## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreword</td>
<td>xiii</td>
</tr>
<tr>
<td>Preface</td>
<td>xvii</td>
</tr>
<tr>
<td>Contributors</td>
<td>xix</td>
</tr>
<tr>
<td>Acronyms</td>
<td>xxi</td>
</tr>
<tr>
<td>1. Fundamentals of Computational Auditory Scene Analysis</td>
<td>1</td>
</tr>
<tr>
<td>DeLiang Wang and Guy J. Brown</td>
<td></td>
</tr>
<tr>
<td>1.1 Human Auditory Scene Analysis</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1 Structure and Function of the Auditory System</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2 Perceptual Organization of Simple Stimuli</td>
<td>4</td>
</tr>
<tr>
<td>1.1.3 Perceptual Segregation of Speech from Other Sounds</td>
<td>5</td>
</tr>
<tr>
<td>1.1.4 Perceptual Mechanisms</td>
<td>8</td>
</tr>
<tr>
<td>1.2 Computational Auditory Scene Analysis (CASA)</td>
<td>11</td>
</tr>
<tr>
<td>1.2.1 What Is CASA?</td>
<td>11</td>
</tr>
<tr>
<td>1.2.2 What Is the Goal of CASA?</td>
<td>12</td>
</tr>
<tr>
<td>1.2.3 Why CASA?</td>
<td>13</td>
</tr>
<tr>
<td>1.3 Basics of CASA Systems</td>
<td>14</td>
</tr>
<tr>
<td>1.3.1 System Architecture</td>
<td>14</td>
</tr>
<tr>
<td>1.3.2 Cochleagram</td>
<td>15</td>
</tr>
<tr>
<td>1.3.3 Correlogram</td>
<td>19</td>
</tr>
<tr>
<td>1.3.4 Cross-Correlogram</td>
<td>21</td>
</tr>
<tr>
<td>1.3.5 Time-Frequency Masks</td>
<td>22</td>
</tr>
<tr>
<td>1.3.6 Resynthesis</td>
<td>23</td>
</tr>
<tr>
<td>1.4 CASA Evaluation</td>
<td>25</td>
</tr>
<tr>
<td>1.4.1 Evaluation Criteria</td>
<td>25</td>
</tr>
<tr>
<td>1.4.2 Corpora</td>
<td>26</td>
</tr>
<tr>
<td>1.5 Other Sound Separation Approaches</td>
<td>28</td>
</tr>
<tr>
<td>1.6 A Brief History of CASA (Prior to 2000)</td>
<td>30</td>
</tr>
<tr>
<td>1.6.1 Monaural CASA Systems</td>
<td>30</td>
</tr>
</tbody>
</table>
2. Multiple F0 Estimation

Alain de Cheveigné

2.1 Introduction

2.2 Signal Models

2.3 Single-Voice F0 Estimation
   2.3.1 Spectral Approach
   2.3.2 Temporal Approach
   2.3.3 Spectrotemporal Approach

2.4 Multiple-Voice F0 Estimation
   2.4.1 Spectral Approach
   2.4.2 Temporal Approach
   2.4.3 Spectrotemporal Approach

2.5 Issues
   2.5.1 Spectral Resolution
   2.5.2 Temporal Resolution
   2.5.3 Spectrotemporal Resolution

2.6 Other Sources of Information
   2.6.1 Temporal and Spectral Continuity
   2.6.2 Instrument Models
   2.6.3 Learning-Based Techniques

2.7 Estimating the Number of Sources

2.8 Evaluation

2.9 Application Scenarios

2.10 Conclusion

Acknowledgments

References

3. Feature-Based Speech Segregation

DeLiang Wang

3.1 Introduction

3.2 Feature Extraction
   3.2.1 Pitch Detection
   3.2.2 Onset and Offset Detection
   3.2.3 Amplitude Modulation Extraction
   3.2.4 Frequency Modulation Detection

3.3 Auditory Segmentation
   3.3.1 What Is the Goal of Auditory Segmentation?
CONTENTS

3.3.2 Segmentation Based on Cross-Channel Correlation and Temporal Continuity 92
3.3.3 Segmentation Based on Onset and Offset Analysis 93
3.4 Simultaneous Grouping 97
  3.4.1 Voiced Speech Segregation 97
  3.4.2 Unvoiced Speech Segregation 102
3.5 Sequential Grouping 106
  3.5.1 Spectrum-Based Sequential Grouping 108
  3.5.2 Pitch-Based Sequential Grouping 108
  3.5.3 Model-Based Sequential Grouping 109
3.6 Discussion 110
Acknowledgments 111
References 111

4. Model-Based Scene Analysis 115
   Daniel P. W. Ellis
   4.1 Introduction 115
   4.2 Source Separation as Inference 115
   4.3 Hidden Markov Models 120
   4.4 Aspects of Model-Based Systems 125
     4.4.1 Constraints: Types and Representations 126
     4.4.2 Fitting Models 130
     4.4.3 Generating Output 136
   4.5 Discussion 139
     4.5.1 Unknown Interference 139
     4.5.2 Ambiguity and Adaptation 140
     4.5.3 Relations to Other Separation Approaches 141
   4.6 Conclusions 143
References 143

