Blind Speech Separation

Blind Speech Separation

edited by

Shoji Makino
NTT Communication Science Labs.
Japan

Te-Won Lee
UC San Diego, La Jolla
USA

and

Hiroshi Sawada
NTT Communication Science Labs.
Japan

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-1-4020-6478-4 (HB)
ISBN 978-1-4020-6479-1 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Printed on acid-free paper

All Rights Reserved
© 2007 Springer
No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Preface

We are surrounded by sounds. Such a noisy environment makes it difficult to obtain desired speech and it is difficult to converse comfortably there. This makes it important to be able to separate and extract a target speech signal from noisy observations for both man–machine and human–human communication.

Blind source separation (BSS) is an approach for estimating source signals using only information about their mixtures observed in each input channel. The estimation is performed without possessing information on each source, such as its frequency characteristics and location, or on how the sources are mixed. The use of BSS in the development of comfortable acoustic communication channels between humans and machines is widely accepted.

Some books have been published on BSS, independent component analysis (ICA), and related subjects. There, ICA-based BSS has been well studied in the statistics and information theory fields, for applications to a variety of disciplines including wireless communication and biomedicine. However, as speech and audio signal mixtures in a real reverberant environment are generally convolutive mixtures, they involve a structurally much more challenging task than instantaneous mixtures, which are prevalent in many other applications.

The goal of this book is to provide a reference to the fascinating topic of BSS for convolved speech mixtures. The editors believe that this book is of particular value, as it comprises cutting edge reports by internationally recognized scientists on their findings and the state of the art. The individual chapters that we selected for this book were designed to be tutorial in nature with specific emphasis on an in-depth treatment of recent important results.

This book is organized into three sections that approximately follow the main areas of BSS.

Part 1 presents overdetermined or critically determined BSS where the number of sources is smaller than or equal to the number of microphones. Here, ICA is the main technology. ICA is a statistical method for extracting mutually independent sources from their mixtures. This approach utilizes spatial diversity to discriminate between desired and undesired components, i.e., it reduces the undesired components by forming a spatial null towards them. It is, in fact, a blind adaptive beamformer realized by unsupervised adaptive filtering.

Part 2 addresses underdetermined BSS, where there are fewer microphones than source signals. This is a challenging problem because of the inherently adverse conditions. ICA cannot be used for this case. Here, the sparseness of speech sources is very useful and time–frequency diversity plays a key role.

Part 3 presents monaural BSS where there is only one microphone. Even with a single microphone, we can separate a mixture by using the harmonicity and temporal structure of the sources. Here, we can build a probabilistic framework by assuming a source model, and separate a mixture by maximizing the a posteriori probability of the sources given the observations.

The authors and editors all hope that this book will serve as a guide for a large audience, inspire many readers, and be the source of new ideas to come. We hope this book will become a useful resource for readers ranging from students and practicing engineers to advanced researchers.

We would like to take this opportunity to express our deep gratitude to each and every one of our contributing authors for making this a unique work. Thanks to their cooperation, editing this book has turned out to be a very pleasant experience. Finally, we are very grateful to Dieter Merkle from Springer and his colleagues, for their encouragement and kind support.

Kyoto, Japan — *Shoji Makino*
San Diego, CA USA — *Te-Won Lee*
July 2007 — *Hiroshi Sawada*

Contents

Part I. Multiple Microphone Blind Speech Separation with ICA

1 Convolutive Blind Source Separation for Audio Signals 3
Scott C. Douglas and Malay Gupta
1.1 Historical Perspective 3
1.1.1 Blind Equalization and Deconvolution 4
1.1.2 Neurological Structures and Information Representation 4
1.2 Mixture and Separation Models 5
1.2.1 Instantaneous Mixture and Separation Models 5
1.2.2 Convolutive Mixture and Separation Model 6
1.3 Ambiguities of BSS 8
1.4 Criteria for BSS in Speech Separation 9
1.5 Time-Domain Convolutive BSS Using Spatial Independence 10
1.5.1 Density Matching BSS Algorithms Using Natural Gradient Adaptation 10
1.5.2 Contrast-Function-based Convolutive BSS 17
1.6 Frequency-Domain Convolutive BSS 25
1.6.1 Criteria for Frequency–Domain Convolutive BSS 27
1.6.2 Permutation and Scaling Ambiguities in Frequency-Domain BSS 29
1.7 Numerical Evaluations 32
1.7.1 Laboratory Setup 33
1.7.2 Effect of Array Configuration and Room Reverberation 35
1.7.3 Effect of Initial Conditions 37
1.7.4 Effect of Signal Power Mismatch 38
1.8 Summary and Conclusions 42
References 42

