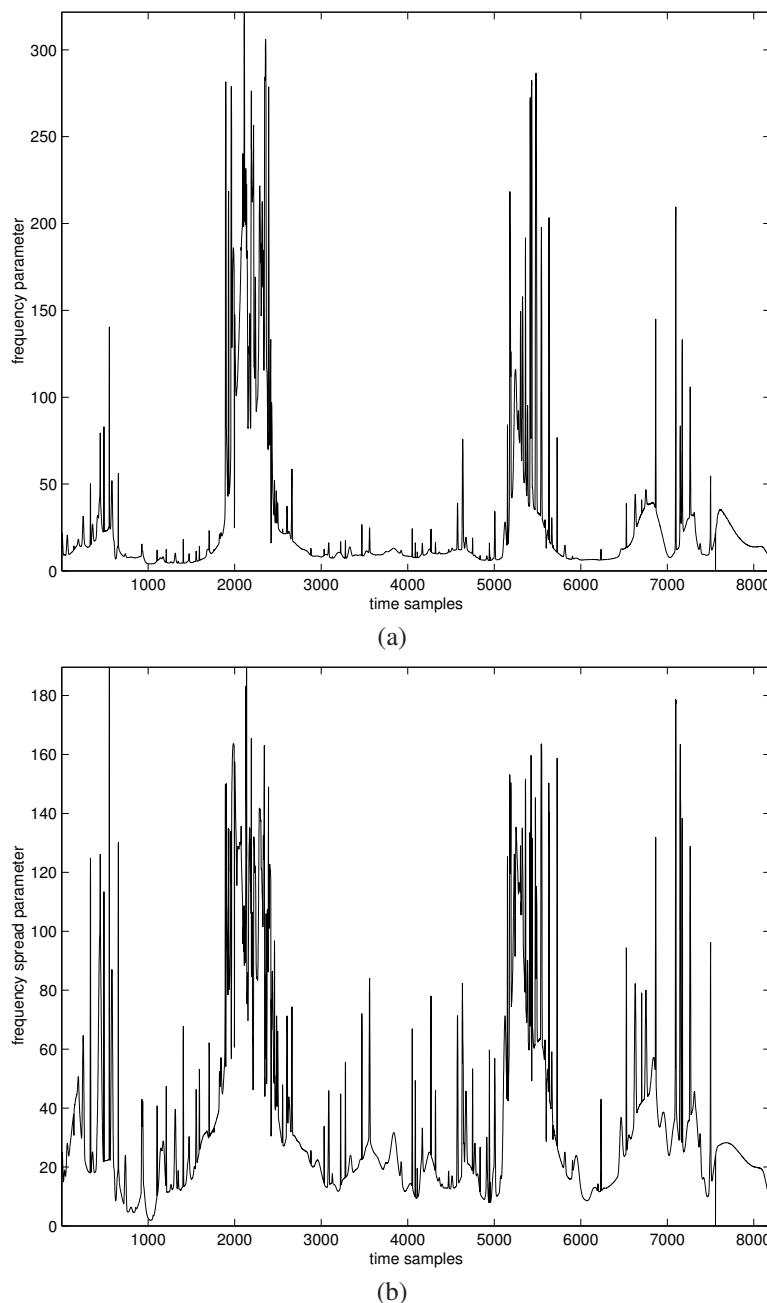


Figure 9.25 (a) FP and (b) FSP obtained from the OMPTFD of the normal VAG signal in Figure 9.22. Reproduced with permission from S. Krishnan, R.M. Rangayyan, G.D. Bell, and C.B. Frank, Adaptive time-frequency analysis of knee joint vibroarthrographic signals for noninvasive screening of articular cartilage pathology, *IEEE Transactions on Biomedical Engineering*, 47(6):773–783, June 2000. ©IEEE.

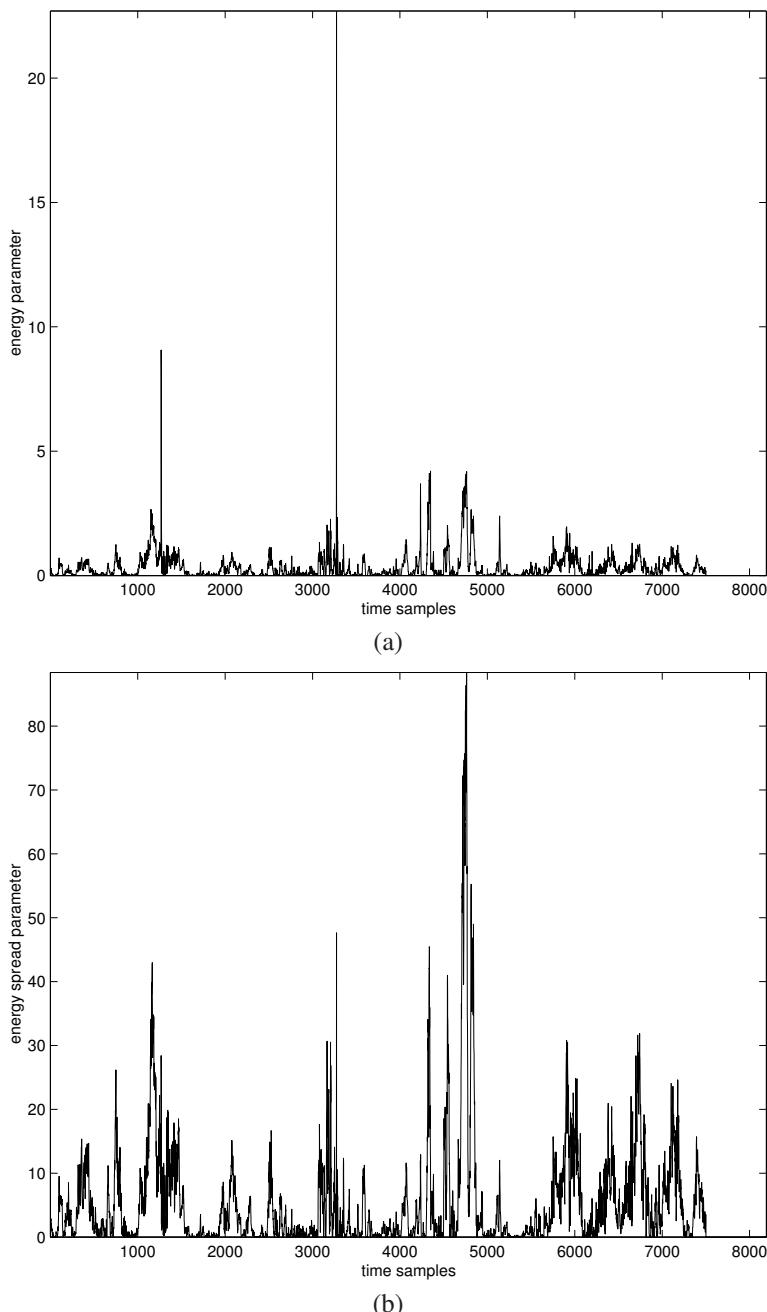


Figure 9.26 (a) EP and (b) ESP obtained from the OMPTFD of the abnormal VAG signal in Figure 9.23. Reproduced with permission from S. Krishnan, R.M. Rangayyan, G.D. Bell, and C.B. Frank, Adaptive time-frequency analysis of knee joint vibroarthrographic signals for noninvasive screening of articular cartilage pathology, *IEEE Transactions on Biomedical Engineering*, 47(6):773–783, June 2000. ©IEEE.

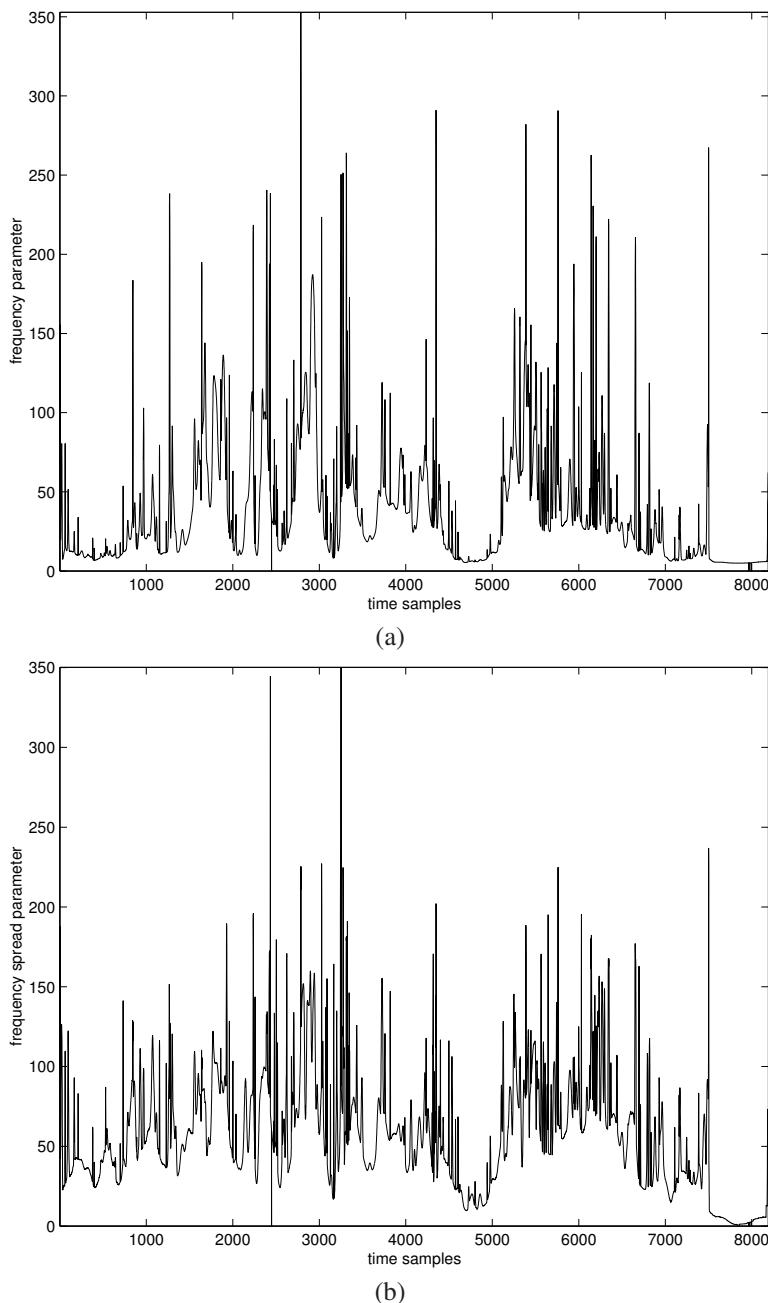


Figure 9.27 (a) FP and (b) FSP obtained from the OMPTFD of the abnormal VAG signal in Figure 9.23. Reproduced with permission from S. Krishnan, R.M. Rangayyan, G.D. Bell, and C.B. Frank, Adaptive time-frequency analysis of knee joint vibroarthrographic signals for noninvasive screening of articular cartilage pathology, *IEEE Transactions on Biomedical Engineering*, 47(6):773–783, June 2000. ©IEEE.

the values of the *CV* of EP and ESP were computed and used with the mean and *SD* of FP and FSP for pattern classification. Using a database consisting of 71 VAG signals, with 51 normal and 20 abnormal signals restricted to chondromalacia patella, an overall accuracy of 77.5% and area under the ROC curve of 0.75 were achieved. The frequency-related parameters were observed to play a major role in the classification task. See Rangayyan and Krishnan [38] for descriptions of methods to detect different types of FM components in TFDs.

It should be noted that, in the analysis of signals using AR coefficients, dominant poles, and other parameters as described in Chapters 7 and 8, each signal is divided into a variable number of quasistationary segments. In TF analysis, each signal in its entirety is represented by a small number of features and a global decision is made on each signal, rather than the segment-by-segment decision made in the other approaches mentioned. The TF method does not require labeling of signal segments as normal or abnormal, and it eliminates the need to estimate the joint angle corresponding to the pathology as observed during arthroscopy. See Krishnan et al. [31] and Kim et al. [104] for further discussion on related topics.

9.10 Application: Detection of T-wave Alternans in ECG Signals

Problem: Propose methods to detect TWA accurately and robustly in ambulatory surface ECGs for the purpose of identification of patients at risk of SCD, while accounting for noise, the presence of P and QRS components, and subtle signal variations associated with the T wave.

Solution: The detection of TWA facilitates risk stratification of heart disease patients who face the possibility of dying suddenly due to ventricular arrhythmia. TWA, often referred to as repolarization alternans, is a heart-rate-dependent phenomenon that appears in the surface ECG as a shift in the T wave's form or amplitude every second heart beat (see Section 9.2.3). TWA has gained prominence as a crucial noninvasive indicator of SCD in people with heart disease. It is difficult to quantify TWA signals in the presence of physiological noise, such as movement, breathing, and changes in heart rate, or PVC, which might have confounding effects.

When TWA is present, the T wave patterns that deviate from the typical pattern — referred to as the “A” pattern — on every other beat are referred to as having the “B” pattern. Heart beats with T waves that have almost comparable amplitudes are referred to as having an A pattern. TWA is characterized by ongoing alternation between the A and B patterns. Figure 9.28 illustrates the TWA phenomenon; the average difference between the T waves in patterns A and B is calculated as the TWA value.

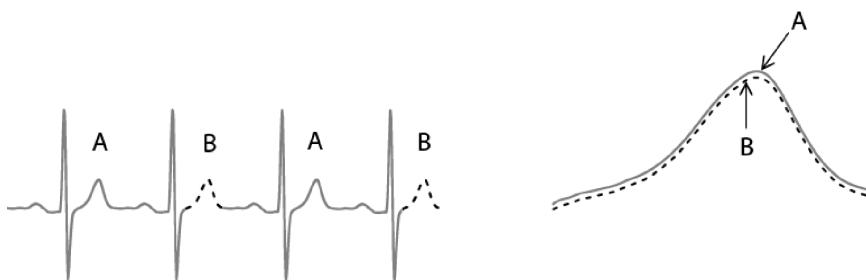


Figure 9.28 Left: TWA pattern illustration with alternate A and B patterns. Right: Average difference between patterns A and B calculated as the TWA amplitude. Reproduced with permission from B. Ghoraani [82].

The SM [5] and MMA [6] are 1D signal analysis methods based on frequency-domain and time-domain representations, respectively. In the SM, the power spectra of T waves are aligned and averaged [5], whereas, in MMA, the TWA magnitude is estimated by taking the average of the

maximum absolute differences between the even and odd heart beat sequences of T waves. The SM and MMA need alignment of T waves, and they are also not adaptable to the nonstationary properties of the ECG signals. To overcome the limitations of the spectral and MMA methods to detect TWA, a quantification framework based on 2D joint TF matrix representations of signals was proposed by Ghoraani et al. [3]. In this method, TFDs are used in conjunction with NMF to improve the robustness of TWA detection in ambulatory ECGs. The technique is referred to as the NMF-adaptive spectral method (NMFASM).

Accurate assessment of TWA magnitude is critical because a larger TWA magnitude is related to an increased risk of SCD. However, because ambulatory ECG recordings are affected by many periodic and nonperiodic noise sources, such as respiration and movement, reliable detection of TWA magnitude is difficult. The approach given by Ghoraani et al. [3] for a TWA pattern detection system is described in the following paragraphs as a viable technique for the detection of TWA; Figure 9.29 illustrates the procedure of their NMFASM technique.

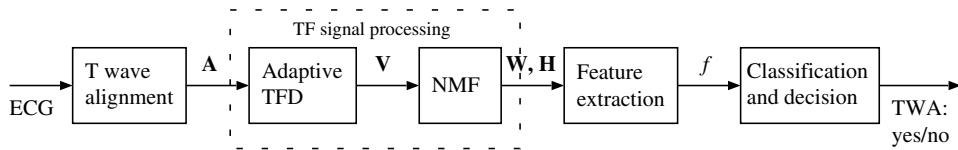


Figure 9.29 The NMF-adaptive spectral method for TWA detection. Adapted from Ghoraani et al. [3].

In the method of Ghoraani et al. [3], the multicomponent ECG signal is first processed to extract T waves, which are then aligned to form a matrix \mathbf{A} . A graphical illustration of the process is shown in Figure 9.30. Each row of \mathbf{A} contains the samples of a T wave. The columns show the beat-to-beat variation in T wave amplitude, and are referred to as the beat domain. All segmented T waves are required to have the same number of samples N . The $M \times N$ matrix \mathbf{A} for M T waves with N samples each is given by

$$\mathbf{A} = \begin{bmatrix} T_1(1) & T_1(2) & \cdots & T_1(N) \\ T_2(1) & T_2(2) & \cdots & T_2(N) \\ T_3(1) & T_3(2) & \cdots & T_3(N) \\ \vdots & \vdots & \cdots & \vdots \\ T_M(1) & T_M(2) & \cdots & T_M(N) \end{bmatrix} = [A_1 \ A_2 \ \cdots \ A_N]. \quad (9.83)$$

An adaptive TFD method is applied to each column of \mathbf{A} , and the average adaptive TFD for the aligned T waves is obtained as

$$\mathbf{V}_{\frac{M}{2} \times M} = \frac{1}{N} \sum_{i=1}^N \mathbf{V}_i, \quad (9.84)$$

where \mathbf{V}_i is the adaptive TF matrix of the i^{th} column in matrix \mathbf{A} . Ghoraani et al. [3] concentrated on decoupling TWA from background noise in the spectral region that may dilute the presence of TWA. As a result, rather than applying the NMF approach to the whole matrix, it was only applied to the last l rows of \mathbf{V} , where l is the number of samples in the spectral bandwidth of 0.36 to 0.5 in normalized frequency units, referred to here as cycles per beat (cpb). NMF is applied on the matrix $\mathbf{V}_{l \times M}$, and the TF matrix is decomposed into two matrices, \mathbf{W}_i and \mathbf{H}_i , and uses the mathematical notation shown in Equation 9.48:

$$\mathbf{V}_{l \times M} = \mathbf{W}_{l \times r} \mathbf{H}_{r \times M} = \sum_{i=1}^r \mathbf{W}_i \mathbf{H}_i, \quad (9.85)$$

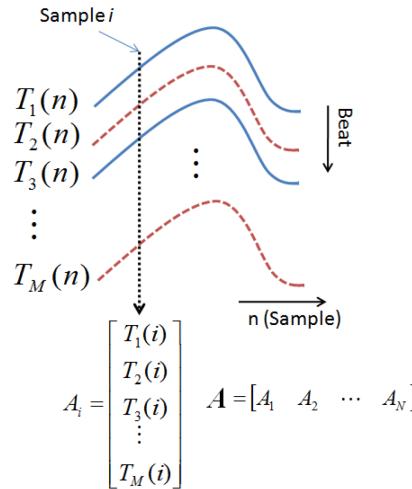


Figure 9.30 Consecutive T waves in an ECG signal extracted and aligned to form the matrix \mathbf{A} . Reproduced with permission from B. Ghoraani [82].

where \mathbf{W}_i is the i^{th} column of the matrix \mathbf{W} , r is the order of the decomposition, and \mathbf{H}_i is the i^{th} row of the matrix \mathbf{H} . $M = 64$, $l = 10$, and $r = 3$ were found to be suitable choices. NMF estimates matrices \mathbf{W} and \mathbf{H} in such a way that the columns of \mathbf{W} contain the spectral components present in the TFD matrix, and the rows of \mathbf{H} contain the corresponding temporal information.

It is assumed that information related to the spectral magnitude of TWA will be concentrated in one column (expressed by \mathbf{W}_t) due to NMF's inherent ability to represent all components with the same spectral behavior in a single column. It is possible to deduce that, by applying NMF to the TFD of the aligned T waveforms, one may isolate the desired components related to TWA from unwanted ECG components related to biological noise or possible T wave misalignment. Therefore, the \mathbf{W}_t vector is employed as a representative feature of the TWA. A feature vector is derived as

$$f_{\text{NMFASM}} = \{\mathbf{W}_t, \alpha\}, \alpha = \text{Real} \left\{ \sqrt{(T - \mu_{\text{noise}})} \right\}, \quad (9.86)$$

where T represents the magnitude of \mathbf{W}_t at 0.5 cpb , and the noise estimate is μ_{noise} calculated over the spectral bandwidth 0.44 to 0.49 cpb .

An illustration of the TFDs obtained by the technique developed by Ghoraani et al. [3] to quantify TWA in ambulatory ECG recordings is provided in Figure 9.31. TWA is simulated here by incrementing alternating beats with a rectangular pulse of $5 \mu\text{V}$. The aligned T waves are shown in Figure 9.31 (a). The reconstructed TFD, $\mathbf{V}_{l \times M}$, is depicted in Figure 9.31 (b), where $l = 10$ corresponds to the point 0.36 cpb . The decomposed matrices \mathbf{W} and \mathbf{H} obtained by NMF with decomposition order three ($r = 3$) are shown in Figure 9.31 (c) and (d). As seen in these images, the columns of the matrix \mathbf{W} correspond to the frequency components contained in the TF matrix of the T waves, whereas the rows of the matrix \mathbf{H} relate to the temporal position of each component. The magnitude of the TWA for each deconstructed component is determined, and the component with the greatest magnitude of the TWA is chosen. As seen in Figure 9.31 (c), the third component, delimited by a dashed box, exhibits the highest variation in T waves. The TFD of the third component ($\mathbf{W}_3 \mathbf{H}_3$) is illustrated in Figure 9.31 (e), and the TFD of the remaining components ($\mathbf{W}_1 \mathbf{H}_1 + \mathbf{W}_2 \mathbf{H}_2$) is illustrated in Figure 9.31 (f). As indicated by Figure 9.31 (e), the decomposed matrix correctly identifies the TWA energy at 0.5 cpb that is obscured by noise in the original TFD.

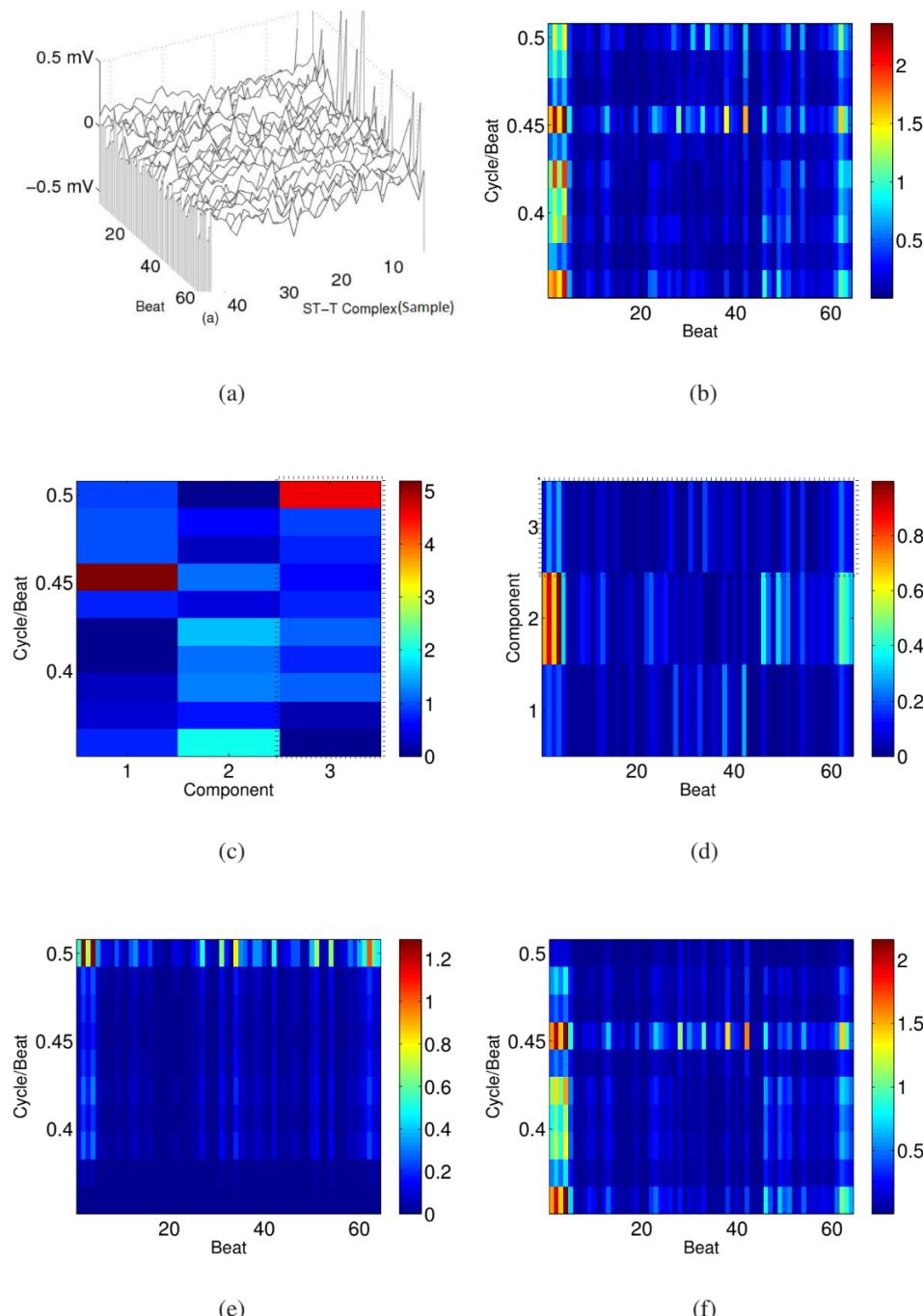


Figure 9.31 Different aspects of the NMF adaptive spectral method are shown. (a) 64 beats of aligned T waves. (b) Adaptive TFD of the T waves. (c) The spectral parts that have been decomposed. (d) The parts of time that have been decomposed. (e) The matrix of the TWA separated from the matrix of the TFD. (f) The unwanted part of the TFD. Reproduced with permission from B. Ghoraani, S. Krishnan, R.J. Selvaraj, and V.S. Chauhan, T wave alternans evaluation using adaptive time-frequency signal analysis and nonnegative matrix factorization, *Medical Engineering and Physics*, 33(6):700–711, 2011. ©Elsevier.

Ghoraani et al. [3] collected ECG recordings with inherent noise from 26 normal participants who completed 24 – 48 hours of 2-channel ambulatory ECG recording at the Toronto General Hospital. Each ECG channel was recorded separately and sampled at 125 Hz. The signals were filtered for baseline correction and QRS onset detection. The noise level was calculated as the *SD* across the first 80 ms of the period after baseline wander was corrected. The ECG was modified with a simulated TWA signal of 5 μV each by consistently raising the T wave amplitude of even beats by 2.5 μV and lowering it by 2.5 μV for odd beats, over each T wave from 40 ms after QRS offset until the end of the T wave. As a result, two sets of ambulatory ECGs were generated: one without simulated TWA and one with simulated TWA.

NMFASM was used to assess TWA on the first 64 beats of each ambulatory ECG channel. A prespecified TWA detection threshold of 5 μV was used because this value approximates the magnitude of TWA recorded in patients with heart disease by Klingenberg et al. [105]. The NMFASM feature set, the TWA, the $K_{\text{score}}(t)$, and the TWA using MMA were computed from the 64-beat data of each ambulatory ECG group. The feature $K_{\text{score}}(t)$ was computed as $\frac{T(t) - \mu_{\text{noise}}(t)}{\sigma_{\text{noise}}(t)}$, where $T(t)$ is the instantaneous amplitude, $\sigma_{\text{noise}}(t)$ is the *SD* of noise, and $\mu_{\text{noise}}(t)$ is the mean value of noise of the SM. The collected characteristics were passed to an LDA classifier, which classified the ECG segments as having or not having TWA. To assess the accuracy of the approaches to detect TWA, the areas under the corresponding ROC curves were calculated; see Figure 9.32. The areas under the ROC curves for NMFASM, the SM, and MMA are 0.92, 0.74, and 0.7, respectively, demonstrating the superiority of NMFASM.

The lack of the TWA signal at particular heart rates is significant in cardiology because the risk of adverse cardiac events is low in such individuals with heart disease and therapy is unnecessary [106]. As a result, it is critical to maintain 100% specificity when using a TWA detection technique. ROC analysis indicated the highest sensitivity for TWA detection while maintaining 100% specificity is 48% for NMFASM and 20% for the SM's K_{score} , indicating a 140% improvement over the commonly used TWA detection method.

9.11 Application: Extraction of the Fetal ECG from Single-channel Maternal ECG

Problem: Develop a single-channel fetal ECG extraction and monitoring system that would be able to achieve energy-efficient transmission and cost-effective fetal healthcare for continuous remote monitoring applications.

Solution: Figure 9.33 illustrates the spectral overlap between fetal and maternal ECG signals. It is evident that linear filters cannot separate the two signals. Separating the ECG of a fetus from multiple channels of ECG from the expectant mother may be viewed as a BSS problem, and several approaches for this purpose have been reported in the literature; see Sections 9.2.1 and 9.7.2. PCA is a method to analyze the correlation between multiple signals by rotating the axis in the direction with the largest covariance (in the present discussion between the maternal ECG and the fetal ECG). Numerous PCA-based fetal ECG extraction techniques have been proposed. Bacharakis et al. [107] proposed an approach based on SVD, in which the decomposed orthogonal matrices contain the left and right singular vectors, on the assumption that the fetal ECG signal may be extracted by linearly combining the multilead abdominal ECG signals. Callaerts et al. [66] proposed an approach based on PCA and generalized SVD.

Due to the fact that PCA alone does not provide good fetal ECG separation, several techniques combining PCA and template subtraction have been described in the literature. Lipponen and Tarvainen [108] used template subtraction in combination with principal component regression (PCR) to remove the maternal ECG from the abdominal ECG signals. The P, T, and QRS complexes were segmented and used as templates in the PCR approach; the templates were dynamically removed from abdominal ECG signals using the most approximate PCR basis vectors. Rather than computing a single template, Christov et al. [109] proposed that an ensemble of maternal ECG complexes

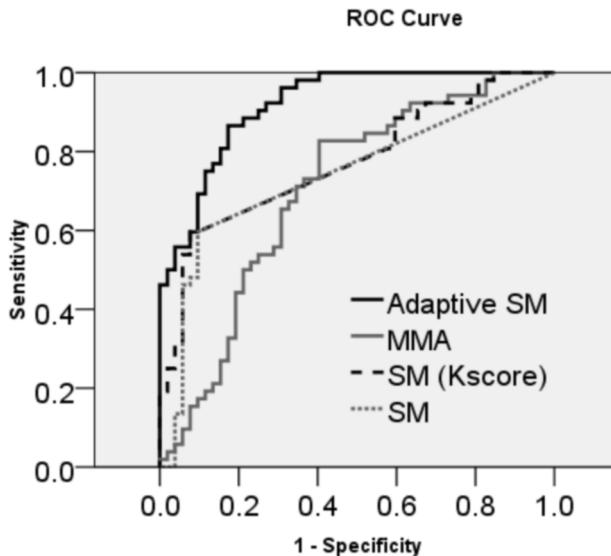


Figure 9.32 ROC curves for the SM, MMA, and NMFASM (adaptive SM). Ambulatory ECGs without TWA are regarded as negative in this study and ECGs with TWA are considered positive. The area under the ROC curve for the NMFASM method and the $K_{\text{score}}(t)$ of the spectral method are 0.92 and 0.77, respectively, with a $p < 0.001$ significance level. The area under the ROC curve for SM and MMA are 0.74 and 0.7, respectively. Reproduced with permission from B. Ghoraani, S. Krishnan, R.J. Selvaraj, and V.S. Chauhan, T wave alternans evaluation using adaptive time-frequency signal analysis and nonnegative matrix factorization, *Medical Engineering and Physics*, 33(6):700–711, 2011. ©Elsevier.

be used. Most of the PCA-based fetal ECG separation techniques require multiple channels of the abdominal ECG signal; see Section 9.7.2.

ICA is a popular approach for BSS; see Section 9.7.2. Several ICA-based techniques are described by Andreotti et al. [110]. Most notably, the FICA method is described by Varanini et al. [111]. The basic idea behind these algorithms is to take a multichannel abdominal ECG as input and separate the mixed sources by optimizing statistical independence under the assumption of non-Gaussian and linearly mixed sources. Andreotti et al. [110] proposed a method that incorporates two approaches to estimate the maternal ECG: the extended Kalman smoother and template adaptation in combination with ICA. Varanini et al. [111] proposed an effective and unsupervised method to extract fetal ECG signals from multichannel abdominal signal recordings. This method extracts the maternal ECG using ICA and cancels it using weighted SVD. Da Poian et al. [112] developed a technique that integrates the ideas of compressive sensing and ICA, in which they applied ICA directly to the abdominal ECG data.

NMF is another technique that has been applied for fetal ECG separation. Mirzal [113] compared the performance of NMF with that of ICA. He and Chen [114] used a single-channel fetal ECG extraction technique based on EMD (see Section 9.4) and NMF, which decomposes a single-channel ECG into many channels using EMD. Because the number of decomposed signals is typically greater than the number of sources, as determined via EMD, real-time implementation is not feasible. Samieinasab and Sameni [115] devised a technique to extract fetal heart sound from acous-

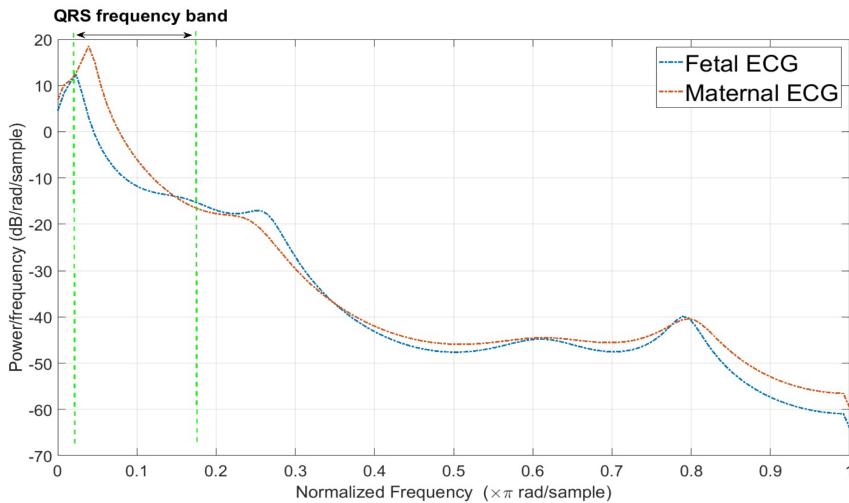


Figure 9.33 Illustration of the spectral overlap between fetal and maternal ECG signals. Reproduced with permission from D. Gurve and S. Krishnan, Separation of fetal-ECG from single-channel abdominal ECG using activation scaled non-negative matrix factorization. *IEEE Journal of Biomedical and Health Informatics*, 24(3):669–680, 2020. ©IEEE.

tic waves recorded from the maternal abdominal surface; the strategy was based on single-channel BSS, combining EMD and NMF to decouple the various sources from signal mixtures.

Most PCA- and ICA-based methods require multilead abdominal ECG signals for fetal ECG extraction, which increases the demand for processing time and transmission power. Single-channel ECG signal acquisition is simple and convenient for at-home monitoring of expectant mothers. There are a few approaches published for fetal ECG separation utilizing a single-lead ECG, including those described by Lamesgin et al. [116], Panigrahy and Sahu [117], Almeida et al. [118], and Su and Wu [119]. Due to the lack of a reference channel required in adaptive filtering applications, BSS of fetal and maternal ECG from a single-channel abdominal ECG recording could be achieved by 2D covariance matrix representation. This approach has been tested under three different simulated scenarios: (i) the original abdominal ECG is used for fetal ECG separation, (ii) the recovered abdominal ECG after compression is used for separation, and (iii) the fetal ECG is extracted from the compressed domain of the abdominal ECG. Gurve and Krishnan [120] proposed the framework shown in Figure 9.34 for fetal ECG extraction from single-lead ECGs using the NMF approach, which is described in the following paragraphs.

The mixed multichannel ECG signal could be expressed as

$$x_{ai}(t) = x_{mi}(t) + x_{fi}(t) + n(t), \quad (9.87)$$

where $x_{ai}(t)$ represents the i^{th} channel of the mixed abdominal ECG signal, the maternal ECG signal is given by $x_{mi}(t)$, the fetal ECG signal is given by $x_{fi}(t)$, and $n(t)$ denotes noise in the i^{th} abdominal ECG signal. In the case of a single-channel abdominal ECG $x_a(t)$, Equation 9.87 becomes

$$x_a(t) = x_m(t) + x_f(t) + n(t). \quad (9.88)$$

The abdominal ECG signal consists of a mixture of the maternal ECG, fetal ECG, and noise; thus, it is important to reduce the noise before the separation procedure. To reduce baseline drift, the raw abdominal ECG signal is processed through a highpass Butterworth filter, with the cutoff

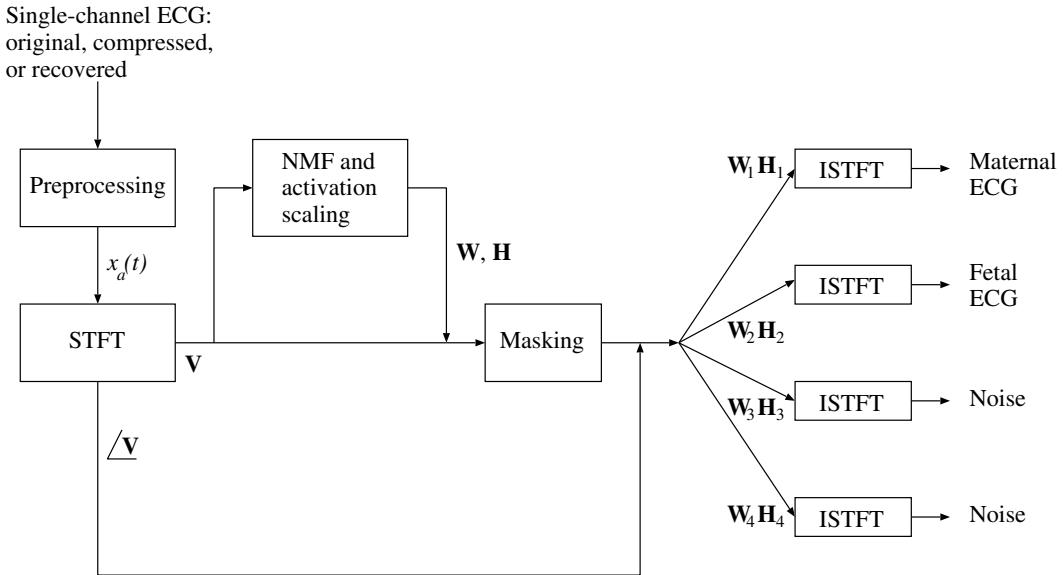


Figure 9.34 Block diagram of the NMF-based method for separation of maternal and fetal ECG. ISTFT = inverse STFT. Reproduced with permission from D. Gurve and S. Krishnan, Separation of fetal-ECG from single-channel abdominal ECG using activation scaled non-negative matrix factorization. *IEEE Journal of Biomedical and Health Informatics*, 24(3):669–680, 2020. ©IEEE.

frequency equal to 2 Hz. The abdominal signal is additionally processed through a notch filter to suppress the 50 Hz power-line interference.

In order to apply NMF to a single-channel abdominal ECG signal $x_a(t)$, it is first converted into an $M \times N$ TF matrix \mathbf{V} using the STFT. For the NMF decomposition, Gurve and Krishnan [120] used the sparse NMF algorithm originally proposed by Eggert and Körner [78]. The update rules in the case of sparse NMF are as follows [78]:

$$H_{ij} \leftarrow H_{ij} \frac{\mathbf{V}_i^T \overline{\mathbf{W}}_j}{\mathbf{R}_i^T \overline{\mathbf{W}}_j + \lambda}, \quad (9.89)$$

$$\mathbf{W}_j \leftarrow \mathbf{W}_j \odot \left(\left\{ \sum_i H_{ij} [\mathbf{V}_i + (\mathbf{R}_i^T \overline{\mathbf{W}}_j) \overline{\mathbf{W}}_j] \right\} \oslash \left\{ \sum_i H_{ij} [\mathbf{R}_i + (\mathbf{R}_i^T \overline{\mathbf{W}}_j) \overline{\mathbf{W}}_j] \right\} \right), \quad (9.90)$$

where H_{ij} denotes the j^{th} element of the i^{th} column of the \mathbf{H} matrix, \mathbf{W}_j is the j^{th} column of the matrix \mathbf{W} , $\overline{\mathbf{W}}_j = \frac{\mathbf{W}_j}{\|\mathbf{W}_j\|}$ denotes the normalized basis vectors, $\mathbf{R}_i = \sum_j H_{ij} \overline{\mathbf{W}}_j$ is the estimated matrix, the sparsity parameter $\lambda \geq 0$, $i = 1, 2, \dots, N$, and $j = 1, 2, \dots, r$. The factorization rank r is a significant parameter in NMF. Gurve and Krishnan [120] used NMF ranks ranging from $r = 3$ to 6 and empirically determined that $r = 4$ provided the best average fetal QRS separation performance.

A key consideration is that the decomposed activation matrix \mathbf{H} can be modified to suit the unmixing process. Typically, the number of QRS peaks in a fetal ECG is higher than the number of QRS peaks in the maternal ECG for a given time duration; thus, the maternal and fetal activations may be selected by using a threshold value T , such as

$$T = \tau A_{\max}, \quad (9.91)$$

and by counting the number of peaks P_r in the activation row (\mathbf{H}_r) that are above the threshold T . Gurve and Krishnan selected A_{\max} as the highest amplitude in a given signal block, $\tau = 0.6$ for the maternal ECG, and $\tau = 0.45$ for the fetal ECG. With the peak counts sorted in ascending order, the minimum number of peaks ($P_r(1)$) corresponds to the maternal activation $\mathbf{H}_{\text{maternal}}$ and the next peak count ($P_r(2)$) corresponds to the fetal activation $\mathbf{H}_{\text{fetal}}$. The remaining peaks [$P_r(3)$ and $P_r(4)$] typically correspond to noise.

To ensure that fetal ECG activation takes precedence over maternal ECG activation, each row of the activation matrix was first normalized, and then a threshold was applied. Figure 9.35 (a) shows the normalized maternal and fetal ECG activations. Figure 9.35 (b) shows the ECG plots with dominant maternal activation in which $\mathbf{H}_{\text{maternal}}$ is multiplied by a number greater than one, and $\mathbf{H}_{\text{fetal}}$ is multiplied by a number less than one. Figure 9.35 (c) shows ECG plots with dominant fetal activation in which $\mathbf{H}_{\text{fetal}}$ is multiplied by a number greater than one, and $\mathbf{H}_{\text{maternal}}$ is multiplied by a number less than one. The signals were reconstructed by computing the inverse STFT after applying a soft mask (\mathbf{M}_s), defined as

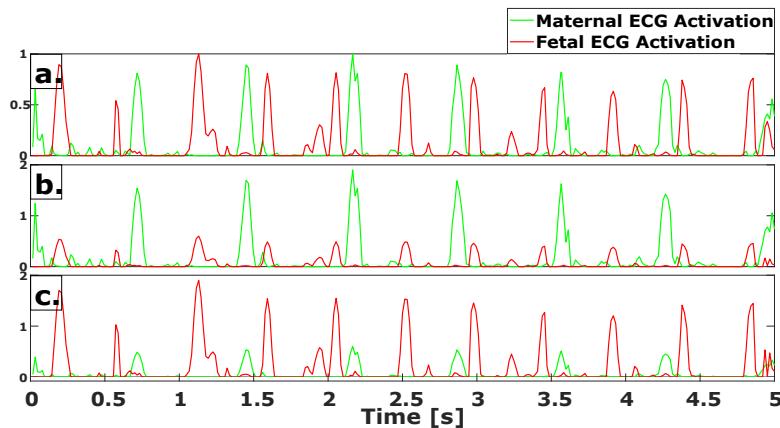


Figure 9.35 (a) Maternal and fetal activations after normalization. (b) Maternal and fetal activations with dominating maternal activations. (c) Maternal and fetal activations with dominating fetal activations. Reproduced with permission from D. Gurve and S. Krishnan, Separation of fetal-ECG from single-channel abdominal ECG using activation scaled non-negative matrix factorization. *IEEE Journal of Biomedical and Health Informatics*, 24(3):669–680, 2020. ©IEEE.

$$\mathbf{M}_s = (\mathbf{W} \mathbf{H}) \oslash \left(\sum_{i=1}^r \mathbf{W}_i \mathbf{H}_i \right). \quad (9.92)$$

The source signals were estimated as

$$|\hat{\mathbf{V}}| = \mathbf{M}_s \odot |\mathbf{V}|. \quad (9.93)$$

Additionally, by calculating the inverse STFT of each $|\hat{\mathbf{V}}|$ and multiplying by $\angle \mathbf{V}$, the separated fetal ECG, maternal ECG, and noise components were extracted. Figure 9.36 illustrates an example of separated fetal and maternal ECG signals using the method of Gurve and Krishnan [120]. Figure 9.36 (a) depicts the first 5 s of the original abdominal ECG signal for record “r01” of the Silesia dataset [121], demonstrating that the fetal ECG peaks are much smaller in magnitude than the maternal ECG peaks. Figure 9.36 (b) depicts the filtered version of the same ECG signal. Figure

9.36 (c) shows the isolated maternal ECG signal; it can be seen that the influence of the fetal ECG has been nearly completely eliminated due to the larger scaling of the maternal activation following factorization. Figure 9.36 (d) depicts a distinct fetal ECG separated from the single-channel abdominal ECG. Again, it is readily apparent that, in the isolated fetal ECG signal, the influence of the maternal ECG has been eliminated, and fetal ECG dominance has been increased.

