

AUDIO SOURCE SEPARATION AND SPEECH ENHANCEMENT

EDITED BY
EMMANUEL VINCENT
TUOMAS VIRTANEN
SHARON GANNOT



WILEY

Audio Source Separation and Speech Enhancement

Audio Source Separation and Speech Enhancement

Edited by

Emmanuel Vincent

Inria
France

Tuomas Virtanen

Tampere University of Technology
Finland

Sharon Gannot

Bar-Ilan University
Israel

WILEY

This edition first published 2018
© 2018 John Wiley & Sons Ltd

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Emmanuel Vincent, Tuomas Virtanen & Sharon Gannot to be identified as authors of the editorial material in this work has been asserted in accordance with law.

Registered Offices

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office

The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist 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.

Library of Congress Cataloging-in-Publication Data

Names: Vincent, Emmanuel (Research scientist), editor. | Virtanen, Tuomas, editor. | Gannot, Sharon, editor.

Title: Audio source separation and speech enhancement / edited by Emmanuel Vincent, Tuomas Virtanen, Sharon Gannot.

Description: Hoboken, NJ : John Wiley & Sons, 2018. | Includes bibliographical references and index. |

Identifiers: LCCN 2018013163 (print) | LCCN 2018021195 (ebook) | ISBN 9781119279884 (pdf) | ISBN 9781119279914 (epub) | ISBN 9781119279891 (cloth)

Subjects: LCSH: Speech processing systems. | Automatic speech recognition.

Classification: LCC TK7882.S65 (ebook) | LCC TK7882.S65 .A945 2018 (print) | DDC 006.4/54—dc23

LC record available at <https://lcn.loc.gov/2018013163>

Cover Design: Wiley

Cover Images: © 45RPM/iStockphoto;

© franckreporter/iStockphoto

Set in 10/12pt WarnockPro by SPi Global, Chennai, India

Contents

List of Authors	<i>xvii</i>
Preface	<i>xxi</i>
Acknowledgment	<i>xxiii</i>
Notations	<i>xxv</i>
Acronyms	<i>xxix</i>
About the Companion Website	<i>xxxi</i>

Part I Prerequisites 1

1	Introduction	3
	<i>Emmanuel Vincent, Sharon Gannot, and Tuomas Virtanen</i>	
1.1	Why are Source Separation and Speech Enhancement Needed?	3
1.2	What are the Goals of Source Separation and Speech Enhancement?	4
1.2.1	Single-Channel vs. Multichannel	4
1.2.2	Point vs. Diffuse Sources	4
1.2.3	Mixing Process	5
1.2.4	Separation vs. Enhancement	6
1.2.5	Typology of Scenarios	6
1.2.6	Evaluation	8
1.3	How can Source Separation and Speech Enhancement be Addressed?	9
1.3.1	General Processing Scheme	9
1.3.2	Converging Historical Trends	10
1.3.3	Typology of Approaches	10
1.4	Outline	11
	Bibliography	12
2	Time-Frequency Processing: Spectral Properties	15
	<i>Tuomas Virtanen, Emmanuel Vincent, and Sharon Gannot</i>	
2.1	Time-Frequency Analysis and Synthesis	15
2.1.1	STFT Analysis	16
2.1.2	STFT Synthesis	17
2.1.3	Time and Frequency Resolution	19
2.1.4	Alternative Time-Frequency Representations	20

2.1.4.1	Nonlinear Frequency Scales	20
2.1.4.2	Computation of Power Spectrum via the STFT	21
2.1.4.3	Computation via a Filterbank	22
2.2	Source Properties in the Time-Frequency Domain	23
2.2.1	Sparsity	23
2.2.2	Structure	23
2.3	Filtering in the Time-Frequency Domain	25
2.3.1	Time-Domain Convolution as Interframe and Interband Convolution	26
2.3.2	Practical Approximations	27
2.4	Summary	28
	Bibliography	28
3	Acoustics: Spatial Properties	31
	<i>Emmanuel Vincent, Sharon Gannot, and Tuomas Virtanen</i>	
3.1	Formalization of the Mixing Process	31
3.1.1	General Mixing Model	31
3.1.2	Microphone Recordings vs. Artificial Mixtures	32
3.2	Microphone Recordings	32
3.2.1	Acoustic Impulse Responses	32
3.2.2	Main Properties of Acoustic Impulse Responses	33
3.3	Artificial Mixtures	36
3.4	Impulse Response Models	37
3.4.1	Narrowband Approximation	38
3.4.1.1	Definition	38
3.4.1.2	Steering Vector – Near Field vs. Far Field	38
3.4.2	Relative Transfer Function and Interchannel Cues	39
3.4.2.1	Definition	39
3.4.2.2	Relative Steering Vector	40
3.4.3	Full-Rank Covariance Model	42
3.4.3.1	Definition	42
3.4.3.2	Parametric Covariance Models	42
3.5	Summary	43
	Bibliography	43
4	Multichannel Source Activity Detection, Localization, and Tracking	47
	<i>Pasi Pertilä, Alessio Brutti, Piergiorgio Svaizer, and Maurizio Omologo</i>	
4.1	Basic Notions in Multichannel Spatial Audio	47
4.1.1	TDOA Estimation	48
4.1.2	GCC-PHAT	49
4.1.3	Beamforming and Acoustic Maps	49
4.2	Multi-Microphone Source Activity Detection	52
4.2.1	Single-Channel Methods and Acoustic Features	52
4.2.2	Multichannel Methods	53
4.2.3	Deep Learning based Approaches	54
4.3	Source Localization	54
4.3.1	Single-Frame Localization of a Static Source	55

