Machine Learning for Speaker Recognition

MAN-WAI MAK  
*Hong Kong Polytechnic University*

JEN-TZUNG CHIEN  
*National Chiao Tung University*
Contents

Preface
List of Abbreviations
Notations

Part I Fundamental Theories

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

2 Learning Algorithms
2.1 Fundamentals of Statistical Learning
  2.1.1 Probabilistic Models
  2.1.2 Neural Networks
2.2 Expectation-Maximization Algorithm
  2.2.1 Maximum Likelihood
  2.2.2 Iterative Procedure
  2.2.3 Alternative Perspective
  2.2.4 Maximum A Posteriori
2.3 Approximate Inference
  2.3.1 Variational Distribution
  2.3.2 Factorized Distribution
  2.3.3 EM versus VB-EM Algorithms
2.4 Sampling Methods
  2.4.1 Markov Chain Monte Carlo
  2.4.2 Gibbs Sampling
2.5 Bayesian Learning
### 3 Machine Learning Models

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Gaussian Mixture Models</td>
<td>55</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Support Vector Machines</td>
<td>64</td>
</tr>
<tr>
<td>3.2.2</td>
<td>GMM Supervectors</td>
<td>74</td>
</tr>
<tr>
<td>3.2.3</td>
<td>GMM-SVM Scoring</td>
<td>75</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Nuisance Attribute Projection</td>
<td>77</td>
</tr>
<tr>
<td>3.3</td>
<td>Factor Analysis</td>
<td>82</td>
</tr>
<tr>
<td>3.4</td>
<td>Probabilistic Linear Discriminant Analysis</td>
<td>89</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Generative Model</td>
<td>89</td>
</tr>
<tr>
<td>3.4.3</td>
<td>PLDA Scoring</td>
<td>92</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Enhancement of PLDA</td>
<td>96</td>
</tr>
<tr>
<td>3.4.5</td>
<td>Alternative to PLDA</td>
<td>96</td>
</tr>
<tr>
<td>3.5</td>
<td>Heavy-Tailed PLDA</td>
<td>96</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Generative Model</td>
<td>97</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Posterior of Latent Variables</td>
<td>97</td>
</tr>
<tr>
<td>3.5.3</td>
<td>Model Parameter Estimation</td>
<td>100</td>
</tr>
<tr>
<td>3.5.4</td>
<td>Scoring in Heavy-Tailed PLDA</td>
<td>101</td>
</tr>
<tr>
<td>3.5.5</td>
<td>Heavy-Tailed PLDA versus Gaussian PLDA</td>
<td>103</td>
</tr>
<tr>
<td>3.6</td>
<td>I-vectors</td>
<td>104</td>
</tr>
<tr>
<td>3.6.1</td>
<td>Generative Model</td>
<td>104</td>
</tr>
<tr>
<td>3.6.2</td>
<td>Posterior Distributions of Total Factors</td>
<td>106</td>
</tr>
<tr>
<td>3.6.3</td>
<td>I-vector Extractor</td>
<td>108</td>
</tr>
<tr>
<td>3.6.4</td>
<td>Relation with MAP Adaptation in GMM-UBM</td>
<td>110</td>
</tr>
<tr>
<td>3.6.5</td>
<td>I-Vector Pre-processing for Gaussian PLDA</td>
<td>111</td>
</tr>
<tr>
<td>3.6.6</td>
<td>Session Variability Suppression</td>
<td>111</td>
</tr>
<tr>
<td>3.6.7</td>
<td>PLDA versus Cosine-Distance Scoring</td>
<td>117</td>
</tr>
<tr>
<td>3.6.8</td>
<td>Effect of Utterance Length</td>
<td>117</td>
</tr>
<tr>
<td>3.6.9</td>
<td>Gaussian PLDA with Uncertainty Propagation</td>
<td>118</td>
</tr>
<tr>
<td>3.6.10</td>
<td>Senone I-Vectors</td>
<td>123</td>
</tr>
<tr>
<td>3.7</td>
<td>Joint Factor Analysis</td>
<td>124</td>
</tr>
<tr>
<td>3.7.1</td>
<td>Generative Model of JFA</td>
<td>125</td>
</tr>
</tbody>
</table>
3.7.2 Posterior Distributions of Latent Factors 126
3.7.3 Model Parameter Estimation 127
3.7.4 JFA Scoring 130
3.7.5 From JFA to I-Vectors 132