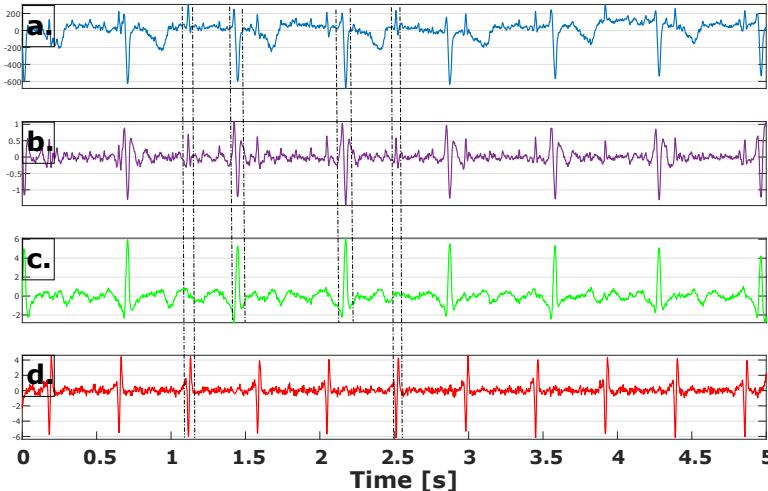


Figure 9.36 (a) First 5 s of the “r01” original abdominal ECG signal from the Silesia dataset [121]. (b) Filtered abdominal ECG signal. (c) Separated maternal ECG signal. (d) Separated fetal ECG signal. Reproduced with permission from D. Gurve and S. Krishnan, Separation of fetal-ECG from single-channel abdominal ECG using activation scaled non-negative matrix factorization. *IEEE Journal of Biomedical and Health Informatics*, 24(3):669–680, 2020. ©IEEE.

9.12 Application: EEG Analysis for Brain–Computer Interfaces

Problem: Develop a BCI system that allows for effective EEG channel selection, without compromising the detection performance of the overall system, so as to achieve fast and accurate performance under hardware complexity constraints.

Solution: Gurve et al. [122] proposed a BCI system for lower limb motor imagery detection, which utilizes techniques of NMF-based EEG channel selection and the neighborhood component feature selection (NCFS) method for feature selection, without requiring any prior knowledge of the nature of the motor imagery task. This facilitates the development of a robust system with a small number of EEG channels, while also maintaining, and in some cases, improving the detection performance of the BCI system.

BCIs provide a link between the human brain and the computer, allowing the user to control a device such as a wheelchair or a neuroprosthesis by just thinking about or imagining the required action. Cortical activations linked to motor imagery can be used to operate a range of rehabilitation equipment, including robotic devices to regain mobility in the lower and upper limbs. Such robotic devices aid in rehabilitation of the lower and upper limbs of stroke victims.

Motor-imagery-based treatment is particularly beneficial for the recovery of movement deficits since it engages the brain areas responsible for the completion of a particular action [123]. However, proper motor imagery classification is crucial for the optimization of the use of motor imagery in EEG-based BCIs for brain rehabilitation. The optimal EEG channels for each individual linked

with a particular cognitive task form critical information for the development of successful motor imagery classification models [124]. The specific EEG channels demonstrate how motor-task-dependent and event-related activity changes. As a result, they shed light on the particular brain location associated with lower-limb motor imagery. Additionally, because the role of specific EEG channels is unpredictable, dealing with motor imagery tasks needs advanced and specialized signal processing techniques for channel selection. EEG channels may include redundant data and noise, making it difficult to extract key properties. The most appropriate EEG channels must be chosen to extract useful feature vectors from multichannel EEG data for motor imagery. The majority of related approaches incorporate all available EEG channels and associated variables into the classifier, resulting in redundancy, increased complexity, and reduced efficiency.

Most of the available methods employ a collection of EEG channels from various brain regions. Because of the large intersubject variability of EEG signal properties, EEG recordings from fixed or predetermined sites may not deliver the best results. The computational complexity, processing time, and spatial complexity of a method rise as the number of input channels and their complexities increase. Overfitting can be compounded by unnecessary channels and redundant features, lowering the efficiency of performance. Most EEG channel selection methods are based on exhaustive search techniques and are time-consuming.

Rahman and Joadder [2] offered several solutions for each step of motor-imagery-based BCI, as well as a comprehensive comparative performance study, which can assist in selecting the optimal methodology for a specific experimental situation. Their study revealed that effective channel and feature selection are critical to enhancement of the performance of a BCI system. Jiang et al. [125] proposed a BCI system that detects the onset of gait by analyzing movement-related brain potentials. A Laplacian spatial filter was employed to increase the SNR of movement-related brain potentials. This technique makes use of ICA to reduce automatically EEG artifacts and improve detection performance. On the other hand, some studies indicate that NMF outperforms ICA [113].

While common spatial patterns (CSP) are often used and have been shown to be effective, the CSP technique has limited applicability in situations with small numbers of signal samples [126]. Another challenge with CSP-based techniques is that several parameters need to be determined to achieve optimal results, such as the frequency range, the time interval between stimuli, and the subset of CSP filters.

Alotaiby et al. [127] published a review on EEG channel selection methods for a number of applications, such as motor imagery classification, sleep staging, seizure detection/prediction, and mental task categorization. Yang et al. [128] suggested a new channel selection approach based on subject-specific temporal and spatial analysis. For critical channel selection, this method employs a wrapper approach with a predefined search strategy. Qiu et al. [129] suggested a sequential floating forward selection approach, in which neighboring EEG channels are considered as features for channel weighting, and EEG channels are selected or eliminated based on the channel weight. Feng et al. [130] suggested a technique to select relevant EEG channels based on motor imagery classification that combines CSP-rank channel selection and multiband signal decomposition approaches. Liu et al. [131] suggested a Fisher's-criterion-based channel selection strategy to find the optimal patient-dependent arrangement of channels for motor imagery detection, based on the Fisher score of the fractal dimension generated from each EEG channel. Gurve et al. [122] proposed NMF-based channel selection approaches, demonstrating that reducing the number of EEG channels reduces computational complexity and processing time while boosting the performance of motor imagery detection. The method detects EEG regions related to a specific cognitive task, such as upper- and lower-limb motion, which could be valuable for neurorehabilitation, allowing for a realistic and real-time connection between the patient and the robotic equipment.

9.12.1 NMF-based channel selection

Gurve et al. [122] proposed an NMF-based channel selection method for motor imagery classification. The procedure begins with the creation of a covariance matrix \mathbf{C}_i from the data \mathbf{X}_i of the i^{th} trial as

$$\mathbf{C}_i = \begin{pmatrix} \text{Var}(\mathbf{x}_1) & \cdots & \text{Cov}(\mathbf{x}_N, \mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \text{Cov}(\mathbf{x}_1, \mathbf{x}_N) & \cdots & \text{Var}(\mathbf{x}_N) \end{pmatrix}. \quad (9.94)$$

Here, $\mathbf{X}_i = \{\mathbf{x}_1(t), \mathbf{x}_2(t), \dots, \mathbf{x}_N(t)\}$ denotes the i^{th} trial that took place at time t . Each trial is represented by an $N \times T$ data matrix where N is the number of EEG channels and T is the number of sample points in the specified time-windowed segment. The matrix \mathbf{C}_i corresponding to the i^{th} trial data \mathbf{X}_i is a covariance matrix in which the diagonal elements of \mathbf{C}_i are the variance values of the individual EEG channels, and the nondiagonal elements are the covariance values of pairs of channels. As the matrix \mathbf{C}_i is a positive-definite matrix, NMF may be applied for further analysis and channel selection.

NMF describes the data in terms of a multiplicative mixture of nonnegative factors that correspond to realistic building blocks in the original data. Gurve et al. [122] used the sparse NMF on \mathbf{C}_i , that is, $\mathbf{V} = \mathbf{C}_i$, and the matrix update rules as described in Section 9.11. One modification performed in the present application to the original sparse NMF algorithm is the normalization of the j^{th} row \mathbf{W}_j of the activation matrix \mathbf{W} , which is expected to help in the selection of the optimal EEG channel weight, as

$$\mathbf{W}_j = \frac{\mathbf{W}_j - \min(\mathbf{W}_j)}{\max(\mathbf{W}_j) - \min(\mathbf{W}_j)}. \quad (9.95)$$

Each normalized row is then compared with \mathbf{W}_{ref} , which is a vector with all elements equal to 0.5, to calculate the *RMS* deviation (*RMSD*) as

$$\text{RMSD}(\mathbf{W}_j, \mathbf{W}_{\text{ref}}) = \sqrt{\frac{1}{N} \sum_{j=1}^N \|\mathbf{W}_j - \mathbf{W}_{\text{ref}}\|^2}. \quad (9.96)$$

In the final step of the EEG channel selection process, the estimated channel weights are multiplied with the corresponding selected EEG channels.

9.12.2 Feature extraction

Gurve et al. [122] applied the NMF-based channel selection/reduction approach to EEG signals from three motor imagery BCI datasets.

Dataset 1 — UFES MI dataset [122]: EEG data were captured over the main and supplementary motor cortex in the frequency range of 0.1 to 100 Hz using BrainNet BNT 36 (EMSA, Brazil). Signals from the channels Fz, FC3, FC1, FCz, FC2, FC4, C5, C3, C1, Cz, C2, C4, C6, CP3, CP1, CPz, CP2, CP4, and Pz were collected. Figure 9.37 shows the locations of the 19 EEG electrodes that were employed using a cap with 64 electrodes ($Ag - AgCl$). The ground electrode was placed between the eye brows, and the reference electrodes were placed on the A1 and A2 earlobes. Because brain activity related to real or imagined motor action is largely generated in the motor cortex, investigators have previously used similar electrode configurations to explore lower-limb motor intention. Surface EMG signals were also acquired from both lateral gastrocnemius muscles and recorded with the frequency range of 10 to 100 Hz. The EMG signals were used to ensure that no muscular contraction occurred while the pedaling motor imagery time intervals were

being annotated. Another channel was used to record the Arduino board's synchronizing signal to demarcate the raw EEG segments that corresponded to the Resting and Pedaling classes. The sampling frequency used was 400 Hz, and a notch filter at 60 Hz was used to remove the power-line interference from the EEG signals.

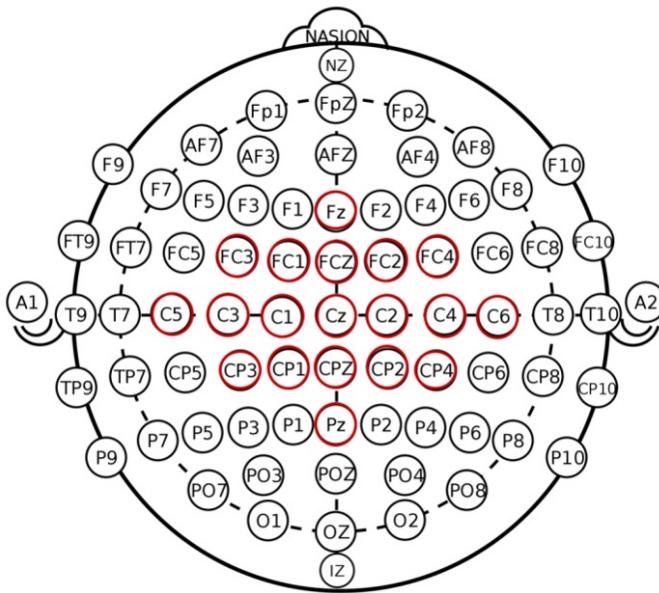


Figure 9.37 EEG electrode positions used for BCI with the 19 relevant electrodes marked in red circles. Reproduced with permission from D. Gurve, D. Delisle-Rodriguez, M. Romero-Laiseca, V. Cardoso, F. Loterio, T. Bastos, and S. Krishnan, Subject-specific EEG channel selection using non-negative matrix factorization for lower-limb motor imagery recognition. *Journal of Neural Engineering*, 17(2):026029, 2020. ©IOP Publishing.

To prepare the training and validation datasets, the process was carried out in two parts. Both stages were made up of six sessions with 12 trials each. A dark screen, a red cue, a yellow cue, and a green cue were used. Figure 9.38 shows the cue sequences that were used to lead subjects through the software.

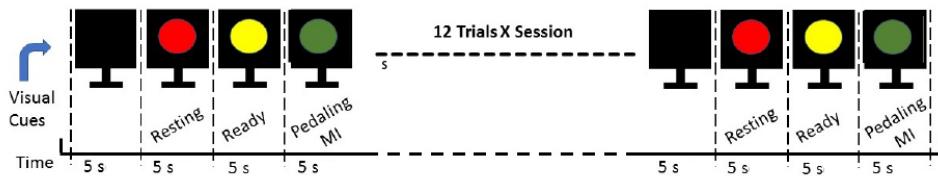


Figure 9.38 Visual cue and timeline of the Gurve et al. [122] protocol. MI: motor imagery. Reproduced with permission from D. Gurve, D. Delisle-Rodriguez, M. Romero-Laiseca, V. Cardoso, F. Loterio, T. Bastos, and S. Krishnan, Subject-specific EEG channel selection using non-negative matrix factorization for lower-limb motor imagery recognition. *Journal of Neural Engineering*, 17(2):026029, 2020. ©IOP Publishing.

For pedaling motor imagery, 10 healthy adults were evaluated using the motor imagery detection method. The Universidade Federal do Espírito Santo (UFES) Ethics Committee approved this study, which included participants without lower-limb injuries or restrictions in movement. The experiment was divided into four stages: ethical form completion, familiarization of protocol, electrode setup, and protocol execution. During the experiment, the participants sat in seats and watched

a monitor. Figure 9.39 shows the experimental setup in detail. The subjects were requested to mentally ride a bicycle while imagining a real-world incident. The subjects were directed to mentally pedal on the motorized pedal after determining the best strategy to execute the pedaling motor imagery.

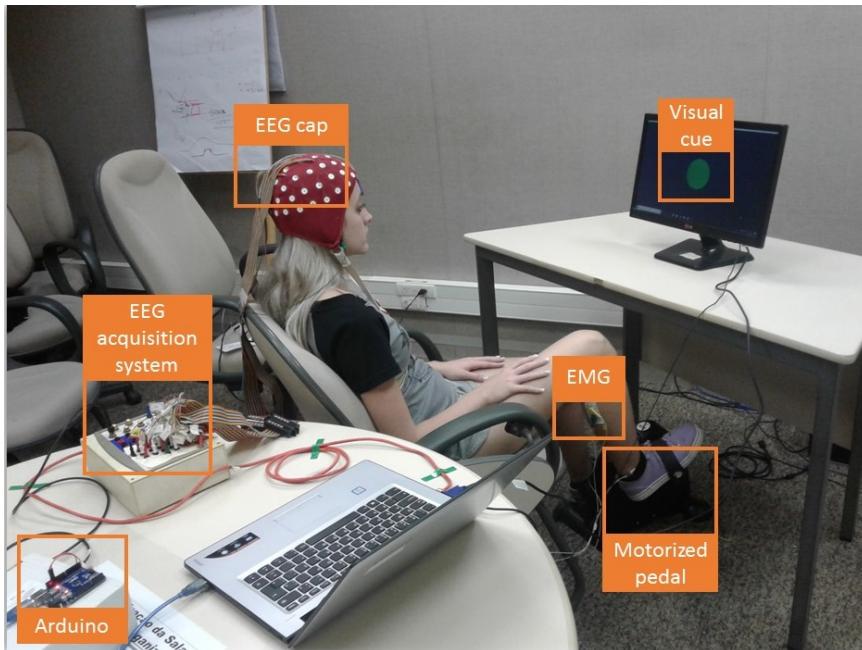


Figure 9.39 Experimental setup for motor imagery data collection. Reproduced with permission from D. Gurve, D. Delisle-Rodriguez, M. Romero-Laiseca, V. Cardoso, F. Loterio, T. Bastos, and S. Krishnan, Subject-specific EEG channel selection using non-negative matrix factorization for lower-limb motor imagery recognition. *Journal of Neural Engineering*, 17(2):026029, 2020. ©IOP Publishing.

Dataset 2 — BCI competition IV motor imagery IIa: This publicly available benchmark dataset [132] has 22 EEG channels from nine subjects (S1 to S9). Motor imagery EEG signals were collected for right- and left-hand movements, as well as for foot and tongue movements. Bandpass filtering of the acquired data between 0.5 and 100 Hz was performed and a 50 Hz notch filter was applied. The data collection took place during two sessions, each with six runs. Figure 9.40 illustrates the cue sequences and timeline used to direct the individuals. There were a total of 288 trials, with each run consisting of 48 trials. The data were recorded at the sampling rate of 250 Hz.

Dataset 3 — BCI competition III dataset IVa: This freely accessible benchmark dataset was created using 118 EEG channels collected from five individuals. The data were collected at the sampling rate of 1,000 Hz and then downsampled to 100 Hz. Motor imagery EEG signals for left-hand, right-hand, and right-foot activities were obtained using 3.5s visual stimuli with a random interval between 1.75 and 2.25 s. For training purposes, the left-hand, right-hand, and right-foot motor imagery data were supplied; however, for assessment purposes, only the right-hand and right-foot motor imagery data were provided.

If S is the total number of selected EEG channels using NMF, then the covariance matrix \mathbf{C}_S of the selected EEG channels can be obtained as in Equation 9.94, changing the number of EEG channels from N to S . As a result, the corresponding spatial covariance matrices \mathbf{C}_{Si} for the decreased \mathbf{X}_i may be constructed as

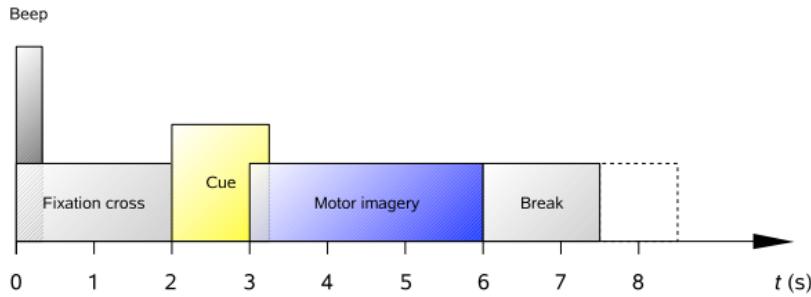


Figure 9.40 The cue sequences and timeline used to guide the subjects for Dataset 2. Reproduced with permission from D. Gurve, D. Delisle-Rodriguez, M. Romero-Laiseca, V. Cardoso, F. Loterio, T. Bastos, and S. Krishnan, Subject-specific EEG channel selection using non-negative matrix factorization for lower-limb motor imagery recognition, *Journal of Neural Engineering*, 17(2):026029, 2020. ©IOP Publishing.

$$\mathbf{C}_{Si} = \frac{\mathbf{X}_i \mathbf{X}_i^T}{(T - 1)}, \quad (9.97)$$

where T is the number of samples in each EEG segment. Examples of the original and reduced covariance matrices for a few subjects are shown in Figure 9.41. It is seen that the reduced covariance matrix has provided a good representation of the covariance entries, which could be further analyzed for channel selection.

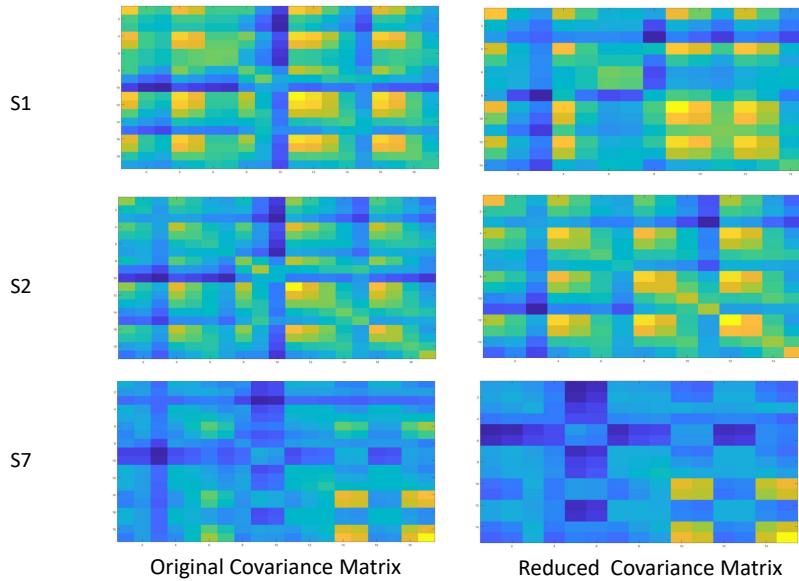


Figure 9.41 Examples of the original and reduced covariance matrices via NMF. Reproduced with permission from D. Gurve [133].

Taking into account the positive-definite nature of the covariance matrix \mathbf{C}_S , Gurve et al. [122] used Riemannian geometrical theory [134] to extract spatial features for the motor imagery classifi-

cation task. The features derived from each reduced covariance matrix $\mathbf{C}_S \in \mathbb{R}^{S \times S}$ were employed for classification. Given that the scalp surfaces from which EEG signals are acquired are not flat surfaces, the commonly used Euclidean geometry does not readily apply for topographic descriptions for BCI. Riemannian geometry, which can be applied to curved surfaces, is a good alternative to apply to topographic surfaces in BCI studies. When projected to the tangent space (a generalization of tangent lines used in 2D planes and higher dimensions), the reference point, or Riemannian mean, is a statistical descriptor that gives a good local approximation of a set of matrices. Specifically, in the Riemannian approach, \mathbf{C}_{Si} is projected on to the tangent space using logarithmic mapping $\log_C(\mathbf{C}_{Si})$ given as [135]

$$\log(\mathbf{C}_{Si}) = \mathbf{C}_R^{1/2} \log(\mathbf{C}_R^{-1/2} \mathbf{C}_{Si} \mathbf{C}_R^{-1/2}) \mathbf{C}_R^{1/2}, \quad (9.98)$$

where \mathbf{C}_R is a reference point in the Riemannian manifold. Given a set $\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_I$ of positive-definite matrices, the reference point \mathbf{C}_R is defined as [135]

$$\mathbf{C}_R = \operatorname{argmin}_{i=1}^I R_d^2(\mathbf{C}_S, \mathbf{C}_{Si}), \quad (9.99)$$

where the Riemannian distance R_d between \mathbf{C}_1 and \mathbf{C}_2 is the arc length of the curve connecting them, given by [135]

$$R_d(\mathbf{C}_1, \mathbf{C}_2) = \|\log(\mathbf{C}_1^{-1/2} \mathbf{C}_2 \mathbf{C}_1^{-1/2})\|_F, \quad (9.100)$$

where $\|\cdot\|_F$ is the Frobenius norm [135].

Due to the fact that the majority of linear classification methods for motor imagery classification need input features in the form of vectors, covariance matrices cannot be utilized directly as input features. The problem may be solved by projecting the covariance matrices to the Riemannian tangent space. Thus, the matrices may be vectorized with a vector dimension of $(Ch + 1)Ch/2$, and is given by

$$vect(\mathbf{C}) = [C_{1,1}; \sqrt{2}C_{1,2}; C_{2,2}; \sqrt{2}C_{1,3}; \dots; C_{Ch,Ch}], \quad (9.101)$$

where Ch is the total number of channels and $C_{i,j}$ denotes the i^{th} row and j^{th} column element of the matrix \mathbf{C} . The *vect* operation vectorizes the upper triangular section of a symmetric matrix by assigning a weight of 1 to diagonal members and $\sqrt{2}$ to nondiagonal elements.

In order to select features of importance, Gurve et al. [122] used the NCFS algorithm. NCFS employs a regularization term to select adaptively the optimal subset of characteristics and enhance motor imagery classification accuracy. Following the feature extraction stage, the motor imagery training sample set is represented as $T_{\text{train}} = \{(\mathbf{f}_1, l_1), (\mathbf{f}_2, l_2), \dots, (\mathbf{f}_i, l_i), \dots, (\mathbf{f}_n, l_n)\}$; here \mathbf{f}_i denotes a feature vector and l_i is the corresponding class label. The NCFS method calculates a weight vector \mathbf{w} , which helps in selecting the feature subset related to the motor intention event. The weighted distance function between two sample vectors \mathbf{f}_i and \mathbf{f}_j is defined as

$$\mathbf{D}_w(\mathbf{f}_i, \mathbf{f}_j) = \sum_{r=1}^n \mathbf{w}_r^2 |\mathbf{f}_{ir} - \mathbf{f}_{jr}|, \quad (9.102)$$

where \mathbf{w}_r is the weight vector associated with the r^{th} feature. The NCFS method selects a random feature vector and the corresponding class label from T_{train} as a reference point. Let p_{ij} denote the probability of selecting a reference feature vector from T_{train} , expressed as

$$p_{ij} = \frac{k(\mathbf{D}_w(\mathbf{f}_i, \mathbf{f}_j))}{\sum_{j=1}^n k(\mathbf{D}_w(\mathbf{f}_i, \mathbf{f}_j))}, \quad (9.103)$$

where $k(z) = \exp\left(\frac{-z}{\sigma}\right)$ is the kernel representing a similarity function, and σ corresponds to the width of the kernel. In terms of correct detection of any motor imagery trial i using T_{train} , the average LOO probability is given as

$$p_i = \sum_{j=1, j \neq i}^n p_{ij} l_{ij}, \quad (9.104)$$

where $l_{ij} = 1$ if $l_i = l_j$ and 0 otherwise. NCFS focuses on maximizing the LOO classification accuracy $F(\mathbf{W})$ by including a regularization term as λ

$$F(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^n p_i - \lambda \sum_{r=1}^n |\mathbf{w}_r|^2. \quad (9.105)$$

The methodology to select the most significant elements based on their weights varies according to the threshold Th , given by

$$Th = \tau \max(\mathbf{w}), \quad (9.106)$$

where τ denotes a tolerance value set manually to be 0.02. If the weights of the features were greater than Th , they were considered to be important features, and the other features were deleted prior to motor imagery detection.

Gurve et al. [122] used the LDA technique to classify the features corresponding to motor imagery detection. LDA maximizes the class variance ratio between the relaxed state and pedaling intention.

With the flexibility to use a range of datasets encompassing upper and lower limb motion, the NMF method's ability to choose EEG regions that are directly involved in the associated motor activity may be evaluated. To evaluate the effectiveness of channel selection methods, a series of experiments on the three datasets were conducted by Gurve et al. [122]. They examined the effect of various threshold settings on classification performance with the rank of the NMF algorithm fixed at $r = 3$ and the three datasets described. Additionally, the accuracy and mean κ values, as well as the true-positive rate (TPR) and false-positive rate (FPR), were utilized to evaluate the effectiveness of channel selection methods.

The classification accuracy and κ score were calculated as

$$\text{Accuracy} = \left(\frac{N_{\text{correct}}}{N_{\text{total}}} \right) \times 100\%, \quad (9.107)$$

where N_{total} corresponds to the number of total samples to be classified in the test dataset and N_{correct} is the number of correctly classified samples, and

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}, \quad (9.108)$$

where $Pr(a)$ is the observed relative agreement and $Pr(e)$ is the probability of hypothetical random agreement.

As stated in the description of the first dataset, 144 trials were gathered for each classification model's training and testing. A total of 144 training trials were utilized to choose channels and features, and to train the classifier. The remaining 144 trials from the validation set were utilized to calculate the evaluation parameters. To assess the channel selection method's performance, the motor imagery detection measures were calculated in three different scenarios: baseline approach, in which all measures were calculated without feature or channel selection; after feature selection with all EEG channels; and after feature and channel selection.

For all of the subjects, the average accuracy with the channel selection technique was greater than that with the baseline strategy. The baseline approach used all 19 EEG channels whereas the channel selection strategy reduced the data required to only 10 EEG channels to get the same result.

Each participant's TPR and FPR values were calculated individually. When utilizing the baseline approach, TPR was 83.03% on average for all subjects but increased to 97.77% when employing the selected EEG channels. The mean FPR when using all of the available EEG channels and the selected EEG channels, respectively, were 16.10% and 4.44%. The channel selection method's utility in boosting motor imagery detection was demonstrated by increased TPR and lower FPR values. The κ value was used because accuracy, TPR, and FPR do not fully use the information provided in the classifier's output. The κ value is a classification performance metric that takes into consideration the impact of random classification accuracy. The baseline technique gave a κ value of 66.94% on average, but the channel selection approach resulted in a κ value of 93.33% on average. This shows that even with a smaller number of EEG channels, the motor imagery identification approach improved the average κ value by 27%.

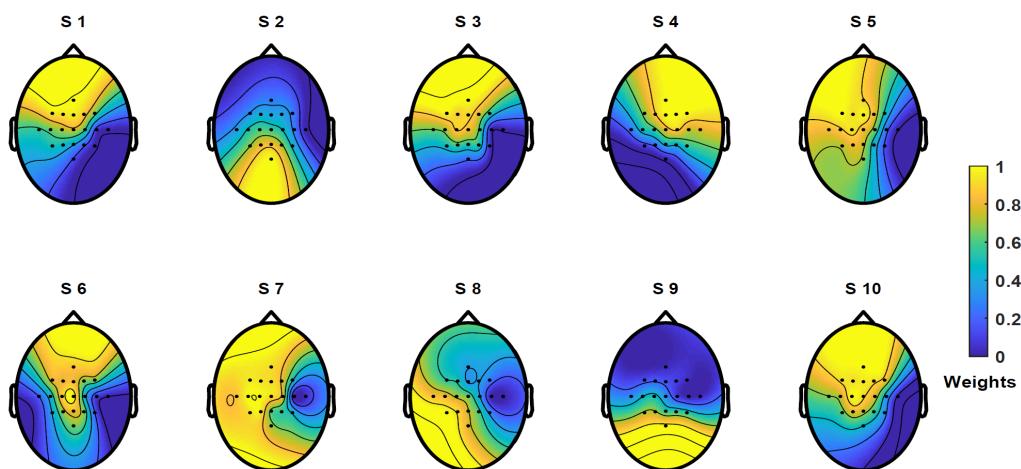


Figure 9.42 Topographic maps of EEG channel weights for 10 subjects. Channels with higher weight are indicated with yellow and channels with lower weight are indicated with blue color. Reproduced with permission from D. Gurve, D. Delisle-Rodriguez, M. Romero-Laiseca, V. Cardoso, F. Loterio, T. Bastos, and S. Krishnan, Subject-specific EEG channel selection using non-negative matrix factorization for lower-limb motor imagery recognition. *Journal of Neural Engineering*, 17(2):026029, 2020. ©IOP Publishing.

Additionally, as shown in Figure 9.42, the EEG channel weights varied from subject to subject. For example, subject S9 required just 10 EEG channels to achieve the same level of accuracy as subject S2, but subject S2 required 15 EEG channels to achieve acceptable results. Thus, the motor imagery identification method based on subject-specific channel selection reduces the channel count while enhancing detection accuracy.

The McNemar test [136, 137] (see Section 10.9.2) was used to perform statistical analysis of the motor imagery detection data obtained using the baseline and channel selection approaches. The resultant p values for the subjects S1, S2, S3, S4, S5, S6, S7, S8, and S10 were 0.001, 0.0385, 0.001, 0, 0, 0.0078, 0, 0, and 0 ($p < 0.05$), demonstrating a statistically significant improvement in motor imagery detection performance. There was no statistical significance in the results for subject S9 with $p = 1$ ($p > 0.05$).

In the process of channel weight computation using NMF analysis, determining the rank r for matrix factorization is a critical issue. As previously said, $r \ll \min[M, N]$ is essential. The rank r must be small enough to decrease computational complexity while being large enough to preserve important information from the covariance of the EEG channel weights. The NMF procedure was

iterated by Gurve et al. [122] by changing the rank r from 1 to 4 and calculating the accuracy for each subject.

The NMF-based EEG channel weighting approach assigns various weights to EEG channels based on differences in neural activity across different brain regions. EEG channels that show a considerably stronger and selective alteration in EEG patterns related to motor imagery are given more weight by the NMF-selected channels. The variance of activations (rows of the activation matrix) can have substantial values only if the connected channels have low channel covariance. The activation distributions give information on the discriminant channels between the two classes since activations are formed from the covariance matrix. If the recorded EEG channels are closely related, NMF chooses just a few EEG channels for further processing. The methods should facilitate improved BCI and rehabilitation.

9.13 Remarks

Analysis of multicomponent, multichannel, and multisource biomedical signals can benefit from adaptive signal decomposition coupled with matrix and image analysis techniques. Specifically, techniques such as matching pursuit and EMD, and classical methods such as PCA, ICA, and NMF can be extended to matrix analysis of signals represented via a TFD. Applications of the techniques to ECG, VAG, and EEG signals were described in this chapter; application to multimodal signals in the context of monitoring Parkinson's disease is presented in Section 10.14.

Assisted by continuous improvements in computer technology and solid-state microsensors, quantitative monitoring of human motor control and movement disorders has been an expanding field of research. Multimodal analysis of various physiological systems in the human body in an unobtrusive and ubiquitous way offers possibilities for many applications, including opportunities for advancements in the field of human behavior modeling, rehabilitation, health, and wellness.

9.14 Study Questions and Problems

1. What are the ways in which (a) a single-channel signal and (b) a multichannel signal could be converted to a matrix representation?
2. What would be the IMFs of a sinusoidal waveform decomposed using the EMD technique? Comment on the properties of an IMF of a signal and how they are related to the other IMFs provided by EMD of the same signal.
3. What are the differences in the decomposition of a signal by (a) the wavelet transform, (b) matching pursuit, and (c) EMD? What makes these techniques suitable for analysis of the nonstationary characteristics of a signal?
4. What makes matching pursuit and EMD suitable for nonstationary signal feature extraction as compared to the wavelet transform and STFT?
5. How are cross-terms avoided in computing an MPTFD? What are the effects of the dictionary elements used on the MPTFD representation?
6. How can selective filtering of components be achieved by NMF of the TFD of a 1D signal?
7. Plot the time–frequency representations of (a) a Gaussian function, (b) its translated version, (c) its scaled version, and (d) its modulated version. Explain how the time–frequency representations indicate the effects of the operations.

9.15 Laboratory Exercises and Projects

Note: Data files related to the exercises are available at the site

<https://github.com/srikrishnan1972/Biomedical-Signal-Analysis>

1. Using the covariance matrix of a synthetic signal, compute the eigenvalues and eigenvectors as provided by PCA decomposition. What do the eigenvectors and eigenvalues signify in this case?
2. Apply ICA to the covariance matrix of a three-source signal, and demonstrate that the independent components obtained are non-Gaussian in nature.
3. Generate a synthetic signal consisting of sinusoidal, transient, and chirp components, and plot the MPTFD of the first five atoms. Compare the MPTFD obtained with WVD and SPWVD, and comment on their time-frequency resolution and cross-term properties.
4. Compute the MPTFD of a PCG signal, and plot its time marginal and frequency marginal functions. Explain how the features of the signal are represented in the TFD and its marginals.
5. Using EMD, obtain the detrended version of an ECG signal, and compare the spectra of the original and detrended ECGs.
6. Demonstrate the denoising performance of matching pursuit, EMD, and wavelet transforms with noisy biomedical signals.
7. Apply NMF to the STFT matrices of the respiratory signal recordings available in the files Resp***.txt. Show that NMF assists in reducing cardiogenic oscillations in the single-channel respiratory signals.

References

- [1] Rangayyan RM. *Biomedical Image Analysis*. CRC Press, Boca Raton, FL, 2005.
- [2] Rahman MKM and Joadder MAM. A review on the components of EEG-based motor imagery classification with quantitative comparison. *Application and Theory of Computer Technology*, 2(2):1–15, 2017.
- [3] Ghoraani B, Krishnan S, Selvaraj RJ, and Chauhan VS. T wave alternans evaluation using adaptive time-frequency signal analysis and non-negative matrix factorization. *Medical Engineering & Physics*, 33(6):700–711, 2011.
- [4] Martínez JP and Olmos S. Methodological principles of T wave alternans analysis: A unified framework. *IEEE Transactions on Biomedical Engineering*, 52(4):599–613, 2005.
- [5] Smith JM, Clancy EA, Valeri CR, Ruskin JN, and Cohen RJ. Electrical alternans and cardiac electrical instability. *Circulation*, 77(1):110–121, 1988.
- [6] Nearing BD and Verrier RL. Modified moving average analysis of T-wave alternans to predict ventricular fibrillation with high accuracy. *Journal of Applied Physiology*, 92(2):541–549, 2002.
- [7] Mallat SG and Zhang Z. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.
- [8] Kovács F, Horváth C, Balogh AT, and Hosszú G. Extended noninvasive fetal monitoring by detailed analysis of data measured with phonocardiography. *IEEE Transactions on Biomedical Engineering*, 58(1):64–70, 2011.
- [9] Kovács F, Horváth C, Balogh AT, and Hosszú G. Fetal phonocardiography — past and future possibilities. *Computer Methods and Programs in Biomedicine*, 104:19–25, 2011.
- [10] Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen NC, Tung CC, and Liu HH. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London A*, 454:903–995, 1998.
- [11] Wu ZH, Huang NE, Long SR, and Peng CK. On the trend, detrending, and variability of nonlinear and nonstationary time series. *Proceedings of the National Academy of Sciences*, 104(38):14889–14894, 2007.
- [12] Wu ZH and Huang NE. Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Advances in Adaptive Data Analysis*, 1(1):1–41, 2009.

- [13] Rehman N and Mandic DP. Multivariate empirical mode decomposition. *Proceedings of the Royal Society of London A*, 466:1291–1302, 2010.
- [14] Echeverria JC, Crowe JA, Woolfson MS, and Hayes-Gill BR. Application of empirical mode decomposition to heart rate variability analysis. *Medical and Biological Engineering and Computing*, 39:471–479, 2001.
- [15] Wu Y, Yang SS, Zheng F, Cai S, Lu M, and Wu MH. Removal of artifacts in knee joint vibroarthrographic signals using ensemble empirical mode decomposition and detrended fluctuation analysis. *Physiological Measurement*, 35:429–439, 2014.
- [16] Kaleem M, Ghoraani B, Guergachi A, and Krishnan S. Pathological speech signal analysis and classification using empirical mode decomposition. *Medical and Biological Engineering and Computing*, 51:811–821, 2013.
- [17] Wu Z and Huang NE. Ensemble empirical mode decomposition: A noise assisted data analysis method. *Advances in Adaptive Data Analysis*, 1(1):1–41, 2009.
- [18] Wang Z, Wu D, Chen J, Ghoneim A, and Hossain MA. A triaxial accelerometer-based human activity recognition via EEMD-based features and game-theory-based feature selection. *IEEE Sensors Journal*, 16(9):3198–3207, 2016.
- [19] Shyu KK, Chiu LJ, Lee PL, Tung TH, and Yang SH. Detection of breathing and heart rates in UWB radar sensor data using FVPIEF-based two-layer EEMD. *IEEE Sensors Journal*, 19(2):774–784, 2018.
- [20] Singh G, Kaur G, and Kumar V. ECG denoising using adaptive selection of IMFs through EMD and EEMD. In *2014 International Conference on Data Science & Engineering (ICDSE)*, pages 228–231. IEEE, 2014.
- [21] Chen X, Xu X, Liu A, McKeown MJ, and Wang ZJ. The use of multivariate EMD and CCA for denoising muscle artifacts from few-channel EEG recordings. *IEEE Transactions on Instrumentation and Measurement*, 67(2):359–370, 2017.
- [22] Bashar SK, Hassan AR, and Bhuiyan MIH. Motor imagery movements classification using multivariate EMD and short time Fourier transform. In *2015 Annual IEEE India Conference (INDICON)*, pages 1–6. IEEE, 2015.
- [23] ur Rehman N, Xia Y, and Mandic DP. Application of multivariate empirical mode decomposition for seizure detection in EEG signals. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1650–1653. IEEE, 2010.
- [24] Flandrin P, Rilling G, and Goncalves P. Empirical mode decomposition as a filter bank. *IEEE Signal Processing Letters*, 11(2):112–114, February 2004.
- [25] Lee J, McManus DD, Merchant S, and Chon KH. Automatic motion and noise artifact detection in Holter ECG data using empirical mode decomposition and statistical approaches. *IEEE Transactions on Biomedical Engineering*, 59(6):1499–1506, June 2012.
- [26] Kaleem MF, Gurve D, Guergachi A, and Krishnan S. Patient-specific seizure detection in long-term EEG using signal-derived empirical mode decomposition (EMD)-based dictionary approach. *Journal of Neural Engineering*, 15(5):056004, 2018.
- [27] Jafari MG and Plumley MD. Fast dictionary learning for sparse representations of speech signals. *IEEE Journal of Selected Topics in Signal Processing*, 5(5):1025–1031, September 2011.
- [28] Kaleem MF, Guergachi A, and Krishnan S. Empirical mode decomposition based sparse dictionary learning with application to signal classification. In *Proceedings of IEEE Digital Signal Processing and Signal Processing Education Workshop (DSP/SPE), Napa, CA*, pages 18–23, 2013.
- [29] Ayenu-Prah A and Attoh-Okine N. A criterion for selecting relevant intrinsic mode functions in empirical mode decomposition. *Advances in Adaptive Data Analysis*, 2(1):1–24, 2010.
- [30] Wu Z and Huang NE. On the filtering properties of the empirical mode decomposition. *Advances in Adaptive Data Analysis*, 2(4):397–414, 2010.
- [31] Krishnan S, Rangayyan RM, Bell GD, and Frank CB. Adaptive time-frequency analysis of knee joint vibroarthrographic signals for noninvasive screening of articular cartilage pathology. *IEEE Transactions on Biomedical Engineering*, 47(6):773–783, June 2000.

- [32] Shore J and Johnson R. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26(1):26–37, 1980.
- [33] Shore J and Johnson R. Properties of cross-entropy minimization. *IEEE Transactions on Information Theory*, 27(4):472–482, 1981.
- [34] Loughlin P, Pitton J, and Atlas L. Construction of positive time-frequency distributions. *IEEE Transactions on Signal Processing*, 42(10):2697–2705, 1994.
- [35] Krishnan S and Rangayyan RM. Automatic de-noising of knee-joint vibration signals using adaptive time-frequency representations. *Medical and Biological Engineering and Computing*, 38(1):2–8, 2000.
- [36] Mallat S. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, CA, 1998.
- [37] Wickerhauser MV. *Adapted Wavelet Analysis from Theory to Software*. IEEE Press, Piscataway, NJ, 1994.
- [38] Rangayyan RM and Krishnan S. Feature identification in the time-frequency plane by using the Hough-Radon transform. *Pattern Recognition*, 34:1147–1158, 2001.
- [39] Krishnan S. *Adaptive Signal Processing Techniques for Analysis of Knee Joint Vibroarthrographic Signals*. PhD thesis, Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB, Canada, June 1999.
- [40] Hall EL. *Computer Image Processing and Recognition*. Academic Press, New York, NY, 1979.
- [41] Gonzalez RC and Woods RE. *Digital Image Processing*. Prentice-Hall, Upper Saddle River, NJ, 2nd edition, 2002.
- [42] Mesin L, Holobar A, and Merletti R. Blind source separation: Application to biomedical signals. In Cerutti S and Marchesi C, editors, *Advanced Methods of Biomedical Signal Processing*, pages 379–409. IEEE and Wiley, New York, NY, 2011.
- [43] Zou H, Hastie T, and Tibshirani R. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.
- [44] Linting M, Meulman JJ, Groenen PJF, and van der Kooij AJ. Nonlinear principal components analysis: Introduction and application. *Psychological Methods*, 12(3):336–358, 2007.
- [45] Hubert M and Engelen S. Robust PCA and classification in biosciences. *Bioinformatics*, 20(11):1728–1736, 2004.
- [46] Ding C and He X. K-means clustering via principal component analysis. In *ACM Proceedings of the Twenty-first International Conference on Machine Learning*, page 29, 2004.
- [47] Jolliffe IJ and Morgan BJT. Principal component analysis and exploratory factor analysis. *Statistical Methods in Medical Research*, 1(1):69–95, 1992.
- [48] Di Franco G. Multiple correspondence analysis: One only or several techniques? *Quality & Quantity*, 50(3):1299–1315, 2016.
- [49] Papoulis A. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, NY, 1965.
- [50] Hyvärinen A and Oja E. Independent component analysis: Algorithms and applications. *Neural Networks*, 13:441–430, 2000.
- [51] Jutten C and Herault J. Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.
- [52] Comon P, Jutten C, and Herault J. Blind separation of sources, Part II: Problems statement. *Signal Processing*, 24(1):11–20, 1991.
- [53] Comon P. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.
- [54] De Lathauwer L, De Moor B, and Vandewalle J. Fetal electrocardiogram extraction by blind source subspace separation. *IEEE Transactions on Biomedical Engineering*, 47(5):567–572, 2000.
- [55] Zarzoso V and Nandi AK. Blind separation of independent sources for virtually any source probability density function. *IEEE Transactions on Signal Processing*, 47(9):2419–2432, 1999.