4.3.2	Effect of Microphone Array Geometry	56
4.3.3	Localization of Moving and Intermittent Sources	56
4.3.4	Towards Localization of Multiple Active Sources	59
4.4	Summary	60
	Bibliography	60

Part II Single-Channel Separation and Enhancement 65

5	Spectral Masking and Filtering	67
	<i>Timo Gerkmann and Emmanuel Vincent</i>	
5.1	Time-Frequency Masking	67
5.1.1	Definition and Types of Masks	67
5.1.2	Oracle Mask	68
5.2	Mask Estimation Given the Signal Statistics	70
5.2.1	Spectral Subtraction	70
5.2.2	Wiener Filtering	71
5.2.3	Bayesian Estimation of Gaussian Spectral Coefficients	72
5.2.4	Estimation of Magnitude Spectral Coefficients	76
5.2.5	Heavy-Tailed Priors	78
5.2.6	Masks Based on Source Presence Statistics	80
5.3	Perceptual Improvements	81
5.4	Summary	82
	Bibliography	83
6	Single-Channel Speech Presence Probability Estimation and Noise Tracking	87
	<i>Rainer Martin and Israel Cohen</i>	
6.1	Speech Presence Probability and its Estimation	87
6.1.1	Speech Presence Probability	88
6.1.2	Estimation of the a Posteriori SNR	90
6.1.3	Estimation of the a Priori SNR	90
6.1.4	Estimation of the Prior Speech Presence Probability	91
6.1.5	SPP Estimation with a Fixed SNR Prior	91
6.2	Noise Power Spectrum Tracking	93
6.2.1	Basic Approaches	93
6.2.2	The Minimum Statistics Approach	95
6.2.3	Minima Controlled Recursive Averaging	97
6.2.4	Harmonic Tunneling and Subspace Methods	99
6.2.5	MMSE Noise Power Estimation	100
6.3	Evaluation Measures	102
6.4	Summary	104
	Bibliography	104
7	Single-Channel Classification and Clustering Approaches	107
	<i>Felix Weninger, Jun Du, Erik Marchi, and Tian Gao</i>	
7.1	Source Separation by Computational Auditory Scene Analysis	108

7.1.1	Auditory Scene Analysis	108
7.1.2	CASA System for Source Separation	108
7.1.2.1	Segmentation	109
7.1.2.2	Grouping	110
7.1.3	Application: Spectral Clustering for Source Separation	110
7.2	Source Separation by Factorial HMMs	111
7.2.1	GMM-HMM and Factorial-Max Architecture	111
7.2.2	MAP Decoding for HMM State Sequence	112
7.2.3	Mask Estimation given State Sequences	113
7.3	Separation Based Training	113
7.3.1	Prerequisites for Separation-Based Training	113
7.3.2	Deep Neural Networks	114
7.3.2.1	Recurrent Neural Networks	118
7.3.2.2	Bidirectional RNNs	120
7.3.2.3	Other Architectures	120
7.3.3	Learning Source Separation as Classification	120
7.3.4	Learning Source Separation as Regression	121
7.3.5	Generalization Capabilities	123
7.3.6	Benchmark Performances	123
7.4	Summary	125
	Bibliography	125
8	Nonnegative Matrix Factorization	131
	<i>Roland Badeau and Tuomas Virtanen</i>	
8.1	NMF and Source Separation	131
8.1.1	NMF Masking	132
8.1.2	Learning-Free Separation	134
8.1.3	Pretrained Basis Vectors	134
8.1.4	Combining Pretrained Basis Vectors and Learning-Free Separation	135
8.2	NMF Theory and Algorithms	137
8.2.1	Criteria for Computing the NMF Model Parameters	138
8.2.2	Probabilistic Frameworks for NMF	138
8.2.2.1	Gaussian Noise Model	139
8.2.2.2	Probabilistic Latent Component Analysis	139
8.2.2.3	Poisson NMF Model	140
8.2.2.4	Gaussian Composite Model	140
8.2.2.5	α -Stable NMF Models	141
8.2.2.6	Choosing a Particular NMF Model	142
8.2.3	Algorithms for NMF	142
8.2.3.1	Multiplicative Update Rules	143
8.2.3.2	The EM Algorithm and its Variants	144
8.2.3.3	Application of the EM Algorithm to PLCA	144
8.2.3.4	Application of the Space-Alternating Generalized EM Algorithm to the Gaussian Composite Model	145
8.3	NMF Dictionary Learning Methods	145
8.3.1	NMF Dictionaries	146
8.3.2	Exemplar-Based Dictionaries	146