2 Frequency-Domain Blind Source Separation 47
Hiroshi Sawada, Shoko Araki, and Shoji Makino
2.1 Introduction 47
2.2 Problem Formulation 49
2.3 Overview of Frequency-Domain BSS 51
2.4 Complex-Valued ICA 55
2.4.1 Source Density Function 56
2.4.2 Whitening 57

2.4.3 FastICA 58
2.4.4 Maximum Likelihood Estimation 60
2.5 Permutation Alignment 62
2.5.1 Various Approaches 62
2.5.2 Method Based on TDOA Estimations 63
2.5.3 Handling Permutations 67
2.6 Time–Frequency Masking 69
2.7 Experiments 70
2.7.1 Basic 3-Source 3-Microphone Case 71
2.7.2 With Many Background Interference Sources 73
2.8 Conclusion 75
References 75

3 Blind Source Separation using Space–Time Independent Component Analysis 79
Mike Davies, Maria Jafari, Samer Abdallah, Emmanuel Vincent, and Mark Plumbley
3.1 Introduction 79
3.2 Review of Convolutive Blind Source Separation 80
3.2.1 Multi-Channel Blind Deconvolution 81
3.2.2 Frequency Domain ICA (FD-ICA) 82
3.2.3 A Unifying Framework 84
3.3 Space–Time ICA 86
3.3.1 Learning the ST-ICA Basis 88
3.3.2 Component Grouping in ST-ICA 89
3.3.3 Clustering using Source DOAs 89
3.3.4 Separation and Reconstruction 91
3.4 An Experiment with Stereo Mixtures 91
3.4.1 Evaluation 93
3.5 Future Work 94
3.5.1 Computation and Data Requirements 94
3.5.2 Noise Reduction 95
3.5.3 Local ST-ICA Bases 95
3.6 Conclusions 96
References 97

4 TRINICON-based Blind System Identification with Application to Multiple-Source Localization and Separation 101
Herbert Buchner, Robert Aichner, and Walter Kellermann
4.1 Introduction 101
4.1.1 Overview 101
4.1.2 Blind Adaptive MIMO Filtering Tasks and Matrix Formulation 103

4.2 Blind MIMO System Identification and Relation to Blind Source Separation . . . 110
4.2.1 Square Case for Two Sources and Two Sensors . . . 111
4.2.2 Relation to SIMO System Identification . . . 113
4.2.3 Ideal Separation Solution in the General Square Case for More than Two Sources and Sensors . . . 115
4.2.4 Ideal Separation Solution and Optimum Separating Filter Length for an Arbitrary Number of Sources and Sensors . . . 115
4.2.5 General Scheme for Blind System Identification . . . 117
4.2.6 Summary . . . 118
4.3 TRINICON – A General Framework for Adaptive MIMO Signal Processing and Application to the Blind Adaptation Problems . . . 119
4.3.1 Cost Function and Gradient-Based Coefficient Optimization . . . 120
4.3.2 Special Cases and Illustration in the Time Domain . . . 125
4.3.3 On Frequency-Domain Realizations . . . 129
4.4 Acoustic Source Localization: An Overview of the Approaches and TRINICON as a Generic Source Localization Scheme . . . 131
4.5 Acoustic Source Localization Based on Time-Differences of Arrival . . . 135
4.5.1 Basic Geometric Considerations . . . 135
4.5.2 Microphone Array Geometry and Grid of Potential Source Positions . . . 136
4.5.3 Estimation of Time Differences of Arrival . . . 139
4.6 Simultaneous Localization of Multiple Sound Sources in Reverberant Environments Using Blind Adaptive MIMO System Identification . . . 141
4.7 Conclusions . . . 143
References . . . 143