- [56] Zarzoso V and Nandi AK. Adaptive blind source separation for virtually any source probability density function. *IEEE Transactions on Signal Processing*, 48(2):477–488, 2000.
- [57] De Vos M, Vergult A, De Lathauwer L, De Clercq W, Van Huffel S, Dupont P, Palmini A, and Van Paesschen W. Canonical decomposition of ictal scalp EEG reliably detects the seizure onset zone. *NeuroImage*, 37(3):844–854, 2007.
- [58] De Vos M, De Lathauwer L, and Van Huffel S. Spatially constrained ICA algorithm with an application in EEG processing. *Signal Processing*, 91:1963–1972, 2011.
- [59] Jiménez-González A and James CJ. De-noising the abdominal phonogram for foetal heart rate extraction: Blind source separation versus empirical filtering. In *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1358–1361, Osaka, Japan, July 2013.
- [60] James CJ and Hesse CW. Independent component analysis for biomedical signals. *Physiological Measurement*, 26:R15–R39, 2005.
- [61] Jiménez-González A and James CJ. Extracting sources from noisy abdominal phonograms: A single-channel blind source separation method. *Medical and Biological Engineering and Computing*, 47(3):655–664, 2009.
- [62] Castells F, Rieta JJ, Millet J, and Zarzoso V. Spatiotemporal blind source separation approach to atrial activity estimation in atrial tachyarrhythmias. *IEEE Transactions on Biomedical Engineering*, 52(2):258–267, 2005.
- [63] Sameni R, Jutten C, and Shamsollahi MB. Multichannel electrocardiogram decomposition using periodic component analysis. *IEEE Transactions on Biomedical Engineering*, 55(8):1935–1940, 2008.
- [64] Farina D, Févotte C, Doncarli C, and Merletti R. Blind separation of linear instantaneous mixtures of nonstationary surface myoelectric signals. *IEEE Transactions on Biomedical Engineering*, 51(9):1555–1567, 2004.
- [65] Zarzoso V and Nandi AK. Noninvasive fetal electrocardiogram extraction: Blind separation versus adaptive noise cancellation. *IEEE Transactions on Biomedical Engineering*, 48(1):12–18, 2001.
- [66] Callaerts D, De Moor B, Vandewalle J, Sansen W, Vantrappen G, and Janssens J. Comparison of SVD methods to extract the foetal electrocardiogram from cutaneous electrode signals. *Medical and Biological Engineering and Computing*, 28(3):217–224, 1990.
- [67] Vanderschoot J, Callaerts D, Sansen W, Vandewalle J, Vantrappen G, and Janssens J. Two methods for optimal MECG elimination and FECG detection from skin electrode signals. *IEEE Transactions on Biomedical Engineering*, 34(3):233–243, 1987.
- [68] Hyvarinen A. Fast ICA for noisy data using Gaussian moments. In *1999 IEEE International Symposium on Circuits and Systems (ISCAS)*, volume 5, pages 57–61. IEEE, 1999.
- [69] Lee DD and Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
- [70] Lee D and Seung HS. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2000.
- [71] Pal S, Beaumont J, Park DH, Amarnath A, Feng S, Chakrabarti C, Kim HS, Blaauw D, Mudge T, and Dreslinski R. Outerspace: An outer product based sparse matrix multiplication accelerator. In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 724–736. IEEE, 2018.
- [72] Agrawal A. *Matrix Multiplication: Inner Product, Outer Product & Systolic Array*. https://www.adityaagrawal.net/blog/architecture/matrix_multiplication, accessed on 2023-04-08.
- [73] López-Serrano P, Dittmar C, Özer Y, and Müller M. NMF toolbox: Music processing applications of nonnegative matrix factorization. In *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx-19), Birmingham, UK*, pages 2–6, 2019.

- [74] International Audio Laboratories Erlangen. *Nonnegative Matrix Factorization (NMF)*. https://www.audiolabs-erlangen.de/resources/MIR/FMP/C8/C8S3_NMFbasic.html, accessed 2023-03-27.
- [75] Van Erven T and Harremos P. Rényi divergence and Kullback–Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.
- [76] Berry MW, Browne M, Langville AN, Pauca VP, and Plemmons RJ. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, 2007.
- [77] Lin CJ. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10):2756–2779, 2007.
- [78] Eggert J and Körner E. Sparse coding and NMF. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, volume 4, pages 2529–2533. IEEE, 2004.
- [79] Choi S. Algorithms for orthogonal nonnegative matrix factorization. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1828–1832. IEEE, 2008.
- [80] O’Grady PD and Pearlmutter BA. Convulsive non-negative matrix factorisation with a sparseness constraint. In *2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, pages 427–432. IEEE, 2006.
- [81] Zhang D, Zhou ZH, and Chen S. Non-negative matrix factorization on kernels. In *Pacific Rim International Conference on Artificial Intelligence*, pages 404–412. Springer, 2006.
- [82] Ghoraani B. *Time–Frequency Feature Analysis*. PhD thesis, Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada, September 2010.
- [83] Donoho D and Stodden V. When does non-negative matrix factorization give a correct decomposition into parts? *Advances in Neural Information Processing Systems*, 16, 2003.
- [84] Wild S, Curry J, and Dougherty A. Improving non-negative matrix factorizations through structured initialization. *Pattern Recognition*, 37(11):2217–2232, 2004.
- [85] Boutsidis C and Gallopoulos E. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008.
- [86] Ghoraani B and Krishnan S. Discriminative base decomposition for time-frequency matrix decomposition. In *2010 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3674–3677. IEEE, 2010.
- [87] Kaleem MF, Guergachi A, and Krishnan S. Comparison of empirical mode decomposition, wavelets, and different machine learning approaches for patient-specific seizure detection using signal-derived empirical dictionary approach. *Frontiers in Digital Health*, 3:738996, 2021.
- [88] Hunyadi B, Signoretto M, van Paesschen M, Suykens JAK, van Huffel S, and De Vos M. Incorporating structural information from the multichannel EEG improves patient-specific seizure detection. *Clinical Neurophysiology*, 123(2012):2352–2361, 2012.
- [89] Liu Y, Zhou W, Yuan Q, and Chen S. Automatic seizure detection using wavelet transform and SVM in long-term intracranial EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(6):749–755, 2012.
- [90] Huang K and Aviyente S. Sparse representation for signal classification. In *Proceedings of Twentieth Annual Conference on Neural Information Processing Systems (NIPS) 2006*, pages 609–616, 2006.
- [91] Mairal J, Bach F, Ponce J, Sapiro G, and Zisserman A. Discriminative learned dictionaries for local image analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008*, pages 1–8, 2008.
- [92] Ramirez I, Sprechmann P, and Sapiro G. Classification and clustering via dictionary learning with structured incoherence and shared features. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*, pages 3501–3508, 2010.

- [93] Jiang Z, Lin Z, and Davis LS. Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011*, pages 1697–1704, 2011.
- [94] Akhtar N, Shafait F, and Mian A. Discriminative Bayesian dictionary learning for classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2374–2388, 2016.
- [95] Nagaraj SB, Stevenson NJ, Marnane WP, Boylan GB, and Lightbody G. Neonatal seizure detection using atomic decomposition with a novel dictionary. *IEEE Transactions on Biomedical Engineering*, 61(11):2724–2732, 2014.
- [96] Sorensen TL, Olsen UL, Conradsen I, Henriksen J, Kjaer TW, Thomsen CE, and Sorensen HBD. Automatic epileptic seizure onset detection using matching pursuit: A case study. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 3277–3280, 2010.
- [97] Shoeb A. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, September 2009.
- [98] Mehla VK, Singhal A, Singh P, and Pachori RB. An efficient method for identification of epileptic seizures from EEG signals using Fourier analysis. *Physical and Engineering Sciences in Medicine*, 44(2021):443–456, 2021.
- [99] Flandrin P and Goncalves P. Empirical mode decompositions as data-driven wavelet-like expansions. *International Journal of Wavelets, Multiresolution and Information Processing*, 2(4):1–20, 2004.
- [100] Faust O, Acharya UR, Adeli H, and Adeli A. Wavelet-based EEG processing for computer-aided seizure detection and epilepsy diagnosis. *Seizure*, 26:56–64, 2015.
- [101] Zabihi M, Kiranyaz S, Jäntti V, Lipping T, and Gabbouj M. Patient-specific seizure detection using nonlinear dynamics and nullclines. *IEEE Journal of Biomedical and Health Informatics*, 24(2):543–555, 2020.
- [102] Stallone A, Cicone A, and Materassi M. New insights and best practices for the successful use of empirical mode decomposition, iterative filtering and derived algorithms. *Scientific Reports*, 10(15161):1–15, 2020.
- [103] Cura OK, Atli SK, Tuere HS, and Akan A. Epileptic seizure classifications using empirical mode decomposition and its derivative. *BioMedical Engineering Online*, 19(10):1–22, 2020.
- [104] Kim KS, Seo JH, Kang JU, and Song CG. An enhanced algorithm for knee joint sound classification using feature extraction based on time-frequency analysis. *Computer Methods and Programs in Biomedicine*, 94:198–206, 2009.
- [105] Klingenbergheben T, Ptaszynski P, and Hohnloser SH. Quantitative assessment of microvolt T-wave alternans in patients with congestive heart failure. *Journal of Cardiovascular Electrophysiology*, 16(6):620–624, 2005.
- [106] Kunavarapu C and Bloomfield DM. Role of noninvasive studies in risk stratification for sudden cardiac death. *Clinical Cardiology*, 27(4):192–197, 2004.
- [107] Bacharakis E, Nandi AK, and Zarzoso V. Foetal ECG extraction using blind source separation methods. In *1996 8th European Signal Processing Conference (EUSIPCO 1996)*, pages 1–4, 1996.
- [108] Lipponen JA and Tarvainen MP. Principal component model for maternal ECG extraction in fetal QRS detection. *Physiological Measurement*, 35(8):1637–1648, 2014.
- [109] Christov I, Simova I, and Abächerli R. Extraction of the fetal ECG in noninvasive recordings by signal decompositions. *Physiological Measurement*, 35(8):1713, 2014.
- [110] Andreotti F, Riedl M, Himmelsbach T, Wedekind D, Wessel Ni, Stepan H, Schmieder C, Jank A, Malberg H, and Zaunseder S. Robust fetal ECG extraction and detection from abdominal leads. *Physiological Measurement*, 35(8):1551–1567, 2014.
- [111] Varanini M, Tartarisco G, Billeci L, Macerata A, Pioggia G, and Balocchi R. An efficient unsupervised fetal QRS complex detection from abdominal maternal ECG. *Physiological Measurement*, 35(8):1607–1619, 2014.

- [112] Da Poian G, Bernardini R, and Rinaldo R. Separation and analysis of fetal-ECG signals from compressed sensed abdominal ECG recordings. *IEEE Transactions on Biomedical Engineering*, 63(6):1269–1279, 2015.
- [113] Mirzal A. NMF versus ICA for blind source separation. *Advances in Data Analysis and Classification*, 11(1):25–48, 2017.
- [114] He P and Chen X. A method for extracting fetal ECG based on EMD-NMF single channel blind source separation algorithm. *Technology and Health Care*, 24(s1):S17–S26, 2016.
- [115] Samieinasab M and Sameni R. Fetal phonocardiogram extraction using single channel blind source separation. In *2015 23rd Iranian Conference on Electrical Engineering*, pages 78–83. IEEE, 2015.
- [116] Lamesgin G, Kassaw Y, and Assefa D. Extraction of fetal ECG from abdominal ECG and heart rate variability analysis. In *Afro-European Conference for Industrial Advancement*, pages 65–76. Springer, 2015.
- [117] Panigrahy D and Sahu PK. Extraction of fetal ECG signal by an improved method using extended Kalman smoother framework from single channel abdominal ECG signal. *Australasian Physical & Engineering Sciences in Medicine*, 40(1):191–207, 2017.
- [118] Almeida R, Gonçalves H, Bernardes J, and Rocha AP. Fetal QRS detection and heart rate estimation: A wavelet-based approach. *Physiological Measurement*, 35(8):1723–1735, 2014.
- [119] Su L and Wu HT. Extract fetal ECG from single-lead abdominal ECG by de-shape short time Fourier transform and nonlocal median. *Frontiers in Applied Mathematics and Statistics*, 3:2, 2017.
- [120] Gurve D and Krishnan S. Separation of fetal-ECG from single-channel abdominal ECG using activation scaled non-negative matrix factorization. *IEEE Journal of Biomedical and Health Informatics*, 24(3):669–680, 2019.
- [121] Jezewski J, Matonia A, Kupka T, Roj D, and Czabanski R. Determination of fetal heart rate from abdominal signals: evaluation of beat-to-beat accuracy in relation to the direct fetal electrocardiogram. *Biomedizinische Technik/Biomedical Engineering*, 57(5):383–394, 2012.
- [122] Gurve D, Delisle-Rodriguez D, Romero-Laiseca M, Cardoso V, Loterio F, Bastos T, and Krishnan S. Subject-specific EEG channel selection using non-negative matrix factorization for lower-limb motor imagery recognition. *Journal of Neural Engineering*, 17(2):026029, 2020.
- [123] Tong Y, Pendy Jr JT, Li WA, Du H, Zhang T, Geng X, and Ding Y. Motor imagery-based rehabilitation: Potential neural correlates and clinical application for functional recovery of motor deficits after stroke. *Aging and Disease*, 8(3):364–371, 2017.
- [124] Zimmermann-Schlatter A, Schuster C, Puhan MA, Siekierka E, and Steurer J. Efficacy of motor imagery in post-stroke rehabilitation: A systematic review. *Journal of Neuroengineering and Rehabilitation*, 5(1):1–10, 2008.
- [125] Jiang N, Gizzzi L, Mrachacz-Kersting N, Dremstrup K, and Farina D. A brain–computer interface for single-trial detection of gait initiation from movement related cortical potentials. *Clinical Neurophysiology*, 126(1):154–159, 2015.
- [126] Wang Y, Gao S, and Gao X. Common spatial pattern method for channel selection in motor imagery based brain-computer interface. In *IEEE Engineering in Medicine and Biology Society 27th Annual Conference*, pages 5392–5395. IEEE, 2006.
- [127] Alotaiby T, Abd El-Samie FE, Alshebeili SA, and Ahmad I. A review of channel selection algorithms for EEG signal processing. *EURASIP Journal on Advances in Signal Processing*, 2015(1):1–21, 2015.
- [128] Yang Y, Bloch I, Chevallier S, and Wiart J. Subject-specific channel selection using time information for motor imagery brain–computer interfaces. *Cognitive Computation*, 8(3):505–518, 2016.
- [129] Qiu Z, Jin J, Lam HK, Zhang Y, Wang X, and Cichocki A. Improved SFFS method for channel selection in motor imagery based BCI. *Neurocomputing*, 207:519–527, 2016.
- [130] Feng JK, Jin J, Daly I, Zhou J, Niu Y, Wang X, and Cichocki A. An optimized channel selection method based on multifrequency CSP-rank for motor imagery-based BCI system. *Computational Intelligence and Neuroscience*, 2019:1–10, 2019.

- [131] Liu YH, Huang S, and Huang YD. Motor imagery EEG classification for patients with amyotrophic lateral sclerosis using fractal dimension and Fisher's criterion-based channel selection. *Sensors*, 17(7):1557–1578, 2017.
- [132] Brunner C, Leeb R, Müller-Putz G, Schlögl A, and Pfurtscheller G. BCI competition 2008–Graz dataset A. *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, 16:1–6, 2008.
- [133] Gurve D. *Signal Analysis Techniques for Resource Optimization in Brain–Computer Interfaces and Other Wearables*. PhD thesis, Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada, September 2020.
- [134] Petersen P. *Riemannian Geometry*, Graduate Texts in Mathematics, volume 171. Springer, Cham, 2006.
- [135] Gallot S, Hulin D, and Lafontaine J. *Riemannian Geometry*, Universitext, volume 2. Springer, Cham, 1990.
- [136] Fleiss JL. *Statistical Methods for Rates and Proportions*. Wiley, New York, NY, 2nd edition, 1981.
- [137] Zar JH. *Biostatistical Analysis*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1984.

CHAPTER 10

COMPUTER-AIDED DIAGNOSIS AND HEALTHCARE

An important purpose or application of biomedical signal analysis is to classify a given signal into one of a few known categories and to arrive at a diagnostic decision regarding the condition of the patient. A physician or medical specialist may achieve this goal via visual or auditory analysis of the signals available: Comparative analysis of the given signals with others of known diagnoses or established protocols and sets of rules assist in such a decision-making process. The basic knowledge, clinical experience, expertise, and intuition of the physician play significant roles in this process. Some measurements may also be made or estimated from a given signal to assist in its analysis, such as the QRS width from an ECG signal plot.

When signal analysis is performed via the application of signal processing techniques and computer algorithms, the typical result, as studied in the present book up to this point, is the extraction of a number of numerical or quantitative features. When the numerical features relate directly to readily comprehensible measures of a signal, such as the QRS width and RR interval of an ECG signal, the clinical specialist may be able to use the features in a logical diagnostic procedure. Even indirect measures, such as the frequency content of PCG signals and murmurs, may find such direct use. However, when parameters such as AR-model coefficients and spectral statistics are derived, a human analyst is not likely to be able to comprehend and analyze the features. Furthermore, as the number of the computed features increases, the associated diagnostic logic may become too complicated and unwieldy for human analysis. Computer methods would then be desirable to realize the classification and decision-making process.

At the outset, it should be borne in mind that a biomedical signal forms but one part of the multifaceted information used in arriving at a diagnosis: The classification of a given signal into one of many categories may assist in the diagnostic or decision-making procedure, but will almost never be the only factor. Regardless, pattern classification based on signal processing is an important aspect of biomedical signal analysis, and forms the theme of this chapter. Remaining within the

realm of CAD as introduced in Figure 1.59 and Section 1.5, it is preferable to design methods so as to assist a medical specialist in arriving at a diagnosis rather than to provide a decision.

Whereas the preceding discussion has concentrated on diagnosis, information derived from biomedical signal analysis may also be used to guide treatment and to measure response to therapy. Attributes derived from biomedical signals may also be used to assess the state of health or well-being of a subject, or a subject's specific organ or system of interest; such methods could be used to derive measures of performance in kinesiology and athletics.

Recent developments have placed emphasis on providing increased control and options to the patient in monitoring for potential medical issues and in managing one's healthcare or well-being. The shrinking size of electronic devices and computers, as well as advances in sensor technology that have facilitated integration of transducers and electronics into clothing items, have led to the development of unobtrusive monitoring systems that can be worn by a subject [1]. Biomedical data acquisition and analysis systems have also been integrated into items of furniture and facilities such as chairs, beds, and toilets. The use of a wearable ECG monitor could help a cardiac patient in taking precautionary measures in the event of an arrhythmic episode. A personalized EEG analysis system to detect signal patterns that precede epileptic seizures could provide advance or early warning to the individual to stop any potentially dangerous activity and stay in a secure place in case a seizure occurs. A real-time continuous glucose monitoring system could issue alerts when blood sugar levels go beyond limits or provide control signals to an insulin pump, all worn by the subject.

The following sections describe techniques for medical decision making using attributes derived from signals and other medical or clinical information. The topic of pattern classification is a broad and sophisticated subject in its own right [2–4]; the purpose of the limited information provided in the present chapter is to lead the reader to appreciate potential end points in applications of biomedical signal and data analysis.

10.1 Problem Statement

A number of measures and features have been derived from a biomedical signal. Explore methods to classify the signal into one of a few specified categories. Investigate the relevance of the features and the classification methods in arriving at a diagnostic decision about the patient.

Observe that the features may have been derived manually or by computer methods. Note the distinction between classifying the given signal and arriving at a diagnosis regarding the patient: The connection between the two tasks or steps may not always be direct. In other words, a pattern classification method may facilitate the labeling of a given signal as being a member of a particular class; arriving at a diagnosis of the condition of the patient will most likely require the analysis of several items of clinical and other information. Although it is common to work with a prespecified number of pattern classes, many problems do exist where the number of classes is not known *a priori*.

The following sections present a few illustrative case studies. A number of methods for pattern classification, decision making, and evaluation of the results of classification are reviewed and illustrated.

10.2 Illustration of the Problem with Case Studies

10.2.1 Diagnosis of bundle-branch block

Bundle-branch block affects the propagation of the excitation pulse through the conduction system of the heart to the ventricles. A block in the left bundle branch results in delayed activation of the left ventricle as compared to the right; a block in the right bundle branch has the opposite effect. Essentially, contraction of the two ventricles becomes asynchronous. The resulting ECG typically

displays a wider-than-normal QRS complex ($100 - 120\text{ ms}$ or more), which could have a jagged or slurred shape as well [5]; see Figure 1.30.

The orientation of the cardiac electromotive forces is affected by bundle-branch block. The initial forces in left bundle-branch block are directed more markedly to the left-posterior, whereas the terminal forces are directed to the superior-left and posterior parts of the left ventricle [5]. Left bundle-branch block results in the loss of Q waves in leads I, V5, and V6. The following logic assists in the diagnosis of incomplete left bundle-branch block [6]:

IF (QRS duration $\geq 105\text{ ms}$ and $\leq 120\text{ ms}$) AND
 (QRS amplitude is negative in leads V1 and V2) AND
 (Q or S duration $\geq 80\text{ ms}$ in leads V1 and V2) AND
 (no Q wave is present in any two of leads I, V5, and V6) AND
 (R duration $> 60\text{ ms}$ in any two of leads I, aVL, V5, and V6) THEN
the patient has incomplete left bundle-branch block.

Incomplete right bundle-branch block is indicated by the following conditions [6]:

IF (QRS duration $\geq 91\text{ ms}$ and $\leq 120\text{ ms}$) AND
 (S duration $\geq 40\text{ ms}$ in any two of leads I, aVL, V4, V5, and V6) AND
 in lead V1 or V2 EITHER
 [(R duration $> 30\text{ ms}$) AND (R amplitude $> 100\text{ }\mu\text{V}$) AND
 (no S wave is present)] OR
 [(R' duration $> 30\text{ ms}$) AND (R' amplitude $> 100\text{ }\mu\text{V}$) AND
 (no S' wave is present)] THEN
the patient has incomplete right bundle-branch block.

[Note: The first positive deflection of a QRS complex is referred to as the R wave and the second positive deflection (if present) is referred to as the R' wave. Similarly, S and S' indicate the first and second (if present) negative deflections, respectively, after the R wave.]

Note that the logic or decision rules given above may be used either by a human analyst or in a computer algorithm after the durations and amplitudes of the various waves mentioned have been measured or computed. Cardiologists with extensive training and experience may arrive at such decisions via visual analysis of an ECG record without resorting to actual measurements.

10.2.2 Normal or ectopic ECG beat?

PVCs caused by ectopic foci could be precursors of more serious arrhythmia, and hence the detection of such beats is important in cardiac monitoring. As illustrated in Sections 5.4.2 and 5.7 as well as in Figures 5.1 and 5.12, PVCs possess shorter preceding RR intervals than normal beats and display bizarre waveshapes that are markedly different from those of the normal QRS complexes of the same subject in the same lead. Therefore, a simple rule to detect PVCs or ectopic beats could be as follows:

IF (the RR interval of the beat is less than the normal at the current heart rate) AND
 (the QRS waveshape is markedly different from the normal QRS of the patient)
 THEN *the beat is a PVC.*

As in the preceding case study of bundle-branch block, the logic above may be easily applied for visual analysis of an ECG signal by a physician or a trained observer. Computer implementation of the first part of the rule relating in an objective or quantitative manner to the RR interval is simple. However, implementation of the second condition on waveshape, being qualitative and subjective, is neither direct nor easy. Regardless, we have seen in Chapter 5 how we may characterize waveshape.

Figures 5.1 and 5.12 illustrate the application of waveshape analysis to quantify the differences between the shapes of normal QRS complexes and ectopic beats. Figure 5.2 suggests how a 2D feature space may be divided by a simple linear decision boundary to categorize beats as normal or ectopic. We study the details of such methods later in this chapter.

10.2.3 Is there an alpha rhythm?

The alpha rhythm appears in an EEG record as an almost-sinusoidal wave (see Figure 1.41); a trained EEG technologist or physician can readily recognize the pattern at a glance from an EEG record plotted at the standard scale. The number of cycles of the wave may be counted over one or two seconds of the plot if an estimate of the dominant frequency of the rhythm is desired.

In computer analysis of EEG signals, the ACF and PSD may be used to detect the presence of the alpha rhythm. We saw in Chapter 4 how these two functions demonstrate peaks at the basic period or dominant frequency of the rhythm, respectively (see Figure 4.11). A peak-detection algorithm may be applied to the ACF, and the presence of a significant peak in the range $75 - 125\text{ ms}$ for the delay or lag may be used as an indication of the existence of the alpha rhythm. If the PSD is available, the fractional power of the signal in the band $8 - 12\text{ Hz}$ (see Equation 6.44) may be computed: A high value of the fraction indicates the presence of the alpha rhythm. Note that the logic described above includes the qualifiers “significant” and “high”; experimentation with a number of signals that have been categorized by experts should assist in assigning a numerical value to represent the significance of the features described.

10.2.4 Is a murmur present?

Detection of the presence of a heart murmur is a fairly simple task for a trained physician or cardiologist: In performing auscultation of a patient with a stethoscope, the cardiologist needs to determine the existence of noise-like, high-frequency sounds between the low-frequency S1 and S2. It is necessary to exercise adequate care to reject high-frequency noise from other sources such as breathing, wheezing, and scraping of the stethoscope against the skin or hair. The cardiologist also has to distinguish between innocent physiological murmurs and those due to cardiovascular defects and diseases. Further discrimination between different types of murmurs requires more careful analysis: Figure 5.6 illustrates a decision tree to classify systolic murmurs based on envelope analysis.

We have seen in Chapters 6 and 7 how we may derive frequency-domain parameters that relate to the presence of murmurs in the PCG signal. Once we have derived such numerical features for a number of signals of known categories of diseases (diagnoses), it becomes possible to design and train classifiers to categorize new signals into one of a few prespecified classes.

10.2.5 Detection of sleep apnea using multimodal biomedical signals

A sleep disorder known as OSA is characterized by recurring bouts of partial or total upper airway obstruction caused by the pharyngeal airway narrowing or collapsing while the person is trying to breathe [7]; see Section 2.4. The laboratory-based PSG sleep study is the current diagnostic standard for OSA; however, it is a time-consuming, labor-intensive technique that makes the patient’s discomfort worse [8]. Typically, signals such as the ECG, EOG, EMG, EEG, PPG, and indicators of abdominal and thoracic movement are simultaneously recorded throughout the PSG data collection process, which could last for about 8 h in an overnight sleep study.

Researchers studying sleep disorders are becoming interested in automatic identification of OSA with the introduction of CAD [9–11]. Techniques in the time domain, frequency domain, and the joint TF domain are frequently used to extract features from PSG signals. The AHI and the signal features are then used to train a classifier to identify certain signal characteristics that could aid in the detection of OSA. Due to the requirement of long-term signal acquisition and the complexity of

instrumentation associated with PSG recording, different home-based wireless wearable devices are being currently investigated to obtain better quality data in the convenience of one's home [12, 13].

The preceding case studies suggest that the classification of patterns in a signal may, in some cases, be based on thresholds applied to quantitative measurements obtained from the signal; in some other cases, it may be based on objective measures derived from the signal that attempt to quantify certain notions regarding the characteristics of signals belonging to various categories. Classification may also be based on the differences between certain measures derived from the signal on hand and those of established examples with known categorization. The succeeding sections of this chapter describe procedures for classification of signals based on the approaches suggested above, and provide several illustrations of application.

10.3 Pattern Classification

Pattern recognition or classification may be defined as categorization of input data into identifiable classes via the extraction of significant features or attributes of the data from a background of irrelevant detail [2–4, 14–17]. In biomedical signal analysis, after quantitative features have been extracted from the given signals, each signal may be represented by a feature vector $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, which is also known as a measurement vector or a pattern vector. When the values x_i are real numbers, \mathbf{x} is a point in an n -dimensional Euclidean space: Vectors of similar objects may be expected to form clusters as illustrated in Figure 10.1.

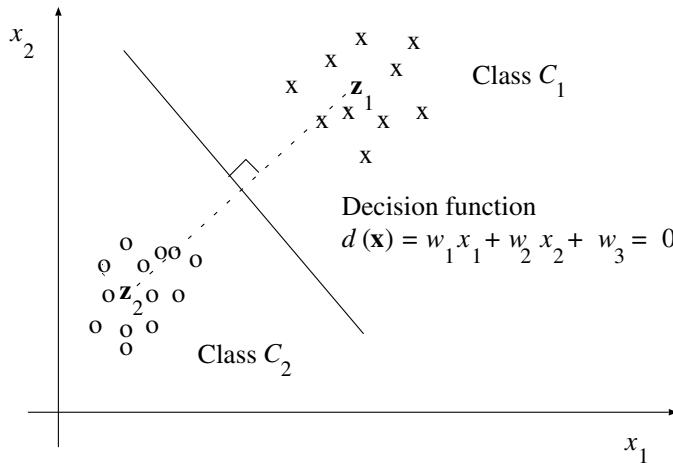


Figure 10.1 Two-dimensional feature vectors of two classes C_1 and C_2 . The prototypes of the two classes are indicated by the vectors \mathbf{z}_1 and \mathbf{z}_2 . The linear decision function shown $d(\mathbf{x})$ (solid line) is the perpendicular bisector of the straight line joining the two class prototypes (dashed line).

For efficient pattern classification, measurements that could lead to disjoint sets or clusters of feature vectors are desired. This point underlines the importance of appropriate design of the pre-processing and feature extraction procedures. Features or characterizing attributes that are common to all patterns belonging to a particular class are known as *intraset* or *intraclass* features. Discriminant features that represent differences between pattern classes are called *interset* or *interclass* features.

The pattern classification problem is that of generating optimal decision boundaries or decision procedures to separate the data into pattern classes based on the feature vectors. Figure 10.1 illustrates a simple linear decision function or boundary to separate 2D feature vectors into two classes.

10.4 Supervised Pattern Classification

Problem: You are provided with a number of feature vectors with classes assigned to them. Propose techniques to characterize the boundaries that separate the classes.

Solution: A given set of feature vectors of known categorization is often referred to as a *training set*. The availability of a training set facilitates the development of mathematical functions that can characterize the separation between the classes. The functions may then be applied to new feature vectors of unknown classes to classify or recognize them. This approach is known as *supervised pattern classification*. An independent set of feature vectors of known categorization that is used to evaluate a classifier designed in this manner is referred to as a *test set*. The following sections describe a few methods that can assist in the development of discriminant and decision functions.

10.4.1 Discriminant and decision functions

A general linear discriminant or decision function is of the form

$$d(\mathbf{x}) = w_1x_1 + w_2x_2 + \cdots + w_nx_n + w_{n+1} = \mathbf{w}^T \mathbf{x}, \quad (10.1)$$

where $\mathbf{x} = (x_1, x_2, \dots, x_n, 1)^T$ is the feature vector augmented by an additional entry equal to unity, and $\mathbf{w} = (w_1, w_2, \dots, w_n, w_{n+1})^T$ is a correspondingly augmented weight vector. A two-class pattern classification problem may be stated as

$$d(\mathbf{x}) = \mathbf{w}^T \mathbf{x} \begin{cases} > 0 & \text{if } \mathbf{x} \in C_1, \\ \leq 0 & \text{if } \mathbf{x} \in C_2, \end{cases} \quad (10.2)$$

where C_1 and C_2 represent the two classes. The discriminant function may be interpreted as the boundary separating the classes C_1 and C_2 , as illustrated in Figure 10.1.

In the general case of an M -class pattern classification problem, we will need M weight vectors and M decision functions to perform the following decisions:

$$d_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} \begin{cases} > 0 & \text{if } \mathbf{x} \in C_i, \\ \leq 0 & \text{otherwise,} \end{cases} \quad (10.3)$$

for $i = 1, 2, \dots, M$, where $\mathbf{w}_i = (w_{i1}, w_{i2}, \dots, w_{in}, w_{i,n+1})^T$ is the weight vector for the class C_i .

Three cases arise in solving this problem [4]:

Case 1: Each class is separable from the rest by a single decision surface:

$$\text{if } d_i(\mathbf{x}) > 0, \text{ then } \mathbf{x} \in C_i. \quad (10.4)$$

Case 2: Each class is separable from every other individual class by a distinct decision surface, that is, the classes are pairwise separable. There are $M(M - 1)/2$ decision surfaces given by $d_{ij}(\mathbf{x}) = \mathbf{w}_{ij}^T \mathbf{x}$, such that

$$\text{if } d_{ij}(\mathbf{x}) > 0 \forall j \neq i, \text{ then } \mathbf{x} \in C_i. \quad (10.5)$$

[Note: $d_{ij}(\mathbf{x}) = -d_{ji}(\mathbf{x})$.]

Case 3: There exist M decision functions $d_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x}$, $k = 1, 2, \dots, M$, with the property that

$$\text{if } d_i(\mathbf{x}) > d_j(\mathbf{x}) \forall j \neq i, \text{ then } \mathbf{x} \in C_i. \quad (10.6)$$

This is a special instance of Case 2. We may define

$$d_{ij}(\mathbf{x}) = d_i(\mathbf{x}) - d_j(\mathbf{x}) = (\mathbf{w}_i - \mathbf{w}_j)^T \mathbf{x} = \mathbf{w}_{ij}^T \mathbf{x}. \quad (10.7)$$

If the classes are separable under Case 3, they are separable under Case 2; the converse, in general, is not true.

When patterns need to be separated into multiple categories, it may be possible to convert the problem into a series of binary decision problems by considering the separation of each class against the set of remaining classes.

Patterns that may be separated by linear decision functions as above are said to be *linearly separable*. In other situations, an infinite variety of complex decision boundaries may be formulated by using generalized decision functions based on nonlinear functions of the feature vectors as

$$d(\mathbf{x}) = w_1 f_1(\mathbf{x}) + w_2 f_2(\mathbf{x}) + \cdots + w_K f_K(\mathbf{x}) + w_{K+1} \quad (10.8)$$

$$= \sum_{i=1}^{K+1} w_i f_i(\mathbf{x}). \quad (10.9)$$

Here, $\{f_i(\mathbf{x})\}$, $i = 1, 2, \dots, K$, are real, single-valued functions of \mathbf{x} ; $f_{K+1}(\mathbf{x}) = 1$.

Whereas the functions $f_i(\mathbf{x})$ may be nonlinear in the n -dimensional space of \mathbf{x} , the decision function may be formulated as a linear function by defining a transformed feature vector $\mathbf{x}^\dagger = [f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_K(\mathbf{x}), 1]^T$. Then, $d(\mathbf{x}) = \mathbf{w}^T \mathbf{x}^\dagger$, with $\mathbf{w} = [w_1, w_2, \dots, w_K, w_{K+1}]^T$. Once evaluated, $\{f_i(\mathbf{x})\}$ is just a set of numerical values, and \mathbf{x}^\dagger is simply a K -dimensional vector augmented by an entry equal to unity.

Extending the function for linear discriminant analysis (LDA) given in Equation 10.1, a quadratic discriminant analysis (QDA) or decision function may be expressed as [4]

$$d(\mathbf{x}) = \sum_{i=1}^n w_{ii} x_i^2 + \sum_{i=1}^{n-1} \sum_{j=i+1}^n w_{ij} x_i x_j + \sum_{i=1}^n w_i x_i + w_{n+1}. \quad (10.10)$$

QDA may perform better than LDA in some applications; however, QDA requires the estimation of a larger number of parameters than LDA, represented by w_{ii} , w_{ij} , and w_i in Equation 10.10.

See Section 10.11.1 for an illustration of application of LDA to an ECG signal with PVCs.

10.4.2 Fisher linear discriminant analysis

Fisher linear discriminant analysis (FLDA) is a technique that projects multidimensional data on to a 1D space or line which allows for improved discrimination [2]; dimensionality reduction is achieved in the process. In a pattern classification problem with two classes, the projection in FLDA is designed to maximize the distance between the means of the feature vectors for the two classes and minimize their variance within each class.

Figure 10.2 illustrates the concept of a linear classifier based on FLDA. Consider two classes, C_1 and C_2 , in which each sample is represented by the features $\mathbf{x} = [x_1, x_2]^T$. If only the feature x_1 is used for classification, it is equivalent to the situation when all of the data points are projected on to the x_1 axis. Such a projection causes overlapping regions containing samples from both classes; thus, x_1 provides poor separation between the classes C_1 and C_2 . The feature x_2 also shows overlap between the two classes and provides poor separation. However, by inspection of Figure 10.2, it is evident that the classes C_1 and C_2 can be clearly separated by a straight line oriented at about 130° to the abscissa. Hence, a classifier can be designed using projections of the feature vectors or samples provided on to the line labeled as F in Figure 10.2. FLDA provides the weights that define the projection as explained above [2, 18].

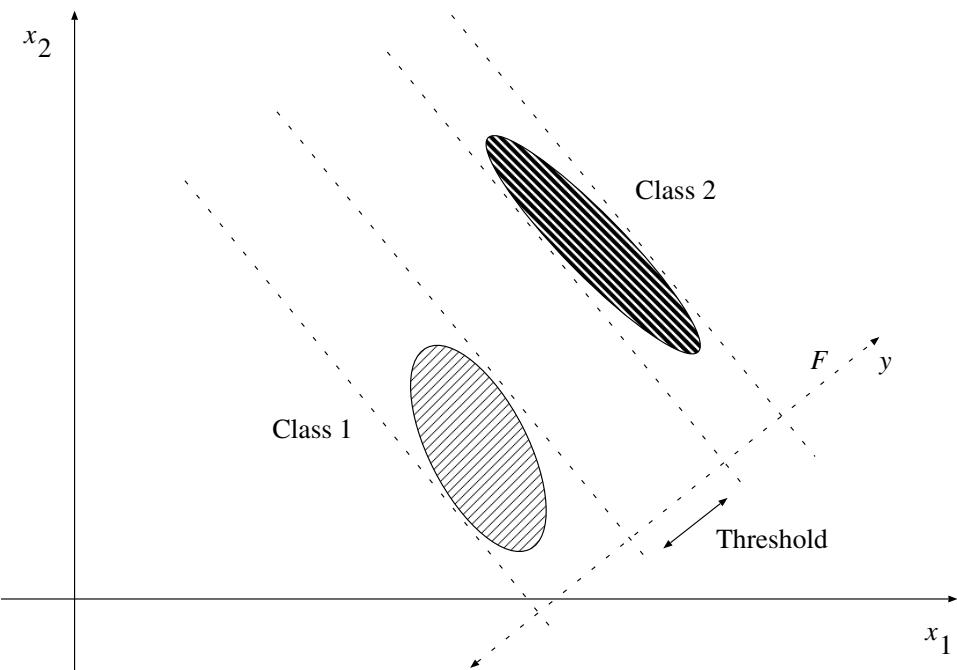


Figure 10.2 Illustration of a classifier based upon FLDA.

Let a given set of feature vectors

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} = \{X_1, X_2\} \quad (10.11)$$

be partitioned into N_1 training samples in subset X_1 , corresponding to class C_1 , and N_2 training samples in subset X_2 , corresponding to class C_2 , with $N_1 + N_2 = N$. The projection of \mathbf{x}_i on to the FLDA discriminant line is

$$y_i = \mathbf{w}^T \mathbf{x}_i. \quad (10.12)$$

The projected value y_i belongs to the subset Y_1 or Y_2 . The weight vector \mathbf{w} defines the direction of the axis of the projected values y_i . The mean of all of the samples belonging to each class before projection is

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{\mathbf{x}_i \in X_k} \mathbf{x}_i, \quad (10.13)$$

where $k = 1, 2$, and $i = 1, 2, \dots, N$. The mean of the projected values in each class is

$$\begin{aligned} \tilde{m}_k &= \frac{1}{N_k} \sum_{y_i \in Y_k} y_i \\ &= \frac{1}{N_k} \sum_{\mathbf{x}_i \in X_k} \mathbf{w}^T \mathbf{x}_i, \\ &= \mathbf{w}^T \mathbf{m}_k, \quad k = 1, 2. \end{aligned} \quad (10.14)$$

The difference between the means of the classes after projection is

$$|\tilde{m}_1 - \tilde{m}_2| = |\mathbf{w}^T \mathbf{m}_1 - \mathbf{w}^T \mathbf{m}_2|. \quad (10.15)$$

The variance or scatter of the projected samples y_i in each class is

$$\begin{aligned} \tilde{s}_k^2 &= \sum_{y_i \in Y_k} (y_i - \tilde{m}_k)^2 \\ &= \sum_{\mathbf{x}_i \in X_k} (\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{m}_k)^2 \\ &= \sum_{\mathbf{x}_i \in X_k} \mathbf{w}^T (\mathbf{x}_i - \mathbf{m}_k) (\mathbf{x}_i - \mathbf{m}_k)^T \mathbf{w}, \quad k = 1, 2 \\ &= \mathbf{w}^T \mathbf{S}_k \mathbf{w}, \end{aligned} \quad (10.16)$$

where

$$\mathbf{S}_k = \sum_{\mathbf{x}_i \in X_k} (\mathbf{x}_i - \mathbf{m}_k) (\mathbf{x}_i - \mathbf{m}_k)^T, \quad k = 1, 2. \quad (10.17)$$

\mathbf{S}_k is known as the within-class or intraclass scatter matrix for class k .

The total within-class scatter of the two classes is

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^T \{\mathbf{S}_1 + \mathbf{S}_2\} \mathbf{w} = \mathbf{w}^T \mathbf{S}_W \mathbf{w}, \quad (10.18)$$

with $\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$.

The FLDA criterion function is defined as

$$J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}. \quad (10.19)$$

Optimal separation between the classes is achieved when $J(\mathbf{w})$ is at its maximum. The denominator term of the function $J(\mathbf{w})$ is given by Equation 10.18. The numerator gives the separation between the means of the projected samples for the two classes and is obtained as follows:

$$\begin{aligned} (\tilde{m}_1 - \tilde{m}_2)^2 &= (\mathbf{w}^T \mathbf{m}_1 - \mathbf{w}^T \mathbf{m}_2)^2 \\ &= \mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2) (\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w} \\ &= \mathbf{w}^T \mathbf{S}_B \mathbf{w}. \end{aligned} \quad (10.20)$$

Here, $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$ is the between-class or interclass scatter matrix. The FLDA criterion is given by

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}. \quad (10.21)$$

The weight vector \mathbf{w}_o that maximizes $J(\mathbf{w})$ can be derived by solving a generalized eigenvalue problem. When a new sample is to be classified, its feature vector \mathbf{x} is projected using Equation 10.12 and two-category classification is performed as

$$\mathbf{x} \in \begin{cases} C_1 & \text{if } y = \mathbf{w}_o^T \mathbf{x} < T_1, \\ C_2 & \text{otherwise,} \end{cases} \quad (10.22)$$

where T_1 is a threshold.

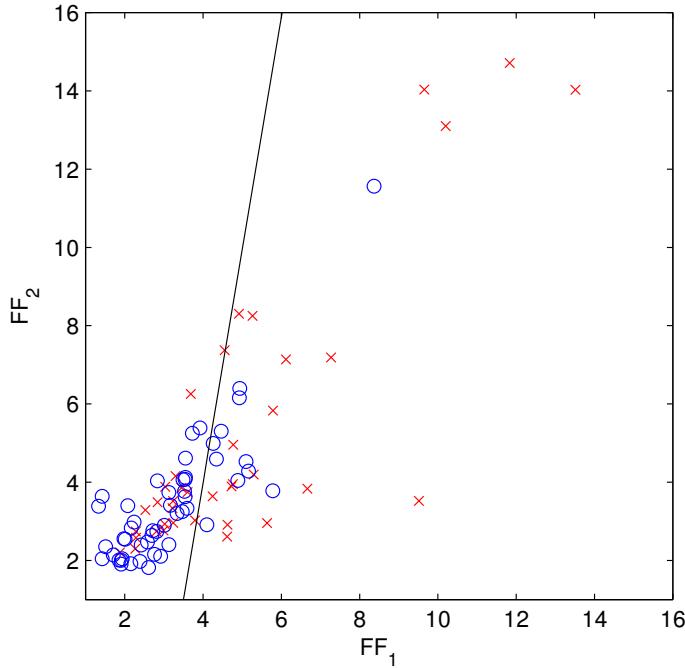


Figure 10.3 Illustration of classification of a dataset of 89 VAG signals using FLDA. The circles represent normal samples of the feature vector $[FF_1, FF_2]^T$ and the crosses represent abnormal samples. The straight line is the FLDA decision function, given by $0.0058 FF_1 - 0.0010 FF_2 - 0.0192 = 0$. Figure courtesy of Tingting Mu, University of Liverpool, Liverpool, UK.

Illustration of application: Rangayyan and Wu [19] applied a number of statistical measures for screening of VAG signals (see Section 5.12.1 for a description of the dataset used). The parameters used included FF computed for the full duration of each swing cycle and for the first and second halves (labeled as FF_1 and FF_2), corresponding to extension and flexion, as well as the entropy (H), skewness (S), and kurtosis (K). The features could not perform screening with high accuracy on their own: the six features FF, FF_1, FF_2, S, K , and H individually provided A_z values of 0.72, 0.73, 0.68, 0.70, 0.61, and 0.60, respectively. Figure 10.3 shows the scatter plot of the feature vector $[FF_1, FF_2]$ for the set of 89 VAG signals. The straight line shown in the figure represents the decision function obtained via FLDA. The equation for the decision boundary is

$$0.0058 FF_1 - 0.0010 FF_2 - 0.0192 = 0. \quad (10.23)$$

Due to substantial overlap of the samples in the two categories, FLDA could correctly classify only 19/38 abnormal signals and 40/51 normal signals, yielding an average classification accuracy of 66%. The use of FLDA with all of the six features listed above and the LOO method for cross-validation did not provide any better results, with $A_z = 0.72$. The results obtained and the illustration in Figure 10.3 indicate that the samples are not linearly separable using the measures derived. Mu et al. [20] obtained much better classification performance with $A_z = 0.95$ using the same features as above but with a genetic algorithm for feature selection and the strict two-surface proximal classifier. See Section 10.8.1 for the results of application of a neural network to the same dataset.

See Figure 5.2 and Section 10.11 for additional illustrations of application of linear decision functions to the classification of ECG signals.