8.3.3	Clustering-Based Dictionary	147
8.3.4	Discriminative Dictionaries	147
8.3.5	Dictionary Adaptation	148
8.3.6	Regularization in Learning Source Models from a Mixture	148
8.4	Advanced NMF Models	148
8.4.1	Regularizations	149
8.4.1.1	Sparsity	149
8.4.1.2	Group Sparsity	150
8.4.1.3	Harmonicity and Spectral Smoothness	150
8.4.1.4	Inharmonicity	151
8.4.2	Nonstationarity	152
8.4.2.1	Time-Varying Fundamental Frequencies	152
8.4.2.2	Time-Varying Spectral Envelopes	152
8.4.2.3	Both Types of Variations	153
8.4.3	Coupled Factorizations	153
8.5	Summary	156
	Bibliography	156
9	Temporal Extensions of Nonnegative Matrix Factorization	161
	<i>Cédric Févotte, Paris Smaragdis, Nasser Mohammadiha, and Gautham J. Mysore</i>	
9.1	Convolutional NMF	161
9.1.1	1D Convolutional NMF	162
9.1.2	Convolutional NMF as a Meta-Model	164
9.1.3	N-D Model	165
9.1.4	Illustrative Examples	166
9.1.4.1	Time-Frequency Component Extraction	167
9.1.4.2	Time-Frequency Dictionaries	167
9.1.4.3	Shift-Invariant Transforms	168
9.2	Overview of Dynamical Models	169
9.3	Smooth NMF	170
9.3.1	Generalities	170
9.3.2	A Special Case	171
9.3.3	Illustrative Example	173
9.4	Nonnegative State-Space Models	174
9.4.1	Generalities	174
9.4.2	A Special Case	175
9.4.2.1	Statistical Model	175
9.4.2.2	Estimation Algorithm	176
9.5	Discrete Dynamical Models	178
9.5.1	Generalities	178
9.5.2	A Special Case	179
9.6	The Use of Dynamic Models in Source Separation	182
9.7	Which Model to Use?	183
9.8	Summary	184
9.9	Standard Distributions	184
	Bibliography	185

Part III Multichannel Separation and Enhancement 189

10 Spatial Filtering 191

Shmulik Markovich-Golan, Walter Kellermann, and Sharon Gannot

- 10.1 Fundamentals of Array Processing 192
 - 10.1.1 Beampattern 193
 - 10.1.2 Directivity 195
 - 10.1.3 Sensitivity 196
- 10.2 Array Topologies 197
- 10.3 Data-Independent Beamforming 199
- 10.4 Data-Dependent Spatial Filters: Design Criteria 202
 - 10.4.1 The Relative Transfer Function 202
 - 10.4.2 General Criterion for the Narrowband Model 203
 - 10.4.3 MWF and SDW-MWF 204
 - 10.4.4 MVDR, Maximum SNR, and LCMV 205
 - 10.4.5 Criteria for Full-Rank Covariance Models 207
 - 10.4.6 Binary Masking and Beamforming 207
 - 10.4.7 Blind Source Separation and Beamforming 208
- 10.5 Generalized Sidelobe Canceler Implementation 209
- 10.6 Postfilters 210
- 10.7 Summary 211
- Bibliography 212

11 Multichannel Parameter Estimation 219

Shmulik Markovich-Golan, Walter Kellermann, and Sharon Gannot

- 11.1 Multichannel Speech Presence Probability Estimators 219
 - 11.1.1 Multichannel Gaussian Model-Based SPP 221
 - 11.1.2 Coherence-Based Prior SPP 224
 - 11.1.3 Multichannel SPP Within GSC Structures 225
 - 11.1.4 Multiple Speakers Position-Based SPP 226
- 11.2 Covariance Matrix Estimators Exploiting SPP 227
- 11.3 Methods for Weakly Guided and Strongly Guided RTF Estimation 228
 - 11.3.1 Single-Speaker Case 228
 - 11.3.2 The Multiple-Speaker Case 230
- 11.4 Summary 231
- Bibliography 231

