Contents

Preface *page* xi
List of Abbreviations xiv
Notations xvi

Part I Fundamental Theories 1

1 Introduction 3
1.1 Fundamentals of Speaker Recognition 3
1.2 Feature Extraction 5
1.3 Speaker Modeling and Scoring 6
1.3.1 Speaker Modeling 7
1.3.2 Speaker Scoring 7
1.4 Modern Speaker Recognition Approaches 7
1.5 Performance Measures 8
1.5.1 FAR, FRR, and DET 9
1.5.2 Decision Cost Function 10

2 Learning Algorithms 13
2.1 Fundamentals of Statistical Learning 13
2.1.1 Probabilistic Models 13
2.1.2 Neural Networks 15
2.2 Expectation-Maximization Algorithm 16
2.2.1 Maximum Likelihood 16
2.2.2 Iterative Procedure 17
2.2.3 Alternative Perspective 19
2.2.4 Maximum *A Posteriori* 21
2.3 Approximate Inference 24
2.3.1 Variational Distribution 25
2.3.2 Factorized Distribution 26
2.3.3 EM versus VB-EM Algorithms 28
2.4 Sampling Methods 29
2.4.1 Markov Chain Monte Carlo 31
2.4.2 Gibbs Sampling 32

2.5 Bayesian Learning 33
2.5.1 Model Regularization 34
2.5.2 Bayesian Speaker Recognition 35

3 Machine Learning Models 36
3.1 Gaussian Mixture Models 36
3.1.1 The EM Algorithm 37
3.1.2 Universal Background Models 40
3.1.3 MAP Adaptation 41
3.1.4 GMM–UBM Scoring 44
3.2 Gaussian Mixture Model–Support Vector Machines 45
3.2.1 Support Vector Machines 45
3.2.2 GMM Supervectors 55
3.2.3 GMM–SVM Scoring 57
3.2.4 Nuisance Attribute Projection 58
3.3 Factor Analysis 62
3.3.1 Generative Model 64
3.3.2 EM Formulation 65
3.3.3 Relationship with Principal Component Analysis 67
3.3.4 Relationship with Nuisance Attribute Projection 68
3.4 Probabilistic Linear Discriminant Analysis 69
3.4.1 Generative Model 69
3.4.2 EM Formulations 70
3.4.3 PLDA Scoring 72
3.4.4 Enhancement of PLDA 75
3.4.5 Alternative to PLDA 75
3.5 Heavy-Tailed PLDA 76
3.5.1 Generative Model 76
3.5.2 Posteriors of Latent Variables 77
3.5.3 Model Parameter Estimation 79
3.5.4 Scoring in Heavy-Tailed PLDA 81
3.5.5 Heavy-Tailed PLDA versus Gaussian PLDA 83
3.6 I-Vectors 83
3.6.1 Generative Model 84
3.6.2 Posterior Distributions of Total Factors 85
3.6.3 I-Vector Extractor 87
3.6.4 Relation with MAP Adaptation in GMM–UBM 89
3.6.5 I-Vector Preprocessing for Gaussian PLDA 90
3.6.6 Session Variability Suppression 91
3.6.7 PLDA versus Cosine-Distance Scoring 96
3.6.8 Effect of Utterance Length 97
3.6.9 Gaussian PLDA with Uncertainty Propagation 97
3.6.10 Senone I-Vectors 102

3.7 Joint Factor Analysis 103
3.7.1 Generative Model of JFA 103
3.7.2 Posterior Distributions of Latent Factors 104
3.7.3 Model Parameter Estimation 106
3.7.4 JFA Scoring 109
3.7.5 From JFA to I-Vectors 111