5 SIMO-Model-Based Blind Source Separation – Principle and its Applications . . . 149
Hiroshi Saruwatari, Tomoya Takatani, and Kiyohiro Shikano
5.1 Introduction . . . 149
5.2 Mixing Process and Conventional BSS . . . 150
5.2.1 Mixing Process . . . 150
5.2.2 Conventional ICA-Based BSS Method . . . 151
5.2.3 Problems in Conventional ICA . . . 152
5.3 SIMO-Model-Based ICA: Algorithm . . . 153
5.4 Application of SIMO-ICA: Binaural Sound Separation . . . 157
5.4.1 Conditions for Experiments . . . 157
5.4.2 Results and Discussion . . . 158

5.5 Application of SIMO-ICA: BSD of MIMO Systems Driven by Colored Sources . . . 159
5.5.1 System Overview . . . 159
5.5.2 Conditions for Experiment . . . 160
5.5.3 Results and Discussion . . . 161
5.6 Conclusion . . . 163
References . . . 167

6 Independent Vector Analysis for Convolutive Blind Speech Separation . . . 169
Intae Lee, Taesu Kim, and Te-Won Lee
6.1 Introduction . . . 169
6.2 Contrast Functions of Independent Component Analysis . . . 171
6.2.1 Mutual Information . . . 171
6.2.2 Likelihood Function . . . 171
6.2.3 Pictorial View in Information Geometry . . . 173
6.3 What is Independent Vector Analysis? . . . 174
6.4 Contrast Functions of IVA (or MICA): Spatially White Data . . . 176
6.4.1 Entropic Contrasts . . . 176
6.4.2 Likelihood Function . . . 179
6.5 Modelling the Frequency Components of Speech . . . 179
6.5.1 L^p-Norm-Invariant Joint Densities . . . 180
6.5.2 Sparse and Spherically Symmetric Joint Densities . . . 180
6.5.3 A Semi-Nonparametric Contrast Function for Spherical Multivariate Sources . . . 182
6.5.4 Sparse and Locally Spherical Joint Densities . . . 183
6.6 Contrast Optimization and Scaling . . . 184
6.6.1 Quadratic Taylor Expansion in Complex Notations . . . 184
6.6.2 Newton's Method . . . 186
6.6.3 Scaling Problem . . . 189
6.7 Experiments . . . 189
6.8 Conclusions . . . 190
References . . . 191

7 Relative Newton and Smoothing Multiplier Optimization Methods for Blind Source Separation . . . 193
Michael Zibulevsky
7.1 Introduction . . . 193
7.1.1 Quasi-ML Blind Source Separation (BSS) . . . 193
7.1.2 Smoothing Method of Multipliers (SMOM) for Sum-Max Problems . . . 194
7.2 Relative Optimization Algorithm . . . 195

7.3 Hessian Evaluation . . . 195
7.3.1 Hessian of $-\log\det W$ . . . 196
7.3.2 Hessian of $\frac{1}{T}\sum_{m,t} h\big(W_m x(t)\big)$ . . . 197
7.4 Newton Method . . . 197
7.5 Relative Newton Method . . . 198
7.5.1 Basic Relative Newton Step . . . 198
7.5.2 Fast Relative Newton Step . . . 199
7.6 Sequential Optimization . . . 200
7.7 Smoothing Method of Multipliers (SMOM) . . . 201
7.7.1 Smoothing the Max-Function . . . 201
7.7.2 Generalized Lagrangian and Augmented Lagrangian . . . 202
7.7.3 SMOM Algorithm . . . 203
7.7.4 Frozen Hessian Strategy . . . 204
7.8 Computational Experiments . . . 204
7.8.1 Relative Newton Method . . . 205
7.8.2 SMOM Combined with the Relative Newton Method . . . 207
7.9 Conclusions . . . 210
References . . . 212