10.4.3 Distance functions

Consider M pattern classes represented by their prototype patterns $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M$. The prototype of a class is typically computed as the average of all of the feature vectors belonging to the class. Figure 10.1 illustrates schematically the prototypes \mathbf{z}_1 and \mathbf{z}_2 of the two classes shown.

The Euclidean distance between an arbitrary pattern vector \mathbf{x} and the i^{th} prototype is given as

$$D_i = \|\mathbf{x} - \mathbf{z}_i\| = \sqrt{(\mathbf{x} - \mathbf{z}_i)^T (\mathbf{x} - \mathbf{z}_i)}. \quad (10.24)$$

A simple rule to classify the pattern vector \mathbf{x} would be to choose that class for which the vector has the smallest distance:

$$\text{if } D_i < D_j \forall j \neq i, \text{ then } \mathbf{x} \in C_i. \quad (10.25)$$

A relationship may be established between discriminant functions and distance functions as follows [4]:

$$\begin{aligned} D_i^2 &= \|\mathbf{x} - \mathbf{z}_i\|^2 = (\mathbf{x} - \mathbf{z}_i)^T (\mathbf{x} - \mathbf{z}_i) \\ &= \mathbf{x}^T \mathbf{x} - 2\mathbf{x}^T \mathbf{z}_i + \mathbf{z}_i^T \mathbf{z}_i = \mathbf{x}^T \mathbf{x} - 2(\mathbf{x}^T \mathbf{z}_i - \frac{1}{2} \mathbf{z}_i^T \mathbf{z}_i). \end{aligned} \quad (10.26)$$

Choosing the minimum of D_i^2 is equivalent to choosing the minimum of D_i (as all $D_i > 0$). Furthermore, from the equation given above, it follows that choosing the minimum of D_i^2 is equivalent to choosing the maximum of $(\mathbf{x}^T \mathbf{z}_i - \frac{1}{2} \mathbf{z}_i^T \mathbf{z}_i)$. Therefore, we may define the decision function

$$d_i(\mathbf{x}) = \mathbf{x}^T \mathbf{z}_i - \frac{1}{2} \mathbf{z}_i^T \mathbf{z}_i, \quad i = 1, 2, \dots, M. \quad (10.27)$$

A decision rule may then be stated as follows:

$$\text{if } d_i(\mathbf{x}) > d_j(\mathbf{x}) \forall j \neq i, \text{ then } \mathbf{x} \in C_i. \quad (10.28)$$

This is a linear discriminant function, which becomes obvious from the following representation: If z_{ij} , $j = 1, 2, \dots, n$, are the components of \mathbf{z}_i , let $w_{ij} = z_{ij}$, $j = 1, 2, \dots, n$; $w_{i,n+1} = -\frac{1}{2} \mathbf{z}_i^T \mathbf{z}_i$; and $\mathbf{x} = [x_1, x_2, \dots, x_n, 1]^T$. Then, $d_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x}$, $i = 1, 2, \dots, M$, where $\mathbf{w}_i = [w_{i1}, w_{i2}, \dots, w_{i,n+1}]^T$. Therefore, distance functions may be formulated as linear discriminant or decision functions.

10.4.4 The nearest-neighbor rule

Suppose that we are provided with a set of N sample patterns $\{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_N\}$ of known classification: Each pattern belongs to one of M classes $\{C_1, C_2, \dots, C_M\}$. We are then given a new feature vector \mathbf{x} whose class needs to be determined. Let us compute a distance measure $D(\mathbf{s}_i, \mathbf{x})$ between the vector \mathbf{x} and each sample pattern. Then, the nearest-neighbor rule states that the vector \mathbf{x} is to be assigned to the class of the sample that is the closest to \mathbf{x} :

$$\mathbf{x} \in C_i \text{ if } D(\mathbf{s}_i, \mathbf{x}) = \min\{D(\mathbf{s}_l, \mathbf{x})\}, \quad l = 1, 2, \dots, N. \quad (10.29)$$

A major disadvantage of the above method is that the classification decision is made based upon a single sample vector of known classification. The nearest neighbor may happen to be an outlier that is not representative of its class. It would be more reliable to base the classification upon several samples: We may consider a certain number k of the nearest neighbors of the sample to be classified, and then seek a majority opinion. This leads to the so-called *k-nearest-neighbor* or *k-NN rule*: Determine the k nearest neighbors of \mathbf{x} , and use the majority of equal classifications in this group as the classification of \mathbf{x} .

10.4.5 The support vector machine

The support vector machine (SVM) belongs to the class of supervised linear classifiers, in which two classes are separated by a hyperplane [2, 21, 22]. An SVM seeks to identify, in a high-dimensional feature space, the hyperplane that best separates the given feature vectors into classes. An SVM determines the hyperplane that maintains the largest margin between the closest feature vectors of each class that are referred to as support vectors; see Figure 10.4. This property enables the SVM to generalize well to new data. Duda et al. [2] suggest consideration of the support vectors of the two classes as the samples that are the most difficult to classify due to their proximity to the decision boundary; for the same reason, they may also be considered to be the most informative samples in the design of a classifier.

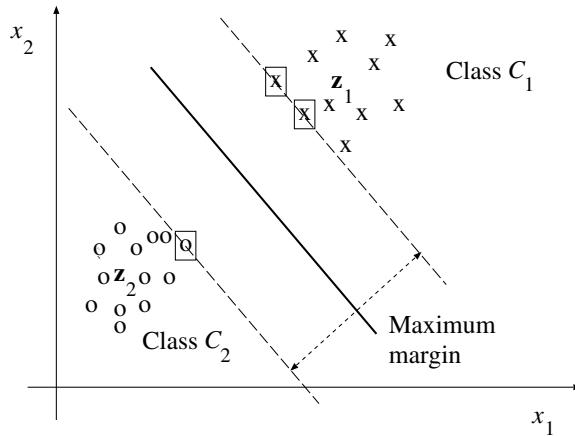


Figure 10.4 Schematic representation of the formulation of an SVM with two linearly separable clusters of feature vectors. The three samples shown boxed are the support vectors. The dashed lines passing through the support vectors form the margin between the two classes [2, 21, 22]. The solid line in the middle of the margin is the SVM decision boundary. Compare with the case depicted in Figure 10.1.

The formulation of the SVM follows a representation that is slightly different from that in Section 10.4.1. Considering the n -dimensional feature vectors \mathbf{x} and the weight vector \mathbf{w} without the extra term of unity or w_{n+1} , the equation to the hyperplane (or straight line in the 2D case) separating the two classes is expressed as

$$\mathbf{w}^T \mathbf{x} + b = 0, \quad (10.30)$$

where b represents a bias. Let us assume that the dataset of samples is normalized. Each sample \mathbf{x}_i is associated with a class label y_i that takes on the values $+1$ or -1 for the two classes. The classification or discrimination task is stated as

$$\mathbf{w}^T \mathbf{x}_i + b = \begin{cases} +1 & \text{if } y_i = +1, \\ -1 & \text{if } y_i = -1. \end{cases} \quad (10.31)$$

Now, if \mathbf{x}_1 and \mathbf{x}_2 are support vectors belonging to the two classes, we have $\mathbf{w}^T \mathbf{x}_1 + b = +1$ and $\mathbf{w}^T \mathbf{x}_2 + b = -1$, which leads to [21] $\mathbf{w}^T (\mathbf{x}_1 - \mathbf{x}_2) = 2$ and furthermore to

$$\frac{\mathbf{w}^T}{\|\mathbf{w}\|} (\mathbf{x}_1 - \mathbf{x}_2) = \frac{2}{\|\mathbf{w}\|}. \quad (10.32)$$

This equation gives the projection of $(\mathbf{x}_1 - \mathbf{x}_2)$ on to the unit vector in the direction of \mathbf{w} , which is orthogonal to the SVM decision boundary. Thus, the width of the margin [21] is $\frac{2}{\|\mathbf{w}\|}$.

The discrimination task in Equation 10.31 may also be stated as the following constraints:

$$\mathbf{w}^T \mathbf{x}_i + b \begin{cases} \geq +1 & \text{if } y_i = +1, \\ \leq -1 & \text{if } y_i = -1. \end{cases} \quad (10.33)$$

No sample is allowed to be present within the margin. The primary objective of an SVM is to maximize the margin between the classes. As indicated by Equation 10.32, to maximize the width of the margin, we could minimize $\|\mathbf{w}\|$. This problem of optimization could be combined with the constraints defined above to state the SVM problem as [21]: minimize $\|\mathbf{w}\|$ subject to the constraint $y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1$ for all i . The parameters of the SVM may then be obtained through a suitable optimization algorithm.

In the preceding discussion, a linear hyperplane was assumed to separate the samples or feature vectors of the two classes. However, in many real situations, the samples may not be separable by a linear hyperplane. In such cases, nonlinear transformations may be used to lead to a linear separator using kernel mapping methods [2, 21, 22]; it is assumed that the samples are linearly separable in a suitably higher dimension. It was shown in Section 10.4.1 that nonlinear functions could be applied to feature vectors prior to the design of a linear classifier; see Equation 10.9. In a comparable manner, polynomial kernels, Gaussian kernels, and radial basis functions (RBFs) are among the commonly used kernel functions with SVMs to form linear boundaries in the transformed space or nonlinear boundaries in the original feature space to separate complex clusters of samples. See Section 10.8.1 for discussions on ANNs with RBFs.

10.5 Unsupervised Pattern Classification

Problem: *We are given a set of feature vectors with no categorization or classes attached to them. No prior training information is available. How may we group the vectors into multiple categories?*

Solution: The design of distance functions and decision boundaries requires a training set of feature vectors of known classes. The functions so designed may then be applied to a new set of feature vectors or samples to perform pattern classification. Such a procedure is known as *supervised* pattern classification due to the initial training step. In some situations, a training step may not be possible, and we may be required to classify a given set of feature vectors into either a prespecified or unknown number of categories. Such a problem is labeled as *unsupervised* pattern classification and may be solved by cluster-seeking methods.

10.5.1 Cluster-seeking methods

Given a set of feature vectors, we may examine them for the formation of inherent groups or clusters. This is a simple task in the case of 2D vectors, where we may plot them, visually identify groups, and label each group with a pattern class; see Figures 10.1 and 10.2. Allowance may have to be made to assign the same class to multiple disjoint groups. Such an approach may be used even when the number of classes is not known at the outset. When the vectors have a dimension higher than three, visual analysis will not be feasible. It then becomes necessary to define criteria to group the given vectors on the basis of similarity, dissimilarity, or distance measures. A few examples of such measures are as follows [4]:

- Euclidean distance

$$D_E^2 = \|\mathbf{x} - \mathbf{z}\|^2 = (\mathbf{x} - \mathbf{z})^T (\mathbf{x} - \mathbf{z}) = \sum_{i=1}^n (x_i - z_i)^2. \quad (10.34)$$

Here, \mathbf{x} and \mathbf{z} are two feature vectors; the latter could be a class prototype, if available. A small value of D_E indicates greater similarity between the two vectors than a large value of D_E .

- Mahalanobis distance

$$D_M^2 = (\mathbf{x} - \mathbf{m})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m}), \quad (10.35)$$

where \mathbf{x} is a feature vector being compared to a pattern class for which \mathbf{m} is the class mean vector and \mathbf{C} is the covariance matrix. Inclusion of the (inverse of the) covariance of the distribution of the class in the distance measure allows consideration of the scatter of the constituent samples of the class: A large scatter of the population diminishes the distance measure as compared to a tighter cluster. A small value of D_M indicates a higher potential membership of the vector \mathbf{x} in the class than a large value of D_M .

- Normalized dot product (cosine of the angle between the vectors \mathbf{x} and \mathbf{z})

$$D_d = \frac{\mathbf{x}^T \mathbf{z}}{\|\mathbf{x}\| \|\mathbf{z}\|}. \quad (10.36)$$

A large dot product value indicates a greater degree of similarity between the two vectors than a small value.

The covariance matrix is defined as

$$\mathbf{C} = E[(\mathbf{y} - \mathbf{m})(\mathbf{y} - \mathbf{m})^T], \quad (10.37)$$

where the expectation operation is performed over all feature vectors \mathbf{y} that belong to the class. The covariance matrix provides the covariance of all possible pairs of the features in the feature vector over all samples belonging to the given class. The elements along the main diagonal of the covariance matrix provide the variance of the individual features that make up the feature vector. The covariance matrix represents the scatter of the features that belong to the given class. The mean and covariance need to be updated as more samples are added to a given class in a clustering procedure.

When the Mahalanobis distance needs to be calculated between a sample vector and a number of classes represented by their mean and covariance matrices, a pooled covariance matrix may be used if the numbers of members in the various classes are unequal and low [15]. For example, if the covariance matrices of two classes are \mathbf{C}_1 and \mathbf{C}_2 , and the numbers of members in the two classes are N_1 and N_2 , the pooled covariance matrix is given by

$$\mathbf{C} = \frac{(N_1 - 1)\mathbf{C}_1 + (N_2 - 1)\mathbf{C}_2}{N_1 + N_2 - 2}. \quad (10.38)$$

Various performance indices may be designed to measure the success of a clustering procedure [4]. A measure of the tightness of a cluster is the sum of the squared errors performance index:

$$J = \sum_{j=1}^{N_c} \sum_{\mathbf{x} \in S_j} \|\mathbf{x} - \mathbf{m}_j\|^2, \quad (10.39)$$

where N_c is the number of cluster domains, S_j is the set of samples in the j^{th} cluster,

$$\mathbf{m}_j = \frac{1}{N_j} \sum_{\mathbf{x} \in S_j} \mathbf{x} \quad (10.40)$$

is the sample mean vector of S_j , and N_j is the number of samples in S_j .

A few other examples of performance indices are:

- Average of the squared distances between the samples in a cluster domain.
- Intracluster variance.

- Average of the squared distances between the samples in different cluster domains.
- Intercluster distances.
- Scatter matrices.
- Covariance matrices.

A simple cluster-seeking algorithm [4]: Suppose we have N sample patterns $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$.

1. Let the first cluster center \mathbf{z}_1 be equal to any one of the samples, for example, $\mathbf{z}_1 = \mathbf{x}_1$.
2. Choose a nonnegative threshold θ .
3. Compute the distance D_{21} between \mathbf{x}_2 and \mathbf{z}_1 . If $D_{21} < \theta$, assign \mathbf{x}_2 to the domain (class) of cluster center \mathbf{z}_1 ; otherwise, start a new cluster with its center as $\mathbf{z}_2 = \mathbf{x}_2$. For the subsequent steps, let us assume that a new cluster with center \mathbf{z}_2 has been established.
4. Compute the distances D_{31} and D_{32} from the next sample \mathbf{x}_3 to \mathbf{z}_1 and \mathbf{z}_2 , respectively. If D_{31} and D_{32} are both greater than θ , start a new cluster with its center as $\mathbf{z}_3 = \mathbf{x}_3$; otherwise, assign \mathbf{x}_3 to the domain of the closer cluster.
5. Continue to apply steps 3 and 4 by computing and checking the distance from *every* new (unclassified) pattern vector to *every* established cluster center and applying the cluster-assignment or cluster-creation rule.
6. Stop when every given pattern vector has been assigned to a cluster.

Note that the procedure does not require knowledge of the number of classes *a priori*. Note also that the procedure does not assign a real-world class to each cluster: it merely groups the given vectors into disjoint clusters. A subsequent step is required to label each cluster with a class related to the actual problem. Multiple clusters may relate to the same real-world class and may have to be merged.

A major disadvantage of the simple cluster-seeking algorithm is that the results depend upon

- the first cluster center chosen for each domain or class,
- the order in which the sample patterns are considered,
- the value of the threshold θ , and
- the geometrical properties (distributions) of the data (or the feature-vector space).

The maximin-distance clustering algorithm [4]: This method is similar to the previous “simple” algorithm, but first identifies the cluster regions that are the farthest apart, as follows.

1. Let \mathbf{x}_1 be the first cluster center \mathbf{z}_1 .
2. Determine the farthest sample from \mathbf{x}_1 , and call it cluster center \mathbf{z}_2 .
3. Compute the distance from each remaining sample to \mathbf{z}_1 and to \mathbf{z}_2 . For every pair of these computations, save the minimum distance, and select the maximum of the minimum distances. If this “maximin” distance is an appreciable fraction of the distance between the cluster centers \mathbf{z}_1 and \mathbf{z}_2 , label the corresponding sample as a new cluster center \mathbf{z}_3 ; otherwise stop forming new clusters and go to Step 5.
4. If a new cluster center was formed in Step 3, repeat Step 3 using a “typical” or the average distance between the established cluster centers for comparison.

5. Assign each remaining sample to the domain of its nearest cluster center.

The term “maximin” refers to the combined use of maximum and minimum distances between the given vectors and the centers of the clusters already formed.

The K -means algorithm [4]: The preceding “simple” and “maximin” algorithms are intuitive procedures. The K -means algorithm is based on iterative minimization of a performance index that is defined as the sum of the squared distances from all points in a cluster domain to the cluster center, as follows.

1. Choose K initial cluster centers $\mathbf{z}_1(1), \mathbf{z}_2(1), \dots, \mathbf{z}_K(1)$. The index in parentheses represents the iteration number.
2. At the k^{th} iterative step, distribute the samples $\{\mathbf{x}\}$ among the K cluster domains, using the relation

$$\mathbf{x} \in S_j(k) \text{ if } \|\mathbf{x} - \mathbf{z}_j(k)\| < \|\mathbf{x} - \mathbf{z}_i(k)\| \forall i = 1, 2, \dots, K, i \neq j, \quad (10.41)$$

where $S_j(k)$ denotes the set of samples whose cluster center is $\mathbf{z}_j(k)$.

3. From the results of Step 2, compute the new cluster centers $\mathbf{z}_j(k+1)$, $j = 1, 2, \dots, K$, such that the sum of the squared distances from all points in $S_j(k)$ to the new cluster center is minimized. In other words, the new cluster center $\mathbf{z}_j(k+1)$ is computed so that the performance index

$$J_j = \sum_{\mathbf{x} \in S_j(k)} \|\mathbf{x} - \mathbf{z}_j(k+1)\|^2, \quad j = 1, 2, \dots, K, \quad (10.42)$$

is minimized. The $\mathbf{z}_j(k+1)$ that minimizes this performance index is simply the sample mean of $S_j(k)$. Therefore, the new cluster center is given by

$$\mathbf{z}_j(k+1) = \frac{1}{N_j} \sum_{\mathbf{x} \in S_j(k)} \mathbf{x}, \quad j = 1, 2, \dots, K, \quad (10.43)$$

where N_j is the number of samples in $S_j(k)$. The name “ K -means” is derived from the manner in which cluster centers are sequentially updated.

4. If $\mathbf{z}_j(k+1) = \mathbf{z}_j(k)$ for $j = 1, 2, \dots, K$, the algorithm has converged: Terminate the procedure. Otherwise, go to Step 2.

The behavior of the K -means algorithm is influenced by:

- the number of cluster centers specified,
- the choice of the initial cluster centers,
- the order in which the sample patterns are considered, and
- the geometrical properties (distributions) of the data (or the feature-vector space).

See Section 10.11.3 for an illustration of application of the K -means method to an ECG signal with PVCs.

10.6 Probabilistic Models and Statistical Decision

Problem: Pattern classification methods such as discriminant functions are dependent upon the set of training samples provided. Their success, when applied to new cases, will depend upon the accuracy of representation of the various pattern classes by the training samples. How can we design pattern classification techniques that are independent of specific training samples and optimal in a broad sense?

Solution: Probability functions and probabilistic models may be developed to represent the occurrence and statistical attributes of classes of patterns. Such functions may be based on large collections of data, historical records, or mathematical models of pattern generation. In the absence of information as above, a training step with samples of known categorization is required to estimate the required model parameters. It is common practice to assume a Gaussian PDF to represent the distribution of the features for each class, and to estimate the required mean and variance parameters from the training sets. When PDFs are available to characterize pattern classes and their features (see Sections 5.12.2 and 5.12.3), optimal decision functions may be designed based on statistical functions and decision theory. The following sections describe a few methods that fall into this category.

10.6.1 Likelihood functions and statistical decision

Let $P(C_i)$ be the probability of occurrence of class C_i , $i = 1, 2, \dots, M$; this is known as the *a priori*, *prior*, or unconditional probability. The *a posteriori* or *posterior* probability that an observed sample pattern \mathbf{x} comes from C_i is expressed as $P(C_i|\mathbf{x})$. If a classifier decides that \mathbf{x} comes from C_j when it actually came from C_i , then the classifier is said to incur a *loss* L_{ij} , with $L_{ii} = 0$ or a fixed operational cost and $L_{ij} > L_{ii} \forall j \neq i$.

Since \mathbf{x} may belong to any of M classes under consideration, the expected loss, known as the *conditional average risk* or *loss*, in assigning \mathbf{x} to C_j is [4]

$$R_j(\mathbf{x}) = \sum_{i=1}^M L_{ij} P(C_i|\mathbf{x}). \quad (10.44)$$

A classifier could compute $R_j(\mathbf{x})$, $j = 1, 2, \dots, M$, for each sample \mathbf{x} and then assign \mathbf{x} to the class with the smallest conditional loss. Such a classifier will minimize the total expected loss over all decisions, and is called the *Bayes classifier*. From a statistical point of view, the Bayes classifier represents an optimal classifier.

According to Bayes formula, we have [4, 14]

$$P(C_i|\mathbf{x}) = \frac{P(C_i) p(\mathbf{x}|C_i)}{p(\mathbf{x})}, \quad (10.45)$$

where $p(\mathbf{x}|C_i)$ is called the *likelihood function* of class C_i or the *state-conditional PDF* of \mathbf{x} , and $p(\mathbf{x})$ is the PDF of \mathbf{x} regardless of class membership (unconditional). [Note: $P(y)$ is used to represent the probability of occurrence of an event y ; $p(y)$ is used to represent the PDF of a random variable y . Probabilities and PDFs involving a multidimensional feature vector are multivariate functions with dimension equal to that of the feature vector.] Bayes formula shows how observing the sample \mathbf{x} changes the *a priori* probability $P(C_i)$ to the *a posteriori* probability $P(C_i|\mathbf{x})$. In other words, Bayes formula provides a mechanism to update the *a priori* probability $P(C_i)$ to the *a posteriori* probability $P(C_i|\mathbf{x})$ due to the observation of the sample \mathbf{x} . Then, we can express the expected loss as [4]

$$R_j(\mathbf{x}) = \frac{1}{p(\mathbf{x})} \sum_{i=1}^M L_{ij} p(\mathbf{x}|C_i) P(C_i). \quad (10.46)$$

As $\frac{1}{p(\mathbf{x})}$ is common for all j , we could modify $R_j(\mathbf{x})$ to

$$r_j(\mathbf{x}) = \sum_{i=1}^M L_{ij} p(\mathbf{x}|C_i) P(C_i). \quad (10.47)$$

In a two-class case with $M = 2$, we obtain the following expressions [4]:

$$r_1(\mathbf{x}) = L_{11} p(\mathbf{x}|C_1) P(C_1) + L_{21} p(\mathbf{x}|C_2) P(C_2). \quad (10.48)$$

$$r_2(\mathbf{x}) = L_{12} p(\mathbf{x}|C_1) P(C_1) + L_{22} p(\mathbf{x}|C_2) P(C_2). \quad (10.49)$$

$$\mathbf{x} \in C_1 \text{ if } r_1(\mathbf{x}) < r_2(\mathbf{x}), \quad (10.50)$$

that is,

$$\begin{aligned} \mathbf{x} \in C_1 \text{ if } & L_{11} p(\mathbf{x}|C_1) P(C_1) + L_{21} p(\mathbf{x}|C_2) P(C_2) \\ & < L_{12} p(\mathbf{x}|C_1) P(C_1) + L_{22} p(\mathbf{x}|C_2) P(C_2), \end{aligned} \quad (10.51)$$

or equivalently,

$$\mathbf{x} \in C_1 \text{ if } (L_{21} - L_{22}) p(\mathbf{x}|C_2) P(C_2) < (L_{12} - L_{11}) p(\mathbf{x}|C_1) P(C_1). \quad (10.52)$$

This expression may be rewritten as [4]

$$\mathbf{x} \in C_1 \text{ if } \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} > \frac{P(C_2)}{P(C_1)} \frac{(L_{21} - L_{22})}{(L_{12} - L_{11})}. \quad (10.53)$$

The LHS of the inequality above, which is a ratio of two likelihood functions, is often referred to as the *likelihood ratio*:

$$l_{12}(\mathbf{x}) = \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)}. \quad (10.54)$$

Then, Bayes decision rule for $M = 2$ is [4]:

1. Assign \mathbf{x} to class C_1 if $l_{12}(\mathbf{x}) > \theta_{12}$, where θ_{12} is a threshold given by $\theta_{12} = \frac{P(C_2)}{P(C_1)} \frac{(L_{21} - L_{22})}{(L_{12} - L_{11})}$.
2. Assign \mathbf{x} to class C_2 if $l_{12}(\mathbf{x}) < \theta_{12}$.
3. Make an arbitrary or heuristic decision if $l_{12}(\mathbf{x}) = \theta_{12}$.

The rule may be generalized to the M -class case as [4]:

$$\mathbf{x} \in C_i \text{ if } \sum_{k=1}^M L_{ki} p(\mathbf{x}|C_k) P(C_k) < \sum_{q=1}^M L_{qj} p(\mathbf{x}|C_q) P(C_q), \quad (10.55)$$

$j = 1, 2, \dots, M$, $j \neq i$.

In most pattern classification problems, the loss is nil for correct decisions. The loss could be assumed to be equal to a certain quantity for all erroneous decisions. Then, $L_{ij} = 1 - \delta_{ij}$, where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \quad (10.56)$$

and

$$\begin{aligned} r_j(\mathbf{x}) &= \sum_{i=1}^M (1 - \delta_{ij}) p(\mathbf{x}|C_i) P(C_i) \\ &= p(\mathbf{x}) - p(\mathbf{x}|C_j) P(C_j), \end{aligned} \quad (10.57)$$

since

$$\sum_{i=1}^M p(\mathbf{x}|C_i) P(C_i) = p(\mathbf{x}). \quad (10.58)$$

The Bayes classifier will assign a pattern \mathbf{x} to class C_i if

$$p(\mathbf{x}) - p(\mathbf{x}|C_i) P(C_i) < p(\mathbf{x}) - p(\mathbf{x}|C_j) P(C_j), \quad j = 1, 2, \dots, M, \quad j \neq i, \quad (10.59)$$

that is,

$$\mathbf{x} \in C_i \text{ if } p(\mathbf{x}|C_i) P(C_i) > p(\mathbf{x}|C_j) P(C_j), \quad j = 1, 2, \dots, M, \quad j \neq i. \quad (10.60)$$

This is nothing more than using the decision functions

$$d_i(\mathbf{x}) = p(\mathbf{x}|C_i) P(C_i), \quad i = 1, 2, \dots, M, \quad (10.61)$$

where a pattern \mathbf{x} is assigned to class C_i if $d_i(\mathbf{x}) > d_j(\mathbf{x}) \forall j \neq i$ for that pattern. Using Bayes formula, we get

$$d_i(\mathbf{x}) = P(C_i|\mathbf{x}) p(\mathbf{x}), \quad i = 1, 2, \dots, M. \quad (10.62)$$

Since $p(\mathbf{x})$ does not depend upon the class index i , this can be reduced to

$$d_i(\mathbf{x}) = P(C_i|\mathbf{x}), \quad i = 1, 2, \dots, M. \quad (10.63)$$

The different decision functions given above provide alternative yet equivalent approaches, depending upon whether $p(\mathbf{x}|C_i)$ or $P(C_i|\mathbf{x})$ is used (or available). Estimation of $p(\mathbf{x}|C_i)$ would require a training set for each class C_i . It is common to assume a Gaussian distribution and estimate its mean and variance using the training set.

10.6.2 Bayes classifier for normal patterns

The univariate normal or Gaussian PDF for a single random variable x is given by

$$p(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left[-\frac{1}{2} \left(\frac{x - m}{\sigma} \right)^2 \right], \quad (10.64)$$

which is completely specified by two parameters: the mean

$$m = E[x] = \int_{-\infty}^{\infty} x p(x) dx, \quad (10.65)$$

and the variance

$$\sigma^2 = E[(x - m)^2] = \int_{-\infty}^{\infty} (x - m)^2 p(x) dx. \quad (10.66)$$

In the case of M pattern classes and pattern vectors \mathbf{x} of dimension n governed by multivariate normal PDFs, we have

$$p(\mathbf{x}|C_i) = \frac{1}{(2\pi)^{n/2} |\mathbf{C}_i|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}_i^{-1} (\mathbf{x} - \mathbf{m}_i) \right], \quad (10.67)$$

$i = 1, 2, \dots, M$, where each PDF is completely specified by its mean vector \mathbf{m}_i and its $n \times n$ covariance matrix \mathbf{C}_i , with

$$\mathbf{m}_i = E_i[\mathbf{x}], \quad (10.68)$$

and

$$\mathbf{C}_i = E_i[(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T]. \quad (10.69)$$

Here, $E_i[\cdot]$ denotes the expectation operator over the patterns belonging to class C_i .

Normal distributions occur frequently in nature and have the advantage of analytical tractability. A multivariate normal PDF reduces to a product of univariate normal PDFs when the elements of \mathbf{x} are mutually independent (then the covariance matrix is a diagonal matrix).

Earlier, we formulated the decision functions

$$d_i(\mathbf{x}) = p(\mathbf{x}|C_i) P(C_i), \quad i = 1, 2, \dots, M. \quad (10.70)$$

Given the exponential in the normal PDF, it is convenient to use

$$d_i(\mathbf{x}) = \ln [p(\mathbf{x}|C_i) P(C_i)] = \ln p(\mathbf{x}|C_i) + \ln P(C_i), \quad (10.71)$$

which is equivalent in terms of classification performance as the natural logarithm \ln is a monotonically increasing function. Then [4],

$$d_i(\mathbf{x}) = \ln P(C_i) - \frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\mathbf{C}_i| - \frac{1}{2} [(\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}_i^{-1} (\mathbf{x} - \mathbf{m}_i)], \quad (10.72)$$

$i = 1, 2, \dots, M$. The second term does not depend upon i ; therefore, we can simplify $d_i(\mathbf{x})$ to

$$d_i(\mathbf{x}) = \ln P(C_i) - \frac{1}{2} \ln |\mathbf{C}_i| - \frac{1}{2} [(\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}_i^{-1} (\mathbf{x} - \mathbf{m}_i)], \quad i = 1, 2, \dots, M. \quad (10.73)$$

The decision functions above are hyperquadrics; hence, the best that a Bayes classifier for normal patterns can do is to place a general second-order (quadratic) decision surface between each pair of pattern classes. In the case of true normal distributions of patterns, the decision functions as above are optimal on an average basis: They minimize the expected loss with the simplified loss function $L_{ij} = 1 - \delta_{ij}$ [4].

If all of the covariance matrices are equal, that is, $\mathbf{C}_i = \mathbf{C}$, $i = 1, 2, \dots, M$, we get

$$d_i(\mathbf{x}) = \ln P(C_i) + \mathbf{x}^T \mathbf{C}^{-1} \mathbf{m}_i - \frac{1}{2} \mathbf{m}_i^T \mathbf{C}^{-1} \mathbf{m}_i, \quad i = 1, 2, \dots, M, \quad (10.74)$$

after omitting terms independent of i . The Bayesian classifier is now represented by a set of linear decision functions.

Before one may apply the decision functions as above, it would be appropriate to verify the Gaussian nature of the PDFs of the variables on hand by conducting statistical tests [3, 23]. Furthermore, it would be necessary to derive or estimate the mean vector and covariance matrix for each class; sample statistics computed from a training set may serve this purpose.

See Section 10.11.2 for an illustration of application of the Bayesian method to an ECG signal with PVCs.

10.7 Logistic Regression Analysis

Logistic classification is a statistical technique based on a logistic regression model that estimates the probability of occurrence of an event [24–26]. The technique is designed for problems where patterns are to be classified into one of two classes. When the response variable is binary, theoretical

and empirical considerations indicate that the response function is often curvilinear. The typical response function is shaped as a forward or backward tilted “S” and is known as a sigmoidal function. The function has asymptotes at 0 and 1.

In logistic pattern classification, an event is defined as the membership of a pattern vector in one of the two classes. The method computes a variable that depends upon the given parameters and is constrained to the range [0, 1] so that it may be interpreted as a probability. The probability of the pattern vector belonging to the second class is simply the difference between unity and the estimated value.

For the case of a single feature or parameter, the logistic regression model is given as

$$P(\text{event}) = \frac{\exp(b_0 + b_1x)}{1 + \exp(b_0 + b_1x)}, \quad (10.75)$$

or equivalently,

$$P(\text{event}) = \frac{1}{1 + \exp[-(b_0 + b_1x)]}, \quad (10.76)$$

where b_0 and b_1 are coefficients estimated from the data, and x is the independent (feature) variable. The relationship between the independent variable and the estimated probability is nonlinear; it follows an S-shaped curve that resembles the integral of a Gaussian function. In the case of an n -dimensional feature vector \mathbf{x} , the model can be written as

$$P(\text{event}) = \frac{1}{1 + \exp(-z)}, \quad (10.77)$$

where z is the linear combination

$$z = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n = \langle \mathbf{b}, \mathbf{x} \rangle, \quad (10.78)$$

that is, z is the dot product of the augmented feature vector \mathbf{x} with a coefficient or weight vector \mathbf{b} .

In linear regression, the coefficients of the model are estimated using the method of least squares; the selected regression coefficients are those that result in the smallest sum of squared distances between the observed and the predicted values of the dependent variable. In logistic regression, the parameters of the model are estimated using the maximum likelihood method [3,24]; the coefficients that make the observed results “most likely” are selected. Since the logistic regression model is nonlinear, an iterative algorithm is necessary for estimation of the coefficients [25, 26]. A training set is required to design a classifier based on logistic regression.

10.8 Neural Networks

In many practical problems, we may have no knowledge of the prior probabilities of patterns belonging to one class or another. No general classification rules may exist for the patterns on hand. Clinical knowledge may not yield symbolic knowledge bases that could be used to classify patterns that demonstrate exceptional behavior. In such situations, conventional pattern classification methods as described in the preceding sections may not be well suited for classification of pattern vectors. Artificial neural networks (ANNs), with the properties of experience-based learning and fault tolerance, should be effective in solving such classification problems [2, 16, 17, 27–30].

Figure 10.5 illustrates a two-layer perceptron with one hidden layer and one output layer for pattern classification. The network learns the similarities among patterns directly from their instances in the training set that is provided initially. Classification rules are inferred from the training data without prior knowledge of the pattern class distributions in the data. Training of an ANN classifier

is typically achieved by the *back-propagation* algorithm [2, 16, 17, 27–30]. The actual output of the ANN y_k is calculated as

$$y_k = f \left(\sum_{j=1}^J w_{jk}^\# x_j^\# - \theta_k^\# \right), \quad k = 1, 2, \dots, K, \quad (10.79)$$

where

$$x_j^\# = f \left(\sum_{i=1}^I w_{ij} x_i - \theta_j \right), \quad j = 1, 2, \dots, J, \quad (10.80)$$

and

$$f(\beta) = \frac{1}{1 + \exp(-\beta)}. \quad (10.81)$$

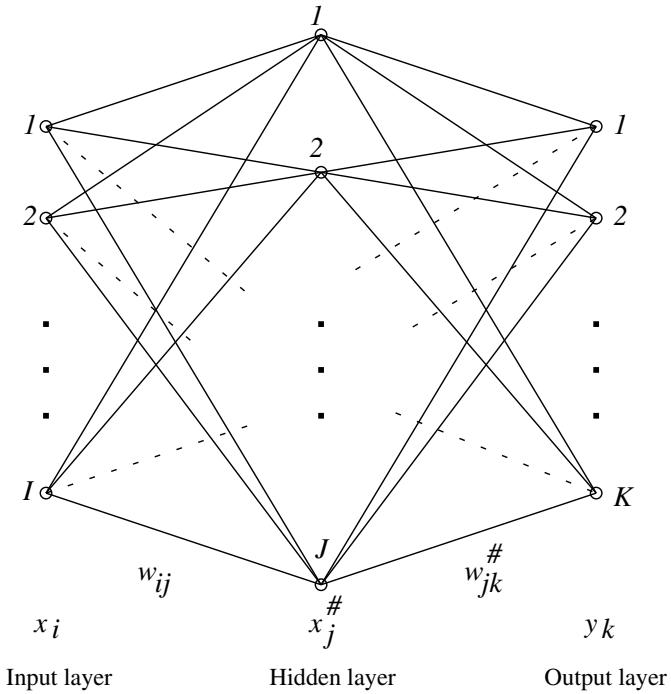


Figure 10.5 A two-layer perceptron.

In the equations given above, θ_j and $\theta_k^\#$ are node offsets; w_{ij} and $w_{jk}^\#$ are node weights; x_i are the elements of the pattern vectors (input parameters); and I , J , and K are the numbers of nodes in the input, hidden, and output layers, respectively. The weights and offsets are updated by

$$w_{jk}^\#(n+1) = w_{jk}^\#(n) + \eta[y_k(1 - y_k)(d_k - y_k)]x_j^\# + \alpha[w_{jk}^\#(n) - w_{jk}^\#(n-1)], \quad (10.82)$$

$$\theta_k^\#(n+1) = \theta_k^\#(n) + \eta[y_k(1 - y_k)(d_k - y_k)](-1) + \alpha[\theta_k^\#(n) - \theta_k^\#(n-1)], \quad (10.83)$$

$$\begin{aligned}
w_{ij}(n+1) &= w_{ij}(n) \\
&+ \eta \left[x_j^\# (1 - x_j^\#) \sum_{k=1}^K \{y_k(1 - y_k)(d_k - y_k)w_{jk}^\#\} \right] x_i \\
&+ \alpha[w_{ij}(n) - w_{ij}(n-1)],
\end{aligned} \tag{10.84}$$

and

$$\begin{aligned}
\theta_j(n+1) &= \theta_j(n) \\
&+ \eta \left[x_j^\# (1 - x_j^\#) \sum_{k=1}^K \{y_k(1 - y_k)(d_k - y_k)w_{jk}^\#\} \right] (-1) \\
&+ \alpha[\theta_j(n) - \theta_j(n-1)],
\end{aligned} \tag{10.85}$$

where d_k are the desired outputs, α is a momentum term, η is a gain term, and n refers to the iteration number. Equations 10.82 and 10.83 represent the back-propagation steps, with $y_k(1 - y_k)x_j^\#$ being the sensitivity of y_k to $w_{jk}^\#$, that is, $\frac{\partial y_k}{\partial w_{jk}^\#}$.

The training algorithm is repeated until the errors between the desired outputs and the actual outputs for the training data are smaller than a predetermined threshold value. Shen et al. [30] present a LOO approach to determine the most suitable values for the parameters J , η , and α .

10.8.1 ANNs with radial basis functions

An ANN with RBF [31] or RBF network (RBFN) with a feed-forward hidden layer (see Figure 10.6) applies a nonlinear transformation from the input space to a high-dimensional hidden space, and then produces responses through a linear output transformation. Cover [32] showed that it is possible to transform a given a set of samples that is not linearly separable into one that is linearly separable by applying a nonlinear transformation to project it into a higher-dimensional space. This approach forms the motivation for the use of nonlinear kernel-based methods in pattern classification and machine learning applications.

Consider a set of N labeled input feature vectors, \mathbf{x}_n , $n = 1, 2, \dots, N$, characterizing the N signals in a study, each of which is an $M \times 1$ vector. Let y_n be the desired output or classification response for the n^{th} signal, represented by its feature vector \mathbf{x}_n . With reference to the RBFN shown in Figure 10.6, we have the output of the network as

$$\widetilde{y}_n = \sum_{j=1}^J w_j \phi(\mathbf{x}_n, \mathbf{c}_j) + w_0, \tag{10.86}$$

where \widetilde{y}_n indicates an estimate of y_n , the RBF ϕ is defined as

$$\phi(\mathbf{x}_n, \mathbf{c}_j) = \exp \left(-\log_e(2) \frac{\|\mathbf{x}_n - \mathbf{c}_j\|^2}{\sigma^2} \right), \tag{10.87}$$

w_j is the weight and \mathbf{c}_j is the center vector for the j^{th} neuron in the hidden layer, J is the number of neurons in the hidden layer, w_0 is the bias, and σ is the spread parameter that determines the width of the area in the input space to which each hidden neuron responds.

The major challenge in the design of an RBFN is the selection of the centers. The selection of the centers in a random fashion commonly leads to a relatively large network with high computational complexity. Rangayyan and Wu [19] applied the orthogonal least-squares (OLS) method [33], a systematic method for center selection which can reduce the size of the RBFN, for screening of VAG signals.

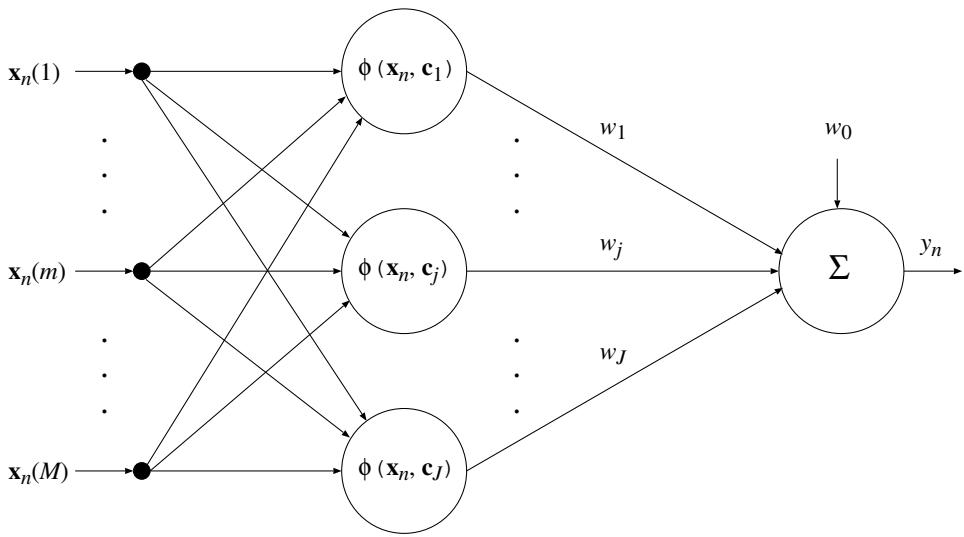


Figure 10.6 Schematic representation of an RBF network. The inputs to the network are the M components of the feature vector \mathbf{x}_n of a signal to be classified. The functions shown as ϕ are the RBFs. The hidden layer has J neurons.

According to Equation 10.86, the mapping performed by the RBFN can be viewed as a regression model, expressed in matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & \phi(\mathbf{x}_1, \mathbf{c}_1) & \cdots & \phi(\mathbf{x}_1, \mathbf{c}_J) \\ 1 & \phi(\mathbf{x}_2, \mathbf{c}_1) & \cdots & \phi(\mathbf{x}_2, \mathbf{c}_J) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \phi(\mathbf{x}_N, \mathbf{c}_1) & \cdots & \phi(\mathbf{x}_N, \mathbf{c}_J) \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_J \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}, \quad (10.88)$$

which is equivalent to

$$\mathbf{y} = \Phi \mathbf{w} + \mathbf{e}, \quad (10.89)$$

where Φ is the $N \times (J + 1)$ regression matrix with the RBFs; \mathbf{y} represents the vectorial form of the corresponding values y_n for $n = 1, 2, \dots, N$; $\mathbf{w} = [w_0, w_1, \dots, w_J]^T$; and \mathbf{e} is the approximation error vector.

The centers of the RBFN are chosen from the set of input feature vectors (a total of N candidates). The task of the OLS method is to perform a systematic selection of fewer than N centers so that the network size can be reduced with minimal degradation of performance during the learning procedure. From Equation 10.88, we can see that there is a one-to-one correspondence between the centers of the RBFN and the coefficients in the regression matrix Φ . At each step of the OLS regression, a new center can be selected in such a manner that the incremental variance of the desired output is maximized. Suppose that there are $Q < N$ centers selected. The OLS solution yielding the weights is given by [33]

$$\tilde{\mathbf{w}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y} = \Phi^+ \mathbf{y}, \quad (10.90)$$

where Φ^+ represents the pseudoinverse of the regression matrix Φ . The output of the RBFN is then expressed as

$$\tilde{\mathbf{y}} = \Phi \tilde{\mathbf{w}} = [\phi_1, \phi_2, \dots, \phi_Q] \tilde{\mathbf{w}}, \quad (10.91)$$

where $\tilde{\mathbf{y}}$ denotes the portion of \mathbf{y} that is within the vectorial space spanned by the columns ϕ_q of the regression matrix Φ .