Part II Advanced Studies 113

4 Deep Learning Models 115
4.1 Restricted Boltzmann Machine 115
4.1.1 Distribution Functions 116
4.1.2 Learning Algorithm 118
4.2 Deep Neural Networks 120
4.2.1 Structural Data Representation 120
4.2.2 Multilayer Perceptron 122
4.2.3 Error Backpropagation Algorithm 124
4.2.4 Interpretation and Implementation 126
4.3 Deep Belief Networks 128
4.3.1 Training Procedure 128
4.3.2 Greedy Training 130
4.3.3 Deep Boltzmann Machine 133
4.4 Stacked Autoencoder 135
4.4.1 Denoising Autoencoder 135
4.4.2 Greedy Layer-Wise Learning 137
4.5 Variational Autoencoder 140
4.5.1 Model Construction 140
4.5.2 Model Optimization 142
4.5.3 Autoencoding Variational Bayes 145
4.6 Generative Adversarial Networks 146
4.6.1 Generative Models 147
4.6.2 Adversarial Learning 148
4.6.3 Optimization Procedure 149
4.6.4 Gradient Vanishing and Mode Collapse 153
4.6.5 Adversarial Autoencoder 156
4.7 Deep Transfer Learning 158
4.7.1 Transfer Learning 159
4.7.2 Domain Adaptation 161
4.7.3 Maximum Mean Discrepancy 163
4.7.4 Neural Transfer Learning 165

5 Robust Speaker Verification 169
5.1 DNN for Speaker Verification 169
5.1.1 Bottleneck Features 169
5.1.2 DNN for I-Vector Extraction 170

5.2 Speaker Embedding 172
5.2.1 X-Vectors 172
5.2.2 Meta-Embedding 173
5.3 Robust PLDA 175
5.3.1 SNR-Invariant PLDA 175
5.3.2 Duration-Invariant PLDA 176
5.3.3 SNR- and Duration-Invariant PLDA 185
5.4 Mixture of PLDA 190
5.4.1 SNR-Independent Mixture of PLDA 190
5.4.2 SNR-Dependent Mixture of PLDA 197
5.4.3 DNN-Driven Mixture of PLDA 202
5.5 Multi-Task DNN for Score Calibration 203
5.5.1 Quality Measure Functions 205
5.5.2 DNN-Based Score Calibration 206
5.6 SNR-Invariant Multi-Task DNN 210
5.6.1 Hierarchical Regression DNN 211
5.6.2 Multi-Task DNN 214

6 Domain Adaptation 217
6.1 Overview of Domain Adaptation 217
6.2 Feature-Domain Adaptation/Compensation 218
6.2.1 Inter-Dataset Variability Compensation 218
6.2.2 Dataset-Invariant Covariance Normalization 219
6.2.3 Within-Class Covariance Correction 220
6.2.4 Source-Normalized LDA 222
6.2.5 Nonstandard Total-Factor Prior 223
6.2.6 Aligning Second-Order Statistics 224
6.2.7 Adaptation of I-Vector Extractor 225
6.2.8 Appending Auxiliary Features to I-Vectors 225
6.2.9 Nonlinear Transformation of I-Vectors 226
6.2.10 Domain-Dependent I-Vector Whitening 227
6.3 Adaptation of PLDA Models 227
6.4 Maximum Mean Discrepancy Based DNN 229
6.4.1 Maximum Mean Discrepancy 229
6.4.2 Domain-Invariant Autoencoder 232
6.4.3 Nuisance-Attribute Autoencoder 233
6.5 Variational Autoencoders (VAE) 237
6.5.1 VAE Scoring 237
6.5.2 Semi-Supervised VAE for Domain Adaptation 240
6.5.3 Variational Representation of Utterances 242
6.6 Generative Adversarial Networks for Domain Adaptation 245

7 Dimension Reduction and Data Augmentation 249
7.1 Variational Manifold PLDA 250
7.1.1 Stochastic Neighbor Embedding 251
7.1.2 Variational Manifold Learning 252
7.2 Adversarial Manifold PLDA 254
7.2.1 Auxiliary Classifier GAN 254
7.2.2 Adversarial Manifold Learning 256
7.3 Adversarial Augmentation PLDA 259
7.3.1 Cosine Generative Adversarial Network 259
7.3.2 PLDA Generative Adversarial Network 262
7.4 Concluding Remarks 265

8 Future Direction 266
8.1 Time-Domain Feature Learning 266
8.2 Speaker Embedding from End-to-End Systems 269
8.3 VAE–GAN for Domain Adaptation 270
8.3.1 Variational Domain Adversarial Neural Network (VDANN) 271
8.3.2 Relationship with Domain Adversarial Neural Network (DANN) 273
8.3.3 Gaussianality Analysis 275

Appendix Exercises 276
References 289
Index 307