Contents

List of Authors  xvii
Preface       xxii
Acknowledgment xxiii
Notations    xxv
Acronyms     xxix
About the Companion Website  xxxi

Part I   Prerequisites  

1 Introduction  3
Emmanuel Vincent, Sharon Gannot, and Tuomas Virtanen
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
Tuomas Virtanen, Emmanuel Vincent, and Sharon Gannot
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
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.2</td>
<td>Effect of Microphone Array Geometry</td>
<td>56</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Localization of Moving and Intermittent Sources</td>
<td>56</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Towards Localization of Multiple Active Sources</td>
<td>59</td>
</tr>
<tr>
<td>4.4</td>
<td>Summary</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>60</td>
</tr>
<tr>
<td><strong>Part II</strong></td>
<td>Single-Channel Separation and Enhancement</td>
<td>65</td>
</tr>
<tr>
<td>5</td>
<td>Spectral Masking and Filtering</td>
<td>67</td>
</tr>
<tr>
<td>5.1</td>
<td>Time-Frequency Masking</td>
<td>67</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Definition and Types of Masks</td>
<td>67</td>
</tr>
<tr>
<td>5.1.2</td>
<td>Oracle Mask</td>
<td>68</td>
</tr>
<tr>
<td>5.2</td>
<td>Mask Estimation Given the Signal Statistics</td>
<td>70</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Spectral Subtraction</td>
<td>70</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Wiener Filtering</td>
<td>71</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Bayesian Estimation of Gaussian Spectral Coefficients</td>
<td>72</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Estimation of Magnitude Spectral Coefficients</td>
<td>76</td>
</tr>
<tr>
<td>5.2.5</td>
<td>Heavy-Tailed Priors</td>
<td>78</td>
</tr>
<tr>
<td>5.2.6</td>
<td>Masks Based on Source Presence Statistics</td>
<td>80</td>
</tr>
<tr>
<td>5.3</td>
<td>Perceptual Improvements</td>
<td>81</td>
</tr>
<tr>
<td>5.4</td>
<td>Summary</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>83</td>
</tr>
<tr>
<td>6</td>
<td>Single-Channel Speech Presence Probability Estimation and Noise Tracking</td>
<td>87</td>
</tr>
<tr>
<td>6.1</td>
<td>Speech Presence Probability and its Estimation</td>
<td>87</td>
</tr>
<tr>
<td>6.1.1</td>
<td>Speech Presence Probability</td>
<td>88</td>
</tr>
<tr>
<td>6.1.2</td>
<td>Estimation of the a Posteriori SNR</td>
<td>90</td>
</tr>
<tr>
<td>6.1.3</td>
<td>Estimation of the a Priori SNR</td>
<td>90</td>
</tr>
<tr>
<td>6.1.4</td>
<td>Estimation of the Prior Speech Presence Probability</td>
<td>91</td>
</tr>
<tr>
<td>6.1.5</td>
<td>SPP Estimation with a Fixed SNR Prior</td>
<td>91</td>
</tr>
<tr>
<td>6.2</td>
<td>Noise Power Spectrum Tracking</td>
<td>93</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Basic Approaches</td>
<td>93</td>
</tr>
<tr>
<td>6.2.2</td>
<td>The Minimum Statistics Approach</td>
<td>95</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Minima Controlled Recursive Averaging</td>
<td>97</td>
</tr>
<tr>
<td>6.2.4</td>
<td>Harmonic Tunneling and Subspace Methods</td>
<td>99</td>
</tr>
<tr>
<td>6.2.5</td>
<td>MMSE Noise Power Estimation</td>
<td>100</td>
</tr>
<tr>
<td>6.3</td>
<td>Evaluation Measures</td>
<td>102</td>
</tr>
<tr>
<td>6.4</td>
<td>Summary</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>104</td>
</tr>
<tr>
<td>7</td>
<td>Single-Channel Classification and Clustering Approaches</td>
<td>107</td>
</tr>
<tr>
<td>7.1</td>
<td>Source Separation by Computational Auditory Scene Analysis</td>
<td>108</td>
</tr>
</tbody>
</table>
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

8 Nonnegative Matrix Factorization 131
Roland Badeau and Tuomas Virtanen
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 $\alpha$-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  
9 Temporal Extensions of Nonnegative Matrix Factorization 161  
9.1 Convolutive NMF 161  
9.1.1 1D Convolutive NMF 162  
9.1.2 Convolutive 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
Hiroshi Sawada and Zbyněk Koldovský
13.1 Convolutive 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
## Contents

14  **Gaussian Model Based Multichannel Separation**  289  
    *Alexey Ozerov and Hirokazu Kameoka*

  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  
    *Emanuël A.P. Habets and Patrick A. Naylor*

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
Bryan Pardo, Antoine Liutkus, Zhiyao Duan, and Gaël Richard
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
Shinji Watanabe, Tuomas Virtanen, and Dorothea Kolossa
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
17.2.2.7 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
Simon Doclo, Sharon Gannot, Daniel Marquardt, and Elior Hadad
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
Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot
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