5. Binaural Sound Localization 147
   Richard M. Stern, Guy J. Brown, and DeLiang Wang
   5.1 Introduction 147
   5.2 Physical and Physiological Mechanisms Underlying Auditory Localization
     5.2.1 Physical Cues 148
     5.2.2 Physiological Estimation of ITD and IID 150
   5.3 Spatial Perception of Single Sources 152
     5.3.1 Sensitivity to Differences in Interaural Time and Intensity 152
     5.3.2 Lateralization of Single Sources 152
     5.3.3 Localization of Single Sources 153
     5.3.4 The Precedence Effect 154
   5.4 Spatial Perception of Multiple Sources 155
     5.4.1 Localization of Multiple Sources 155
5.4.2 Binaural Signal Detection 156
5.5 Models of Binaural Perception 158
  5.5.1 Classical Models of Binaural Hearing 158
  5.5.2 Cross-Correlation-Based Models of Binaural Interaction 160
  5.5.3 Some Extensions to Cross-Correlation-Based Binaural Models 164
5.6 Multisource Sound Localization 168
  5.6.1 Estimating Source Azimuth from Interaural Cross-Correlation 169
  5.6.2 Methods for Resolving Azimuth Ambiguity 172
  5.6.3 Localization of Moving Sources 175
5.7 General Discussion 175
Acknowledgments 177
References 178

6. Localization-Based Grouping 187
   Albert S. Feng and Douglas L. Jones
6.1 Introduction 187
6.2 Classical Beamforming Techniques 188
  6.2.1 Fixed Beamforming Techniques 188
  6.2.2 Adaptive Beamforming Techniques 189
  6.2.3 Independent Component Analysis Techniques 190
  6.2.4 Other Localization-Based Techniques 191
6.3 Location-Based Grouping Using Interaural Time Difference Cue 191
6.4 Location-Based Grouping Using Interaural Intensity Difference Cue 199
6.5 Location-Based Grouping Using Multiple Binaural Cues 200
6.6 Discussion and Conclusions 202
Acknowledgments 202
References 203

7. Reverberation 209
   Guy J. Brown and Kalle J. Palomäki
7.1 Introduction 209
7.2 Effects of Reverberation on Listeners 211
  7.2.1 Speech Perception 211
  7.2.2 Sound Localization 213
  7.2.3 Source Separation and Signal Detection 215
  7.2.4 Distance Perception 219
  7.2.5 Auditory Spatial Impression 219
7.3 Effects of Reverberation on Machines 220
7.4 Mechanisms Underlying Robustness to Reverberation in Human Listeners 224
  7.4.1 The Role of Slow Temporal Modulations in Speech Perception 224
8. Analysis of Musical Audio Signals 251

Masataka Goto

8.1 Introduction 251
8.2 Music Scene Description 252
8.2.1 Music Scene Descriptions 253
8.2.2 Difficulties Associated with Musical Audio Signals 255
8.3 Estimating Melody and Bass Lines 256
8.3.1 PreFEst-front-end: Forming the Observed Probability Density Functions 258
8.3.2 PreFEst-core: Estimating the F0’s Probability Density Function 258
8.3.3 PreFEst-back-end: Sequential F0 Tracking by Multiple-Agent Architecture 262
8.3.4 Other Methods 263
8.4 Estimating Beat Structure 267
8.4.1 Estimating Period and Phase 268
8.4.2 Dealing with Ambiguity 270
8.4.3 Using Musical Knowledge 271
8.5 Estimating Chorus Sections and Repeated Sections 275
8.5.1 Extracting Acoustic Features and Calculating Their Similarity 278
8.5.2 Finding Repeated Sections 281
8.5.3 Grouping Repeated Sections 282
8.5.4 Detecting Modulated Repetition 284
8.5.5 Selecting Chorus Sections 285
8.5.6 Other Methods 285
8.6 Discussion and Conclusions 286
8.6.1 Importance 286
8.6.2 Evaluation Issues 287
9. Robust Automatic Speech Recognition

Jon Barker

9.1 Introduction 297
9.2 ASA and Speech Perception in Humans 299
  9.2.1 Speech Perception and Simultaneous Grouping 299
  9.2.2 Speech Perception and Sequential Grouping 302
  9.2.3 Speech Schemes 306
  9.2.4 Challenges to the ASA Account of Speech Perception 309
  9.2.5 Interim Summary 310
9.3 Speech Recognition by Machine 311
  9.3.1 The Statistical Basis of ASR 311
  9.3.2 Traditional Approaches to Robust ASR 313
  9.3.3 CASA-Driven Approaches to ASR 315
9.4 Primitive CASA and ASR 316
  9.4.1 Speech and Time-Frequency Masking 316
  9.4.2 The Missing-Data Approach to ASR 318
  9.4.3 Marginalization-Based Missing-Data ASR Systems 321
  9.4.4 Imputation-Based Missing-Data Solutions 325
  9.4.5 Estimating the Missing-Data Mask 328
  9.4.6 Difficulties with the Missing-Data Approach 330
9.5 Model-Based CASA and ASR 333
  9.5.1 The Speech Fragment Decoding Framework 334
  9.5.2 Coupling Source Segregation and Recognition 337
9.6 Discussion and Conclusions 340
9.7 Concluding Remarks 343
References 343

10. Neural and Perceptual Modeling

Guy J. Brown and DeLiang Wang

10.1 Introduction 351
10.2 The Neural Basis of Auditory Grouping 352
  10.2.1 Theoretical Solutions to the Binding Problem 352
  10.2.2 Empirical Results on Binding and ASA 353
10.3 Models of Individual Neurons 354
  10.3.1 Relaxation Oscillators 354
  10.3.2 Spike Oscillators 355
  10.3.3 A Model of a Specific Auditory Neuron 357
10.4 Models of Specific Perceptual Phenomena 359
  10.4.1 Perceptual Streaming of Tone Sequences 359
  10.4.2 Perceptual Segregation of Concurrent Vowels with
      Different F0s 367