Part II Advanced Studies 135

4 Deep Learning Models 137
4.1 Restricted Boltzmann Machine 137
  4.1.1 Distribution Functions 138
  4.1.2 Learning Algorithm 140
4.2 Deep Neural Networks 143
  4.2.1 Structural Data Representation 143
  4.2.2 Multilayer Perceptron 145
  4.2.3 Error Backpropagation Algorithm 146
  4.2.4 Interpretation and Implementation 149
4.3 Deep Belief Networks 151
  4.3.1 Training Procedure 151
  4.3.2 Greedy Training 153
  4.3.3 Deep Boltzmann Machine 156
4.4 Stacked Autoencoder 158
  4.4.1 Denoising Autoencoder 159
  4.4.2 Greedy Layer-wise Learning 161
4.5 Variational Autoencoder 164
  4.5.1 Model Construction 164
  4.5.2 Model Optimization 166
  4.5.3 Autoencoding Variational Bayes 169
4.6 Generative Adversarial Networks 170
  4.6.1 Generative Models 171
  4.6.2 Adversarial Learning 173
  4.6.3 Optimization Procedure 174
  4.6.4 Gradient Vanishing and Mode Collapse 178
  4.6.5 Adversarial Autoencoder 181
4.7 Deep Transfer Learning 183
  4.7.1 Transfer Learning 184
  4.7.2 Domain Adaptation 186
  4.7.3 Maximum Mean Discrepancy 188
  4.7.4 Neural Transfer Learning 190

5 Robust Speaker Verification 194
5.1 DNN for Speaker Verification 194
  5.1.1 Bottleneck Features 194
  5.1.2 DNN for I-Vector Extraction 195
5.2 Speaker Embedding 196
## 5.2 X-vectors
- 5.2.1 X-vectors
- 5.2.2 Meta-Embedding

## 5.3 Robust PLDA
- 5.3.1 SNR-Invariant PLDA
- 5.3.2 Duration-invariant PLDA
- 5.3.3 SNR- and Duration-invariant PLDA

## 5.4 Mixture of PLDA
- 5.4.1 SNR-Independent Mixture of PLDA
- 5.4.2 SNR-Dependent Mixture of PLDA
- 5.4.3 DNN-Driven Mixture of PLDA

## 5.5 Multi-Task DNN for Score Calibration
- 5.5.1 Quality Measure Functions
- 5.5.2 DNN-based Score Calibration

## 5.6 SNR-Invariant Multi-Task DNN
- 5.6.1 Hierarchical Regression DNN
- 5.6.2 Multi-Task DNN

## 6 Domain Adaptation
- 6.1 Overview of Domain Adaptation
- 6.2 Feature-Domain Adaptation/Compensation
  - 6.2.1 Inter-dataset Variability Compensation
  - 6.2.2 Dataset-Invariant Covariance Normalization
  - 6.2.3 Within-Class Covariance Correction
  - 6.2.4 Source-Normalized LDA
  - 6.2.5 Non-standard Total-Factor Prior
  - 6.2.6 Aligning Second-Order Statistics
  - 6.2.7 Adaptation of I-Vector Extractor
  - 6.2.8 Appending Auxiliary Features to I-vectors
  - 6.2.9 Nonlinear Transformation of I-Vectors
  - 6.2.10 Domain-Dependent I-vector Whitening
- 6.3 Adaptation of PLDA Models
- 6.4 Maximum Mean Discrepancy Based DNN
  - 6.4.1 Maximum Mean Discrepancy
  - 6.4.2 Domain-invariant Autoencoder
  - 6.4.3 Nuisance-attribute Autoencoder
- 6.5 Variational Autoencoders (VAE)
  - 6.5.1 VAE Scoring
  - 6.5.2 Semi-supervised VAE for Domain Adaptation
  - 6.5.3 Variational Representation of Utterances
- 6.6 Generative Adversarial Networks for Domain Adaptation