By using Gram–Schmidt orthogonalization [31], the regression matrix can be decomposed as

$$\Phi = \mathbf{B} \mathbf{A} = [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_Q] \begin{bmatrix} 1 & a_{11} & a_{12} & \cdots & a_{1Q} \\ 0 & 1 & a_{23} & \cdots & a_{2Q} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \quad (10.92)$$

where \mathbf{A} is a $Q \times Q$ upper-triangular matrix with 1s on the main diagonal, and \mathbf{B} is an $N \times Q$ matrix with mutually orthogonal columns \mathbf{b}_q such that

$$\mathbf{B}^T \mathbf{B} = \mathbf{H} = \text{diag}[h_1, h_2, \dots, h_Q], \quad (10.93)$$

where the $Q \times Q$ matrix \mathbf{H} is a diagonal matrix with elements h_k given by

$$h_k = \mathbf{b}_k^T \mathbf{b}_k = \sum_{q=1}^N b_{kq}^2. \quad (10.94)$$

By substituting Equation 10.92 into Equation 10.89, we obtain

$$\mathbf{y} = \mathbf{B} \mathbf{A} \mathbf{w} + \mathbf{e} = \mathbf{B} \mathbf{g} + \mathbf{e}, \quad (10.95)$$

where $\mathbf{g} = \mathbf{A} \mathbf{w}$. In Equation 10.95, the desired output vector \mathbf{y} is expressed as a linear combination of the mutually orthogonal columns of the matrix \mathbf{B} . The OLS solution for the coordinate vector \mathbf{g} is given by

$$\tilde{\mathbf{g}} = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{y} = \mathbf{B}^+ \mathbf{z} = \mathbf{H}^{-1} \mathbf{B}^T \mathbf{y}. \quad (10.96)$$

The k^{th} component of the vector $\tilde{\mathbf{g}}$ is given by

$$g_k = \frac{\mathbf{b}_k^T \mathbf{y}}{\mathbf{b}_k^T \mathbf{b}_k}. \quad (10.97)$$

Because Gram–Schmidt orthogonalization ensures the orthogonality between the approximation error \mathbf{e} and $\mathbf{B} \mathbf{g}$ in Equation 10.95, we have

$$\mathbf{y}^T \mathbf{y} = \mathbf{g}^T \mathbf{B}^T \mathbf{B} \mathbf{g} + \mathbf{e}^T \mathbf{e} = \mathbf{g}^T \mathbf{H} \mathbf{g} + \mathbf{e}^T \mathbf{e} = \sum_{k=1}^Q h_k g_k^2 + \mathbf{e}^T \mathbf{e}. \quad (10.98)$$

Because there is a one-to-one correspondence between the elements of the regression vector \mathbf{g} and the RBF centers \mathbf{c}_j , each term in the summation above reflects the contribution of each of the RBF centers. We can, therefore, define an error reduction ratio (ϵ) with respect to the j^{th} RBF center as [33]

$$\epsilon_j = \frac{h_j g_j^2}{\mathbf{y}^T \mathbf{y}}. \quad (10.99)$$

The error reduction ratio offers an effective criterion for the selection of RBF centers in a regression model. At each step of the regression, an RBF center is selected so as to maximize the error reduction ratio toward a tolerance value.

The input layer of the RBFN used by Rangayyan and Wu [19] contained $M = 5$ nodes to accept the set of five features (FF_1, FF_2, S, K, H) extracted from each VAG signal (see Sections 5.12 and 10.4.2). The spread parameter σ was varied over the range [1, 6], and the number of hidden nodes J was varied over the range [1, 30]. The resulting output values were used to derive ROC curves and the associated A_z values. The highest A_z value obtained was 0.82 using the RBFN classifier with $\sigma = 6$ and $J = 23$ hidden nodes using $N = 89$ VAG signals. In another work, Rangayyan and Wu [34] used an RBFN with the turns count computed for the first and second halves of each VAG signal, and obtained better classification performance with $A_z = 0.92$.

Regardless of the good results they provide and their popularity, RBFNs pose problems related to the derivation of optimal network parameters and generalization from a training set to a test set.

10.8.2 Deep learning

Recently, deep learning has gained popularity due to the availability of large datasets and high computational capabilities with graphical processing units and tensor processing units [35]. The forerunner of deep learning architectures is the ANN. As seen in the preceding sections, an ANN has an input layer, an output layer, and a few hidden layers. Neurons in each layer perform specialized functions. The term “hidden layer” refers to the layers between the input and output layers; there could be one or more hidden layers. In deep learning, the neurons are referred to as nodes, and each node applies a mathematical modification or some form of adaptive filtering to its input, resulting in a reduction in dimensionality. Deep neural networks contain multiple hidden layers, and convolutional neural networks (CNNs) are commonly used deep learning models [35]. Deep learning techniques require substantial computing time and power.

A typical deep learning network includes numerous hidden layers, and each hidden layer node functions as a learnable kernel or adaptive filter that applies a particular mathematical operation to the input it receives. The network gradually learns the many levels of feature representation in the data by feeding the output of one layer as an input to the next. The final layer, known as the output layer, functions as a fundamental classifier, converting the processed data into decision outputs in accordance with a predetermined threshold. Deep learning has several benefits, including the ability to model complicated nonlinear properties in signals and compatibility with many nonstationary, nonlinear biomedical applications.

In deep learning, the number of layers raises the level of abstraction and the risk of overfitting the data. In order to prevent overfitting, a variety of regularization strategies, including straightforward approaches such as data augmentation techniques, could be used. Deep learning’s difficulty in being explained or interpreted is frequently cited as one of its drawbacks. This is particularly true in the medical field, where CAD necessitates the use of knowledgeable judgment by specialists in the field of healthcare. Deep learning could also be extended to other applications, such as analysis of electronic medical records, behavioral analytics, bioinformatics, computational audio analysis, computer vision, genome sequencing, and natural language processing, which could all be incorporated in biomedical signal analysis tasks to provide a multimodal framework for clinical decision support systems.

10.9 Measures of Diagnostic Accuracy and Cost

Pattern recognition or classification decisions that are made in the context of medical diagnosis have implications that go beyond statistical measures of accuracy and validity. We need to provide a clinical or diagnostic interpretation of statistical or rule-based decisions made with signal pattern vectors.

Consider the simple situation of *screening*, which represents the use of a test to detect the presence or absence of a specific disease in a certain study population. The decision to be made is

binary: A given subject has or does not have the disease in question. The presence or absence of other diseases or abnormalities is not considered. Let us represent by A the event that a subject has the particular pathology, and by N the event that the subject does not have the disease. Let the prior probabilities $P(A)$ and $P(N)$ represent the fractions of subjects with the disease and the subjects without the disease, respectively, in the test population. Let T^+ represent a positive screening test result (indicative of the presence of the disease) and T^- a negative result. The following possibilities arise [36]:

- A *true positive* (TP) is the situation when the test is positive for a subject with the disease (also known as a *hit*). The true-positive fraction (*TPF*) or *sensitivity* S^+ is given as $P(T^+|A)$ or

$$S^+ = \frac{\text{number of TP decisions}}{\text{number of subjects with the disease}}. \quad (10.100)$$

The sensitivity of a test represents its capability to detect the presence of the disease of concern.

- A *true negative* (TN) represents the case when the test is negative for a subject who does not have the disease. The true-negative fraction (*TNF*) or *specificity* S^- is given as $P(T^-|N)$ or

$$S^- = \frac{\text{number of TN decisions}}{\text{number of subjects without the disease}}. \quad (10.101)$$

The specificity of a test indicates its accuracy in identifying the absence of the disease of concern.

- A *false negative* (FN) is said to occur when the test is negative for a subject who has the disease of concern; that is, the test has missed the case. The probability of this error, known as the false-negative fraction (*FNF*), is $P(T^-|A)$.
- A *false positive* (FP) is defined as the case where the result of the test is positive when the individual being tested does not have the disease. The probability of this type of error or false alarm, known as the false-positive fraction (*FPF*), is $P(T^+|N)$.

Table 10.1 summarizes the classification possibilities. Note that

- $FNF + TPF = 1$,
- $FPF + TNF = 1$,
- $S^- = 1 - FPF = TNF$, and
- $S^+ = 1 - FNF = TPF$.

A summary measure of accuracy may be defined as [36]

$$\text{accuracy} = S^+ P(A) + S^- P(N), \quad (10.102)$$

where $P(A)$ is the fraction of the study population that actually has the disease (that is, the prevalence of the disease) and $P(N)$ is the fraction of the study population that is actually free of the disease.

If the prior probabilities are not available, the accuracy of classification can be estimated as

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (10.103)$$

In the same way, we have

$$TPF = \frac{TP}{TP + FN} \quad (10.104)$$

Actual Group	Predicted Group	
	Without the Disease	With the Disease
Without the Disease	$S^- = TNF$	FPF
With the Disease	FNF	$S^+ = TPF$

Table 10.1 Schematic representation of a classification matrix. S^- denotes the specificity (true-negative fraction or TNF), FPF denotes the false-positive fraction, FNF denotes the false-negative fraction, and S^+ denotes the sensitivity (true-positive fraction or TPF).

and

$$TNF = \frac{TN}{TN + FP}. \quad (10.105)$$

The efficiency of a test may also be indicated by its predictive values. The *positive predictive value* PPV of a test, defined as

$$PPV = \frac{TP}{TP + FP}, \quad (10.106)$$

represents the fraction of the cases labeled as positive by the test that are actually positive. The *negative predictive value* NPV , defined as

$$NPV = \frac{TN}{TN + FN}, \quad (10.107)$$

represents the fraction of the cases labeled as negative by the test that are actually negative. PPV is also known as precision and TPF is also known as recall. The harmonic mean of precision and recall is known as the $F1$ score, and is given by

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + (FP + FN)/2}. \quad (10.108)$$

A method gets a high $F1$ score only if both precision and recall are high.

When a new test or method of diagnosis is being developed and tested, it is necessary to use another previously established method as a reference to confirm the presence or absence of the disease. Such a reference method is often called the *gold standard*, and its results are referred to as the *ground truth*. When computer-based methods need to be tested, it is common practice to use the diagnosis or classification provided by an expert in the field as the gold standard. Results of biopsy, other established laboratory or investigative procedures, or long-term clinical follow-up in the case of normal subjects may also serve this purpose. The term “actual group” in Table 10.1 indicates the result of the gold standard, and the term “predicted group” refers to the result of the test conducted.

Healthcare professionals (and the general public) would be interested in knowing the probability that a subject with a positive test result actually has the disease: This is given by the conditional probability $P(A|T^+)$. The question could be answered by using Bayes theorem [3] as

$$P(A|T^+) = \frac{P(A) P(T^+|A)}{P(A)P(T^+|A) + P(N)P(T^+|N)}. \quad (10.109)$$

Note that $P(T^+|A) = S^+$ and $P(T^+|N) = 1 - S^-$. In order to determine the posterior probability as above, the sensitivity and specificity of the test and the prior probabilities of negative cases and positive cases (the rate of prevalence of the disease) should be known.

A cost matrix may be defined, as in Table 10.2, to reflect the overall cost effectiveness of a test or method of diagnosis. The cost of conducting the test and arriving at a TN decision is indicated

by C_N ; this is the cost of subjecting a normal person to the test for the purposes of screening for a disease. The cost of the test when a TP is found is shown as C_A ; this might include the costs of further tests, treatment, and follow-up, which are secondary to the test itself but part of the screening and healthcare program. The value C^+ indicates the cost of an FP result; this represents the cost of erroneously subjecting an individual without the disease to further tests or therapy. Whereas it may be easy to identify the costs of clinical tests or treatment procedures, it is difficult to quantify the traumatic and psychological effects of an FP result and the consequent procedures on a normal subject. The cost C^- is the cost of an FN result: The presence of the disease in a patient is not diagnosed, the condition worsens with time, the patient faces more complications of the disease, and the healthcare system or the patient has to bear the costs of further tests and delayed therapy.

A loss factor due to misclassification may be defined as

$$L = FPF \times C^+ + FNF \times C^- \quad (10.110)$$

The total cost of the screening program may be computed as

$$C_S = TPF \times C_A + TNF \times C_N + FPF \times C^+ + FNF \times C^- \quad (10.111)$$

Metz [36] provides more details on the computation of the costs of diagnostic tests.

Actual Group	Predicted Group	
	Without the Disease	With the Disease
Without the Disease	C_N	C^+
With the Disease	C^-	C_A

Table 10.2 Schematic representation of the cost matrix of a diagnostic method.

10.9.1 Receiver operating characteristics

Measures of overall correct classification of patterns as percentages provide limited indications of the accuracy of a diagnostic method. The provision of separate percentage correct classification for each category, such as sensitivity and specificity, can facilitate improved analysis. However, these measures do not indicate the dependence of the results upon the decision threshold. Furthermore, the effect of the rate of incidence or prevalence of the particular disease is not considered.

From another perspective, it is desirable to have a screening or diagnostic test that is both highly sensitive and highly specific. In reality, however, such a test is usually not achievable. Most tests are based on clinical measurements that can assume limited ranges of a variable (or a few variables) with an inherent trade-off between sensitivity and specificity. The relationship between sensitivity and specificity is illustrated by the ROC curve, which facilitates improved analysis of the classification accuracy of a diagnostic method [36–39].

Consider the situation illustrated in Figure 10.7. For a given diagnostic test with the decision variable z , we have predetermined state-conditional PDFs of the decision variable z for actually negative cases indicated as $p(z|N)$ and for actually positive indicated as $p(z|A)$. As indicated in Figure 10.7, the two PDFs will almost always overlap, given that no method can be perfect. The user or operator needs to determine a decision threshold (indicated by the vertical line) so as to strike a compromise between sensitivity and specificity. Lowering the decision threshold will increase TPF at the cost of increased FPF . (Note: TNF and FNF may be derived easily from FPF and TPF , respectively.)

An ROC curve is a graph on a unit square that plots (FPF, TPF) points obtained for a range of decision threshold or cut points of the decision method (see Figure 10.8). The cut point could

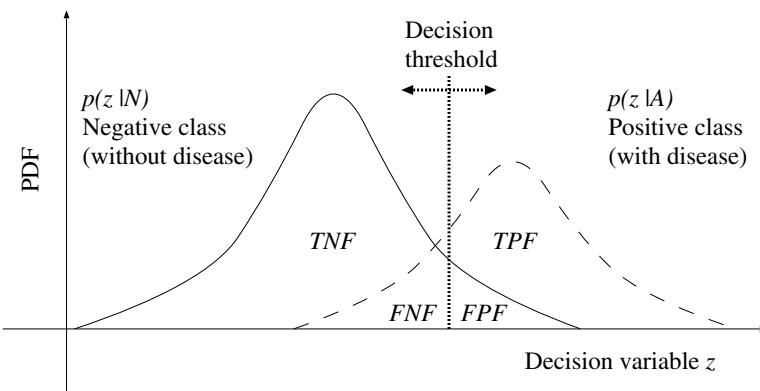


Figure 10.7 State-conditional PDFs of a diagnostic decision variable z for negative and positive cases. The vertical line represents the decision threshold. Based on a similar figure in T.M. Cabral and R.M. Rangayyan, *Fractal Analysis of Breast Masses in Mammograms*, Morgan & Claypool, 2012.

correspond to the threshold of the probability of prediction. By varying the decision threshold, we get different decision fractions, within the range $[0, 1]$. An ROC curve describes the inherent detection (diagnostic or discriminant) characteristics of a test or method: A receiver (user) may choose to operate at any point along the curve. The ROC curve is independent of the prevalence of the disease or disorder being investigated as it is based on normalized decision fractions. As all cases may be simply labeled as negative or all may be labeled as positive, an ROC curve has to pass through the points $(0, 0)$ and $(1, 1)$.

In a diagnostic situation where a human operator or specialist is required to provide the diagnostic decision, ROC analysis is usually conducted by requiring the specialist to rank each case as one of five possibilities [36]:

1. definitely or almost definitely negative,
2. probably negative,
3. possibly positive,
4. probably positive,
5. definitely or almost definitely positive.

Item 3 above may be replaced by “indeterminate,” if appropriate. Various values of TPF and FPF are then calculated by varying the decision threshold from level 5 to level 1 according to the decision items listed above. The resulting (FPF, TPF) points are then plotted to form an ROC curve. The maximum likelihood estimation method [40] is commonly used to fit a binormal curve to data as above.

A summary measure of effectiveness of a test is given by the area under the ROC curve, traditionally labeled as A_z . It is clear from Figure 10.8 that A_z is limited to the range $[0, 1]$. A test that gives a larger area under the ROC curve indicates a better method than one with a smaller area: in Figure 10.8, the method corresponding to ROC3 is better than the method corresponding to ROC2; both are better than the method represented by ROC1 with $A_z = 0.5$. An ideal method will have an ROC curve that follows the vertical line from $(0, 0)$ to $(0, 1)$ and then the horizontal line from $(0, 1)$ to $(1, 1)$, with $A_z = 1$: The method has $TPF = 1$ with $FPF = 0$, which is ideal. (Note: This would require the PDFs represented in Figure 10.7 not to overlap.) The results of several applications presented in the book are quoted in terms of A_z .

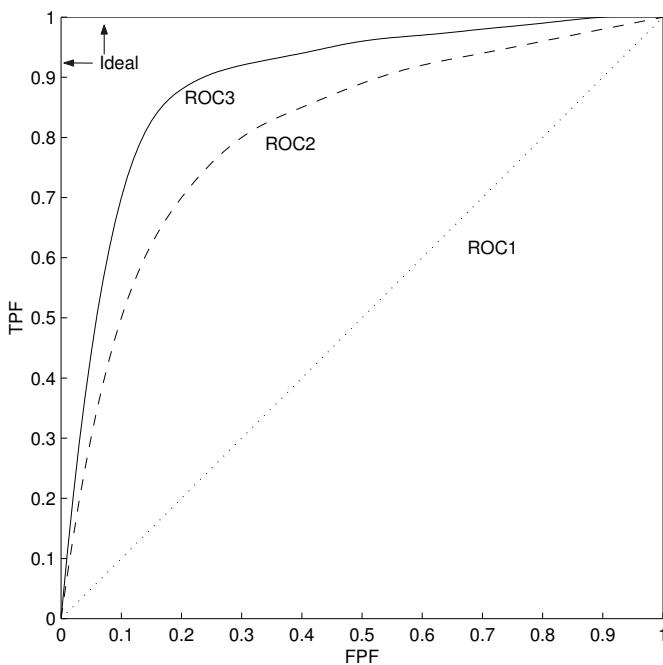


Figure 10.8 Examples of receiver operating characteristic curves.

10.9.2 McNemar's test of symmetry

Problem: Suppose we have two methods to perform a certain diagnostic test. How may we compare the classification performance of one against that of the other?

Solution: Measures of overall classification accuracies such as a percentage of correct classification or the area under the ROC curve provide simple measures to compare two or more diagnostic methods. If more details are required as to how the classifications of groups of cases vary from one method to another, McNemar's test of symmetry [41, 42] would be an appropriate tool.

McNemar's test is based on the construction of contingency tables that compare the results of two classification methods. The rows of a contingency table represent the outcomes of one of the methods used as the reference, possibly a gold standard (labeled as Method A in Table 10.3); the columns represent the outcomes of the other method, which is usually a new method (Method B) to be evaluated against the gold standard. The entries in the table are counts that correspond to particular diagnostic categories, which in Table 10.3 are labeled as normal, indeterminate, and abnormal. A separate contingency table should be prepared for each true category of the patterns; for example, normal and abnormal. (The class "indeterminate" may not be applicable as a true category.) The true category of each case may have to be determined by a third method (for example, biopsy or surgery).

In Table 10.3, the variables a , b , c , d , e , f , g , h , and i denote the counts in each cell, and the numbers in parentheses denote the cell number. The variables $C1$, $C2$, and $C3$ denote the total numbers of counts in the corresponding columns; $R1$, $R2$, and $R3$ denote the total numbers of counts in the corresponding rows. The total number of cases in the true category represented by the table is $N = C1 + C2 + C3 = R1 + R2 + R3$.

Each cell in a contingency table represents a paired outcome. For example, in evaluating the diagnostic efficiency of Method B versus Method A, cell number 3 will contain the number of

		Method B			
Method A		Normal	Indeterminate	Abnormal	Total
Normal		a (1)	b (2)	c (3)	$R1$
Indeterminate		d (4)	e (5)	f (6)	$R2$
Abnormal		g (7)	h (8)	i (9)	$R3$
Total		$C1$	$C2$	$C3$	N

Table 10.3 Schematic representation of a contingency table for McNemar's test of symmetry. Adapted from Krishnan et al. [43] and Krishnan [44].

samples that were classified as normal by Method A but as abnormal by Method B. The row totals ($R1$, $R2$, and $R3$) and the column totals ($C1$, $C2$, and $C3$) may be used to determine the sensitivity and specificity of the methods.

High values along the main diagonal (a, e, i) of a contingency table (see Table 10.3) indicate no change or difference in diagnostic performance with Method B as compared to Method A. In a contingency table for truly abnormal cases, a high value in the upper-right portion (cell number 3) will indicate an improvement in diagnosis (higher sensitivity) with Method B as compared to Method A. In evaluating a contingency table for truly normal cases, Method B will have a higher specificity than Method A if a large value is found in cell 7. McNemar's method may be used to perform detailed statistical analysis of improvement in performance based on contingency tables if large numbers of cases are available in each category [41,42].

Illustration of application: Krishnan et al. [43] proposed two methods for auditory display and sonification of processed VAG signals. Table 10.4 shows the contingency table for 18 abnormal VAG signals comparing their classification by listening to direct playback of the signal data or after a sonification technique based on the *IMF* and envelope of a given VAG signal. The presence of knee-joint pathology was confirmed by arthroscopy in the cases labeled as abnormal. The table shows substantial improvement in sensitivity with the sonification method (15/18) as compared to direct playback (4/18). Ten abnormal signals that were called normal with direct playback were correctly classified as abnormal with the sonification method.

		Sonification			
Direct Playback		Normal	Indeterminate	Abnormal	Total
Normal		2	0	10	12
Indeterminate		1	0	1	2
Abnormal		0	0	4	4
Total		3	0	15	18

Table 10.4 Contingency table for a method of sonification of VAG signals versus direct playback with 18 abnormal signals. Adapted from Krishnan et al. [43] and Krishnan [44].

Table 10.5 shows the contingency table for 19 normal VAG signals comparing their classification by listening to direct playback of the signal data or after sonification. The normal nature of the knee-joint in the cases labeled as normal was confirmed by clinical history and physical examination. The results indicate poorer specificity (8/19) of the results of sonification as compared to direct playback (13/19). The results indicate that the higher sensitivity of the sonification method was gained at the expense of a decrease in specificity. See Krishnan et al. [43] for further details.

		Sonification			
Direct Playback		Normal	Indeterminate	Abnormal	Total
Normal	8	2	3	13	
Indeterminate	0	0	1	1	
Abnormal	0	0	5	5	
Total	8	2	9	19	

Table 10.5 Contingency table for a method of sonification of VAG signals versus direct playback with 19 normal signals. Adapted from Krishnan et al. [43] and Krishnan [44].

10.10 Reliability of Features, Classifiers, and Decisions

In most practical applications of biomedical signal analysis, the researcher is presented with the problem of designing a system for pattern classification and decision making using a small number of training samples (signals), with no knowledge of the distributions of the features or parameters computed from the signals. The size of the training set, relative to the number of features used in the pattern classification system, determines the accuracy and reliability of the decisions made [45, 46]. One should not increase the number of features to be used without a simultaneous increase in the number of training samples, as the two quantities together affect the bias and variance of the classifier. On the other hand, when the training set has a fixed number of samples, the addition of more features beyond a certain limit will lead to poorer performance of the classifier: this is known as the “curse of dimensionality.” It is desirable to be able to analyze the bias and variance of a classification rule while isolating the effects of the functional form of the distributions of the features used.

Raudys and Jain [46] give a rule-of-thumb table for the number of training samples required in relation to the number of features used in order to remain within certain limits of classification errors for five pattern classification methods. When the available features are ordered in terms of their individual classification performance, the optimal number of features to be used with a certain classification method and training set may be determined by obtaining unbiased estimates of the classification accuracy with the number of features increased one at a time in order. A point is reached when the performance deteriorates, which will indicate the optimal number of features to be used. This method, however, cannot take into account the joint performance of various combinations of features: Exhaustive combinations of all features may have to be evaluated to take this aspect into consideration.

Durand et al. [47] reported on the design and evaluation of several pattern classification systems for the assessment of bioprosthetic heart valves based on 18 features computed from PCG spectra (see Section 6.5). Based on the rule of thumb that the number of training samples should be five or more times the number of features used, and with the number of training samples limited to data from 20 normal and 28 degenerated valves, exhaustive combinations of the 18 features taken 2, 3, 4, 5, and 6 at a time were used to design and evaluate pattern classification systems. The Bayes method was seen to provide the best performance (98% correct classification) with six features; as many as 511 combinations of the 18 features taken six at a time provided correct classification between 90% and 98%. The nearest-neighbor algorithm with the Mahalanobis distance provided 94% correct classification with only three features and did not perform any better with more features.

10.10.1 Separability of features

Normalized distance between PDFs: Consider a feature x that has the means m_1 and m_2 and SD values σ_1 and σ_2 for the two classes C_1 and C_2 . Assume that the PDFs $p(x|C_1)$ and $p(x|C_2)$ overlap; then, the area of overlap is related to the error of classification; see Figure 10.7. If the SD values σ_1 and σ_2 remain constant, the overlap between the PDFs decreases as $|m_1 - m_2|$ increases. If the means remain constant, the overlap increases as σ_1 and σ_2 increase (the dispersion of the features increases). These observations are captured by the normalized distance between the means, defined as [48]

$$d_n = \frac{|m_1 - m_2|}{\sigma_1 + \sigma_2}. \quad (10.112)$$

The measure d_n provides an indicator of the statistical separability of the PDFs. A limitation of d_n is that $d_n = 0$ if $m_1 = m_2$ regardless of σ_1 and σ_2 .

See Section 5.12.3 for a discussion on the KLD .

Divergence: Let us rewrite the likelihood ratio in Equation 10.54 as

$$l_{ij}(\mathbf{x}) = \frac{p(\mathbf{x}|C_i)}{p(\mathbf{x}|C_j)}. \quad (10.113)$$

Applying the logarithm, we get

$$l'_{ij}(\mathbf{x}) = \ln[l_{ij}(\mathbf{x})] = \ln[p(\mathbf{x}|C_i)] - \ln[p(\mathbf{x}|C_j)]. \quad (10.114)$$

The divergence D_{ij} between the PDFs $p(\mathbf{x}|C_i)$ and $p(\mathbf{x}|C_j)$ is defined as [48]

$$D_{ij} = E[l'_{ij}(\mathbf{x})|C_i] + E[l'_{ji}(\mathbf{x})|C_j], \quad (10.115)$$

where

$$E[l'_{ij}(\mathbf{x})|C_i] = \int_{\mathbf{x}} l'_{ij}(\mathbf{x}) p(\mathbf{x}|C_i) d\mathbf{x}. \quad (10.116)$$

Divergence has the following properties [48]:

- $D_{ij} > 0$;
- $D_{ii} = 0$;
- $D_{ij} = D_{ji}$; and
- if the individual features x_1, x_2, \dots, x_n are statistically independent, then $D_{ij}(x_1, x_2, \dots, x_n) = \sum_{k=1}^n D_{ij}(x_k)$.

It follows that adding more features that are statistically independent of one another will increase divergence and statistical separability.

In the case of multivariate Gaussian PDFs, we have [48]

$$\begin{aligned} D_{ij} &= \frac{1}{2} \operatorname{Tr}[(\mathbf{C}_i - \mathbf{C}_j)(\mathbf{C}_j^{-1} - \mathbf{C}_i^{-1})] \\ &+ \frac{1}{2} \operatorname{Tr}[(\mathbf{C}_i^{-1} + \mathbf{C}_j^{-1})(\mathbf{m}_i - \mathbf{m}_j)(\mathbf{m}_i - \mathbf{m}_j)^T]. \end{aligned} \quad (10.117)$$

The second term in the equation given above is similar to the normalized distance d_n as defined in Equation 10.112 and is zero for PDFs with identical means; however, due to the first term, $D_{ij} \neq 0$ unless the covariance matrices are identical.

In the case of the existence of multiple classes $C_i, i = 1, 2, \dots, m$, the pairwise divergence values may be averaged to obtain a single measure across all of the m classes as [48]

$$D_{\text{av}} = \sum_{i=1}^m \sum_{j=1}^m p(C_i) p(C_j) D_{ij}. \quad (10.118)$$

A limitation of both d_n and D_{ij} is that they increase without an upper bound as the separation between the means increases. On the other hand, the error of classification is limited to the range $[0, 100]\%$ or $[0, 1]$.

Jeffries–Matusita distance: The Jeffries–Matusita (JM) distance provides an improved measure of the separation between PDFs as compared to the normalized distance and divergence. The JM distance between the PDFs $p(\mathbf{x}|C_i)$ and $p(\mathbf{x}|C_j)$ is defined as [48]

$$J_{ij} = \left\{ \int_{\mathbf{x}} \left[\sqrt{p(\mathbf{x}|C_i)} - \sqrt{p(\mathbf{x}|C_j)} \right]^2 d\mathbf{x} \right\}^{1/2}. \quad (10.119)$$

In the case of multivariate Gaussian PDFs, we have [48]

$$J_{ij} = \sqrt{2[1 - \exp(-\alpha)]}, \quad (10.120)$$

where

$$\begin{aligned} \alpha &= \frac{1}{8} (\mathbf{m}_i - \mathbf{m}_j)^T \left(\frac{\mathbf{C}_i + \mathbf{C}_j}{2} \right)^{-1} (\mathbf{m}_i - \mathbf{m}_j) \\ &\quad + \frac{1}{2} \ln \left[\frac{|(\mathbf{C}_i + \mathbf{C}_j)/2|}{(|\mathbf{C}_i| |\mathbf{C}_j|)^{1/2}} \right], \end{aligned} \quad (10.121)$$

where $|\mathbf{C}_i|$ is the determinant of \mathbf{C}_i .

An advantage of the JM distance is that it is limited to the range $[0, \sqrt{2}]$. $J_{ij} = 0$ when the means of the PDFs are equal and the covariance matrices are zero matrices. Pairwise JM distances may be averaged over multiple classes similar to the averaging of divergence as in Equation 10.118. The JM distance determines the upper and lower bounds on the error of classification [48].

It should be observed that divergence and JM distance are defined for a given feature vector \mathbf{x} . The measures would have to be computed for all combinations of features in order to select the best feature set for a particular pattern classification problem.

See Section 5.12.3 for a description of the *KLD* between PDFs.

The *t*-test and the *p*-value: The *t*-test is used to assess whether the means of a feature, x , for two groups or classes are statistically different from each other [18, 49–51]. The test is performed by assessing the difference between the means of the two groups of the features relative to their spread or variability. If it is assumed that the variance, σ^2 , is the same for the two groups, the *t*-statistic, t_s , is computed as

$$t_s = \frac{\mu_1 - \mu_2}{\sigma_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad (10.122)$$

where μ_1 and n_1 are the mean and size of the first group, μ_2 and n_2 are the mean and size of the second group, and σ_p is the pooled *SD*, defined as

$$\sigma_p = \sqrt{\frac{(n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2}{n_1 + n_2 - 2}}. \quad (10.123)$$

If the variance is not the same for the two groups, t_s is computed as

$$t_s = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \quad (10.124)$$

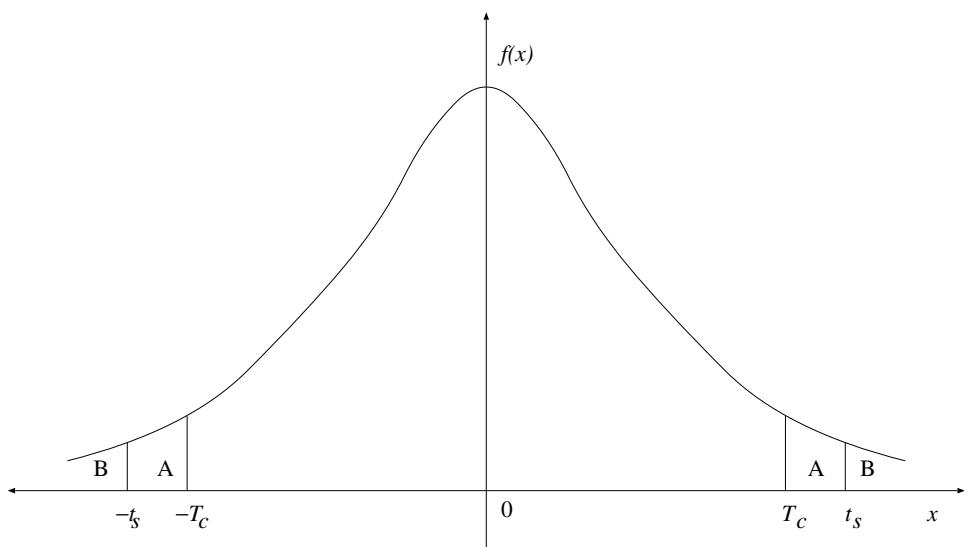


Figure 10.9 The t -distribution curve. Reproduced with permission from T.M. Cabral and R.M. Rangayyan, *Fractal Analysis of Breast Masses in Mammograms*, Morgan & Claypool, 2012. ©Morgan & Claypool.

where σ_1^2 is the variance of the first group and σ_2^2 is the variance of the second group. The number of degrees of freedom is equal to $n_1 + n_2 - 2$.

The significance level is usually set at 0.05; this means that, five times out of 100, the difference between the means is concluded to be statistically significant when it is not actually so. This is the probability of being incorrect if the null hypothesis (which assumes that $\mu_1 = \mu_2$) is rejected. Given the degrees of freedom and significance level, the T-critical value, T_c , can be determined using the t -distribution look-up table. When t_s , computed with either Equation 10.124 or Equation 10.122, is larger than T_c , the null hypothesis is rejected.

The t -distribution curve shown in Figure 10.9 illustrates the significance level and p -value in terms of the area under the t -distribution curve. The significance level is the sum of the area under the curve to the right of $+T_c$ and the area to the left of $-T_c$ (all of the areas labeled as A and B in Figure 10.9). The p -value is the sum of the area under the curve to the right of $+t_s$ and the area to the left of $-t_s$ (all of the areas labeled as B in Figure 10.9).

The p -value is a statistical measure of the probability that the results observed in a study could have occurred by chance. A small p -value is a rejection of the null hypothesis in favor of the alternative hypothesis (which assumes that $\mu_1 \neq \mu_2$) because it indicates how unlikely it is that a test statistic as extreme as or more extreme than that given by the data will be observed from the population if the null hypothesis were true. For example, a p -value of 0.01 indicates that there is a one out of a hundred chance that the result occurred by chance. A test resulting in a p -value less than 0.05 is considered to indicate that the difference between the means is statistically significant. A p -value less than 0.01 is taken to indicate that the difference between the means is statistically highly significant.

10.10.2 Feature selection

In a practical application of pattern classification, several parameters may be required to discriminate between multiple classes. Because most features lead to limited discrimination between classes due to the overlap in the ranges of their values for the various classes, it is common to use several features. However, various costs are associated with the measurement, derivation, and computation

of each feature. It would be advantageous to be able to assess the contribution of each feature to the task of discriminating between the classes of interest, and to select the feature set that provides the best separation between classes and the lowest classification error. The notion of statistical separability of features between classes is useful in addressing these concerns [48]. Various parameters related to the separation of features between classes may be used to select features, such as A_z [36, 37, 39], JM distance, and the p -value [51].

Popular methods for feature selection [2, 52–54] include sequential forward selection, sequential backward selection, and stepwise regression. In sequential forward selection, starting with an empty set, features are sequentially included in the set until the inclusion of additional features does not improve the classification performance. To begin, the best single feature in terms of separation of the classes of interest, such as the p -value, is determined and included in the set of selected features. The selected feature is then used in combination with each remaining feature, in turn, to create pairs. The classification performance of each pair is evaluated and the best pair is selected for the next iteration. The process of including more and more features and evaluation of performance is continued until a prespecified number of features is selected or the classification performance does not improve much with the inclusion of more features.

Sequential backward selection starts with the full set of features and eliminates features until the removal of more features deteriorates the classification performance. To begin with, the feature whose elimination leads to the largest separation of classes is identified and removed. The process is continued by removing one feature at a time until a prespecified number of features remain or any further removal of features causes deterioration in the classification performance.

Stepwise regression [54, 55] may be seen as a variant of forward selection. At each stage, a new feature is added and the obtained model is checked for any possible elimination of the selected features without substantially increasing an error measure. The process is continued until the error is minimized or the improvement achieved is below a limit. The method starts with an initial model and then evaluates the discriminating power of incrementally larger and smaller models at each iteration [54, 55]. At each iteration, the p -values of an F -statistic are computed to evaluate the models with and without a potential feature. The feature is included if sufficient evidence is found to support its inclusion; otherwise, it is removed.

When a specific pattern classification method and its performance measures related to the selected features are integrated into the feature selection process, the procedure is referred to as a wrapper method [56, 57]. The classifier is an integral part of the process of feature selection and provides the performance metrics to select the best set of features.

Advanced methods for feature selection include genetic algorithms and genetic programming [20, 58]. Instead of selecting a subset of a given set of features, one could also take the approach of reducing the dimension of the given feature vector via PCA (see Section 9.7.1). PCA helps in removing redundancy in the given features. It should be noted that the result of PCA does not include any of the given features directly but provides transformed values derived from the given feature vector. The transform itself is derived by using statistical measures of a population of feature vectors.

In general, it is expected that selecting a reduced set of features as above will lead to higher classification accuracy than the case with the full set of features. A smaller set of features will also facilitate the design and implementation of more efficient classifiers at lower cost than a larger set. See Banik et al. [54, 59], Nandi et al. [58], and Mu et al. [20, 60] for additional discussions on feature selection and examples of application.

10.10.3 The training and test steps

In the situation when the number of available sample vectors with known classification is limited, questions arise as to how many of the samples may be used to design or train a classifier, with the understanding that the classifier so designed needs to be tested using an independent set of samples of known classification as well. (See Sections 9.5 and 9.8 for related discussions.) When a suffi-

ciently large number of samples are available, they may be randomly split into two approximately equal sets, one for use as the training set and the other to be used as the test set. The random splitting procedure may be repeated a number of times to generate several classifiers. Finally, one of the classifiers so designed may be selected based on its performance in both the training and test steps.

When the available number of labeled samples is small, the bootstrap method [61, 62] may be used. In this procedure, a training set is created by drawing at random a large number of samples from the available pool of labeled samples with replacement; that is, a given sample may be drawn and included a number of times in the training set. Similarly, a test set is created by random drawing of samples with replacement. The resampling procedure may be repeated a number of times to obtain multiple estimates of various measures of classification performance.

In the procedure known as k -fold cross-validation, the original labeled dataset is partitioned randomly into k subsets of equal size with no overlap. For example, in 10-fold cross-validation with $k = 10$, the given set of samples is randomly divided into 10 subsets with no overlap. Out of the k subsets, one subset is set aside to test the classifier being designed, and the remaining $(k - 1)$ subsets are used together as the training data. The training and testing process, also known as cross-validation, is repeated k times, with each of the k subsets used only once as the test data. The k results from the k -fold cross-validation process may be averaged to obtain an overall measure of performance of the features used and classification method chosen. In k -fold cross-validation, all data samples are used for both training and testing, and each sample is used only once for testing; this cannot be guaranteed in the random resampling procedure described in the preceding paragraphs.

Another option used to split the available dataset in the development of a classifier is to create three subsets: training, validation, and test sets. The validation set, which is independent of the training set, is used to decide when to stop training, validate the classifier designed, and tune the related parameters and configurations. The validation step is expected to reduce the possibility of overfitting the classifier to the training set and failure to generalize to new data.

When a dataset with multiple classes of samples is partitioned as described above, it is important to ensure that all classes are equally or adequately represented in each partition.

The leave-one-out method: The LOO method [3] is suitable for the estimation of the classification accuracy of a pattern classification technique, particularly when the number of available samples is small. In this method, one of the available samples is excluded, the classifier is designed with the remaining samples, and then the classifier is applied to the excluded sample. The validity of the classification so performed is noted. This procedure is repeated with each available sample: if N training samples are available, N classifiers are designed and tested. The training and test sets for any one classifier so designed and tested are independent. However, whereas the training set for each classifier has $N - 1$ samples, the test set has only one sample. This procedure is equivalent to k -fold cross-validation with $k = N$. In the final analysis, every sample will have served $(N - 1)$ times as a training sample but only once as a test sample. An average classification accuracy is then computed using all of the test results.

Let us consider a simple case in which the covariances of the sample sets of two classes are equal. Assume that two sample sets, $S_1 = \{\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, \dots, \mathbf{x}_{N_1}^{(1)}\}$ and $S_2 = \{\mathbf{x}_1^{(2)}, \mathbf{x}_2^{(2)}, \dots, \mathbf{x}_{N_2}^{(2)}\}$ from classes C_1 and C_2 , respectively, are given. Here, N_1 and N_2 are the numbers of samples in the sets S_1 and S_2 , respectively. Assume also that the prior probabilities of the two classes are equal to each other. Then, according to the Bayes classifier and assuming \mathbf{x} to be governed by a multivariate Gaussian PDF, a sample \mathbf{x} is assigned to class C_1 if

$$(\mathbf{x} - \mathbf{m}_1)^T(\mathbf{x} - \mathbf{m}_1) - (\mathbf{x} - \mathbf{m}_2)^T(\mathbf{x} - \mathbf{m}_2) > \theta, \quad (10.125)$$

where θ is a threshold, and the sample mean $\tilde{\mathbf{m}}_i$ is given by

$$\tilde{\mathbf{m}}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{x}_j^{(i)}. \quad (10.126)$$

In the LOO method, one sample $\mathbf{x}_k^{(i)}$ is excluded from the training set and then used as the test sample. The mean estimate for class C_i without $\mathbf{x}_k^{(i)}$, labeled as $\tilde{\mathbf{m}}_{ik}$, may be computed as

$$\tilde{\mathbf{m}}_{ik} = \frac{1}{N_i - 1} \left[\sum_{j=1}^{N_i} \mathbf{x}_j^{(i)} - \mathbf{x}_k^{(i)} \right], \quad (10.127)$$

which leads to

$$\mathbf{x}_k^{(i)} - \tilde{\mathbf{m}}_{ik} = \frac{N_i}{N_i - 1} (\mathbf{x}_k^{(i)} - \tilde{\mathbf{m}}_i). \quad (10.128)$$

Then, testing a sample $\mathbf{x}_k^{(1)}$ from C_1 can be carried out as

$$\begin{aligned} & (\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_{1k})^T (\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_{1k}) - (\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_2)^T (\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_2) \\ &= \left(\frac{N_1}{N_1 - 1} \right)^2 (\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_1)^T (\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_1) - (\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_2)^T (\mathbf{x}_k^{(1)} - \tilde{\mathbf{m}}_2) > \theta. \end{aligned} \quad (10.129)$$

Note that when $\mathbf{x}_k^{(1)}$ is tested, only $\tilde{\mathbf{m}}_1$ is changed and $\tilde{\mathbf{m}}_2$ is not changed. Likewise, when a sample $\mathbf{x}_k^{(2)}$ from C_2 is tested, the decision rule is

$$\begin{aligned} & (\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_1)^T (\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_1) - (\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_{2k})^T (\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_{2k}) \\ &= (\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_1)^T (\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_1) - \left(\frac{N_2}{N_2 - 1} \right)^2 (\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_2)^T (\mathbf{x}_k^{(2)} - \tilde{\mathbf{m}}_2) < \theta. \end{aligned} \quad (10.130)$$

The LOO method provides the least biased (practically unbiased) estimate of the classification accuracy of a given classification method for a given training set; it is useful when the number of samples available with known classification is small. The LOO approach may be applied on a sample (signal or feature vector) basis or on a subject basis if multiple signals are included in the dataset for each subject.

10.11 Application: Normal versus Ectopic ECG Beats

We have seen the distinctions between normal and ectopic (or PVC) beats in the ECG in several different contexts (see Sections 1.2.5, 5.4.2, 5.7, and 10.2.2, as well as Figures 5.1 and 5.12). We shall now see how we can put together several of the topics we have studied so far for the purpose of detecting PVCs in an ECG signal.

10.11.1 Classification with a linear discriminant function

Training step: Figure 10.10 shows the ECG signal of a patient with several ectopic beats, including episodes of bigeminy (alternating normal beats and PVCs). The beats in the portion of the signal in Figure 10.10 were manually labeled as normals (“o” marks) or PVCs (“x” marks) and were used to train a pattern classification system. The training set includes 121 normal beats and 39 PVCs.

The following procedure was applied to the signal to detect each beat, compute features, and develop a pattern classification rule:

1. The signal was filtered with a Butterworth lowpass filter of order 8 and cutoff frequency 70 Hz to remove noise (see Section 3.7.1); the sampling rate is 200 Hz.
2. The Pan–Tompkins algorithm was applied to detect each beat (see Section 4.3.2).

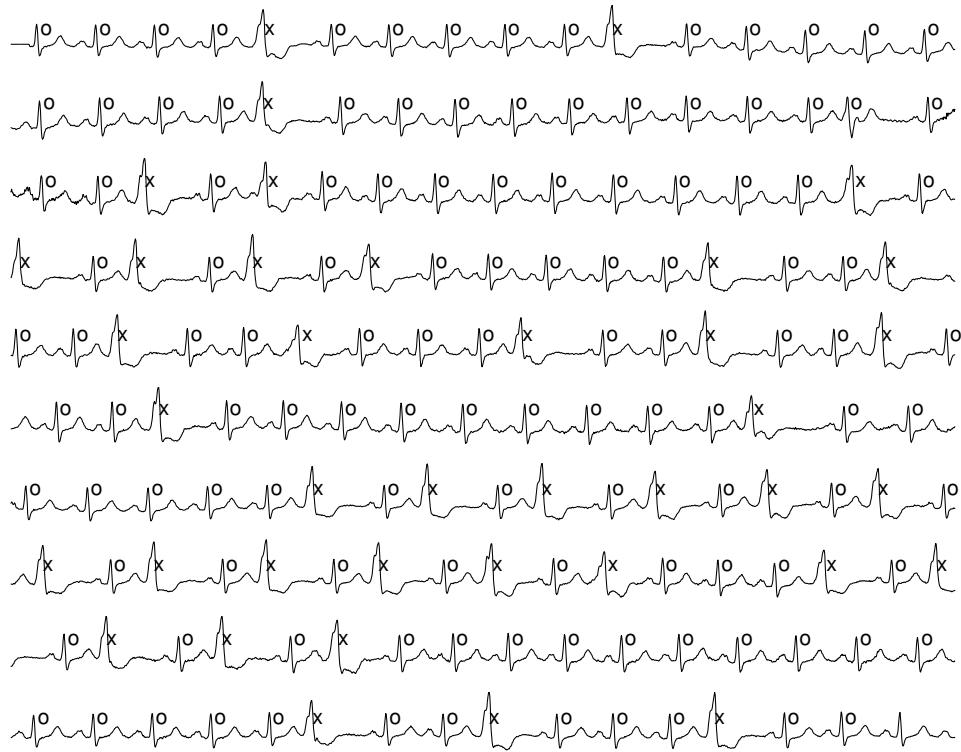


Figure 10.10 The ECG signal of a patient (male, 65 years) with PVCs (training set). Each strip is of duration 10 s; the signal continues from top to bottom. The second half of the seventh strip and the first half of the eighth strip illustrate an episode of bigeminy. Each beat was manually labeled as normal (“o”) or PVC (“x”). The last beat was not processed.