12 Multichannel Clustering and Classification Approaches 235

Michael I. Mandel, Shoko Araki, and Tomohiro Nakatani

- 12.1 Two-Channel Clustering 236
 - 12.1.1 Wideband Clustering with Simple IPD to ITD Mapping 237
 - 12.1.2 Wideband Clustering with Latent ITD Variable 238
 - 12.1.3 Incorporating Pitch into Localization-Based Clustering 243
- 12.2 Multichannel Clustering 244
 - 12.2.1 Generalization of Wideband Clustering to more than Two Channels 244
 - 12.2.2 Narrowband Clustering Followed by Permutation Alignment 246

12.2.2.1	Feature Extraction	248
12.2.2.2	Narrowband Clustering	248
12.2.2.3	Permutation Alignment	249
12.2.2.4	Time-Frequency Masking	250
12.2.3	Source Number Estimation	250
12.3	Multichannel Classification	251
12.3.1	Two-Channel Classification	252
12.3.2	Generalization to More than Two Channels	253
12.3.3	Generalization in Classification Systems	254
12.4	Spatial Filtering Based on Masks	255
12.4.1	Mask-Based Beamforming using Covariance Subtraction	256
12.4.2	Mask-Based Multichannel Wiener Filtering	256
12.4.3	Mask-Based Maximum SNR Beamforming	257
12.4.4	Classification-Based Multichannel Wiener Filtering	257
12.5	Summary	257
	Bibliography	258
13	Independent Component and Vector Analysis	263
	<i>Hiroshi Sawada and Zbyněk Koldovský</i>	
13.1	Convulsive Mixtures and their Time-Frequency Representations	264
13.2	Frequency-Domain Independent Component Analysis	265
13.2.1	ICA Principle	266
13.2.2	Nongaussianity-Based Separation	266
13.2.3	Modeling the Signal Probability Distributions	268
13.2.4	Alternative Models	270
13.2.4.1	Nonstationarity	270
13.2.4.2	Nonwhiteness	271
13.2.4.3	Hybrid Models	271
13.2.5	ICA Algorithms	272
13.2.5.1	Natural Gradient	272
13.2.5.2	FastICA	273
13.2.5.3	JADE	274
13.2.6	A Comparative Experiment	274
13.2.7	Required Post-Processing	275
13.2.8	Scaling Ambiguity	276
13.2.9	Permutation Problem	276
13.2.9.1	Activity Sequence Clustering	277
13.2.9.2	TDOA Clustering	278
13.3	Independent Vector Analysis	279
13.3.1	Formulation	279
13.3.2	Algorithms	279
13.3.2.1	Natural Gradient	279
13.3.2.2	FastIVA	280
13.4	Example	280
13.5	Summary	284
	Bibliography	284

14	Gaussian Model Based Multichannel Separation	289
	<i>Alexey Ozerov and Hirokazu Kameoka</i>	
14.1	Gaussian Modeling	289
14.1.1	Joint Spectral-Spatial Local Gaussian Modeling	289
14.1.2	Source Separation: Main Steps	292
14.1.2.1	Mixing Models	292
14.1.2.2	Source Spectral Models	293
14.1.2.3	Spatial Models	294
14.1.2.4	Parameter Estimation Schemes	294
14.1.2.5	Source Signal Estimation Schemes	294
14.2	Library of Spectral and Spatial Models	295
14.2.1	Spectral Models	296
14.2.1.1	GMM, Scaled GMM, HMM	296
14.2.1.2	NMF, NTF	297
14.2.1.3	AR and Variants	298
14.2.1.4	Composite Models and DNN	299
14.2.2	Spatial Models	300
14.3	Parameter Estimation Criteria and Algorithms	300
14.3.1	Parameter Estimation Criteria	300
14.3.2	Parameter Estimation Algorithms	302
14.3.2.1	EM Algorithm	302
14.3.2.2	MM Algorithm	303
14.3.2.3	VB Algorithm	305
14.3.3	Categorization of Existing Methods	305
14.4	Detailed Presentation of Some Methods	305
14.4.1	IS Multichannel NTF EM Algorithm	306
14.4.2	IS Multichannel NMF MM Algorithm	308
14.4.3	Other Algorithms for Demixing Filter Estimation	311
14.5	Summary	312
	Acknowledgment	312
	Bibliography	312
15	Dereverberation	317
	<i>Emanuël A.P. Habets and Patrick A. Naylor</i>	
15.1	Introduction to Dereverberation	317
15.2	Reverberation Cancellation Approaches	319
15.2.1	Signal Models	319
15.2.2	Identification and Equalization Approaches	321
15.2.2.1	Cross-Relation Based Blind System Identification	321
15.2.2.2	Noise Subspace Based Blind System Identification	322
15.2.2.3	Multichannel Equalization for Dereverberation	323
15.2.3	Identification and Estimation Approaches	326
15.2.4	Multichannel Linear Prediction Approaches	326
15.3	Reverberation Suppression Approaches	329
15.3.1	Signal Models	329
15.3.2	Early Signal Component Estimators	330