## 7 Dimension Reduction and Data Augmentation
- 7.1 Variational Manifold PLDA
  - 7.1.1 Stochastic Neighbor Embedding
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1.2 Variational Manifold Learning</td>
<td>280</td>
</tr>
<tr>
<td>7.2 Adversarial Manifold PLDA</td>
<td>282</td>
</tr>
<tr>
<td>7.2.1 Auxiliary Classifier GAN</td>
<td>283</td>
</tr>
<tr>
<td>7.2.2 Adversarial Manifold Learning</td>
<td>284</td>
</tr>
<tr>
<td>7.3 Adversarial Augmentation PLDA</td>
<td>287</td>
</tr>
<tr>
<td>7.3.1 Cosine Generative Adversarial Network</td>
<td>288</td>
</tr>
<tr>
<td>7.3.2 PLDA Generative Adversarial Network</td>
<td>291</td>
</tr>
<tr>
<td>7.4 Concluding Remarks</td>
<td>293</td>
</tr>
<tr>
<td>8 Future Direction</td>
<td>295</td>
</tr>
<tr>
<td>8.1 Time-Domain Feature Learning</td>
<td>295</td>
</tr>
<tr>
<td>8.2 Speaker Embedding from End-to-End Systems</td>
<td>297</td>
</tr>
<tr>
<td>8.3 VAE–GAN for Domain Adaptation</td>
<td>298</td>
</tr>
<tr>
<td>8.3.1 Variational Domain Adversarial Neural Network (VDANN)</td>
<td>300</td>
</tr>
<tr>
<td>8.3.2 Relationship with Domain Adversarial Neural Network (DANN)</td>
<td>303</td>
</tr>
<tr>
<td>8.3.3 Gaussianity Analysis</td>
<td>303</td>
</tr>
<tr>
<td>Appendix Exercises</td>
<td>305</td>
</tr>
<tr>
<td>References</td>
<td>318</td>
</tr>
<tr>
<td>Index</td>
<td>337</td>
</tr>
</tbody>
</table>
Preface

In the last 10 years, many methods have been developed and deployed for real-world biometric applications and multimedia information systems. Machine learning has been playing a crucial role in these applications where the model parameters could be learned and the system performance could be optimized. As for speaker recognition, researchers and engineers have been attempting to tackle the most difficult challenges: noise robustness and domain mismatch. These efforts have now been fruitful, leading to commercial products starting to emerge, e.g., voice authentication for e-banking and speaker identification in smart speakers.

Research in speaker recognition has traditionally been focused on signal processing (for extracting the most relevant and robust features) and machine learning (for classifying the features). Recently, we have witnessed the shift in the focus from signal processing to machine learning. In particular, many studies have shown that model adaptation can address both robustness and domain mismatch. As for robust feature extraction, recent studies also demonstrate that deep learning and feature learning can be a great alternative to traditional signal processing algorithms.

This book has two perspectives: machine learning and speaker recognition. The machine learning perspective gives readers insights on what makes state-of-the-art systems perform so well. The speaker recognition perspective enables readers to apply machine learning techniques to address practical issues (e.g., robustness under adverse acoustic environments and domain mismatch) when deploying speaker recognition systems. The theories and practices of speaker recognition are tightly connected in the book.