3. The QRS – T portion of each beat was segmented by selecting the interval from the sample 160 ms before the peak of the Pan–Tompkins output to the sample 240 ms after the peak (see Figure 5.12).
4. The RR interval and form factor FF were computed for each beat (see Sections 5.6.4 and 5.7 and Figure 5.12). Figure 10.11 (a) illustrates the feature vector plot for the training set.
5. The prototype (mean) feature vectors were computed for the normal and PVC groups in the training set. The prototype vectors are $[RR, FF] = [0.66, 1.58]$ and $[RR, FF] = [0.45, 2.74]$ for the normal and PVC classes, respectively.
6. The equations of the straight line joining the two prototype vectors and its normal bisector were determined; the latter is a linear decision function (see Section 10.4.1 and Figure 10.1). Figure 10.11 illustrates the two lines.
7. The equation of the linear decision function was obtained as $RR - 5.56FF + 11.44 = 0$. The decision rule was derived as

$$\text{if } RR - 5.56FF + 11.44 \left\{ \begin{array}{ll} > 0 & \text{normal beat,} \\ \leq 0 & \text{PVC.} \end{array} \right. \quad (10.131)$$

All of the beats in the training set were correctly classified by the decision rule in Equation 10.131.

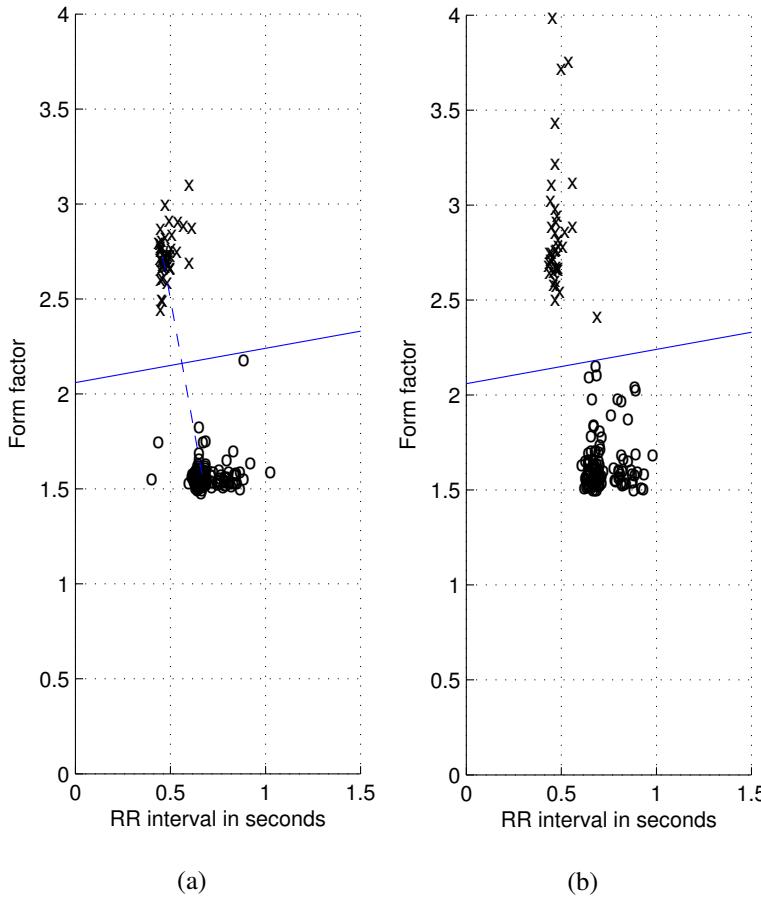


Figure 10.11 (a) The $[RR, FF]$ feature-vector space corresponding to the ECG in Figure 10.10 (training set). Normal: “ \circ ”, PVC: “ \times ”. The straight line joining the two prototype vectors (dashed) and its normal bisector (solid) are also shown; the latter is a linear decision function. (b) $[RR, FF]$ feature-vector space corresponding to the ECG in Figure 10.12 (test set). The straight line is the linear decision function given in Equation 10.131. The “ \times ” mark closest to the decision boundary with $[RR, FF] = [0.66, 2.42]$ corresponds to an FP classification.

Observe from Figure 10.11 (a) that a simple threshold on FF alone can effectively separate the PVCs from the normals in the training set. A viable classification rule to detect PVCs may also be stated in a manner similar to that in Section 10.2.2. The example given here is intended to serve as a simple illustration of the design of a 2D linear decision function.

Test step: Figure 10.12 illustrates an ECG segment immediately following that in Figure 10.10. The same procedure as described above was applied to detect the beats in the signal in Figure 10.12 and to compute their features, which were used as the test set. The decision rule in Equation 10.131 was applied to the feature vectors and the beats in the signal were automatically classified as normal or PVC. Figure 10.11 (b) illustrates the feature-vector space of the beats in the test set, along with the decision boundary given by Equation 10.131. Figure 10.12 shows the automatically applied labels for each beat: All of the 37 PVCs were correctly classified, and only one of the 120 normal beats was misclassified as a PVC (that is, there was one FP).

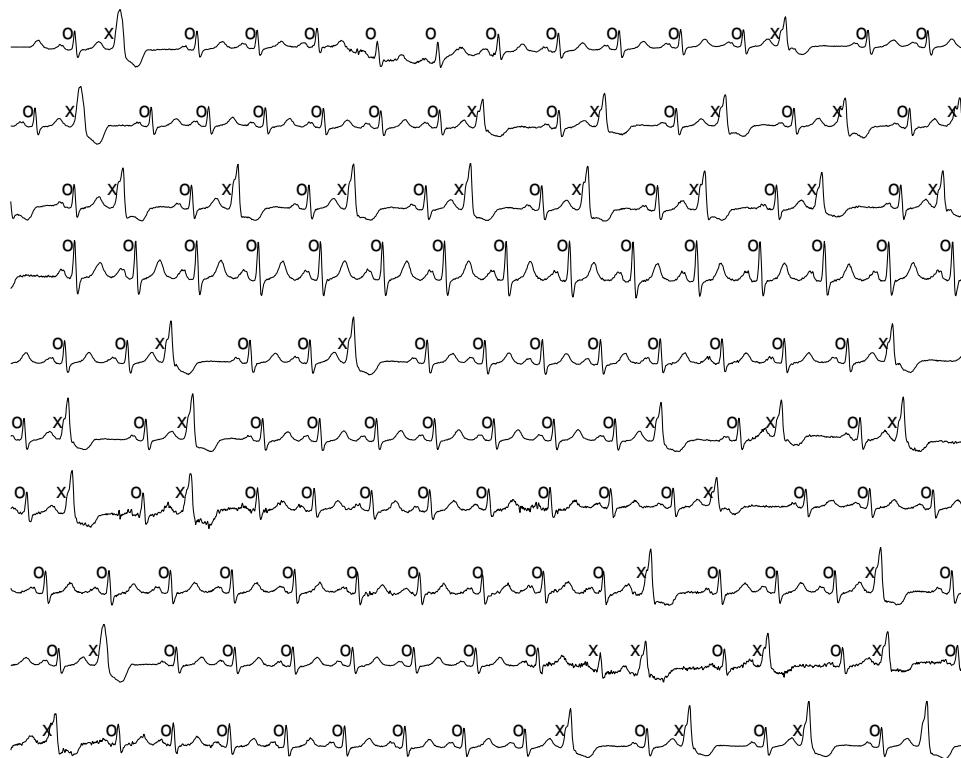


Figure 10.12 The ECG signal of a patient with PVCs (test set); this portion immediately follows that in Figure 10.10. Each strip is of duration 10 s; the signal continues from top to bottom. Each beat was automatically labeled as normal (“o”) or PVC (“x”) by the decision rule stated in Equation 10.131. The 10th beat in the 9th strip with $[RR, FF] = [0.66, 2.42]$ was misclassified. The last beat was not processed.

It should be observed that a PVC has, by definition, an *RR* interval that is less than that for a normal beat (at the same heart rate). However, the heart rate of a subject will vary over time, and the reference *RR* interval to determine the prematurity of PVCs needs to be updated periodically. A decision rule as in Equation 10.131 cannot be applied on a continuing basis even to the same subject. Note that the proposed method can be extended for the identification of sinus beats (originating from the SA node) that meet the prematurity condition due to sinus arrhythmia but are, nevertheless, normal in waveshape.

The *FF* values will depend upon the waveshape of each ECG beat, which will vary from one ECG lead to another. Therefore, the same decision rule based on waveshape cannot be applied to all ECG leads of even the same subject. Furthermore, a given subject may have PVCs originating from various ectopic foci resulting in widely different waveshapes even in the same ECG lead. A shape factor to be used for pattern classification must be capable of maintaining different values for (a) PVCs of various waveshapes as one group and (b) normal beats as another group.

The preceding illustration is intended to serve as a simple example of the design of a pattern classification system; in practice, more complex decision rules based on more than two features are required. Furthermore, it should be observed that a pattern classification procedure as described above provides beat-by-beat labeling; the overall diagnosis of the patient’s condition requires many other items of clinical information and the expertise of a cardiologist.

10.11.2 Application of the Bayes classifier

In another classification experiment with the same ECG signal as in the preceding section, the waveform for each beat was segmented using the Pan–Tompkins algorithm. Each signal was normalized by subtracting its mean and dividing by the maximum value of the result. The area under each segmented and normalized QRS–T wave, referred to as $QRSTA$, was computed as follows (see Section 5.4.3). The baseline value of each beat was obtained as the average of its starting and ending sample values. The baseline was subtracted and the signal was rectified, that is, its absolute value was obtained. $QRSTA$ was computed as the sum of all of the rectified values multiplied by the sampling interval. The feature vector $[QRSTA, FF]^T$ was computed for each beat.

The mean and SD of each feature were computed using the samples in the training set with 121 normal beats and 39 PVCs. Each feature value was normalized by dividing by its SD . Using the training set only, the 1D conditional and posterior probability functions were estimated. Figures 10.13 (a) and 10.14 (a) show the estimated functions using equal prior probabilities for the normal and PVC classes. In a real ECG signal, the number of PVCs could be expected to be far lower than the number of normal beats. Figures 10.13 (b) and 10.14 (b) show the estimated functions using prior probabilities of 0.999 and 0.001 for the normal and PVC classes, respectively. The figures listed above show how the features vary in their probabilities across the two classes and also how the assumed prior probabilities can affect the posterior probabilities derived.

The 2D scatter of the feature vectors $[QRSTA, FF]^T$ for the training set is shown in Figure 10.15 (a). Assuming the prior probabilities for the two classes to be equal, the Bayesian classifier was derived (see Section 10.6.1). Ellipses are shown for each cluster (normal or PVC) to show the boundaries of the 2D Gaussian functions estimated at σ , 2σ , and 3σ in Figure 10.15 (a); the thick black contour shown is the Bayesian decision boundary, which misclassifies only one normal beat as a PVC.

Figure 10.15 (b) demonstrates the application of the Bayesian decision function shown in Figure 10.15 (a) to the test set of ECG signals with 183 normal beats and 53 PVCs. It is seen that the classifier correctly recognizes all PVCs but fails to identify a small number of normal beats. However, inspection of the two scatter plots indicates that the samples are linearly separable; a linear classifier without the need to assume prior probabilities could possibly lead to better results.

10.11.3 Classification using the K -means method

The K -means clustering method (see Section 10.5.1) was also applied to the ECG signal described in the preceding sections. Figure 10.16 shows the evolution of class means and the separating boundary over four iterations. It is seen that, after the fourth and final iteration, all of the samples are correctly classified.

10.12 Application: Detection of Knee-joint Cartilage Pathology

Moussavi et al. [63], Krishnan et al. [64], and Rangayyan et al. [65] proposed a series of adaptive segmentation, modeling, and pattern classification techniques for the detection of knee-joint cartilage pathology using VAG signals (see Sections 1.2.14, 5.12, 6.6, 8.2.3, 8.6, and 9.9). In consideration of the fact that VAG signals are nonstationary, each VAG signal was first divided into locally stationary segments using the RLS or the RLSL algorithm (see Sections 8.6.1 and 8.6.2). Each segment was considered as a separate signal and modeled by the forward–backward linear prediction or the Burg-lattice method (see Section 8.6.2). The model coefficients or poles were used as parameters for pattern classification.

A striking difference that may be appreciated visually and aurally between normal and abnormal VAG signals is that abnormal signals are much more variable in amplitude across a swing cycle than normal signals. However, this difference is lost in the process of dividing the signals into

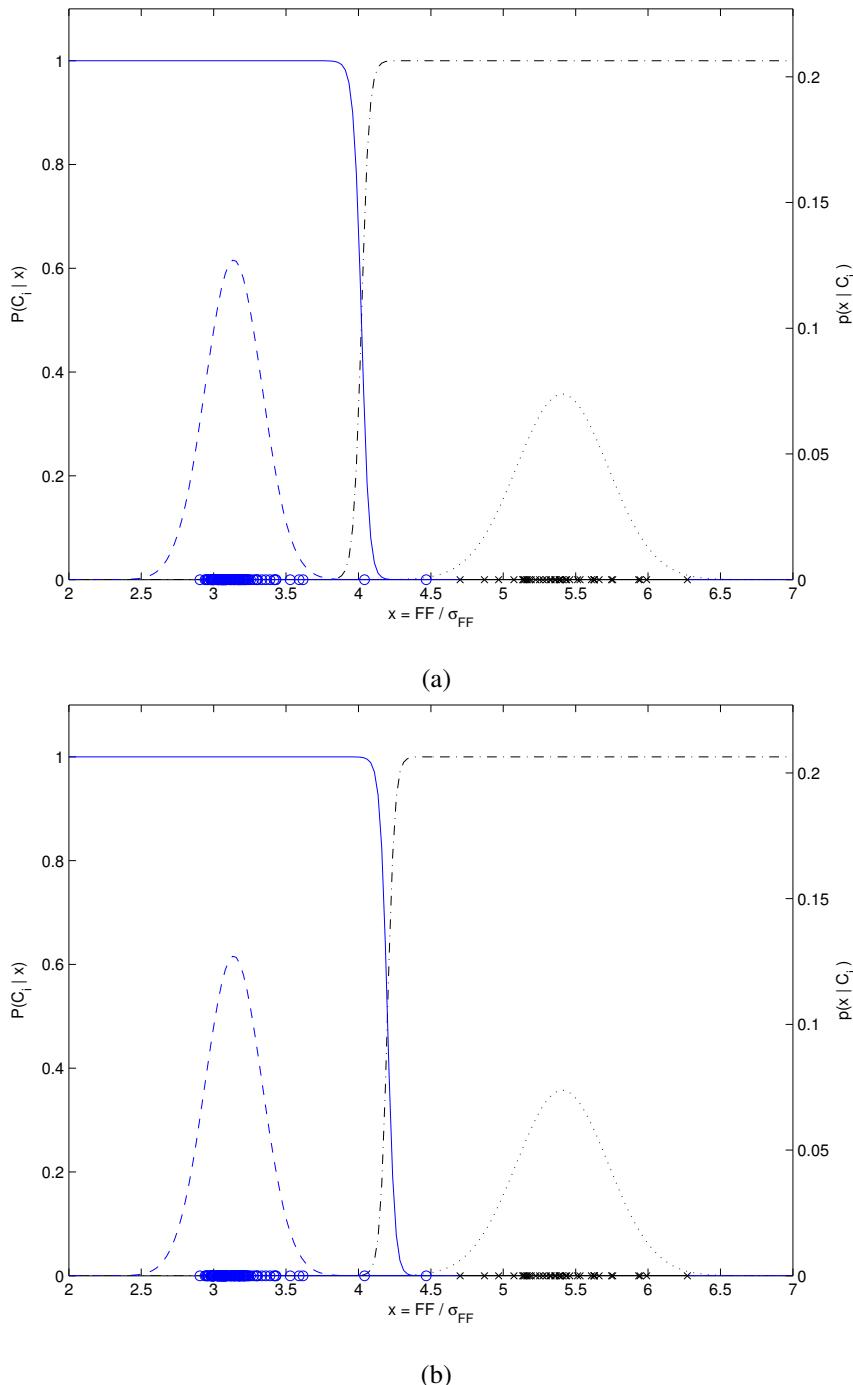


Figure 10.13 Conditional and posterior probability functions estimated for the feature FF for the training set of normal ECG signals (“ \circ ”) and PVCs (“ \times ”). (a) The prior probabilities using the two classes were assumed to be equal. (b) The prior probabilities of 0.999 and 0.001 were assumed for the normal and PVC classes, respectively. Figure courtesy of Fábio José Ayres, University of Calgary, Calgary, Alberta, Canada.

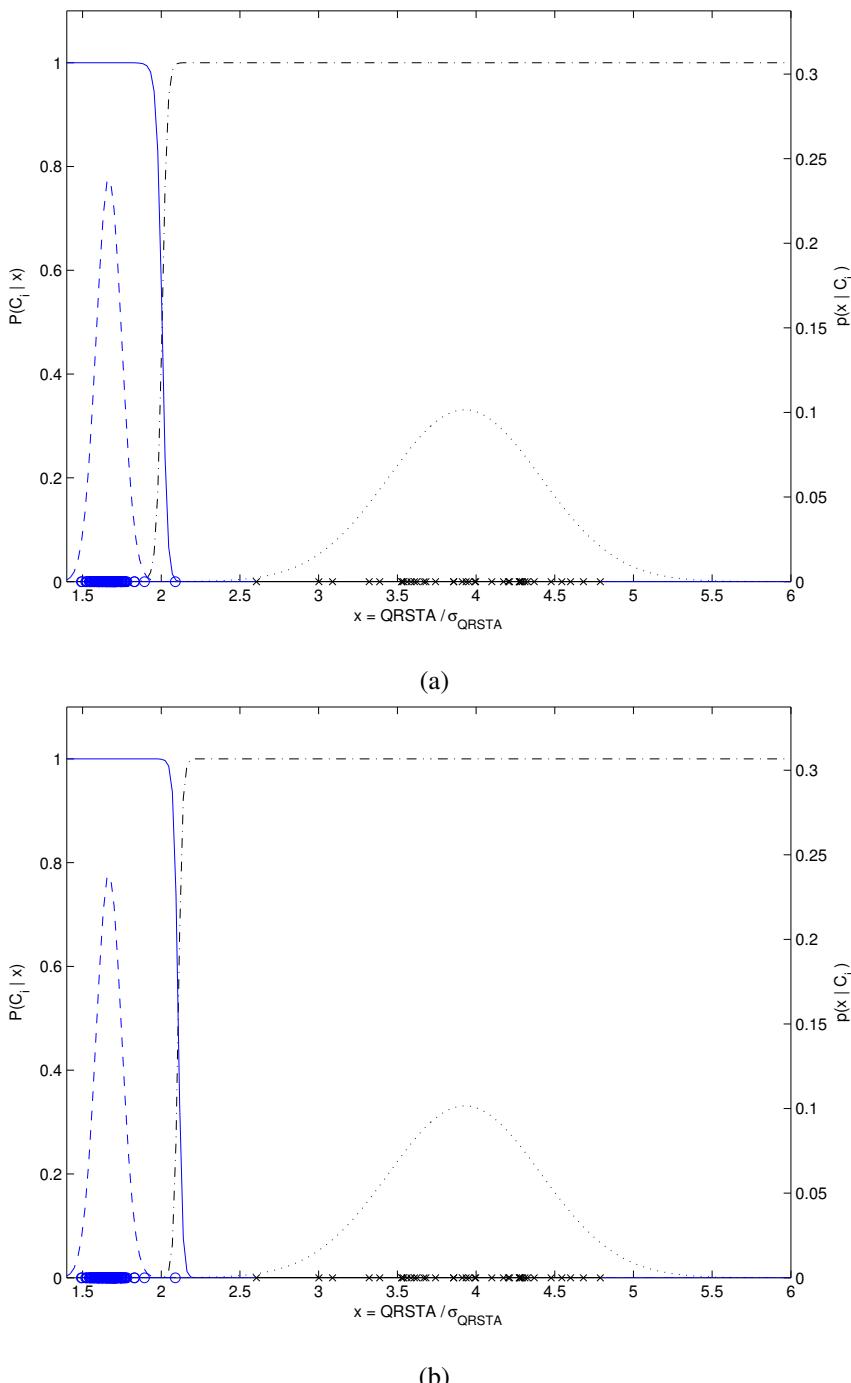


Figure 10.14 Conditional and posterior probability functions estimated for the feature $QRSTA$ using the training set of normal ECG signals (“ \circ ”) and PVCs (“ \times ”). (a) The prior probabilities for the two classes were assumed to be equal. (b) The prior probabilities of 0.999 and 0.001 were assumed for the normal and PVC classes, respectively. Figure courtesy of Fábio José Ayres, University of Calgary, Calgary, Alberta, Canada.

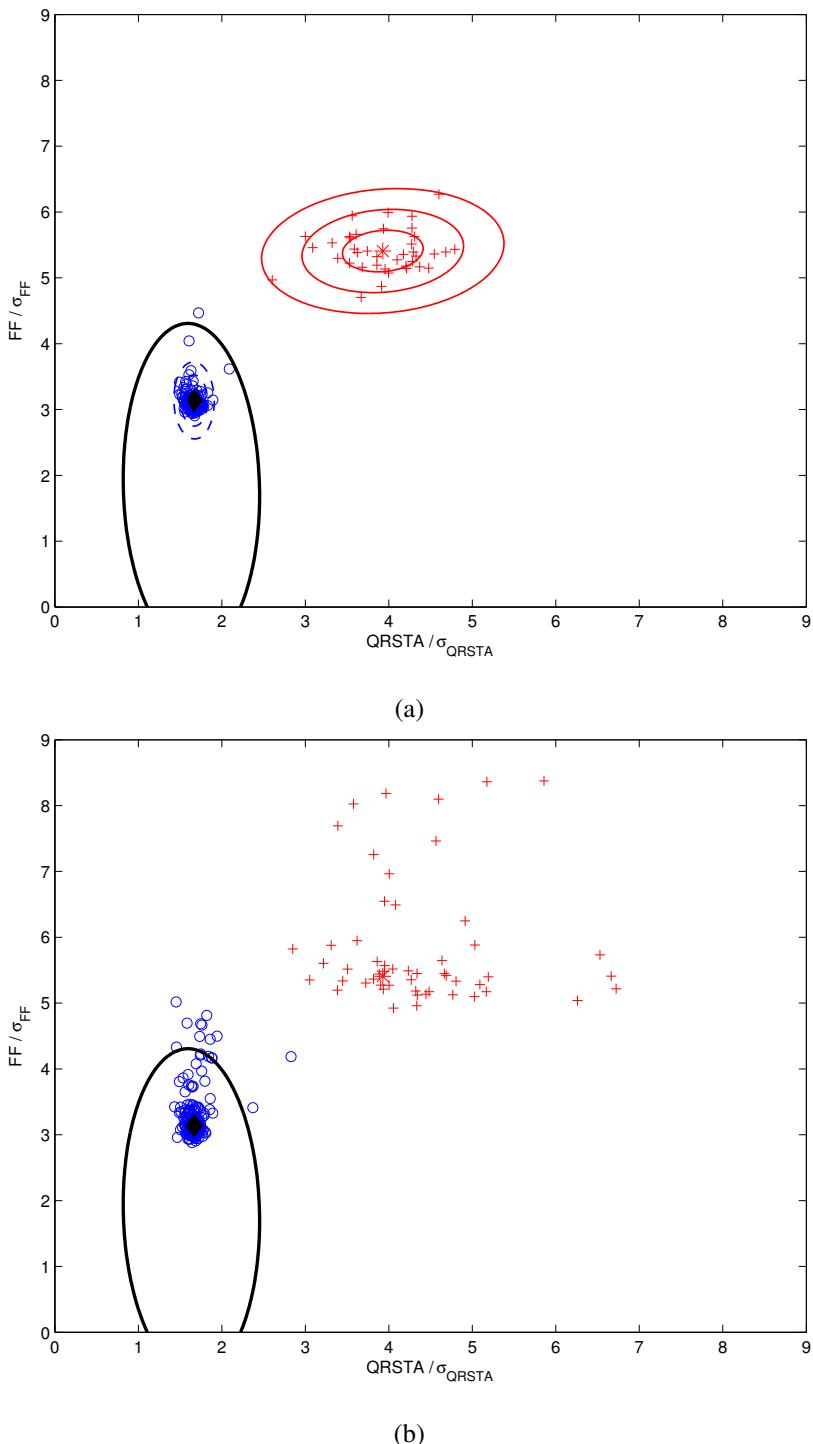


Figure 10.15 2D scatter plot of the feature vectors $[QRSTA, FF]^T$ for (a) the training set and (b) the test set of ECG signals. The circles represent normal beats and the plus marks represent PVCs; the diamond and star symbols represent the corresponding average or prototype vectors. The ellipses show the boundaries of the 2D Gaussian functions estimated at σ , 2σ , and 3σ for each class. The partial elliptical contour is the Bayesian decision boundary. Figure courtesy of Fábio José Ayres, University of Calgary, Calgary, Alberta, Canada.

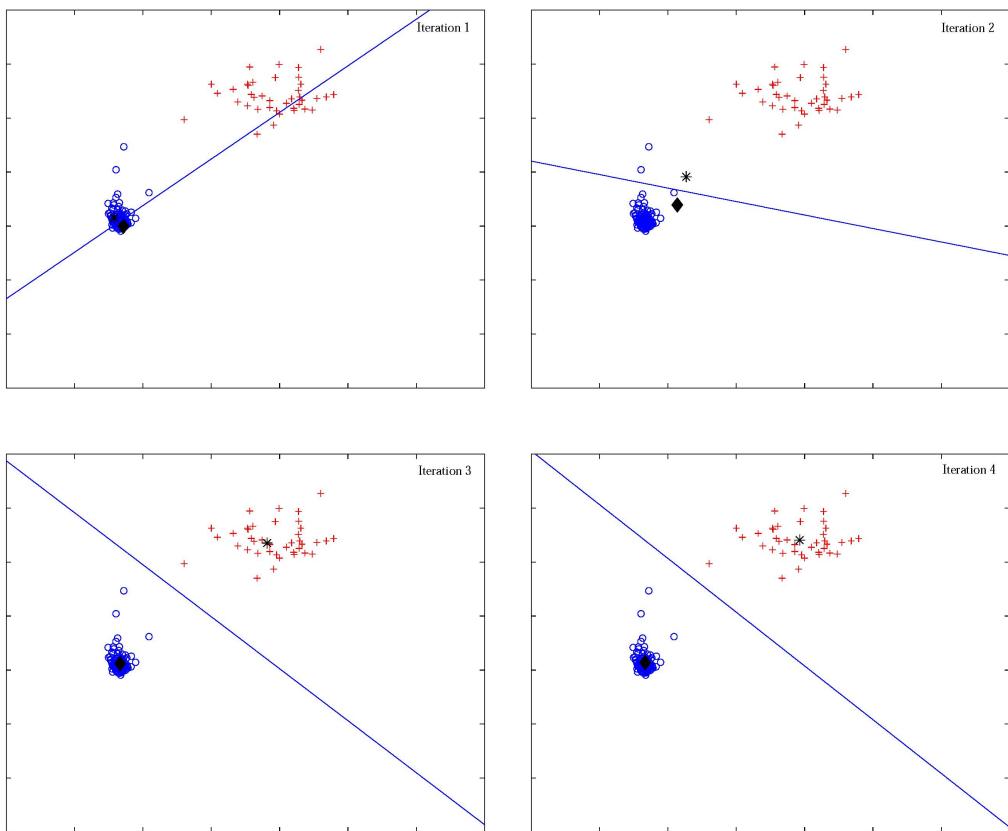


Figure 10.16 2D scatter plot of the feature vectors $[QRSTA, FF]^T$ for the training set of ECG signals with the results of the K -means method after the first, second, third, and fourth (final) iterations. Each feature was normalized by its SD . The range of each axis is $[0, 7]$. The axis labels have been removed to reduce clutter. The circles represent normal beats and the plus marks represent PVCs; the diamond and star symbols represent the corresponding mean vectors being modified at each iteration. The straight line is the separating boundary. Figure courtesy of Fábio José Ayres, University of Calgary, Alberta, Canada.

segments and considering each segment as a separate signal. To overcome this problem, the means (time averages) of the segments of each subject's signal were computed, and then the variance of the means was computed across the various segments of the same signal. The variance of the means represents the above-mentioned difference, and was used as one of the discriminant features. (The MS or RMS values of VAG segments may be more suitable for this purpose than their mean values.)

In addition to quantitative parameters derived from VAG signal analysis, clinical parameters (to be described in the following paragraphs) related to the subjects were also investigated for possible discriminant capabilities. At the outset, as shown in Figure 10.17, knee joints of the subjects in the study were categorized into two groups: normal and abnormal. The normal group was divided into two subgroups: normal-silent and normal-noisy. If no sound was heard during auscultation, a normal knee was considered to be normal-silent; otherwise, it was considered to be normal-noisy. All knees in the abnormal group used were examined by arthroscopy (see Section 8.2.3 and Figure 8.2) and divided into two groups: arthroscopically normal and arthroscopically abnormal.

Labeling of VAG signal segments was achieved by comparing the auscultation and arthroscopy results of each patient with the corresponding segmented VAG and joint angle signals. Localization of the pathology was performed during arthroscopy and the joint angle ranges where the affected areas could come into contact with other joint surfaces were estimated. These results were then compared with the auscultation reports to determine whether the joint angle(s) at which pathology existed corresponded to the joint angle(s) at which sound was heard. For example, if it was found from the arthroscopy report of a patient that the abnormal parts of the patient's knee could cause contact in the range $30^\circ - 90^\circ$, VAG signal segments of the subject corresponding to the angle range of $30^\circ - 90^\circ$ were labeled as arthroscopically abnormal; the rest of the segments of the signal were labeled as arthroscopically normal.

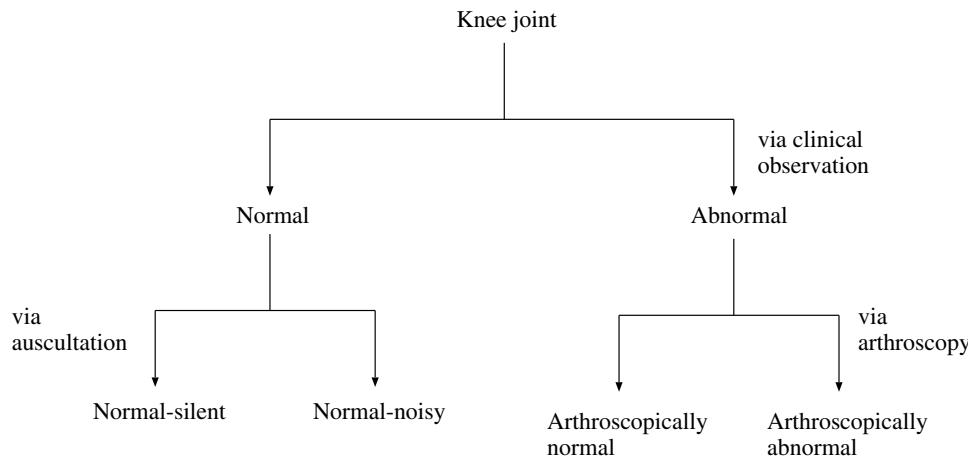


Figure 10.17 Categorization of knee joints based on auscultation and arthroscopy.

Categorization into four groups as above was done based on the presumptions that normal-noisy and arthroscopically abnormal signals might be distinguishable in their characteristics and that normal-silent and arthroscopically normal knees would also be distinguishable. The possibilities of arthroscopically normal knees being associated with sounds, normal-noisy knees not having any associated pathology, and normal-silent knees having undetermined pathologies were also admitted. Krishnan et al. [64] further subdivided the arthroscopically normal and arthroscopically abnormal categories into silent and noisy categories, thereby having a total of six categories; this is not shown in Figure 10.17.

Based on clinical reports and auscultation of knee joints, the following clinical parameters were chosen as features (in addition to AR-model parameters) for classification:

Sound: The sound heard by auscultation during flexion and extension movement of the knee joint was coded as

- 0— silent,
- 1— click,
- 2— pop,
- 3— grinding, or
- 4— a mixture of the above-mentioned sounds or other sounds.

Each segment of the VAG signals was labeled with one of the above codes.

Activity level: The activity level of each subject was coded as

- 1— exercising once per week or less,
- 2— exercising two or three times per week, or
- 3— exercising more than three times per week.

Age: The age of the subject in years.

Gender: The gender of the subject, which was coded as

- 0— female, or
- 1— male.

Among the parameters mentioned above, gender may not be a discriminant parameter; however, it is customary to record gender in clinical analysis. Note that among the four clinical parameters listed above, only the first one can vary between the different segments of a given subject's VAG signal.

Moussavi et al. [63] compared the performance of various sets of features in the classification of VAG signals into two groups and four groups (see Figure 10.17) with random selections of cases. Using a set of 540 segments obtained from 20 normal subjects and 16 subjects with cartilage pathology, different numbers of segments were randomly selected for use in the training step of designing a discriminant function, and finally the selection which provided the best result was chosen for the final classification system. Two-group classification accuracies in the range 77 – 91% and four-group classification accuracies in the range 65 – 88% were obtained.

By combining the steps of classification into two groups and four groups, a two-step method was proposed by Moussavi et al. [63]; a flowchart of this method is illustrated in Figure 10.18. The algorithm first uses training sets to design classifiers for two and four groups. The resulting discriminant functions are used as Classifier 1 (two groups) and Classifier 2 (four groups), respectively. An unknown signal, which has been adaptively divided into segments, enters Classifier 1. If segments spanning more than 90% of the duration of the signal are classified as being normal, the signal (or subject) is considered to be normal. On the other hand, if more than 90% of the duration of the signal is classified as being abnormal, the signal (or subject) is considered to be abnormal. If more than 10% but less than 90% of the signal duration is classified as abnormal, the signal goes to Classifier 2, which classifies the signal into four groups (see Figure 10.17). In the second step, if more than 10% of the duration of the signal is classified as being arthroscopically abnormal, the signal is considered to be abnormal; otherwise, it is considered to be normal. At this stage, information on the numbers of segments belonging to the four categories shown in Figure 10.17 is available, but the final decision is on the normality of the whole signal (subject or knee joint).

The two-step diagnosis method was trained with 262 segments obtained from 10 normal subjects and eight subjects with cartilage pathology, and it was tested with 278 segments obtained from a different set of 10 normal subjects and eight subjects with cartilage pathology but without any restriction on the kind of abnormality. Except for one normal signal which was indicated as being abnormal over 12% of its duration, all of the signals were correctly classified. The results also showed that all of the abnormal signals including signals associated with chondromalacia grades I to IV (see Section 8.2.3 and Figure 8.2) were classified correctly. Based on this result, it was indicated that the method has the ability to detect chondromalacia patella at its early stages as well as advanced stages. Krishnan et al. [64] and Rangayyan et al. [65] reported on further work along these directions.

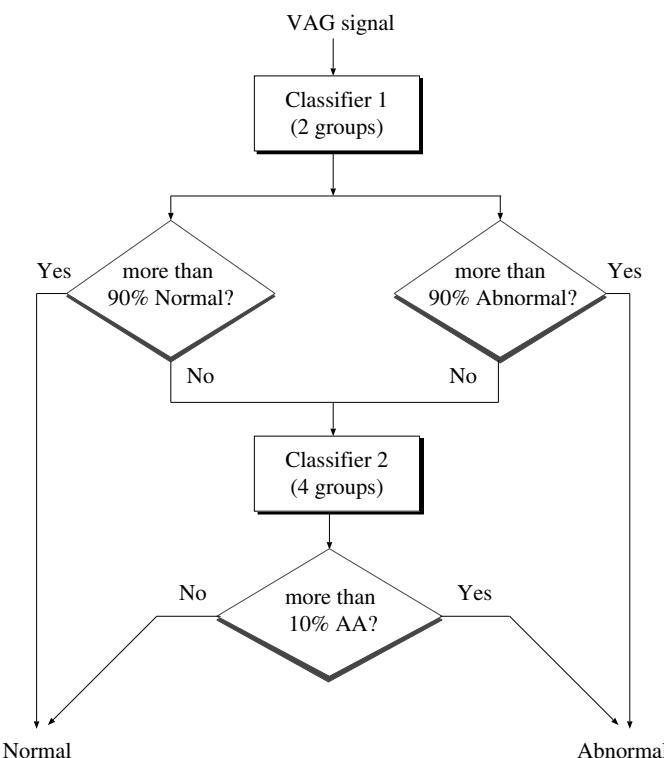


Figure 10.18 A two-step classification method for the diagnosis of cartilage pathology. AA, arthroscopically abnormal. See also Figure 10.17. Reproduced with permission from Z.M.K. Moussavi, R.M. Rangayyan, G.D. Bell, C.B. Frank, K.O. Ladly, and Y.T. Zhang, Screening of vibroarthrographic signals via adaptive segmentation and linear prediction modeling, *IEEE Transactions on Biomedical Engineering*, 43(1):15–23, 1996. ©IEEE.

10.13 Application: Detection of Sleep Apnea

In order to maintain good health and well-being, sleep is crucial. Rest provided by deep stages of sleep is essential for good physical, mental, and emotional health. The body grows and strengthens bones, regenerates tissues, and strengthens the immune system during sleep. To ensure that the body and the brain are functioning properly, it is crucial to obtain adequate sleep. Adults should strive for 7 – 9 hours of sleep each night, according to the National Sleep Foundation [66]; teenagers and children may require more.

Short-term and long-term effects result from sleep deprivation, whether it is caused by lifestyle choices or sleep disorders. The immediate result is reduced attention and focus, lower quality of life, higher absenteeism rates with lower output when present, and accidents at work, at home, or while driving a vehicle. The long-term effects of lack of sleep include an increase in morbidity and mortality from accidents, stroke, heart failure, coronary artery disease, obesity, depression, and memory impairment [67].

REM and non-REM (NREM) states of sleep alternate in a cycle (a total of four to six cycles have been noted during sleep in adults), with each cycle lasting on average from 90 to 110 minutes [66]. Sleep is divided into these two states based on EEG, EOG, and EMG signals. Each state has independent functions and controls. In an adult human, slow-wave sleep predominates during the first third of sleep while REM sleep predominates during the final third period. Adults spend between 75 and 80% of their sleep duration in NREM sleep, which is classified into four stages (NREM

stages 1 through 4). Normal wakefulness, NREM sleep, and REM sleep all include a number of physiological and behavioral changes [67].

The International Classification of Sleep Disorders classifies various sleep disorders into seven groups [67]; see Section 2.4. The second category, sleep-related breathing disorders, includes OSA, CSA, and sleep-related hypoxemia and hypoventilation. The most prevalent disorder in this group, OSA, is defined by intermittent or persistent partial or total obstruction of the upper airway and collapse of the upper airway, which impairs ventilation during sleep. Excessive daytime sleepiness brought on by nonrestorative sleep is one of the signs of OSA. The gold standard for diagnosing OSA is PSG, which uses many sensors to record body position, breathing patterns, respiratory motion, EEG, EOG, and EMG activity. The severity of sleep apnea can be determined using the AHI. The number of apnea and hypopnea episodes per hour of sleep serves as a proxy for AHI. The total absence of airflow through the mouth and nose is known as apnea. The airway partially collapses during hypopnea, which makes breathing difficult. An episode of apnea (pause in breathing) is required to be at least 10 s long and be linked to a drop in blood oxygenation in order to be taken into account. There are three levels of OSA severity: mild (5 to 15 AHI events/h), moderate (15 to 30 AHI events/h), and severe (AHI greater than 30 events/h) [67].

PSG facilitates accurate determination of AHI and diagnosis of OSA, but the procedure is time-consuming and expensive because the patient must spend several hours (usually overnight) in a sleep laboratory under the observation of a trained technician. The test could also be carried out in the patient's home using portable PSG equipment, but doing so would still be uncomfortable due to the requirement of several sensors [8]. Clinical reports are produced by manually scoring the recorded signal data. Alternative gadgets that monitor patients at home with fewer sensors and autonomous diagnostic algorithms have been developed in an effort to address these difficulties [68]. However, it is generally accepted that the evaluation accuracy of such algorithms is insufficient for a reliable medical diagnosis.

Numerous techniques to detect OSA have been proposed, but none of them has been implemented on a practically feasible system. In some cases, this is likely because the complexity of the algorithms utilized makes it difficult for them to be executed effectively on systems with limited resources.

Based on analysis of the HRV, Quiceno-Manrique et al. [69] extracted features from Cohen's class TFDs using spectral subband centroids and cepstral coefficients. The kNN method was used to perform categorization. Figure 10.19 shows the SPWVD of normal and abnormal HRV signals. It is seen that the SPWVD of the normal HRV signal has a constant frequency component around 0.23 Hz, which could be due to respiratory activity. In the case of the abnormal HRV signal related to apnea, the SPWVD does not show any constant frequency component.

By using PDFs to map each long-term RR interval into a disease state space, Chen and Zhang [70] were able to analyze how the RR interval signal changed states based on changes in the Gamma distribution's parameters. A multistate cumulative sum method based on the Gamma distribution was used to calculate shifts in RR intervals. The exponential likelihood ratio test and backward elimination were used to remove the spurious state change points that were observed. Based on the state change points and using a generic formula, a severity index that reflects the disease's severity was created. The severity index was used as a feature with three classifiers: LDA, SVM, and logistic regression. The results showed that logistic regression produced the best outcomes in diagnosing OSA.

Daubechies wavelets were employed by Khandoker et al. [71] to separate the respiratory signals derived from RR intervals obtained from the ECG. The HRV and ECG-derived respiratory signals' wavelet decomposition features were sent to SVM, kNN, probabilistic neural network, and LDA classifiers as inputs. They found the SVM classifier provided the best results compared to the three other classifiers.

PPG signals obtained by pulse oximetry can be analyzed at home with ease and comfort. There are many published works on the topic of employing PPG signals for OSA detection and classi-

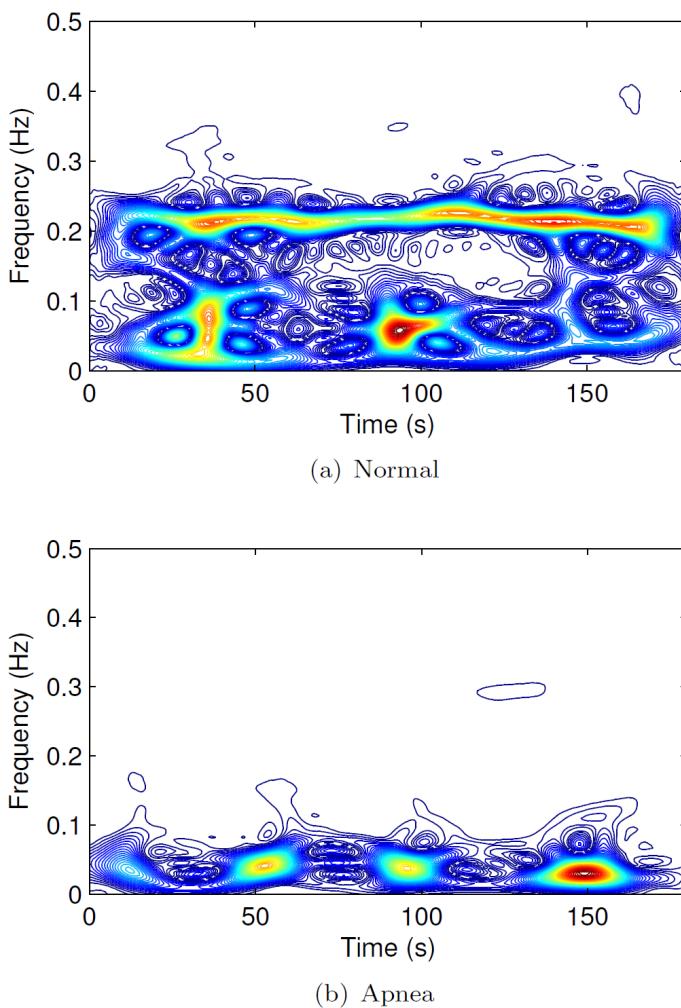


Figure 10.19 SPWVD of (a) a normal HRV signal and (b) an abnormal HRV signal related to apnea. Reproduced with permission from A.F. Quiceno-Manrique, J.B. Alonso-Hernandez, C.M. Travieso-Gonzalez, M.A. Ferrer-Ballester, and G. Castellanos-Dominguez, Detection of obstructive sleep apnea in ECG recordings using time-frequency distributions and dynamic features. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp 5559–5562, 2009. ©IEEE.

fication. Marcos et al. [72] examined four statistical pattern recognition methods: kNN, QDA, LDA, and logistic regression. LDA with spectral characteristics yielded the best results, with an A_z value of 0.925. SVM and kNN were compared for OSA classification by Morales et al. [73]; kNN provided higher accuracy.