15.3.3	Single-Channel Spectral Variance Estimators	333
15.3.4	Multichannel Spectral Variance Estimators	333
15.4	Direct Estimation	335
15.4.1	Synthesizing a Clean Residual Signal	335
15.4.2	Linear Prediction Residual Processing	335
15.4.3	Deep Neural Networks	336
15.5	Evaluation of Dereverberation	336
15.6	Summary	337
	Bibliography	337

Part IV Application Scenarios and Perspectives 345

16	Applying Source Separation to Music	347
	<i>Bryan Pardo, Antoine Liutkus, Zhiyao Duan, and Gaël Richard</i>	
16.1	Challenges and Opportunities	348
16.1.1	Challenges	348
16.1.2	Opportunities	348
16.2	Nonnegative Matrix Factorization in the Case of Music	349
16.2.1	Shift-Invariant NMF	349
16.2.2	Constrained and Structured NMF	350
16.2.2.1	Exploiting Music Instrument Models	351
16.2.2.2	Exploiting Music Signal Models	353
16.3	Taking Advantage of the Harmonic Structure of Music	354
16.3.1	Pitch-Based Harmonic Source Separation	354
16.3.2	Modeling Timbre	355
16.3.3	Training and Adapting Timbre Models	356
16.3.4	Score-Informed Source Separation	357
16.4	Nonparametric Local Models: Taking Advantage of Redundancies in Music	358
16.4.1	HPSS: Harmonic-Percussive Source Separation	359
16.4.2	REPET: Separating Repeating Background	360
16.4.3	REPET-Sim: Exploiting Self-Similarity	361
16.4.4	KAM: Nonparametric Modeling for Spectrograms	361
16.5	Taking Advantage of Multiple Instances	363
16.5.1	Common Signal Separation	363
16.5.2	Multireference Bleeding Separation	365
16.5.3	A General Framework: Reference-Based Separation	366
16.6	Interactive Source Separation	367
16.7	Crowd-Based Evaluation	367
16.8	Some Examples of Applications	368
16.8.1	The Good Vibrations Problem	368
16.8.2	Reducing Drum Leakage: Drumatom	369
16.8.3	Impossible Duets Made Real	370
16.9	Summary	370
	Bibliography	370

17	Application of Source Separation to Robust Speech Analysis and Recognition	377
	<i>Shinji Watanabe, Tuomas Virtanen, and Dorothea Kolossa</i>	
17.1	Challenges and Opportunities	377
17.1.1	Challenges	377
17.1.2	Opportunities	378
17.2	Applications	380
17.2.1	Automatic Speech Recognition	380
17.2.1.1	Feature Extraction	381
17.2.1.2	Acoustic Model	382
17.2.1.3	GMM	383
17.2.1.4	DNN	383
17.2.1.5	Other Network Architectures	384
17.2.1.6	Training Objectives	384
17.2.1.7	Decoding	385
17.2.2	Speaker and Language Recognition	385
17.2.3	Paralinguistic Analysis	387
17.2.4	Audiovisual Analysis	389
17.3	Robust Speech Analysis and Recognition	390
17.3.1	Application of Single-Channel Source Separation	391
17.3.1.1	Matrix Factorization	391
17.3.1.2	Deep-Learning-Based Enhancement	392
17.3.2	Application of Multichannel Source Separation	393
17.3.3	Feature Extraction and Acoustic Models	393
17.3.3.1	Robust Feature Extraction	394
17.3.3.2	Feature Normalization	394
17.3.3.3	Feature Transformation	394
17.3.4	Acoustic Model	395
17.4	Integration of Front-End and Back-End	397
17.4.1	Uncertainty Modeling and Uncertainty-Based Decoding	397
17.4.1.1	Observation Uncertainties in the GMM-HMM Framework	397
17.4.1.2	Observation Uncertainties in the DNN-HMM Framework	399
17.4.2	Joint Training Frameworks	401
17.5	Use of Multimodal Information with Source Separation	403
17.5.1	Localization-Based Multimodal Source Separation	403
17.5.2	Voice Activity Detection Based Multimodal Source Separation	403
17.5.3	Joint Model-Based Multimodal Source Separation	403
17.6	Summary	404
	Bibliography	405
18	Binaural Speech Processing with Application to Hearing Devices	413
	<i>Simon Doclo, Sharon Gannot, Daniel Marquardt, and Elior Hadad</i>	
18.1	Introduction to Binaural Processing	413
18.2	Binaural Hearing	415
18.3	Binaural Noise Reduction Paradigms	416
18.3.1	Paradigm 1: Binaural Spectral Postfiltering	417
18.3.2	Paradigm 2: Binaural Spatial Filtering	418