This book covers different components in speaker recognition including front-end feature extraction, back-end modeling, and scoring. A range of learning models are detailed, from Gaussian mixture models, support vector machines, joint factor analysis, and probabilistic linear discriminant analysis (PLDA) to deep neural networks (DNN). The book also covers various learning algorithms, from Bayesian learning, unsupervised learning, discriminative learning, transfer learning, manifold learning, and adversarial learning to deep learning. A series of case studies and modern models based on PLDA and DNN are addressed. In particular, different variants of deep models and their solutions to different problems in speaker recognition are presented. In addition, the book highlights some of the new trends and directions for speaker recognition based on deep
learning and adversarial learning. However, due to space constraints, the book has overlooked many promising machine learning topics and models, such as reinforcement learning, recurrent neural networks, etc. To those numerous contributors, who deserve many more credits than are given here, the authors wish to express their most sincere apologies.

The book is divided into two parts: fundamental theories and advanced studies.

1 **Fundamental theories**: This part explains different components and challenges in the construction of a statistical speaker recognition system. We organize and survey speaker recognition methods according to two categories: learning algorithms and learning models. In learning algorithms, we systematically present the inference procedures from maximum likelihood to approximate Bayesian for probabilistic models and error backpropagation algorithm for DNN. In learning models, we address a number of linear models and non-linear models based on different types of latent variables, which capture the underlying speaker and channel characteristics.

2 **Advanced studies**: This part presents a number of deep models and case studies, which are recently published for speaker recognition. We address a range of deep models ranging from DNN and deep belief networks to variational auto-encoders and generative adversarial networks, which provide the vehicle to learning representation of a true speaker model. In case studies, we highlight some advanced PLDA models and i-vector extractors that accommodate multiple mixtures, deep structures, and sparsity treatment. Finally, a number of directions and outlooks are pointed out for future trend from the perspectives of deep machine learning and challenging tasks for speaker recognition.

In the Appendix, we provide exam-style questions covering various topics in machine learning and speaker recognition.

In closing, *Machine Learning for Speaker Recognition* is intended for one-semester graduate-school courses in machine learning, neural networks, and speaker recognition. It is also intended for professional engineers, scientists, and system integrators who want to know what state-of-the-art speaker recognition technologies can provide. The prerequisite courses for this book are calculus, linear algebra, probabilities, and statistics. Some explanations in the book may require basic knowledge in speaker recognition, which can be found in other textbooks.
Acknowledgments

This book is the result of a number of years of research and teaching on the subject of neural networks, machine learning, speech and speaker recognition, and human–computer interaction. The authors are very much grateful to their students for their questions on and contribution to many examples and exercises. Some parts of the book are derived from the dissertations of several postgraduate students and their joint papers with the authors. We wish to thank all of them, in particular Dr. Eddy Zhili Tan, Dr. Ellen Wei Rao, Dr. Na Li, Mr. Wei-Wei Lin, Mr. Youzhi Tu, Miss Xiamin Pang, Mr. Qi Yao, Miss Ching-Huai Chen, Mr. Cheng-Wei Hsu, Mr. Kang-Ting Peng, and Mr. Chun-Lin Kuo. We also thank Youzhi Tu for proofreading the earlier version of the manuscript.

We have benefited greatly from the enlightening exchanges and collaboration with colleagues, particularly Prof. Helen Meng, Prof. Brian Mak, Prof. Tan Lee, Prof. Koichi Shinoda, Prof. Hsin-min Wang, Prof. Sadaoki Furui, Prof. Lin-shan Lee, Prof. Sun-Yuan Kung, and Prof. Pak-Chung Ching. We have been very fortunate to have worked with Ms. Sarah Strange, Ms. Julia Ford and Mr. David Liu at Cambridge University Press, who have provided the highest professional assistance throughout this project. We are grateful to the Department of Electronic and Information Engineering at The Hong Kong Polytechnic University and the Department of Electrical and Computer Engineering at the National Chiao Tung University for making available such a scholarly environment for both teaching and research.