Part II. Underdetermined Blind Speech Separation with Sparseness

8 The DUET Blind Source Separation Algorithm . . . 217
Scott Rickard
8.1 Introduction . . . 217
8.2 Assumptions . . . 219
8.2.1 Anechoic Mixing . . . 219
8.2.2 W-Disjoint Orthogonality . . . 219
8.2.3 Local Stationarity . . . 220
8.2.4 Microphones Close Together . . . 221
8.2.5 Different Spatial Signatures . . . 221
8.3 DUET . . . 222
8.3.1 Main Observation and Outline . . . 222
8.3.2 Two-Dimensional Smoothed Weighted Histogram . . . 224
8.3.3 Separating the Sources . . . 226
8.3.4 The DUET BSS Algorithm . . . 228
8.4 Big Delay DUET . . . 229
8.4.1 Method One: Differential . . . 229
8.4.2 Method Two: Tiling . . . 230
8.5 Approximate W-Disjoint Orthogonality . . . 231
8.5.1 Approximate W-Disjoint Orthogonality of Speech . . . 233
8.6 Conclusions . . . 234
References . . . 235

9 *K*-means Based Underdetermined Blind Speech Separation 243
Shoko Araki, Hiroshi Sawada, and Shoji Makino
9.1 Introduction 243
9.2 Separation Procedures 245
9.3 Validity of Sparseness Assumption and Anechoic Assumption 250
9.3.1 Sparseness Evaluation of Speech Signals 250
9.3.2 Sparsest Representation with STFT 252
9.3.3 Sparseness of Reverberant Speech 254
9.3.4 Validity of Anechoic Assumption 254
9.4 Features for K-means Clustering Algorithm 256
9.4.1 K-means Clustering for Features 257
9.4.2 Feature Vectors for Multiple Sensors 258
9.5 Speech Separation Experiments 262
9.5.1 Experimental Conditions 262
9.5.2 Separation Results 263
9.5.3 Weight Parameter 265
9.5.4 Discussion 266
9.6 Conclusion 266
A Weight Parameter for Modified Feature 267
B Initial Values for the $\boldsymbol{K}$-means Clustering 268
References 269

10 Underdetermined Blind Source Separation of Convolutive Mixtures by Hierarchical Clustering and L_1-Norm Minimization 271
Stefan Winter, Walter Kellermann, Hiroshi Sawada, and Shoji Makino
10.1 Introduction 271
10.2 Underdetermined BSS for Acoustic Sources in Reverberant Environments 273
10.2.1 Definitions 273
10.2.2 Two-Stage Approach to Underdetermined BSS 278
10.3 Blind System Identification 280
10.3.1 Hierarchical Clustering 280
10.3.2 Permutation Indeterminacy 284
10.3.3 Scaling Indeterminacy 285
10.4 Blind Source Recovery 286
10.4.1 Sparseness-Based Source Model 286
10.4.2 Constrained ℓ_1-Norm Minimization 287
10.5 Experiments 292
10.5.1 Experimental Conditions 292
10.5.2 Performance Measures 293
10.5.3 Results 294

10.6 Conclusions 298
References 299

11 Bayesian Audio Source Separation 305
Cédric Févotte
11.1 Introduction 305
11.1.1 The Bayesian Setting 305
11.1.2 Sparse Models: Analysis vs Synthesis 307
11.2 Bayesian Sparse Source Separation 309
11.2.1 Sparse Source Priors 309
11.2.2 Noise and Mixing Matrix Priors 312
11.2.3 Markov Chain Monte Carlo Inference 313
11.3 Audio-Specific Models 320
11.3.1 MDCT Representation 321
11.3.2 Frequency-Dependent Models 321
11.3.3 Structural Constraints 322
11.3.4 Modified Inference Steps 323
11.4 Separation of a Stereophonic Recording with Three Sources 325
11.4.1 Experimental Setup 325
11.4.2 Discussion 327
11.5 Conclusions 330
A Appendices 331
A.1 Standard Distributions 331
A.2 M–H Update of the Degrees of Freedom Parameter α_i 331
A.3 Prior Weight of Horizontal Markov Chains 331
A.4 Update of Markov Transition Probabilities 332
References 333