18.4	The Binaural Noise Reduction Problem	420
18.4.1	Acoustic Scenario and Signal Definitions	420
18.4.2	Performance Measures and Binaural Cues	422
18.4.3	Binaural MWF and Binaural MVDR Beamformer	423
18.5	Extensions for Diffuse Noise	425
18.5.1	Binaural MWF with Partial Noise Estimation	426
18.5.2	Binaural MWF with Interaural Coherence Preservation	427
18.5.3	Psychoacoustically Optimized Tradeoff Parameters	428
18.5.4	Experimental Results	429
18.6	Extensions for Interfering Sources	431
18.6.1	Binaural MWF with Interference RTF Constraint	431
18.6.2	Binaural MWF with Interference Reduction Constraint	432
18.6.3	Special Case: Binaural MWF-IR for $\delta = 0$	433
18.6.4	Simulations with Measured Acoustic Transfer Functions	434
18.6.5	Simulations with Noisy Speech Signals	436
18.7	Summary	437
	Bibliography	437
19	Perspectives	443
	<i>Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot</i>	
19.1	Advancing Deep Learning	443
19.1.1	DNN Design Choices	443
19.1.2	End-to-End Approaches	445
19.1.3	Unsupervised Separation	445
19.2	Exploiting Phase Relationships	447
19.2.1	Phase Reconstruction and Joint Phase-Magnitude Estimation	447
19.2.2	Interframe and Interband Filtering	448
19.2.3	Phase Models	449
19.3	Advancing Multichannel Processing	450
19.3.1	Dealing with Moving Sources and Microphones	450
19.3.2	Manifold Learning	451
19.4	Addressing Multiple-Device Scenarios	453
19.4.1	Synchronization and Calibration	453
19.4.2	Distributed Algorithms	455
19.4.3	Multimodal Source Separation and Enhancement	455
19.5	Towards Widespread Commercial Use	455
19.5.1	Practical Deployment Constraints	455
19.5.2	Quality Assessment	456
19.5.3	New Application Areas	456
	Acknowledgment	457
	Bibliography	457
	Index	465

List of Authors

Shoko Araki

NTT Communication Science
Laboratories
Japan

Roland Badeau

Institut Mines-Télécom
France

Alessio Brutti

Fondazione Bruno Kessler
Italy

Israel Cohen

Technion
Israel

Simon Doclo

Carl von Ossietzky-Universität
Oldenburg
Germany

Jun Du

University of Science and Technology
of China
China

Zhiyao Duan

University of Rochester
NY
USA

Cédric Févotte

CNRS
France

Sharon Gannot

Bar-Ilan University
Israel

Tian Gao

University of Science and Technology
of China
China

Timo Gerkmann

Universität Hamburg
Germany

Emanuël A.P. Habets

International Audio Laboratories
Erlangen
Germany

Elior Hadad

Bar-Ilan University
Israel

Hirokazu Kameoka

The University of Tokyo
Japan

Walter Kellermann

Friedrich-Alexander Universität
Erlangen-Nürnberg
Germany

Zbyněk Koldovský

Technical University of Liberec
Czech Republic

Dorothea Kolossa

Ruhr-Universität Bochum
Germany

Antoine Liutkus

Inria
France

Michael I. Mandel

City University of New York
NY
USA

Erik Marchi

Technische Universität München
Germany

Shmulik Markovich-Golan

Bar-Ilan University
Israel

Daniel Marquardt

Carl von Ossietzky-Universität
Oldenburg
Germany

Rainer Martin

Ruhr-Universität Bochum
Germany

Nasser Mohammadiha

Chalmers University of Technology
Sweden

Gautham J. Mysore

Adobe Research
CA
USA

Tomohiro Nakatani

NTT Communication Science
Laboratories
Japan

Patrick A. Naylor

Imperial College London
UK

Maurizio Omologo

Fondazione Bruno Kessler
Italy

Alexey Ozerov

Technicolor
France

Bryan Pardo

Northwestern University
IL
USA

Pasi Pertilä

Tampere University of Technology
Finland

Gaël Richard

Institut Mines-Télécom
France

Hiroshi Sawada

NTT Communication Science
Laboratories
Japan

Paris Smaragdis

University of Illinois at
Urbana-Champaign
IL
USA

Piergiorgio Svaizer

Fondazione Bruno Kessler
Italy

Emmanuel Vincent

Inria
France

Tuomas Virtanen

Tampere University of Technology
Finland

Felix Weninger

Nuance Communications
Germany

Shinji Watanabe

Johns Hopkins University
MD
USA

Preface

Source separation and speech enhancement are some of the most studied technologies in audio signal processing. Their goal is to extract one or more source signals of interest from an audio recording involving several sound sources. This problem arises in many everyday situations. For instance, spoken communication is often obscured by concurrent speakers or by background noise, outdoor recordings feature a variety of environmental sounds, and most music recordings involve a group of instruments. When facing such scenes, humans are able to perceive and listen to individual sources so as to communicate with other speakers, navigate in a crowded street or memorize the melody of a song. Source separation and speech enhancement technologies aim to empower machines with similar abilities.