Maali and Al-Jumaily employed airflow along with thoracic and abdominal respiratory motion signals for detection of OSA [74]. In order to create features to feed an SVM with a polynomial kernel, wavelet transformation was applied to the signals. A genetic algorithm was used to choose the best feature subset and training data in an interactive manner. They demonstrated that the use of a small number of biomedical signals, such as air flow and respiratory signals, could provide good results. Ng et al. [75] modeled snoring signals using LP coding and derived formant frequencies from the model spectrum. Apnea and regular breathing sounds were distinguished using a threshold value.

Given the value of different signal modalities studied and applied for OSA detection, it should be possible to combine them to improve the classification accuracy of machine learning approaches. For instance, Madhav et al. [76] used EMD to obtain respiratory signals derived from both the PPG and the ECG. Each windowed signal was subjected to EMD and fitted with an AR model of order 15 to detect OSA. They obtained an accuracy rate of 94% to 100% on the signals of 90 patients in intensive care units, demonstrating the simplicity and usefulness of ECG and PPG over the use of EEG, EMG, and EOG signals for OSA identification.

Al-Angari and Sahakian [77] integrated ECG signals with oximetry data as well as abdominal and thoracic breathing effort signals. Signal features were derived from oxygen saturation, HRV computed from the ECG signals by applying the Pan–Tompkins method, and the phase and magnitude of the respiratory effort signals. Phase-locking value, a nonlinear synchronization indicator, was computed from the effort signals, and statistical analysis of the oxygen saturation data and HRV was performed. Utilizing an SVM with a polynomial kernel, the data were categorized. A classification accuracy rate of 82.4% was obtained using a dataset of 50 normal and 50 OSA signals.

In further research on home-based monitoring and diagnostic systems for OSA, consideration should be given to algorithm complexity factors and suitability for hardware implementation. According to Krishnan [1], wearable sensors produce signals with “4N” aspects (nonstationary, nonlinear, non-Gaussian, and non-short-term) and “4V” qualities (volume, velocity, veracity, and variety); such signals could benefit from analysis using “4G” (four generations) algorithms to achieve device-level, gateway-level, and cloud-based computing implementations. Long-term remote monitoring of patients with sleep and other disorders may soon be a reality with rapid advancements in wearable sensors, information processing, machine learning, communication, and cybersecurity techniques and technologies [1].

10.14 Application: Monitoring Parkinson's Disease Using Multimodal Signal Analysis

Parkinson's disease is a degenerative brain disorder that affects aging individuals; it is well recognized for causing delayed movements, tremors, and balance problems, among other symptoms [78]. The majority of cases are caused by unidentified reasons; however, some are inherited. Symptoms typically manifest gradually and worsen with time. As the disease progresses, people with Parkinson's disease may experience difficulty walking and speaking. They may also have behavioral and mental abnormalities, depression, memory loss, fatigue, and sleep disturbances.

When nerve cells in the region of the brain that controls movement, the basal ganglia, are damaged or die, the most noticeable symptoms and signs of Parkinson's disease arise. These nerve cells, or neurons, ordinarily produce dopamine, an important neurotransmitter in the brain. When neurons are diseased or damaged, they produce less dopamine, resulting in mobility problems associated with the condition. People affected by Parkinson's disease also lose nerve endings that release norepinephrine, the principal chemical messenger of the SNS, which regulates several internal functions such as BP and heart rate. Some of the symptoms of Parkinson's disease not related to movement, such as fatigue, fluctuating BP, decreased food movement through the digestive tract, and a sudden drop in BP when rising from a sitting or lying position, may be attributable to reduced norepinephrine.

Typically, Parkinson's disease is diagnosed by a clinical approach in which a healthcare professional analyzes symptoms, asks questions, and investigates medical history. There are a few diagnostic and laboratory tests available; however, they are commonly used to rule out other diseases or conditions. Figure 10.20 shows the various motor and nonmotor symptoms associated with Parkinson's disease. The symptoms present a multimodal context for the use of biomedical signals in a diagnostic application, including speech, EMG, ECG, and EEG.

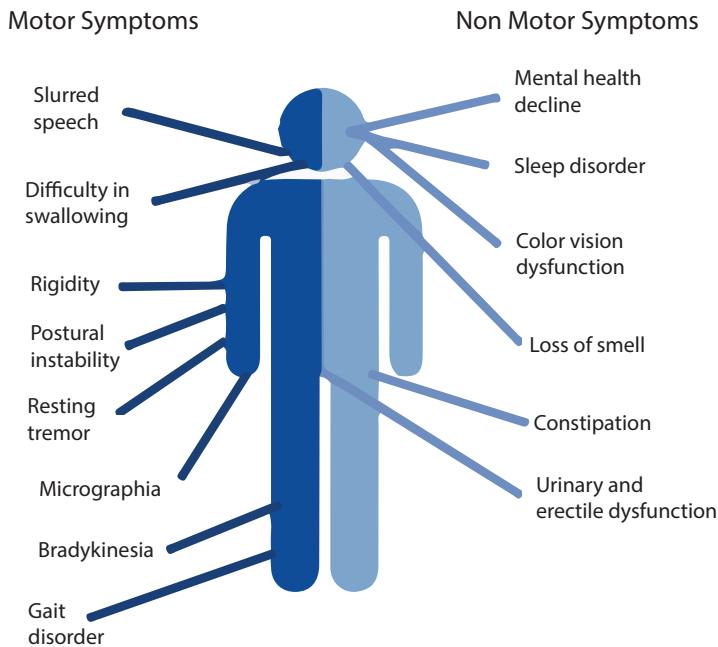


Figure 10.20 Various symptoms of Parkinson’s disease. Reproduced with permission from A. Rueda [79].

Figure 10.21 shows examples of the speech signal for sustained utterance of the vowel /a/ with normal and abnormal characteristics, such as breathy, low-resonance, unstable, and hoarse nature caused by neurological disorders including Parkinson’s disease. From the spectrograms of the speech samples, it is seen that the breathy, low-resonance, and hoarse voices have variable and disorderly distributions of energy over the TF plane. An unstable voice fluctuates across frequency over time. The differences in energy spread in the spectrograms could be captured using TF features computed through signal and matrix decomposition approaches (see Sections 9.3, 9.4, 9.6, 9.7.3).

Ghoraani and Krishnan [80] applied NMF to the MPTFD of pathological speech signals. Features extracted from the joint TF representation and NMF matrix decomposition approaches provided good clustering characteristics, as shown in Figure 10.22. The abnormality measure was calculated as the ratio of the total number of abnormal feature vectors to the total number of feature vectors in the voice sample.

Another signal of interest could be from hand tremor, which is one of the main symptoms of Parkinson’s disease. To obtain hand tremor signals, an inertial measurement unit (IMU) could be used. An IMU sensor contains an accelerometer, a gyroscope, and a magnetometer; the three components provide triaxial measurements in the x , y , and z directions. Figure 10.23 illustrates the hand tremor data collection protocol developed at the Centre for Innovation and Technology Assessment in Health at the Federal University of Uberlândia, Uberlândia, Minas Gerais, Brazil [81]. Typically, for tremor measurements, it is sufficient to place one sensor at the back of the hand; however, in the protocol developed by the Uberlândia group, two IMU sensors were placed on the back of the hand and the forearm.

Figure 10.24 shows plots of the signals collected from a subject with the wrist flexed and at rest. The multimodal signals from the two IMU sensors and the pulse signal are shown; also, the triaxial signals from the gyroscope, the accelerometer, and the magnetometer are shown for each

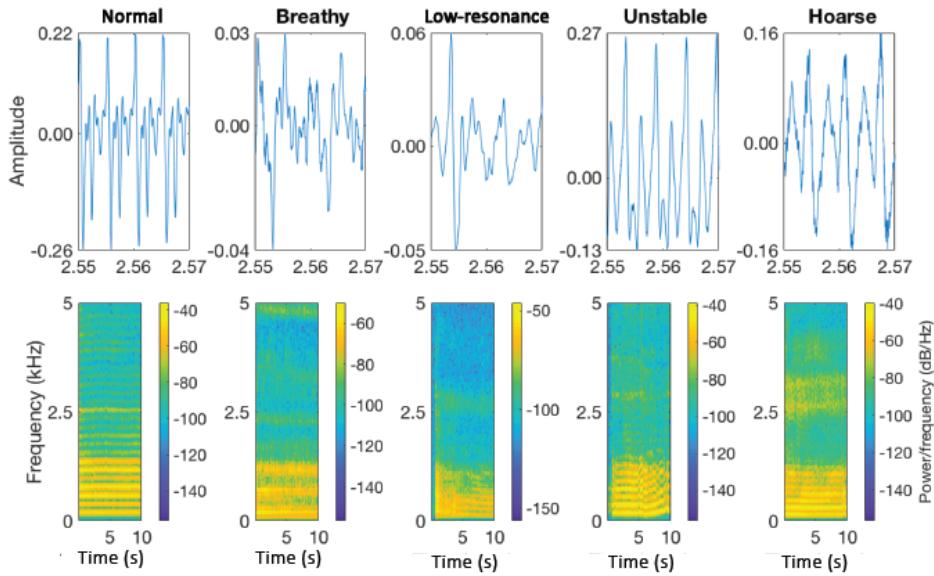


Figure 10.21 Examples of speech signals and the corresponding spectrograms for sustained utterance of the vowel /a/ illustrating, from left to right, normal, breathy, low-resonance, unstable, and hoarse voices. Reproduced with permission from A. Rueda and S. Krishnan. Augmenting dysphonia voice using Fourier-based synchrosqueezing transform for a CNN classifier. In Proceedings of the 2019 International Conference of Acoustics, Speech, and Signal Processing, pp 6415–6419, May 2019. ©IEEE.

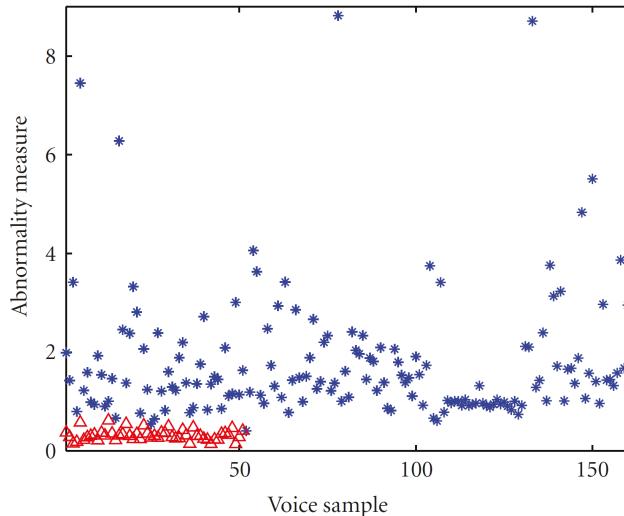


Figure 10.22 Clustering of NMF features of normal and pathological speech. Asterisks (blue) denote pathological samples and triangles (red) denote normal samples. Reproduced with permission from B. Ghoraani and S. Krishnan. A joint time-frequency and matrix decomposition feature extraction methodology for pathological voice classification. *EURASIP Journal on Advances in Signal Processing*, 1:11, 2009.

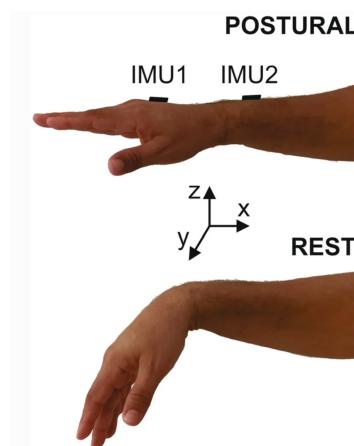


Figure 10.23 Tremor signal acquisition using triaxial inertial sensors placed at the hand (IMU1) and forearm (IMU2). The protocol developed at the Federal University of Uberlândia included a movement sequence of the reference task starting at (1) the resting position, (2) the hand levelled, and (3) moving the hand towards the chest. Reproduced with permission from A. de Oliveira Andrade, A.P.S. Paixão, A.M. Cabral, A.G. Rabelo, L.M.D. Luiz, V.C. Dionísio, M.F. Vieira, J.M. Pereira, A. Rueda, S. Krishnan, and A.A. Pereira, Task-specific tremor quantification in a clinical setting for Parkinson's disease. *Journal of Medical and Biological Engineering*, 40(6):821–850, 2020. ©Springer.

IMU. Signal features including the median of the RMS value of signal amplitude and weighted mean energy values from four spectral bands of 0 to 2.5 Hz, 2.5 to 5.0 Hz, 5.0 to 10.0 Hz, and 10.0 to 15.0 Hz were extracted from the multimodal tremor signals following signal preprocessing steps. Figure 10.25 shows the feature plots during rest and postural states for a subject with Parkinson's disease. The higher values of the features for the rest condition indicate the severity of the tremor due to Parkinson's disease. Similar trends were also observed with the 10 other subjects that participated in the study [81].

To enable diagnostic decision support for health conditions such as Parkinson's disease, there are three possible ways of multimodal data fusion: early stage at the data level, intermediate stage at the feature level, and later stage at the classifier level. Signals generated from a diverse range of sources may be converted to the same information space using ICA or NMF techniques. Intermediate fusion combines the attributes that distinguish each type of data to generate a more accurate representation than early fusion; it is more expressive than the individual signal-space representations from which the data originated. In later fusion, many models are trained, each of which corresponds to an incoming data source, and an ensemble approach is used to make an overall decision about the disease condition and progression. Simultaneous recording of the speech, EMG, IMU, ECG, and EEG signals from patients with Parkinson's disease could facilitate the application of multimodal data fusion methods for monitoring [82].

10.15 Strengths and Limitations of CAD

As we come to the end of the book, it is appropriate and desirable to review and continue the initial discussion on CAD in Section 1.5.

Technical advances and improvements in systems for acquisition and analysis of biomedical signals as well as their expanding use in routine clinical work have led to increases in the scope and complexity of related problems, requiring further advanced techniques for their solution. This has led to further research and development in biomedical instrumentation, biomedical signal analysis,

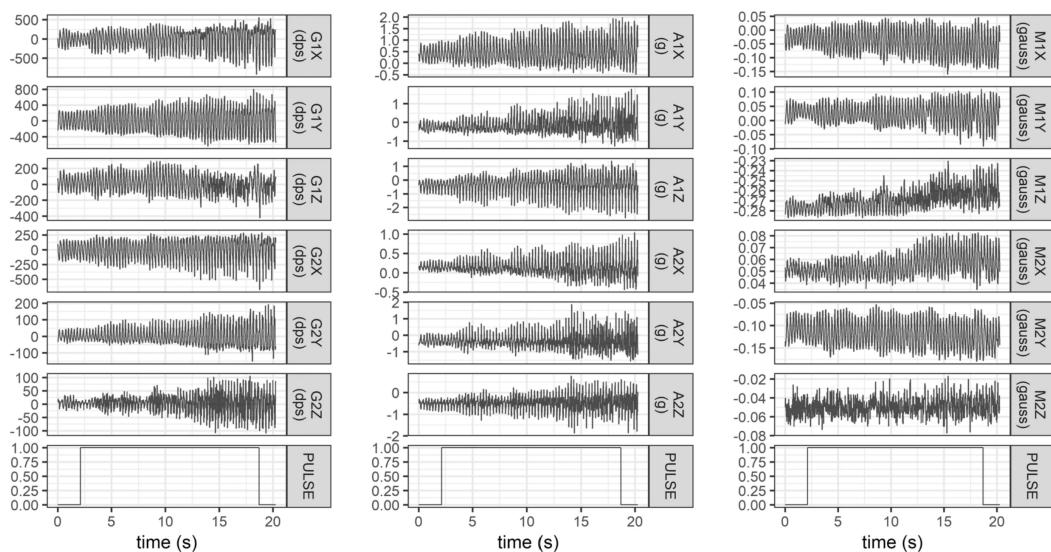


Figure 10.24 Multimodal tremor signals obtained using IMU sensors and a pulse sensor. G in the plots denotes a gyroscope signal, A denotes an accelerometer signal, and M denotes a magnetometer signal in the triaxial directions x (anterior–posterior), y (medial–lateral), and z (vertical). Reproduced with permission from A. de Oliveira Andrade, A.P.S. Paixão, A.M. Cabral, A.G. Rabelo, L.M.D. Luiz, V.C. Dionísio, M.F. Vieira, J.M. Pereira, A. Rueda, S. Krishnan, and A.A. Pereira, Task-specific tremor quantification in a clinical setting for Parkinson’s disease. *Journal of Medical and Biological Engineering*, 40(6):821–850, 2020. ©Springer.

medical informatics, and CAD [83]. CAD is described as diagnosis made by a physician or medical specialist by using the output of a computerized scheme for signal or image analysis as a diagnostic aid [84]. Two variations in CAD have been recognized: CADe for computer-aided detection of abnormal segments or events of interest and CADx for computer-aided diagnosis with labeling of the detected parts in terms of the presence or absence of a certain disease or abnormality. (*Note:* The text and tables in this section have been adapted from “Preface,” pp xv–xx, and “Concluding Remarks,” pp 505–507, by P.M. Azevedo-Marques, A. Mencattini, M. Salmeri, and R.M. Rangayyan, in “Medical Image Analysis and Informatics: Computer-Aided Diagnosis and Therapy,” Edited by P.M. Azevedo-Marques, A. Mencattini, M. Salmeri, and R.M. Rangayyan, CRC Press, Boca Raton, FL. © 2018 by Imprint. Reproduced by permission of Taylor & Francis Group.)

Typically, a clinician using a CAD system may make an initial decision and then consider the result of the CAD system as a second opinion; traditionally, such an opinion would have been obtained from another medical specialist. The clinician may or may not change the initial decision after receiving the second opinion, be it from a CAD system or another clinician. In such a situation, the CAD system need not be better than or even comparable to clinicians in terms of diagnostic accuracy or efficiency. If the CAD system is designed to be complementary to the clinician, the symbiotic and synergistic combination of the clinician with the CAD system can improve the accuracy of diagnosis [84].

In a more radical manner, one may apply a CAD system for initial screening of all cases to be assessed, and then send to the clinician or medical specialist only those cases that merit attention at an advanced level; the remaining cases may be analyzed by other suitably trained personnel. While this process may be desirable when the patient population is large and the number of available medical experts is disproportionately small, it places heavier reliance on the CAD system. Not all

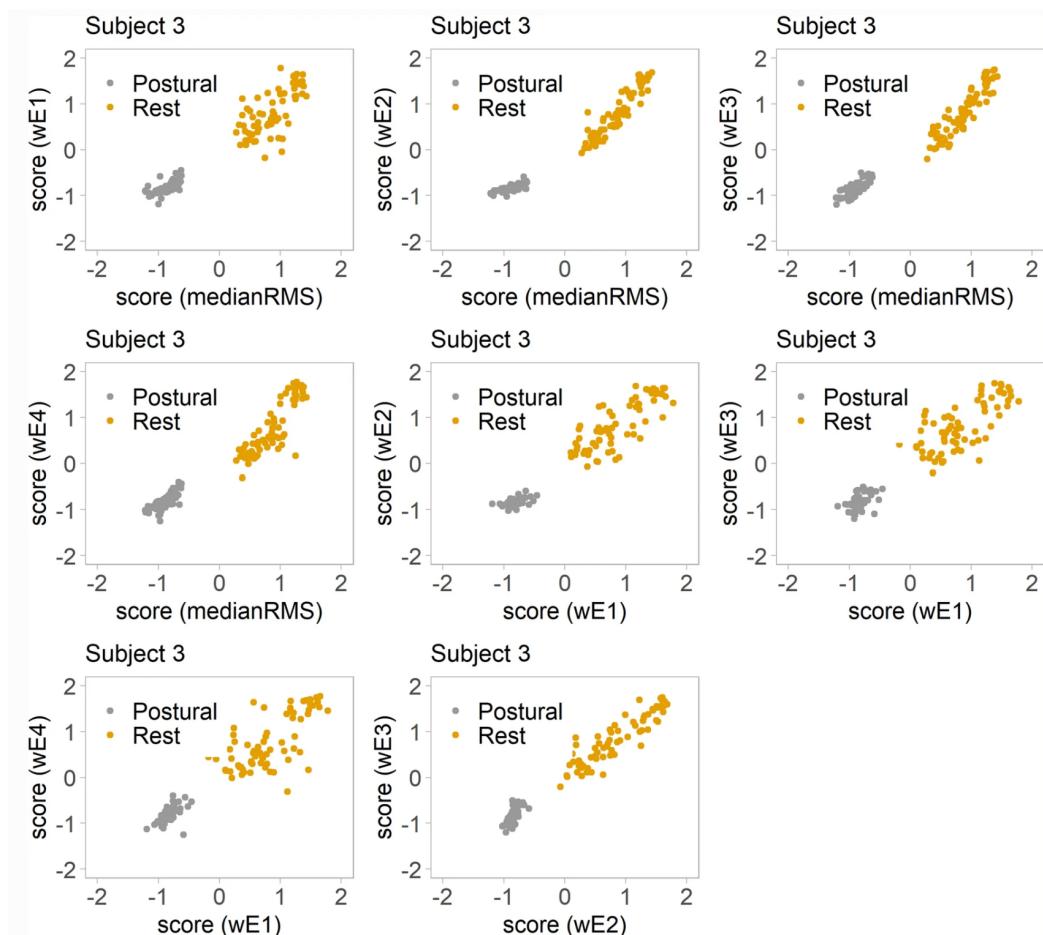


Figure 10.25 Pairwise feature plots of tremor signals associated with postural and rest conditions of Subject 3 with Parkinson's disease. Reproduced with permission from A. de Oliveira Andrade, A.P.S. Paixão, A.M. Cabral, A.G. Rabelo, L.M.D. Luiz, V.C. Dionísio, M.F. Vieira, J.M. Pereira, A. Rueda, S. Krishnan, and A.A. Pereira, Task-specific tremor quantification in a clinical setting for Parkinson's disease. *Journal of Medical and Biological Engineering*, 40(6):821–850, 2020. ©Springer.

societies and jurisdictions may accept such an application where a computational procedure is used to make a diagnostic decision.

Medical informatics deals with the design of methods and procedures to improve the analysis, efficiency, accuracy, usability, and reliability of medical data for healthcare. CAD and content-based retrieval (CBR) are two important applications in medical image informatics. CBR systems are designed to bring relevant and clinically established cases from a database when presented with a current case as a query. The features and diagnoses associated with the retrieved cases are expected to assist the clinician or medical specialist in diagnosing the current case. Even though CBR systems may not suggest a diagnosis, they rely on several of the same techniques that are used by CAD systems and share some similarities [83, 85].

Regardless of our interest in and enthusiasm for CAD, it is important to recognize the need for the application of computers in the analysis of biomedical signals as well as the associated strengths and limitations [83]. Clinicians and medical specialists are highly trained professionals. Why, when,

and what for would they need the assistance of computers? Data related to biomedical signals could be voluminous and bear intricate details. Some of the details may pose challenges in perception due to the limited capabilities of the human visual, auditory, and other sensory systems [86–88]. More often than not, normal cases in a clinical setting or the various parts within a given signal overwhelmingly outnumber abnormal cases or details. Regardless of the level of expertise and experience of a medical specialist, visual analysis of biomedical signals is prone to several types of errors, some of which are listed in Table 10.6. The application of computational techniques could address some of these limitations, as indicated in Table 10.7.

Causes of error in manual analysis of signals	Type of error
Subjective and qualitative analysis	Inconsistency
Differences in knowledge, learning, and training; differences in opinion; variations in personal preferences and judgment	Interobserver error
Inconsistent application of knowledge; lack of diligence; environmental effects and distraction; fatigue and boredom due to workload and repetitive tasks	Intraobserver error

Table 10.6 Causes of various types of errors in manual analysis of biomedical signals. Adapted from Azevedo-Marques et al. [83] and Doi [84].

The typical steps of a CAD system are as follows [83]:

1. Preprocess the given signal for further analysis by filtering to remove noise and artifacts.
2. Detect events or epochs and divide the signal into relevant segments.
3. Extract measures or features for quantitative analysis.
4. Select an optimal set of features.
5. Train classifiers and develop decision rules.
6. Perform pattern classification to achieve diagnostic decision.

Table 10.8 shows a generic and simplified scheme to overcome some of the limitations of manual or visual analysis by applying computational procedures.

The paths and procedures shown in Table 10.8 are not simple and straightforward; furthermore, they are not free of problems and limitations. In spite of the efforts of several researchers, the development and clinical application of CAD systems encounter several difficulties, some of which are listed below [83]:

- Difficulty in translating methods of visual and other manual analysis into computational procedures.
- Difficulty in translating clinical observations into numerical features or attributes.
- Difficulty in dealing with large numbers of features in a classification rule, known as the curse of dimensionality.

Manual analysis of signals	Computer analysis of signals
Inconsistencies in identifying and labeling segments of signals, events, or epochs	Consistent application of established rules and methods
Errors in localization of segments of signals, events, or epochs due to limited manual dexterity	High numerical precision and computational accuracy
Extensive time and effort for manual measurement of intricate details, such as time periods and signal amplitudes	High speed of computation and efficiency with large amounts of signal data
Limitations in the precision and reproducibility of manual measurements and calculations	High accuracy, repeatability, and reproducibility
Effects of distraction, fatigue, and boredom	Immunity to effects of work environment, fatigue, and boredom

Table 10.7 Comparison of various aspects of manual versus computer analysis of biomedical signals. Adapted from Azevedo-Marques et al. [83] and Doi [84].

Move from	Via	To
Qualitative analysis	Computation of measures and features using DSP techniques	Quantitative analysis
Subjective analysis	Development of rules for diagnostic decision making using pattern classification and machine learning techniques	Objective analysis
Inconsistent analysis	Implementation of established rules and robust procedures as computational algorithms and computer programs	Consistent analysis
Interobserver and intraobserver errors	Biomedical signal analysis, biomedical informatics, expert systems, and CAD	Improved diagnostic accuracy

Table 10.8 Techniques and means to move from manual to computer analysis of biomedical signals. Adapted from Azevedo-Marques et al. [83].

- Substantial requirements of computational resources and time.
- Need for large sets of annotated signals and labeled cases to train and test a CAD system.
- Large numbers of false alarms or false positives.
- Difficulty in integrating CAD systems into established clinical workflows and protocols.

Despite recent advances, there are many difficulties and challenges in improving the utilization of information and communication technologies (ICT) in the healthcare environment. Different ways to use diverse technologies, lack of widely adopted data communication standards, and the intersection of multiple domains of knowledge are some of the issues that must be overcome in order to improve healthcare. Various standards have been developed by the International Standards Organization (ISO) and other organizations, including HL7: Health Level-7, OSI: Open Systems Interconnection, and EHR: Electronic Health Record. Integrating the Healthcare Enterprise (IHE) is an effort to promote the inclusion of CAD and several other technological developments into clinical workflow and the Hospital Information System (HIS).

Some of the essentials of engineering that should be integrated into CAD are:

- Scientific investigation and analysis;
- Quantitative and objective analysis;
- Mathematical modeling;
- Optimal design of components, systems, and processes;
- Synthesis of multidisciplinary principles and methodologies;
- Project management;
- Efficient solutions to practical problems;
- Innovation and creativity.

An important point to note is that quantitative analysis becomes possible by the application of computational procedures to biomedical signals. The logic of medical or clinical diagnosis via analysis of biomedical signals and data can then be objectively encoded and consistently applied in routine or repetitive tasks. However, we emphasize that the end goal of computerized biomedical signal analysis should be *computer-aided* diagnosis and not automated diagnosis. A physician, clinician, or medical specialist typically uses a large amount of information in addition to signals and measurements, including the general physical and mental state of the patient, family history, and socioeconomic factors affecting the patient, many of which are not amenable to quantification and logical rule-based processes. Biomedical signals are, at best, indirect indicators of the state of the patient; the results of signal analysis need to be integrated with other clinical signs, symptoms, and results of laboratory tests, as well as nonmedical information by a physician or a clinical expert. The general background knowledge, expertise, and intuition of the medical specialist play important roles in arriving at the final diagnosis. Keeping in mind the realms of practice of various licensed and regulated professions, liability, and legal factors, the final diagnostic decision and communication with the patient are best left to the clinician, physician, or medical specialist. It is expected that quantitative and objective analysis facilitated by the application of biomedical signal analysis, medical informatics, CBR, and CAD will lead to improved diagnostic decision by the physician and improved healthcare for the patient.

It is evident from the material presented in this book that DSP techniques can assist in quantitative analysis of biomedical signals, that pattern recognition and classification techniques can facilitate

CAD, and that CAD systems can assist in achieving efficient diagnosis. Furthermore, CAD systems can also assist in designing optimal treatment protocols, in analyzing the effects of or response to treatment, and in clinical management of various abnormal conditions.

Several areas of application of computers in medicine exist beyond CAD in the quest to improve healthcare, some of which are:

- Computer-aided therapy and surgery;
- Computer-aided analysis of response to therapy;
- Computer-aided prognosis;
- Computer-aided risk assessment;
- Computer-aided patient management;
- Computer-aided clinical management;
- Computer-aided treatment protocol;
- Computer-aided personalized medicine.

Personalized medicine has been defined as healthcare that is informed and guided by an individual patient's unique clinical, genetic, genomic, and environmental information [89–91]; see related discussions in Section 7.8.4. Personalized medicine has been focused on understanding of normal physiology as well as pathology based on the “omics” to enhance preventive healthcare strategies and to guide therapeutic procedures. (*Note:* The omics include areas such as genomics, proteomics, and metabolomics.) The stated goal of personalized medicine is to optimize healthcare and the ensuing results for each patient; that is, to customize patient care. The requirement of multidisciplinary approaches to diagnosis and delivery of healthcare in order to achieve this goal has been emphasized, with particular attention not only to the omics but also to biomedical sensors, instrumentation, data analysis, computational methods, and diagnostic decision support systems. With the inclusion of a broad array of multidisciplinary techniques, it is evident that biomedical signal and image analysis can make important contributions to the realization of personalized medicine.

To emphasize the role of features and measures derived from radiological (medical) images in diagnosis and therapy, the term *radiomics* has been introduced [92–94]. The intent with this terminology is to promote the combined use of quantitative information derived from medical images and the various omics fields in diagnosis; it is also intended to establish a corresponding formal field of study and investigation. In a similar vein, we suggest the term *sinalomics* or *gramomics* to promote the incorporation of quantitative information derived from biomedical signals — the various -grams — in medical diagnosis. Indeed, biomedical signals and images have been used and are being applied to larger and larger extents in diagnosis and therapy. The terminology mentioned above could lead to formalization of the practice and further development of the related fields.

Biomedical signal analysis, biomedical image analysis, medical informatics, and CAD are proven techniques for improved healthcare. Biomedical engineering is an exciting multidisciplinary field that provides abundant opportunities to learn new areas of application of engineering, collaborate with professionals in other fields of research and investigation, develop broad perspectives and problem-solving skills, and, above all, contribute to the well-being of people!

10.16 Remarks

The subject of pattern classification is a vast area by itself. The topics presented in this chapter provide a brief introduction to the subject with several illustrations of application to biomedical signals.

We have now seen how biomedical signals may be processed and analyzed to extract quantitative features that may be used to classify the signals as well as to design diagnostic decision functions. Practical development of such techniques is usually hampered by a number of limitations related to the extent of discriminant information present in the signals selected or available for analysis, as well as the limitations of the features designed and computed. Artifacts inherent in the signal or caused by the signal acquisition systems impose further limitations.

A pattern classification system that is designed with limited data and information about the chosen signals and features will provide results that should be interpreted with due care. Above all, it should be borne in mind that the final diagnostic decision requires far more information than that provided by biomedical signals and their analysis: this aspect is best left to the physician or healthcare specialist in the spirit of computer-*aided* diagnosis.

10.17 Study Questions and Problems

1. The prototype vectors of two classes of signals are specified as Class 1: $[1, 0.5]$, and Class 2: $[3, 3]$. A new sample vector is given as $[2, 1]$. Plot the feature-vector space and describe your observations. Give the equations for two measures of similarity or dissimilarity, compute the measures for the sample vector, and classify the sample as Class 1 or Class 2 using each measure.

2. In a three-class pattern classification problem, the three decision boundaries are $d_1(\mathbf{x}) = -x_1 + x_2$, $d_2(\mathbf{x}) = x_1 + x_2 - 5$, and $d_3(\mathbf{x}) = -x_2 + 1$.

Draw the decision boundaries on a sheet of graph paper.

Classify the sample pattern vector $\mathbf{x} = [6, 5]$ using the decision functions.

3. Two pattern class prototype vectors are given to you as $\mathbf{z}_1 = [3, 4]$ and $\mathbf{z}_2 = [10, 2]$. Classify the sample pattern vector $\mathbf{x} = [4, 5]$ using (a) the normalized dot product, and (b) the Euclidean distance. Plot the feature-vector space and describe your observations.

4. A researcher makes two measurements per sample on a set of 10 normal and 10 abnormal samples. The set of feature vectors for the normal samples is

$$\{[2, 6], [22, 20], [10, 14], [10, 10], [24, 24], [8, 10], [8, 8], [6, 10], [8, 12], [6, 12]\}.$$

The set of feature vectors for the abnormal samples is

$$\{[4, 10], [24, 16], [16, 18], [18, 20], [14, 20], [20, 22], [18, 16], [20, 20], [18, 18], [20, 18]\}.$$

Plot the scatter diagram of the samples in both classes in the feature-vector space on a sheet of graph paper. Draw a linear decision function to classify the samples with the least error of misclassification. Write the decision function as a mathematical rule.

How many (if any) samples are misclassified by your decision function? Mark the misclassified samples on the plot.

Two new observation sample vectors are provided to you as $\mathbf{x}_1 = [12, 15]$ and $\mathbf{x}_2 = [14, 15]$. Classify the samples using your decision rule.

Now, classify the samples \mathbf{x}_1 and \mathbf{x}_2 using the k -nearest-neighbor method, with $k = 7$. Measure distances graphically on your graph paper plot and mark the neighbors used in this decision process for each sample.

Comment upon the results — whether the two methods resulted in the same classification result or not — and provide reasons.

5. A researcher makes measurements of RR intervals (in seconds) and FF for a number of ECG beats including (i) normal beats, (ii) PVCs, and (iii) normal beats with a compensatory pause (NBCP). The values (training set) are given in Table 10.9.

(a) Plot the $[RR, FF]$ feature-vector points for the three classes of beats. (b) Compute the prototype vectors for each class as the class means. Indicate the prototypes on the plot. (c) Derive linear discriminant functions (or decision functions) as the perpendicular bisectors of the straight lines joining the prototypes. State the decision rule(s) for each type of beat. (d) Three new beats are observed to have the parameters listed in Table 10.10. Classify each beat using the decision functions derived in part (c).

Normal Beats		PVCs		NBCPs	
RR	FF	RR	FF	RR	FF
0.700	1.5	0.600	5.5	0.800	1.2
0.720	1.0	0.580	6.1	0.805	1.1
0.710	1.2	0.560	6.4	0.810	1.6
0.705	1.3	0.570	5.9	0.815	1.3
0.725	1.4	0.610	6.3	0.790	1.4

Table 10.9 Training set of $[RR, FF]$ feature vectors.

Beat No.	RR	FF
1	0.650	5.5
2	0.680	1.9
3	0.820	1.8

Table 10.10 Test set of $[RR, FF]$ feature vectors.

6. For the training data given in the preceding problem, compute the mean and covariance matrices of the feature vectors for each class, as well as the pooled covariance matrix. Design a classifier based upon the Mahalanobis distance using the pooled covariance matrix.
7. You have won a contract to design a software package for CAD of cardiovascular diseases using the heart sound signal (PCG) as the main source of information. The main task is to identify the presence of murmurs in systole and/or diastole. You may use other signals for reference.
Propose a signal processing system to (i) acquire the required signals; (ii) preprocess them as required; (iii) extract at least two features for classification; and (iv) classify the PCG signals as: class 1, normal (no murmurs); class 2, systolic murmur; class 3, diastolic murmur; or class 4, systolic and diastolic murmur. Provide a block diagram of the complete procedure. Explain the reason behind the application of each step and state the expected results or benefits. Provide algorithmic details and/or mathematical definitions for at least two major steps in your procedure.
Draw a schematic plot of the feature-vector space and indicate where samples from the four classes listed above would fall. Propose a framework of decision rules to classify an incoming signal as belonging to one of the four classes.

10.18 Laboratory Exercises and Projects

Note: Data files related to the exercises are available at the site

<https://github.com/srikrishnan1972/Biomedical-Signal-Analysis>

1. The data file ecgpvc.dat contains the ECG signal of a patient with PVCs (see Figures 10.10 and 10.12). Refer to the file ecgpvc.m for details. Use the first 40% of the signal as training data to develop a PVC detection system (see Section 10.11). Develop code to segment the QRS – T portion of each beat using the Pan–Tompkins method (see Section 4.3.2), and compute the RR interval, QRS width (see Figure 4.6), and FF for each beat (see Section 5.6.4). Design linear discriminant functions using (i) RR and QRS width and (ii) QRS width and FF as the features; see Figure 10.11. Analyze the results in terms of TPF and FPF .
Code the decision function into your program as a classification rule. Test the pattern classifier program with the remaining 60% of the signal as the test signal. Compute the test-stage classification accuracy in terms of TPF and FPF .

2. Repeat the previous exercise by replacing the linear discriminant function with the k -nearest-neighbor method, with $k = 1, 3, 5$, and 7 . Evaluate the method with feature sets composed as (a) RR and QRS width, (b) QRS width and FF , and (c) RR , FF , and QRS width.
Compare the results of the three classifiers and provide reasons for any differences between them.
3. For the PVC detection system using a linear discriminant function classifier in Problem 1, prepare the confusion matrix. Derive the values of sensitivity, specificity, and $F1$ score. Repeat the experiment with a logistic regression classifier, and compare the sensitivity, specificity, and $F1$ score obtained with those of the linear discriminant function.
4. The VAG dataset (in the file VAG_Signals89Share.zip) consists of 50 normal cases and 39 abnormal cases. Develop code to derive a selection of the features described in Sections 5.12 and 6.6. Obtain classification results using 5-fold cross-validation with an SVM method with at least two different kernel functions, and comment on their performance. Compare the results of 5-fold cross-validation with results using the LOO method.
5. A set of pathological and normal voice signals is available in the file Voice_Patho.zip. Using the TFD and NMF methods as described in Section 9.7.3, derive sets of time-domain features, spectral-domain features, and joint TF features. Apply a classifier of your choice (for example, LDA) to the various sets of features with and without the inclusion of a method for feature selection. Compare the classification accuracies obtained using the various sets of features and methods.

References

- [1] Krishnan S. *Biomedical Signal Analysis for Connected Healthcare*. Academic Press, New York, NY, 2021.
- [2] Duda RO, Hart PE, and Stork DG. *Pattern Classification*. Wiley, New York, NY, 2nd edition, 2001.
- [3] Fukunaga K. *Introduction to Statistical Pattern Recognition*. Academic, San Diego, CA, 2nd edition, 1990.
- [4] Tou JT and Gonzalez RC. *Pattern Recognition Principles*. Addison-Wesley, Reading, MA, 1974.
- [5] Rushmer RF. *Cardiovascular Dynamics*. WB Saunders, Philadelphia, PA, 4th edition, 1976.
- [6] GE-Marquette Medical Systems, Inc., Milwaukee, WI. *Physician's Guide to Resting ECG Analysis Program, 12SL-tm*, 1991.
- [7] Strollo Jr PJ and Rogers RM. Obstructive sleep apnea. *New England Journal of Medicine*, 334(2):99–104, 1996.
- [8] Douglas NJ, Thomas S, and Jan MA. Clinical value of polysomnography. *The Lancet*, 339(8789):347–350, 1992.
- [9] Meyer TJ, Eveloff SE, Kline LR, and Millman RP. One negative polysomnogram does not exclude obstructive sleep apnea. *Chest*, 103(3):756–760, 1993.
- [10] Portier F, Portmann A, Czernichow P, Vascaut L, Devin E, Benhamou D, Cuvelier A, and Muir JF. Evaluation of home versus laboratory polysomnography in the diagnosis of sleep apnea syndrome. *American Journal of Respiratory and Critical Care Medicine*, 162(3):814–818, 2000.
- [11] Ahmadi N, Shapiro GK, Chung SA, and Shapiro CM. Clinical diagnosis of sleep apnea based on single night of polysomnography vs. two nights of polysomnography. *Sleep and Breathing*, 13(3):221–226, 2009.
- [12] Mendonça F, Mostafa SS, Ravelo-García AG, Morgado-Dias F, and Penzel T. Devices for home detection of obstructive sleep apnea: A review. *Sleep Medicine Reviews*, 41:149–160, 2018.
- [13] Zancanella E, do Prado LF, de Carvalho LB, Machado Júnior AJ, Crespo AN, and do Prado GF. Home sleep apnea testing: An accuracy study. *Sleep and Breathing*, 26(1):117–123, 2022.
- [14] Duda RO and Hart PE. *Pattern Classification and Scene Analysis*. Wiley, New York, NY, 1973.
- [15] Johnson RA and Wichern DW. *Applied Multivariate Statistical Analysis*. Prentice-Hall, Englewood Cliffs, NJ, 3rd edition, 1992.

- [16] Schürmann J. *Pattern Classification — A Unified View of Statistical and Neural Approaches*. Wiley, New York, NY, 1996.
- [17] Micheli-Tzanakou E. *Supervised and Unsupervised Pattern Recognition*. CRC Press, Boca Raton, FL, 2000.
- [18] Cabral TM and Rangayyan RM. *Fractal Analysis of Breast Masses in Mammograms*. Morgan & Claypool and Springer, 2012.
- [19] Rangayyan RM and Wu YF. Screening of knee-joint vibroarthrographic signals using statistical parameters and radial basis functions. *Medical and Biological Engineering and Computing*, 46(3):223–232, 2008.
- [20] Mu T, Nandi AK, and Rangayyan RM. Screening of knee-joint vibroarthrographic signals using the strict 2-surface proximal classifier and genetic algorithm. *Computers in Biology and Medicine*, 38(10):1103–1111, 2008.
- [21] Hearst MA, Dumais ST, Osuna E, Platt J, and Scholkopf B. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998.
- [22] Cortes C and Vapnik V. Support-vector networks. *Machine Learning*, 20:273–297, 1995.
- [23] Bendat JS and Piersol AG. *Random Data: Analysis and Measurement Procedures*. Wiley, New York, NY, 2nd edition, 1986.
- [24] Neter J, Kutner MH, Nachtsheim CJ, and Wasserman W. *Applied Linear Statistical Models*. Irwin, Chicago, IL, 4th edition, 1990.
- [25] SPSS Inc., Chicago, IL. *SPSS Advanced Statistics User's Guide*, 1990.
- [26] SPSS Inc., Chicago, IL. *SPSS Base System User's Guide*, 1990.
- [27] Pao YH. *Adaptive Pattern Recognition and Neural Networks*. Addison-Wesley, Reading, MA, 1989.
- [28] Lippmann RP. An introduction to computing with neural nets. *IEEE Signal Processing Magazine*, pages 4–22, April 1987.
- [29] Nigrin A. *Neural Networks for Pattern Recognition*. MIT Press, Cambridge, MA, 1993.
- [30] Shen L, Rangayyan RM, and Desautels JEL. Detection and classification of mammographic calcifications. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(6):1403–1416, 1993.
- [31] Haykin S. *Neural Networks: A Comprehensive Foundation*. Prentice-Hall PTR, Englewood Cliffs, NJ, 2nd edition, 2002.
- [32] Cover TM. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14(10):326–334, 1965.
- [33] Chen S, Cowan CFN, and Grant PM. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, 2(2):302–309, 1991.
- [34] Rangayyan RM and Wu YF. Analysis of vibroarthrographic signals with features related to signal variability and radial basis functions. *Annals of Biomedical Engineering*, 37(1):156–163, 2009.
- [35] LeCun Y, Bengio Y, and Hinton G. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [36] Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, VIII(4):283–298, 1978.
- [37] Metz CE. ROC methodology in radiologic imaging. *Investigative Radiology*, 21:720–733, 1986.
- [38] Swets JA and Pickett RM. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic, New York, NY, 1982.
- [39] The University of Chicago, Chicago, IL, <http://metz-roc.uchicago.edu/MetzROC/software>, accessed on 2023-04-26. *Metz ROC Software*.
- [40] Dorfman DD and Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals — Rating method data. *Journal of Mathematical Psychology*, 6:487–496, 1969.
- [41] Fleiss JL. *Statistical Methods for Rates and Proportions*. Wiley, New York, NY, 2nd edition, 1981.
- [42] Zar JH. *Biostatistical Analysis*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1984.