Part III. Single Microphone Blind Speech Separation

12 Monaural Source Separation 339
Gil-Jin Jang and Te-Won Lee
12.1 Introduction 339
12.2 Adapting Basis Functions and Model Parameters 341
12.2.1 A Model for Signal Representation 341
12.2.2 Learning the Basis Functions 343
12.2.3 The Generalized Gaussian Distributions 344
12.3 Maximum Likelihood Source Inference 345
12.3.1 For Single Filter and Single Source 346
12.3.2 Combining Individual Inference Rules 349
12.3.3 Updating Scaling Factors 350

12.4 Evaluations . . . 352
12.4.1 Simulation Data . . . 353
12.4.2 Practical Considerations . . . 354
12.4.3 Separation Results of Simulated Mixtures . . . 354
12.4.4 Comparison to Wiener Filtering . . . 357
12.4.5 Experiments with Real Recordings . . . 358
12.5 Discussions . . . 359
12.5.1 Frequency Domain Approaches to Single Channel Separation Techniques . . . 359
12.5.2 Time Domain Approaches to Single Channel Separation Techniques . . . 360
12.5.3 Comparison to Blind Signal Separation . . . 361
12.5.4 Separability . . . 361
12.5.5 Extensions . . . 361
12.6 Conclusions . . . 362
References . . . 362

13 Probabilistic Decompositions of Spectra for Sound Separation . . . 365
Paris Smaragdis
13.1 Spectral Components for Separation . . . 365
13.2 Probabilistic Latent Component Analysis (PLCA) . . . 367
13.2.1 Sparsity . . . 370
13.2.2 Shift-Invariant Form . . . 372
13.3 Source Separation using PLCA . . . 374
13.3.1 Unsupervised . . . 374
13.3.2 Supervised . . . 377
13.3.3 Semi-Supervised . . . 380
13.4 Discussion . . . 382
13.5 Conclusions . . . 384
References . . . 385

14 Sparsification for Monaural Source Separation . . . 387
Hiroki Asari, Rasmus K. Olsson, Barak A. Pearlmutter, and Anthony M. Zador
14.1 Introduction . . . 387
14.2 Problem Formulation . . . 389
14.2.1 Dictionary Learning . . . 390
14.3 Sparse Representation by Linear Programming . . . 393
14.3.1 Basics . . . 394
14.3.2 Automatic Differentiation . . . 395
14.3.3 Dictionaries Optimized for Sparsity . . . 397
14.4 Source Separation Example . . . 398
14.4.1 Dictionary Learning . . . 398
14.4.2 Source Separation . . . 399

14.5 Convolutional Mixing and Head-Related Transfer Function . . . 400
14.5.1 Head-Related Transfer Function . . . 400
14.5.2 Reformulation . . . 401
14.6 Separation of Convolutive Sources . . . 403
14.6.1 Dictionary Learning . . . 403
14.6.2 Separation with HRTF . . . 404
14.7 Conclusion . . . 406
References . . . 407

15 Monaural Speech Separation by Support Vector Machines: Bridging the Divide Between Supervised and Unsupervised Learning Methods . . . 411
Sepp Hochreiter and Michael C. Mozer
15.1 Introduction . . . 411
15.2 Viewing Sparse Decomposition as ϵ-SVR . . . 414
15.2.1 The Sparse-Decomposition Method . . . 414
15.2.2 ϵ-Support Vector Regression . . . 416
15.2.3 The Relationship Between the Sparse-Decomposition Method and ϵ-SVR . . . 418
15.3 Nonlinear Formulation . . . 419
15.3.1 Nonlinear Mixtures . . . 419
15.3.2 Nonlinear Transformation of Atomic Sources . . . 420
15.3.3 Nonlinear Transformed Atomic Sources: Inhomogeneous Case . . . 421
15.4 Experiments . . . 422
15.4.1 Nonlinear Approximation of the Linear Correlations . . . 422
15.4.2 Transformed Atomic Sources . . . 423
15.5 Conclusions . . . 425
References . . . 426

Index . . . 429