These technologies are already present in our lives today. Beyond “clean” single-source signals recorded with close microphones, they allow the industry to extend the applicability of speech and audio processing systems to multi-source, reverberant, noisy signals recorded with distant microphones. Some of the most striking examples include hearing aids, speech enhancement for smartphones, and distant-microphone voice command systems. Current technologies are expected to keep improving and spread to many other scenarios in the next few years.

Traditionally, *speech enhancement* has referred to the problem of segregating speech and background noise, while *source separation* has referred to the segregation of multiple speech or audio sources. Most textbooks focus on one of these problems and on one of three historical approaches, namely sensor array processing, computational auditory scene analysis, or independent component analysis. These communities now routinely borrow ideas from each other and other approaches have emerged, most notably based on deep learning.

This textbook is the first to provide a comprehensive overview of these problems and approaches by presenting their shared foundations and their differences using common language and notations. Starting with prerequisites (Part I), it proceeds with single-channel separation and enhancement (Part II), multichannel separation and enhancement (Part III), and applications and perspectives (Part IV). Each chapter provides both introductory and advanced material.

We designed this textbook for people in academia and industry with basic knowledge of signal processing and machine learning. Thanks to its comprehensiveness, we hope it will help students select a promising research track, researchers leverage the acquired cross-domain knowledge to design improved techniques, and engineers and developers

choose the right technology for their application scenario. We also hope that it will be useful for practitioners from other fields (e.g., acoustics, multimedia, phonetics, musicology) willing to exploit audio source separation or speech enhancement as a pre-processing tool for their own needs.

May 2017

Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot

Acknowledgment

We would like to thank all the chapter authors, as well as the following people who helped with proofreading: Sebastian Braun, Yaakov Buchris, Emre Cakir, Aleksandr Diment, Dylan Fagot, Nico Gößling, Tomoki Hayashi, Jakub Janský, Ante Jukić, Václav Kautský, Martin Krawczyk-Becker, Simon Leglaive, Bochen Li, Min Ma, Paul Magron, Zhong Meng, Gaurav Naithani, Zhaoheng Ni, Aditya Arie Nugraha, Sanjeel Parekh, Robert Rehr, Lea Schönherr, Georgina Tryfou, Ziteng Wang, and Mehdi Zohourian

May 2017

Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot

Notations

Linear algebra

x	scalar
\mathbf{x}	vector
$[x_i]_i$	vector with entries x_i
$(\mathbf{x})_i$	i th entry of vector \mathbf{x}
$\mathbf{0}_I$	$I \times 1$ vector of zeros
$\mathbf{1}_I$	$I \times 1$ vector of ones
\mathbf{X}	matrix
$[x_{ij}]_{ij}$	matrix with entries x_{ij}
$(\mathbf{X})_{ij}$	(i, j) th entry of matrix \mathbf{X}
\mathbf{I}_I	$I \times I$ identity matrix
\mathcal{X}	tensor/array (with three or more dimensions) or set
$\{x_{ijk}\}_{ijk}$	tensor with entries x_{ijk}
$\text{Diag}(\mathbf{x})$	diagonal matrix whose entries are those of vector \mathbf{x}
$\mathbf{X} \circ \mathbf{Y}$	entrywise product of matrices \mathbf{X} and \mathbf{Y}
$\text{tr}(\mathbf{X})$	trace of matrix \mathbf{X}
$\det(\mathbf{X})$	determinant of matrix \mathbf{X}
\mathbf{x}^T	transpose of vector \mathbf{x}
\mathbf{x}^H	conjugate-transpose of vector \mathbf{x}
x^*	conjugate of scalar x
$\Re(x)$	real part of scalar x
j	imaginary unit

Statistics

$p(x)$	probability distribution of continuous random variable x
$p(x y)$	conditional probability distribution of x given y
$P(x)$	probability value of discrete random variable x
$P(x y)$	conditional probability value of x given y
$\mathbb{E}\{x\}$	expectation of random variable x
$\mathbb{E}\{x y\}$	conditional expectation of x
$\mathbb{H}\{x\}$	entropy of random variable x
$\mathcal{N}(\mathbf{x} \boldsymbol{\mu}, \boldsymbol{\Sigma})$	real Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$\mathcal{N}_c(\mathbf{x} \boldsymbol{\mu}, \boldsymbol{\Sigma})$	complex Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
\hat{x}	estimated value of random variable x (e.g., first-order statistics)