- [43] Krishnan S, Rangayyan RM, Bell GD, and Frank CB. Auditory display of knee-joint vibration signals. *Journal of the Acoustical Society of America*, 110(6):3292–3304, 2001.
- [44] Krishnan S. *Adaptive Signal Processing Techniques for Analysis of Knee Joint Vibroarthrographic Signals*. PhD thesis, Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB, Canada, June 1999.
- [45] Fukunaga K and Hayes RR. Effects of sample size in classifier design. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8):873–885, 1989.
- [46] Raudys SJ and Jain AK. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):252–264, 1991.
- [47] Durand LG, Blanchard M, Cloutier G, Sabbah HN, and Stein PD. Comparison of pattern recognition methods for computer-assisted classification of spectra of heart sounds in patients with a porcine bioprosthetic valve implanted in the mitral position. *IEEE Transactions on Biomedical Engineering*, 37(12):1121–1129, 1990.
- [48] Swain PH. Fundamentals of pattern recognition in remote sensing. In Swain PH and Davis SM, editors, *Remote Sensing: The Quantitative Approach*, pages 136–187. McGraw-Hill, New York, NY, 1978.
- [49] Bailar III JC and Mosteller F, editors. *Medical Uses of Statistics*. NEJM Books, Boston, MA, 2nd edition, 1992.
- [50] Walpole RE, Myers RH, and Myers SL, editors. *Probability and Statistics for Engineers and Scientists*. Prentice Hall, Upper Saddle River, NJ, sixth edition, 1998.
- [51] Ware JH, Mosteller F, Delgado F, Donnelly C, and Ingelfinger JA. P values. In Bailar III JC and Mosteller F, editors, *Medical Uses of Statistics*, pages 181–200. NEJM Books, Boston, MA, second edition, 1992.
- [52] Sahiner B, Chan HP, Petrick N, Wagner RF, and Hadjiiski L. Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size. *Medical Physics*, 27(7):1509–1522, July 2000.
- [53] Ramsey FL and Schafer DW. *The Statistical Sleuth — A Course in Methods of Data Analysis*. Wadsworth Publishing Company, Belmont, CA, 1997.
- [54] Banik S, Rangayyan RM, and Desautels JEL. *Computer-aided Detection of Architectural Distortion in Prior Mammograms of Interval Cancer*. Morgan & Claypool and Springer, 2013.
- [55] Draper NR and Smith H. *Applied Regression Analysis*. Wiley, Hoboken, NJ, 1998.
- [56] Kohavi R and John GH. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.
- [57] Farmer ME and Jain AK. A wrapper-based approach to image segmentation and classification. *IEEE Transactions on Image Processing*, 14(12):2060–2072, 2005.
- [58] Nandi RJ, Nandi AK, Rangayyan RM, and Scutt D. Classification of breast masses in mammograms using genetic programming and feature selection. *Medical and Biological Engineering and Computing*, 44:683–694, 2006.
- [59] Banik S, Rangayyan RM, and Desautels JEL. Digital image processing and machine learning techniques for the detection of architectural distortion in prior mammograms. In Suzuki K, editor, *Machine Learning in Computer-aided Diagnosis: Medical Imaging Intelligence and Analysis*, pages 23–65. IGI Global, Hershey, PA, 2012.
- [60] Mu T, Nandi AK, and Rangayyan RM. Classification of breast masses using selected shape, edge-sharpness, and texture features with linear and kernel-based classifiers. *Journal of Digital Imaging*, 21(2):153–169, June 2008.
- [61] Efron B and Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.
- [62] Liu K, Smith MR, Fear EC, and Rangayyan RM. Evaluation and amelioration of computer-aided diagnosis with artificial neural networks utilizing small-sized sample sets. *Biomedical Signal Processing and Control*, 8:255–262, 2013.

- [63] Moussavi ZMK, Rangayyan RM, Bell GD, Frank CB, Ladly KO, and Zhang YT. Screening of vibroarthrographic signals via adaptive segmentation and linear prediction modeling. *IEEE Transactions on Biomedical Engineering*, 43(1):15–23, 1996.
- [64] Krishnan S, Rangayyan RM, Bell GD, Frank CB, and Ladly KO. Adaptive filtering, modelling, and classification of knee joint vibroarthrographic signals for non-invasive diagnosis of articular cartilage pathology. *Medical and Biological Engineering and Computing*, 35(6):677–684, 1997.
- [65] Rangayyan RM, Krishnan S, Bell GD, Frank CB, and Ladly KO. Parametric representation and screening of knee joint vibroarthrographic signals. *IEEE Transactions on Biomedical Engineering*, 44(11):1068–1074, 1997.
- [66] Simon KC, Nadel L, and Payne JD. The functions of sleep: A cognitive neuroscience perspective. *Proceedings of the National Academy of Sciences*, 119(44):e2201795119, 2022.
- [67] Mendonça F, Mostafa SS, Ravelo-García AG, Morgado-Dias F, and Penzel T. A review of obstructive sleep apnea detection approaches. *IEEE Journal of Biomedical and Health Informatics*, 23(2):825–837, 2018.
- [68] SagaTech Electronics Inc., Calgary, Alberta, Canada, www.sagatech.ca, accessed on 2023-04-26. *Remmers Sleep Recorder*.
- [69] Quiceno-Manrique AF, Alonso-Hernandez JB, Travieso-Gonzalez CM, Ferrer-Ballester MA, and Castellanos-Dominguez G. Detection of obstructive sleep apnea in ECG recordings using time-frequency distributions and dynamic features. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5559–5562. IEEE, 2009.
- [70] Chen L and X Zhang. State-based general gamma CUSUM for modeling heart rate variability using electrocardiography signals. *IEEE Transactions on Automation Science and Engineering*, 14(2):1160–1171, 2015.
- [71] Khandoker AH, Karmakar CK, and Palaniswami M. Automated recognition of patients with obstructive sleep apnoea using wavelet-based features of electrocardiogram recordings. *Computers in Biology and Medicine*, 39(1):88–96, 2009.
- [72] Marcos JV, Hornero R, Alvarez D, del Campo F, and Zamarrón C. Assessment of four statistical pattern recognition techniques to assist in obstructive sleep apnoea diagnosis from nocturnal oximetry. *Medical Engineering and Physics*, 31(8):971–978, 2009.
- [73] Morales JF, Varon C, Deviaene M, Borzée P, Testelmans D, Buyse B, and Van Huffel S. Sleep apnea hypopnea syndrome classification in SpO₂ signals using wavelet decomposition and phase space reconstruction. In *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 43–46. IEEE, 2017.
- [74] Maali Y and Al-Jumaily A. Automated detecting sleep apnea syndrome: A novel system based on genetic SVM. In *2011 11th International Conference on Hybrid Intelligent Systems (HIS)*, pages 590–594. IEEE, 2011.
- [75] Ng AK, San Koh T, Baey E, Lee TH, Abeyratne UR, and Puvanendran K. Could formant frequencies of snore signals be an alternative means for the diagnosis of obstructive sleep apnea? *Sleep Medicine*, 9(8):894–898, 2008.
- [76] Madhav KV, Krishna EH, and Reddy KA. Detection of sleep apnea from multiparameter monitor signals using empirical mode decomposition. In *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, pages 1–6. IEEE, 2017.
- [77] Al-Angari HM and Sahakian AV. Automated recognition of obstructive sleep apnea syndrome using support vector machine classifier. *IEEE Transactions on Information Technology in Biomedicine*, 16(3):463–468, 2012.
- [78] Poewe W, Seppi K, Tanner CM, Halliday GM, Brundin P, Volkmann J, Schrag A, and Lang AE. Parkinson disease. *Nature Reviews Disease Primers*, 3(1):1–21, 2017.
- [79] Rueda A. *Non-linear and Non-stationary Speech Analysis of Parkinson's Disease Using Empirical Mode Decomposition*. PhD thesis, Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON, Canada, October 2021.

- [80] Ghoraani B and Krishnan S. A joint time-frequency and matrix decomposition feature extraction methodology for pathological voice classification. *EURASIP Journal on Advances in Signal Processing*, Vol. 2009:1–11, 2009.
- [81] de Oliveira Andrade A, Paixão APS, Cabral AM, Rabelo AG, Luiz LMD, Dionísio VC, Vieira MF, Pereira JM, Rueda A, Krishnan S, and Pereira AA. Task-specific tremor quantification in a clinical setting for Parkinson’s disease. *Journal of Medical and Biological Engineering*, 40(6):821–850, 2020.
- [82] Junaid M, Ali S, Eid F, El-Sappagh S, and Abuhmed T. Explainable machine learning models based on multimodal time-series data for the early detection of Parkinson’s disease. *Computer Methods and Programs in Biomedicine*, 234:107495, 2023.
- [83] Azevedo-Marques PM, Mencattini A, Salmeri M, and Rangayyan RM. Preface. In Azevedo-Marques PM, Mencattini A, Salmeri M, and Rangayyan RM, editors, *Medical Image Analysis and Informatics: Computer-Aided Diagnosis and Therapy*, pages xv–xx. CRC Press, Boca Raton, FL, 2018.
- [84] Doi K. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Proceedings of the IEEE*, 31(4–5):198–211, 2007.
- [85] Azevedo-Marques PM and Rangayyan RM. *Content-based Retrieval of Medical Images: Landmarking, Indexing, and Relevance Feedback*. Morgan & Claypool and Springer, 2013.
- [86] Feigen LP. Physical characteristics of sound and hearing. *The American Journal of Cardiology*, 28(2):130–133, 1971.
- [87] Butterworth JS and Reppert EH. Auscultatory acumen in the general medical population. *Journal of the American Medical Association*, 174(1):32–34, 1960.
- [88] Dobrow RJ, Calatayud JB, Abraham S, and Caceres CA. A study of physician variation in heart-sound interpretation. *Medical Annals of the District of Columbia*, 33:305–308, 1964.
- [89] Chan IS and Ginsburg GS. Personalized medicine: Progress and promise. *Annual Review of Genomics and Human Genetics*, 12:217–244, 2011.
- [90] Alyass A, Turcotte M, and Meyre D. From big data analysis to personalized medicine for all: Challenges and opportunities. *BMC Medical Genomics*, 8(1):1–12, 2015.
- [91] Ahmed MU, Saaem I, Wu PC, and Brown AS. Personalized diagnostics and biosensors: A review of the biology and technology needed for personalized medicine. *Critical Reviews in Biotechnology*, 34(2):180–196, 2014.
- [92] Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, and Bellomi M. Radiomics: The facts and the challenges of image analysis. *European Radiology Experimental*, 2(1):1–8, 2018.
- [93] Van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, and Baessler B. Radiomics in medical imaging – “how-to” guide and critical reflection. *Insights into Imaging*, 11(1):1–16, 2020.
- [94] Shur JD, Doran SJ, Kumar S, Ap Dafydd D, Downey K, O’Connor JPB, Papanikolaou N, Messiou C, Koh DM, and Orton MR. Radiomics in oncology: A practical guide. *Radiographics*, 41(6):1717–1732, 2021.

Index

- accuracy, 620, 621
action potential, 5, 7, 8
 cardiac myocyte, 8, 405
 ECG, 23
 model, 405
 neuron, 9, 11, 406
 propagation, 10, 405
activity, 283, 286, 437
ad hoc filter, 171
adaptive decomposition, 515
adaptive filter, 180, 193, 197, 199
 segmentation, 452
adaptive noise canceler, 181, 197, 199
adaptive segmentation, 445, 447, 452, 454, 460, 463, 485
airflow, 292
Akaike criterion, 377
all or nothing, 7
alpha rhythm, *see* EEG, 33
alpha-trimmed mean filter, 170
amplitude modulation, 278
analytic signal, 281, 521
apnea, 79
arrhythmia, 22, 75
artifact, 91
 physiological, 57, 98
 power-line, 98
 random, 92
 structured, 98
artificial intelligence, 59
atrial electrogram, 73
autocorrelation, 96, 100, 174, 187, 321, 437
 distance, 450, 452
 estimation, 321, 330
autocorrelation method, 371
averaging of envelopes, 280
averaging of power spectral density, 325
back-propagation algorithm, 616
bandpass filter, 148, 319
Bartlett method, 325, 328, 332
Bartlett window, 324–326
baseline drift, 102
 removal, 145, 149, 152, 161, 196, 289
Bayes classifier, 611, 613, 632, 637
Bayes formula, 611
beta rhythm, *see* EEG, 33
bias, 323–325, 339
bilinear TFDs, 479
bilinear transformation, 154, 155
bioacoustic signals, 52
biphasic, 15
Blackman window, 327
blind source separation, 539, 542
blood pressure, 2, 44, 45
bootstrap, 632
brain
 anatomy, 29, 32
 waves, *see* EEG
brain-computer interface, 518, 577
brain-machine interface, 472, 505
bundle-branch block, 257, 268, 596
Burg-lattice method, 457
Butterworth highpass filter, 161, 196
Butterworth lowpass filter, 154, 196, 218, 220, 253, 282, 285, 633

- CAD
 limitations, 650
 strengths, 650
 why use, 58
- cardiac cycle, 21, 78, 214
- cardiac rhythm, 21, 358
- cardiorespiratory interaction, 75
- carotid pulse, 40, 43, 215, 228, 255
 dicrotic notch, 41, 43, 73, 215, 255
 detection, 228
 filtering, 158
 relation to ECG, 37, 78
 relation to PCG, 37, 73, 78
- catheter-tip signals, 43
- causality, 109
- cell, 5, 6, 8, 9, 405
- centroidal time, 270, 272
- cepstral prediction, 393
- cepstrum
 complex, 247, 271, 393
 power, 250
 real, 250, 272, 376
 relation to autoregressive model, 384
- chirp, 528, 535, 561
- chondromalacia patella, 49, 434, 435
- clinical parameters, 641, 642
- cluster seeking, 607
 K-means, 610
 maximin-distance, 609
- coefficient of variation, 94
- coherence, 235
- coherent averaging, *see* synchronized averaging
- comb filter, 162, 196
- compensatory pause, 288
- complexity, 286, 302
- computer-aided diagnosis, *see* CAD, 595
- concurrent processes, 71
- conduction velocity, 10–12
- conjugate-symmetry, 130, 350
- contingency table, 625
- convolution, 107–109, 112, 113, 131, 132
 circular, 131–134
 linear, 109, 133, 134
 periodic, 131–134
 using the DFT, 131
- coronary artery
 disease, 421
 sounds, 400
 stenosis, 400
 tree, 400
- correlated processes, 71
- correlation analysis, 228
- correlation coefficient, 96, 98, 137, 229, 270, 289, 292
- correlation filter, 237, 479
- cost analysis, 620
- coupled processes, 71
- covariance, 96, 98, 538
- covariance matrix, 608
- covariance method, 374
- CPR
 wavelet analysis, 494
- cross-correlation, 98, 174, 187, 225, 229, 241
- cross-spectral density, 176, 235
- cross-validation, 632
 k-fold, 632
 leave-one-out, 632
- cry melody, 493
- cumulants, 438
- cyclostationary signal, 101, 325, 332, 358
- decision function, 600, 633, 634
- decision making, 596
- decision rule, 597, 634
- decision tree, 281
- decomposition, 125
 adaptive, 515
- deconvolution, 244
 speech, 251
- deep learning, 59, 620
- defibrillation, 494
- delta function, 107–109
 sifting property, 108
- delta rhythm, *see* EEG, 33
- demodulation
 amplitude, 278
 asynchronous, 279
 complex, 279, 282
 synchronous, 279
- depolarization, 6
- derivative, 146, 218, 219, 222, 227, 228, 285, 286
 three-point central difference, 148, 218
- deterministic signal, 92
- DFT, 126, 130
- diagnostic accuracy, 620
- diagnostic decision, 59, 595, 596, 657
- diastole, 20, 38, 78, 254
- dicrotic notch, *see* carotid pulse: dicrotic notch
- dictionary learning, 523, 553
- difference equation, 124, 139, 155, 366
- differentiation, *see* derivative
- difficulties in signal analysis, 55
- digital twins, 415
- Dirac delta function, 107
- discriminant function, 275, 600, 601, 633
- dispersion of energy, 270
- distance, 605, 607
 between probability density functions, 628
 Euclidean, 195, 607
 Jeffries–Matusita, 629
 Kullback–Leibler, 302
 Mahalanobis, 608
- divergence, 628
- dominant frequency map, 501, 504
- dot product, 128, 174, 229, 615
 normalized, 608
- Durbin’s method, 373
- dynamic system, 44, 56, 93, 99, 100, 192, 435, 452, 463
- ECG, 20
 12-lead, 23, 30, 31
 action potential, 23, 405
 bigeminy, 23, 288, 633
 bundle-branch block, 23, 26, 268, 596

- classification, 633
 Einthoven's triangle, 24
 electrode positions, 24, 27, 28
 exercise, 288
 fetal, 105, 197, 541, 572
 filtering, 145, 148, 152, 161, 165, 177, 196, 197
 form factor, 288
 interval series, 359
 introduction, 20
 leads, 23
 maternal, 105, 197, 541, 572
 cancellation, 197
 myocardial ischemia, 43, 268, 290
 P wave, 21, 214
 detection, 224, 226
 power spectral density, 104, 196, 574
 PQ segment, 22, 214
 PVC, 22, 43, 268, 273, 287, 597
 classification, 633
 compensatory pause, 74
 QRS wave, 22, 214
 detection, 218, 220, 226
 relation to atrial electrogram, 73
 relation to carotid pulse, 37, 78
 relation to PCG, 37, 72, 78
 rhythm analysis, 253, 358, 416
 RR interval, 73, 253, 288, 317, 358, 416
 ST segment, 22, 214, 288
 synchronized averaging, 136
 T wave, 22, 214
 detection, 226
 template matching, 136
 trigger, 136, 280
 waveform analysis, 273, 274
 wavelet analysis, 499
 waves, 21, 214
 Wilson's terminal, 24
 echo, 240, 249, 250
 ectopic beats, *see* ECG: PVC
 EEG, 29
 alpha rhythm, 33, 228, 346, 598
 detection, 230, 237
 power spectral density, 328, 378
 autoregressive model, 378
 BCI, 577
 coherence analysis, 235
 correlation analysis, 228
 delta rhythm, 345
 description of a record, 217, 485, 486
 dictionary learning, 553
 electrode positions, 30
 form factor, 286
 introduction, 29
 power spectral density, 486
 rhythms, 33, 215, 234, 236
 detection, 228
 segmentation, 433, 447, 452, 485
 seizure, 234, 503, 553
 sleep, 345
 spectral analysis, 345
 spike, 217
 spike-and-wave, 217, 488
 detection, 231, 241
 state diagram, 486
 theta rhythm, 345
 transients, 215
 detection, 486
 wavelet analysis, 503
 waves, 215
 EGG, 36
 Einthoven's triangle, 24
 electrocardiogram, *see* ECG
 electroencephalogram, *see* EEG
 electrogastrogram, *see* EGG
 electromyogram, *see* EMG
 electroneurogram, *see* ENG
 EMD, 520
 EMD variants, 521
 EMG, 12, 14
 envelope, 285, 294
 fractal analysis, 305
 interference pattern, 17, 268, 362
 introduction, 12
 motor unit firing pattern, 358
 muscle force, 18, 19, 290, 291, 294, 302
 point process model, 360
 principal component analysis, 539
 relation to VMG, 77
 respiration, 292
 root-mean-squared value, 285, 295
 spectral analysis, 362
 turns count, 285
 zero-crossing rate, 285
 empirical mode decomposition, 520, 553
 energy distribution, 270, 277, 284
 moments of, 270
 ENG, 10
 ensemble averages, 95, 96, 99, 135, 326, 361
 ensemble averaging, *see* synchronized averaging
 ensemble EMD, 522
 entropy, 94, 299
 Shannon, 95
 envelopogram, 281
 envelope, 269, 287
 EMG, 285, 294
 extraction, 277
 PCG, 280
 spectral, 362, 364, 376
 epoch, 213
 ergodic proces, 99
 ERP, 35, 102, 193, 269
 synchronized averaging, 96, 136, 193
 errors
 interobserver, 59, 653
 intraobserver, 59, 653
 types of, 653
 estimation error, 173
 ethics approval, 54, 58, 297
 Euclidean distance, 195, 607
 even part, 135, 272
 even-symmetry, 134, 176, 179, 326, 330, 335, 350
 event detection, 213
 event-related potential, *see* ERP
 events

- in ECG, 214
- in EEG, 215
- in PCG, 215
- evoked potential, *see* ERP
- expectation operator, 94
- expert system, 59
- exponential sequences, 247
- exponential signals, 393
- F1 score, 622
- false negative, 621
- false positive, 621
- fatigue, 17, 57, 58, 195, 306
- feature selection, 630
 - sequential backward selection, 631
 - sequential forward selection, 631
 - stepwise regression, 631
- feature vector, 599
- fetal ECG, 517
 - extraction, 541, 572
- FFT, 130
- filter, 106, 109, 110
 - ad hoc, 171
 - adaptive, 180, 193, 197, 199
 - alpha-trimmed mean, 170
 - bandpass, 148
 - Butterworth highpass, 161, 196
 - Butterworth lowpass, 154, 196, 218, 220
 - comb, 162, 196
 - correlation, 237
 - derivative-based, 145
 - finite impulse response, 141
 - frequency-domain, 153, 193
 - fundamentals, 106
 - generalized linear, 244
 - Hann, 140
 - highpass, 146, 150, 161, 221
 - homomorphic, 242, 395
 - infinite impulse response, 150, 155
 - inverse, 375, 385
 - Kalman, 463
 - L, 170
 - least-mean-squares, 184, 199
 - linear shift-invariant, 107
 - lowpass, 141, 144, 154, 220, 221
 - matched, 237
 - max, 169
 - mean, 171, 172
 - median, 171, 172
 - min, 169
 - min/max, 170
 - moving-average, 114, 139, 192, 219, 277, 326
 - relation to integration, 143
 - nonlinear, 169
 - nonrecursive, 141
 - notch, 162, 218
 - optimal, 171, 181, 185, 186, 463
 - order-statistic, 169
 - ramp, 114–116
 - recursive, 141
 - recursive least-squares, 185, 199, 452
 - recursive least-squares lattice, 456, 460
 - selection of, 190
 - specifications, 152
 - time-domain, 139
 - transform domain, 117
 - whitening, 375
 - Wiener, 171, 185, 193, 198, 371, 391

finite impulse response filter, 141

Fisher linear discriminant analysis, 601

folding frequency, 121

forgetting factor, 185, 186

form factor, 286, 287, 299

formants, 46, 360, 394, 443

forward–backward prediction, 456

Fourier transform, 117, 122, 125

 - basis functions, 127–129
 - convolution, 131
 - discrete, 126
 - fast, 130
 - properties, 133
 - short-time, 438, 439

fractal analysis, 302, 339

 - box-counting, 304
 - EMG, 305
 - frequency domain, 339
 - Higuchi's method, 304
 - physiological signals, 304
 - ruler method, 303
 - VAG, 339

fractal dimension, 303, 339

fractals, 302

fractional Brownian motion, 339

frequency modulation, 529, 535, 561

frequency response, 117, 124, 141, 143, 144, 146, 150, 151, 154, 158, 161, 165, 176, 197

frequency-domain analysis, 317

frequency-domain filter, 153, 193

Gabor atoms, 519

gamma rhythm, 33

Gaussian model, 300, 301

Gaussian noise, 113–115

Gaussian probability density function, 613

generalized likelihood ratio, 450, 452

generalized linear filtering, 244

glottal waveform, 245

gold standard, 622

gramomics, 656

ground, 23, 136

ground truth, 622

group delay, 482, 485

habituation, 194

Hamming window, 327

Hann filter, 140

Hann window, 327

heart

 - anatomy, 20
 - electrical system, 21
 - model, 410
 - prosthetic valves, 337
 - sounds, *see* PCG
 - valvular defects, 319

- heart rate, 21, 224, 253
 spectrogram, 490
 variability, 82, 83, 359, 416, 490
- heart-rate variability
 sleep apnea, 645
- higher-order moments, 438
- higher-order statistics, 335, 437
- highpass filter, 146, 150, 161, 221, 365
- Hilbert transform, 281, 501, 521
- histogram, 297, 300, 301
- Hodgkin–Huxley model, 406
- homomorphic deconvolution, 244
- homomorphic filtering, 242, 395
- homomorphic prediction, 393
- human–instrument system, 54
- Hurst coefficient, 340
- impulse function, 107–109, 528, 535
- impulse response, 107, 108, 110, 116, 140, 142, 179, 180
- impulsive noise, 169–171
- independent component analysis, 539
 comparison, 546
- infinite impulse response filter, 150, 155
- inner product, *see* dot product
- innervation ratio, 13
- instantaneous mean frequency, 482, 501, 561
- instrumentation, 52
- integration, 143, 222
- interference, 91, 101
 maternal, 105
 cancellation, 197
 muscle-contraction, 77, 105
 removal, 185, 189, 199
- nonstationary, 181
- physiological, 98
- power-line, 98
- interference pattern, 17
- interobserver error, 59, 653
- interpulse interval, 358, 360, 363
- interval series, 359, 416
- intraobserver error, 59, 653
- intrinsic mode function, 520, 557, 559
- inverse filter, 375, 385
- inverse linear prediction, 389
- inverse of a signal, 271
- isoelectric segment, 21, 22, 177, 214, 227, 276, 290
- isovolumic contraction, 37, 319
- Jeffries–Matusita distance, 629
- k-fold cross-validation, 632
- K-means classifier, 637
- K-means clustering, 610
- k-nearest-neighbor, 605
- Kalman filter, 463, 505
 control of prostheses, 505
- Karhunen–Loëve transform, 475, 533
- knee joint, 48
 anatomy, 48, 433
 arthroscopy, 434, 641
 cartilage pathology, 433
- detection, 637
- crepitus, 402
- sound generation model, 402
- sounds, *see* VAG
- Kullback–Leibler distance, 302
- kurtosis, 94, 299, 336, 541
- L filter, 170
- lacunarity, 503
- Laplace domain, 118, 122, 154
- Laplace transform, 117
- least-mean-squares filter, 184, 199
- least-squares method, 370
- leave-one-out method, 632
- length transformation, 226
- Levinson–Durbin algorithm, 373, 459
- likelihood function, 611
- limitations of manual analysis, 654
- linear discriminant analysis, 275, 600, 633
- linear prediction, 368
 inverse, 389
 inverse model, 395
- linear shift-invariant filters, 107
- logistic regression, 614
- loss function, 611
- lowpass filter, 141, 144, 154, 220, 221, 277
- machine learning, 59
- magnitude response, 117
- Mahalanobis distance, 608
- manual to computer analysis, 654
- manual versus computer analysis, 654
- matched filter, 237, 479
- matching pursuit, 518, 526
 time–frequency distribution, 526
- max filter, 169
- maximin-distance clustering, 609
- maximum-phase component, 249, 271
- maximum-phase signal, 249, 385, 393
- McNemar’s test of symmetry, 625
- mean, 93, 99, 322, 325, 436
- mean filter, 171, 172
- mean frequency, 335, 485
 instantaneous, 482, 561
- mean-squared error, 538
- mean-squared value, 94
- measures of accuracy, 620
- median filter, 171, 172
- median frequency, 335
- Mexican hat wavelet, 478
- min filter, 169
- min/max filter, 170
- minimum phase, 271
- minimum-phase component, 249, 271
- minimum-phase correspondent, 270, 272
- minimum-phase signal, 249, 385, 393
- mixed phase, 271
- mixed-phase signal, 249, 385, 393
- mixing matrix, 541
- mobility, 286
- mode, 93, 94
- model

- action potential, 405
- all-pole, 369
- autoregressive, 367, 369
 - EEG, 378
 - optimal order, 377
 - parameters, 377
 - PCG, 378, 418
 - relation to cepstrum, 384
- autoregressive moving-average, 367, 385
- electromechanical, 395
- electrophysiological, 404
- Hodgkin–Huxley, 406
- linear prediction, 368
- linear system, 366
- moving-average, 367
- piecewise linear, 177
- pole-zero, 368
 - speech, 394
- respiration, 396
- time-varying, 436, 453, 454, 456
- modeling, 357
- modulation
 - amplitude, 278
 - frequency, 529, 535, 561
- moments, 94, 270, 272, 299, 480, 560
 - first-order, 94, 99
 - higher-order, 437
 - of energy distribution, 270
 - of power spectral density, 334
 - second-order, 94, 100
- monophasic, 15
- Morlet function, 477
- morphological analysis, 269
- motion artifact, 102
- motor unit, 12–14
 - firing pattern, 358
 - recruitment, 16, 17
- motor-unit action potential, *see* MUAP
- moving-average filter, 114, 139, 192, 219, 277, 326
 - relation to integration, 143
- MUAP, 13, 15, 16, 358, 360
 - biphasic, 13, 362
 - polyphasic, 15
 - triphasic, 13
- multichannel signal, 71, 515
- multicomponent signal, 432, 477, 482, 515, 528, 561
- multimodal, 336
- multiorgan analysis, 82
- multisource signal, 515
- multivariate analysis, 82
- multivariate EMD, 523
- muscle contraction, 16, 294
 - VMG, 52
- muscle sounds, *see* VMG
- muscle-contraction interference, 57, 77, 105
 - removal, 185, 189, 199
- myocardial elasticity, 318
- myocardial infarction, 268, 319
- myocardial ischemia, 268, 290
- myocyte, 7, 8, 405
- nearest-neighbor rule, 605
- negative predictive value, 622
- nerve
 - action potential, 9, 11
 - conduction velocity, 11, 12
- neural decoding, 505
- neural networks, 59, 615
- neuron, 9, 10
 - action potential, 9, 11, 406
- Newton–Raphson procedure, 386
- NMF, 542
 - algorithms, 542
 - comparison, 546
- noise, 91
 - random, 92
 - Gaussian, 113, 115
 - high-frequency, 102
 - removal, 154, 196
 - impulsive, 169–171
 - in ECG, 102, 136, 177
 - in ERP, 102, 136
 - low-frequency, 102
 - removal, 145, 149, 152, 161
 - physiological, 98
 - power, 195
 - removal, 135, 139, 176, 181, 190
 - shot, 170, 171
 - structured, 98
 - uniform, 172
 - nondeterministic signal, 92
 - nonlinear filters, 169
 - nonnegative matrix factorization, 542, 569, 577
 - nonrecursive filter, 141
 - nonstationary process, 44, 50, 56, 93, 99, 106, 181, 185, 331, 431
 - normal equation, 175, 187, 371–373, 385
 - notch filter, 162, 218
 - Nyquist frequency, 121
 - objective analysis, 59, 654
 - objectives of signal analysis, 52
 - odd part, 135
 - odd-symmetry, 134, 350
 - off-line analysis, 59
 - on-line analysis, 59
 - optimal filter, 171, 181, 185, 186, 463
 - order-statistic filters, 169
 - orthogonal, 96
 - osteoarthritis, 434
 - otoacoustic emission, 52
 - outer product, 174, 230
 - outliers, 170
 - p-value, 629
 - Pan–Tompkins algorithm, 220, 253, 255, 282, 288, 633
 - parallel
 - systems in, 116, 120, 121
 - parametric analysis, 423
 - parametric modeling, 357, 365
 - Parkinson’s disease, 647
 - Parseval’s theorem, 134, 334, 375
 - Parzen window, 297
 - pattern classification, 595, 596, 599

- ECG, 633
- reliability, 627
- supervised, 600
- test set, 600, 635
- test step, 631, 635
- training set, 600, 627, 633
- training step, 631, 633
- unsupervised, 607
- VAG, 295, 339, 637
- PCG, 37
 - aortic component, 256
 - aortic stenosis, 321, 337
 - autoregressive model, 378, 418, 421
 - bundle-branch block, 257
 - coronary artery disease, 421
 - envelope, 269, 280
 - first heart sound, 37, 85, 215, 254, 318
 - detection, 72, 255, 420
 - introduction, 37
 - murmur, 38, 215, 280, 284, 285, 337, 598
 - decision tree, 281
 - spectral analysis, 378, 418
 - myocardial elasticity, 318
 - myocardial infarction, 319
 - power spectral density, 319, 337, 421, 439
 - prosthetic heart valves, 337
 - pulmonary component, 256
 - relation to carotid pulse, 37, 73, 78
 - relation to ECG, 37, 72, 78
 - second heart sound, 37, 215, 254, 256, 318
 - detection, 73, 255, 420
 - split, 39, 257, 282
 - segmentation, 72, 73, 78, 254, 332, 419, 432, 489
 - spectral analysis, 318, 319, 332, 378, 418
 - spectrogram, 320, 439
 - synchronized averaging, 280
 - zero-crossing rate, 285
 - peak frequency, 336, 337
 - peak searching, 220, 223
 - perceptron, 615
 - performance criterion, 173, 174, 186
 - periodogram, 323
 - permeability, 5
 - personalized medicine, 415, 656
 - phase
 - linear component, 134, 247
 - unwrapping, 247, 384
 - phase map, 501, 504
 - phase response, 117
 - phonocardiogram, *see* PCG
 - photoplethysmogram, *see* PPG
 - photoplethysmography, *see* PPG
 - physiological interference, 98
 - physiological system, 1
 - pitch, 46, 360
 - point process, 358, 360
 - pole-zero model, 368, 384, 387
 - pole-zero plot, 118, 123, 124, 143, 144, 151, 158, 164, 165, 453, 454
 - poles, 117, 141, 149, 154, 368, 387, 393, 452
 - dominant, 378, 418
 - EEG, 378
 - PCG, 378, 418
 - polyphasic, 16
 - polysomnography, 80
 - positive predictive value, 622
 - power law, 303–305, 339
 - power spectral density, 98, 100, 176, 235, 241, 318, 437
 - averaging, 325
 - Bartlett estimate, 325, 328, 332
 - ECG, 196
 - EEG, 328, 486
 - EMG, 362
 - estimation, 321
 - heart rate, 490
 - moments, 334
 - parameters, 333, 338
 - PCG, 319, 332, 439
 - point process, 363
 - VAG, 365
 - Welch estimate, 326, 329
 - power-line interference, 98, 102, 103, 177, 197
 - removal, 162, 196
 - PPG, 41
 - introduction, 41
 - wavelet analysis, 493
 - pressure, 2
 - in coronary artery, 400
 - intracardiac, 44
 - principal component analysis, 533
 - comparison, 546
 - probabilistic models, 611
 - Bayes formula, 611
 - conditional probability, 611
 - likelihood function, 611
 - posterior probability, 611
 - prior probability, 611
 - probability density function, 93
 - divergence, 628
 - estimation of, 297
 - Gaussian, 613
 - normalized distance, 628
 - of VAG signals, 300, 301
 - projection, 125, 128, 229
 - prostheses
 - control, 505
 - prototype, 97, 605, 634
 - PVC, *see* ECG: PVC
 - QRS, *see* ECG: QRS wave
 - quadratic discriminant analysis, 601
 - qualitative analysis, 58, 654
 - quantitative analysis, 59, 654
 - quasiperiodic process, 44, 56, 136, 358
 - quasistationary process, 100, 438
 - radial basis functions, 617
 - radiomics, 656
 - ramp filter, 114–116
 - random noise, 92
 - random signal, 92, 371
 - randomness

- test for, 92, 285
- rational function, 123, 393
- real-time analysis, 59
- receiver operating characteristics, 623
- recruitment of motor units, 16, 17
- rectangular window, 324, 327
- rectification, 225, 277, 279, 280, 285, 294
- recursive filter, 141
- recursive least-squares filter, 185, 199, 452
- recursive least-squares lattice filter, 456, 460
- reflection coefficient, 373, 459
- refractory period, 7
- repolarization, 6
- resonance, 336, 397
- resonance frequency, 317, 378, 394, 400, 418, 422, 452
- respiration
 - analysis of, 292
 - model, 396
- respiratory signal, 43
- resting potential, 5
- rhythm
 - alpha, 33
 - beta, 33
 - delta, 33
 - gamma, 33
 - mu, 217
 - spike-and-wave, 217
 - theta, 33
- rhythm analysis, 73, 317
- rhythms
 - in EEG, 215, 236, 345
- risk–benefit analysis, 54, 58
- Rogers–McCulloch model, 413
- root-mean-squared value, 95, 283, 287, 295
- rotor mechanism, 499, 501
- s-plane, 118, 122, 154
- saltatory conduction, 10
- sample statistics, 95
- scale space, 476
- Schwarz’s inequality, 238
- screening, 620
- search-back procedure, 224
- segmentation
 - adaptive, 445, 452
 - EEG, 447, 452, 485
 - fixed, 438
 - PCG, 489
 - VAG, 456, 463
- seizure, 236, 503, 553
- self-affinity, 340
- self-similarity, 302–305, 339–341
- sensitivity, 621
- SEP, *see* ERP
- separability of features, 628
- separation of signals, 531
- series
 - systems in, 115, 120, 121
- Shanks’ method, 387, 394
- shift
 - circular, 131
- linear, 111
- periodic, 131
- short-time analysis, 100, 284–286, 418, 438, 439, 441, 493
 - window, 441
- short-time Fourier transform, 439
- shot noise, 170, 171
- sifting property, 108
- signal length, 270, 272
- signal-flow diagram, 139, 140, 150, 151, 155, 156, 173, 370
- significance
 - statistical, 630
- signum function, 282
- sinalomics, 656
- single-motor-unit action potential, *see* MUAP
- skewness, 94, 299, 336
- sleep apnea, 79, 598, 644
- sleep EEG, 345
- SMUAP, *see* MUAP
- snoring, 80
- SNR, 95, 135, 183, 195
- somatosensory evoked potentials, *see* ERP
- spatiotemporal recruitment, 16, 17
- specificity, 621
- spectral analysis
 - EEG, 345, 378
 - EMG, 362
 - PCG, 378, 418
 - point process, 363
 - VAG, 363, 560
- spectral contours, 419
- spectral envelope, 376
- spectral error measure, 445, 446, 452
- spectral leakage, 324–326
- spectral matching, 374
- spectral parameters, 333, 338, 374
- spectral power ratios, 321, 337
- spectral resolution, 324–326
- spectral tracking, 420
- spectral windows, 326, 328
- spectrogram, 439, 493
 - cry melody, 494
 - heart rate, 490
 - PCG, 320, 439
 - speech, 100, 442
 - VAG, 190, 199, 456
- speech, 44, 93, 99, 285
 - formants, 359, 394
 - homomorphic filtering, 251
 - in Parkinson’s disease, 648
 - introduction, 44
 - phoneme, 44
 - pitch, 359
 - pole–zero model, 394
 - spectrogram, 100, 442
 - voiced, 245, 251
- speech system, 46
- spike, 33, 217
- spike-and-wave, 33, 217
- standard deviation, 94
- state-space model, 463

- stationary process, 99, 175
 statistical analysis, 93, 295, 334, 436
 statistical decision, 611
 statistical significance, 630
 statistically independent processes, 96, 176, 184
 statistics
 ensemble, 95
 filter, 169
 measures, 93
 order, 169
 sample, 95
 temporal, 97, 98
 steepest descent, 184
 Steiglitz-McBride method, 389
 step function, 107, 109, 110, 121
 stick-slip model, 404
 structured noise, 98
 subjective analysis, 58, 654
 support vector machine, 606
 SVM, 606
 synchronized averaging, 135, 192, 193, 258, 280, 282, 331
 system identification, 389
 systole, 20, 38, 78, 254
 t-test, 629
 T-wave alternans, 518, 568
 detection, 568
 tap weights, 139, 141, 173, 183, 187, 452
 temperature, 1
 template matching, 98, 136, 225, 230, 231, 241, 270
 temporal statistics, *see* time averages
 test set, 632, 635, 636
 test step, 631, 635
 theta rhythm, 33
 three-point central difference, 148
 thresholding, 223, 225, 475
 time averages, 97, 99, 139
 time shift, 134
 time-bandwidth product, 441
 time-frequency analysis, 474, 560
 time-frequency atoms, 477, 519, 559
 time-frequency distribution, 439, 479, 493
 adaptive, 526
 time-scale space, 476
 time-domain filter, 139
 time-invariant system, 366
 time-variant system, 359, 435, 436, 453, 454, 456
 training set, 632, 634, 635, 640, 641
 training step, 631, 633
 transfer function, 117, 119, 124, 140, 154, 155, 368, 370
 transform-domain analysis, 117
 triangular window, 324–327
 triphasic, 15
 true negative, 621
 true positive, 621
 turning points, 93, 285, 286
 turns count, 285–287, 299
 uncertainty principle, 441
 uncorrelated, 96, 98
 uniform noise, 172
 unimodal, 93, 336
 VAG, 48, 105, 433
 classification, 295, 339, 604, 620, 626, 637
 crepitus, 360, 363
 empirical mode decomposition, 521
 filtering, 185, 189, 199
 fractal analysis, 339
 introduction, 48
 muscle-contraction interference, 50, 105
 point process model, 363
 probability density function, 300, 301
 relation to VMG, 77
 segmentation, 456, 463
 signal acquisition, 50, 297
 spectrogram, 190, 199, 456
 statistical analysis, 295
 time-frequency analysis, 560
 validation set, 632
 validation step, 632
 variance, 94, 99, 283, 286, 323, 326, 335, 436
 ventricular fibrillation, 494
 vibroarthrogram, *see* VAG
 vibromyogram, *see* VMG
 VMG, 52, 105
 introduction, 52
 muscle force, 294
 relation to EMG, 77
 relation to VAG, 77
 root-mean-squared value, 295
 vocal tract, 44, 45
 vocal-tract response, 245, 251
 waveform analysis
 ECG, 273, 274, 287
 waveform complexity, 267, 283
 wavelet analysis, 503
 wavelet denoising, 493
 wavelet packet, 477, 528
 wavelet transform, 475
 wavelets, 474, 478–480, 553
 waveshape, 267
 weighting function, 252
 Welch method, 326, 329
 whitening filter, 375
 Widrow-Hoff algorithm, 185
 Wiener filter, 171, 185, 193, 198, 371, 373, 391
 Wiener-Hopf equation, 175, 184, 186
 Wigner-Ville distribution, 482
 Wilson's terminal, 24
 window
 Bartlett, 324–327
 Blackman, 327
 frequency response, 327, 329
 Hamming, 327
 Hann, 327, 329
 in short-time analysis, 441
 rectangular, 324, 327, 328, 332
 triangular, 324–327
 wrapper, 631
 z-domain, 122–124
 z-transform, 119
 inverse, 247, 251, 390
 zero padding, 133, 252
 zero-crossing rate, 283, 285, 286
 zeros, 117, 140, 143, 148, 162, 368, 385, 387, 393

IEEE Press Series in Biomedical Engineering

The focus of our series is to introduce current and emerging technologies to biomedical and electrical engineering practitioners, researchers, and students. This series seeks to foster interdisciplinary biomedical engineering education to satisfy the needs of the industrial and academic areas. This requires an innovative approach that overcomes the difficulties associated with traditional textbooks and edited collections.

Series Editor, Metin Akay, University of Houston, TX

1. *Time Frequency and Wavelets in Biomedical Signal Processing*
Metin Akay
2. *Neural Networks and Artificial Intelligence for Biomedical Engineering*
Donna L. Hudson and Maurice E. Cohen
3. *Physiological Control Systems: Analysis, Simulation, and Estimation*
Michael C.K. Khoo
4. *Principles of Magnetic Resonance Imaging: A Signal Processing Perspective*
Zhi-Pei Liang and Paul C. Lauterbur
5. *Nonlinear Biomedical Signal Processing, Volume 1, Fuzzy Logic, Neural Networks, and New Algorithms*
Metin Akay
6. *Fuzzy Control and Modeling: Analytical Foundations and Applications*
Hao Ying
7. *Nonlinear Biomedical Signal Processing, Volume 2, Dynamic Analysis and Modeling*
Metin Akay
8. *Biomedical Signal Analysis: A Case-Study Approach*
Rangaraj M. Rangayyan
9. *System Theory and Practical Applications of Biomedical Signals*
Gail D. Baura
10. *Introduction to Biomedical Imaging*
Andrew Webb
11. *Medical Image Analysis*
Atam P. Dhawan

12. *Identification of Nonlinear Physiological Systems*
David T. Westwick and Robert E. Kearney
13. *Electromyography: Physiology, Engineering, and Non-Invasive Applications*
Roberto Merletti and Philip A. Parker
14. *Nonlinear Dynamic Modeling of Physiological Systems*
Vasilis Z. Marmarelis
15. *Genomics and Proteomics Engineering in Medicine and Biology*
Metin Akay
16. *Handbook of Neural Engineering*
Metin Akay
17. *Medical Image Analysis, Second Edition*
Atam P. Dhawan
18. *Advanced Methods of Biomedical Signal Processing*
Sergio Cerutti and Carlo Marchesi
19. *Epistemology of the Cell: A Systems Perspective on Biological Knowledge*
Edward R. Dougherty and Michael L. Bittner
20. *Micro and Nanotechnologies in Engineering Stem Cells and Tissues*
Murugan Ramalingam, Esmaiel Jabbari, Seeram Ramakrishna,
and Ali Khademhosseini
21. *Introduction to Neural Engineering for Motor Rehabilitation*
Dario Farina, Winnie Jensen, and Metin Akay
22. *Introduction to Tissue Engineering: Applications and Challenges*
Ravi Birla
23. *Handbook of Biomedical Telemetry*
Konstantina S. Nikita
24. *Biomedical Signal Analysis, Second Edition*
Rangaraj M. Rangayyan
25. *Models and Algorithms for Biomolecules and Molecular Networks*
Bhaskar Das Gupta and Jie Liang
26. *Surface Electromyography: Physiology, Engineering and Applications*
Roberto Merletti and Dario Farina
27. *m-Health: Fundamentals and Applications*
Robert Istepanian and Bryan Woodward

28. *Physiological Control Systems: Analysis, Simulation, and Estimation*
Second Edition
Michael C.K. Khoo
29. *Computational Modeling and Simulation Examples in Bioengineering*
Nenad D. Filipovic
30. *Introduction to Biomedical Imaging, Second Edition*
Andrew Webb
31. *Multiscale Modelling in Biomedical Engineering*
Dimitrios I. Fotiadis, Antonis I. Sakellarios, and Vassiliki T. Potsika
32. *Biomedical Signal Analysis, Third Edition*
Rangaraj M. Rangayyan and Sridhar Krishnan

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.