σ_x^2	variance of random variable x
$\hat{\sigma}_x^2$	estimated second-order statistics of random variable x
$\Sigma_{\mathbf{x}}$	autocovariance of random vector \mathbf{x}
$\hat{\Sigma}_{\mathbf{x}}$	estimated second-order statistics of random vector \mathbf{x}
$\Sigma_{\mathbf{xy}}$	covariance of random vectors \mathbf{x} and \mathbf{y}
$\hat{\Sigma}_{\mathbf{xy}}$	estimated second-order statistics of random vectors \mathbf{x} and \mathbf{y}
$C^{\text{cost}}(\boldsymbol{\theta})$	cost function to be minimized w.r.t. the vector of parameters $\boldsymbol{\theta}$
$\mathcal{M}^{\text{objective}}(\boldsymbol{\theta})$	objective function to be maximized w.r.t. the vector of parameters $\boldsymbol{\theta}$
$\mathcal{Q}(\boldsymbol{\theta}, \cdot)$	auxiliary function to be minimized or maximized, depending on the context

Common indexes

I	number of microphones or channels
i	microphone or channel index in $\{1, \dots, I\}$
J	number of sources
j	source index in $\{1, \dots, J\}$
T	number of time-domain samples
t	sample index in $\{0, \dots, T-1\}$
L	time-domain filter length
τ	tap index in $\{0, \dots, L-1\}$
N	number of time frames
n	time frame index in $\{0, \dots, N-1\}$
F	number of frequency bins
f	frequency bin index in $\{0, \dots, F-1\}$
ν_f	frequency in Hz corresponding to frequency bin f
$x(t)$	time-domain signal x
$x(n, f)$	complex-valued STFT coefficient of signal x

Signals

x_i	input signal recorded at microphone i
\mathbf{x}	$I \times 1$ multichannel input signal, e.g. $\mathbf{x}(t) = [x_1(t), \dots, x_I(t)]^T$
\mathbf{X}	matrix of input signals, e.g. $\mathbf{X} = [x_i(t)]_{it}$ or $\mathbf{X} = [x(n, f)]_{fn}$
$ \mathbf{X} $	input magnitude spectrogram, i.e. $ \mathbf{X} = [x(n, f)]_{fn}$
\mathcal{X}	tensor/array/set of input signals, e.g. $\mathcal{X} = [x_i(n, f)]_{ifn}$
s_j	point source signal
\mathbf{s}	$J \times 1$ vector of source signals, e.g. $\mathbf{s}(t) = [s_1(t), \dots, s_J(t)]^T$
\mathbf{S}	matrix of source signals, e.g. $\mathbf{S} = [s_j(t)]_{jt}$
c_{ij}	spatial image of source j as recorded on microphone i
\mathbf{c}_j	$I \times 1$ spatial image of source j on all microphones
\mathbf{C}	tensor/array/set of spatial source image signals, e.g. $\mathbf{C} = [c_{ij}(n, f)]_{ijfn}$
a_{ij}	acoustic impulse response (or transfer function) from source j to microphone i
\mathbf{a}_j	$I \times 1$ vector of acoustic impulse responses (or transfer functions) from source j , mixing vector

A	$I \times J$ matrix of acoustic impulse responses (or transfer functions), mixing matrix
u	$I \times 1$ noise signal

Filters

★	convolution operator
w	single-output single-channel filter (mask), e.g. $\hat{s} = w^* x$
w	single-output multichannel filter (beamformer), e.g. $\hat{s} = \mathbf{w}^H \mathbf{x}$
W	multiple-output multichannel filter, e.g. $\hat{\mathbf{s}} = \mathbf{W}^H \mathbf{x}$

Nonnegative matrix factorization

b_k	k th nonnegative basis spectrum
B	matrix of nonnegative basis spectra
$h_k(n)$	k th activation coefficient in time frame n
h(n)	vector of activation coefficients in time frame n
H	matrix of activation coefficients

Deep learning

H	number of layers
h	layer index in $\{1, \dots, H\}$
K_h	number of neurons in layer h
k	neuron index in $\{1, \dots, K_h\}$
Z_h	matrix of weights and biases in layer h
g_h	activation function in layer h
g_Z	multivariate nonlinear function encoded by the full DNN

Geometry

m_i	3D location of microphone i with respect to the array origin
$\ell_{ii'}$	distance between microphones i and i'
p_j	3D location of source j with respect to the array origin
r_{ij}	distance between source j and microphone i
θ_j	azimuth of source j
φ_j	elevation of source j
c	speed of sound in air
$\Delta_{ii'j}$	time difference of arrival of source j between microphones i and i'