We are pleased to acknowledge that the work presented in this book was in part supported by the Research Grants Council, Hong Kong Special Administrative Region (Grant Nos. PolyU 152117/14E, PolyU 152068/15E, PolyU 152518/16E, and PolyU 152137/17E); and The Ministry of Science and Technology, Taiwan (Grant Nos. MOST 107-2634-F-009-003 and MOST 108-2634-F-009-003).

We would like to thank the researchers who have contributed to the field of neural networks, machine learning, and speaker recognition. The foundation of this book is based on their work. We sincerely apologize for the inevitable overlooking of many important topics and references because of time and space constraints.

Finally, the authors wish to acknowledge the kind support of their families. Without their full understanding throughout the long writing process, this project would not have been completed so smoothly.
References


References

References


D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification,” in Proc. of An-


[184] Q. Hong, L. Li, M. Li, L. Huang, L. Wan, and J. Zhang, “Modified-prior PLDA and score calibration for duration mismatch compensation in speaker recognition
system,” in Proc. of Annual Conference of International Speech Communication Association (INTERSPEECH), 2015.


[212] I. Viñals, A. Ortega, J. Villalba, A. Miguel, and E. Lleida, “Domain adaptation of PLDA models in broadcast diarization by means of unsupervised speaker clus-


G. Bhattacharya, J. Monteiro, J. Alam, and P. Kenny, “Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification,”


### Index

- Adversarial augmentation learning, 287
- Adversarial autoencoder, 181
- Adversarial learning, 173
- Adversarial manifold learning, 284
- Autoencoding variational Bayes, 169
- Bayesian learning, 51, 164
- Bayesian speaker recognition, 53
- Bottleneck features, 194
- Center loss, 263
- Conditional independence, 44, 49
- Conjugate prior, 38
- Contrastive divergence, 141
- Cross entropy error function, 173, 179, 182
- Data augmentation, 277
- Deep belief network, 151
- Deep Boltzmann machine, 156
- Deep learning, 137
- Deep neural network, 32, 143
- Error backpropagation algorithm, 146, 175
- Feedforward neural network, 146
- Multilayer perceptron, 145
- Rectified linear unit, 146
- Training strategy, 153
- Deep transfer learning, 183
- Covariate shift, 186
- Distribution matching, 188
- Domain adaptation, 186
- Feature-based domain adaptation, 188
- Instance-based domain adaptation, 186
- Maximum mean discrepancy, 188
- Multi-task learning, 185, 190
- Neural transfer learning, 190
- Transfer learning, 184
- Denoising autoencoder, 159
- Dimension reduction, 277
- Domain adaptation, 244
  - Dataset-invariant covariance normalization, 246
  - Inter-dataset variability compensation, 246
  - Within-class covariance correction, 248
- Domain adversarial training
  - Adversarial I-vector transformer, 275
- Domain adversarial neural network (DANN), 300
- Variational domain adversarial neural network (VDANN), 300
- Domain-invariant autoencoder, 259
- EM algorithm, 34, 38
- End-to-end, 297
- Evidence lower bound, 32, 43
- Expectation-maximization algorithm, 33
- Factor analysis, 82
- Data likelihood, 85
- E-step, 86
- EM algorithm, 87
- EM formulation, 84
- Generative model, 83
- M-step, 85
- Posterior density of latent factors, 86
- Relationship with NAP, 88
- Relationship with PCA, 87
- Factorized posterior distribution, 44
- Factorized variational inference, 43
- Feature learning, 295
  - Variational representation of utterances, 270
- Feature-domain adaptation, 245
- Gaussian mixture models, 55
  - Auxiliary function, 58
  - EM algorithm, 56
  - EM steps, 59
  - Gaussian density function, 55, 279, 281
  - Incomplete data, 57
  - Log-likelihood, 56
  - Mixture posteriors, 57
  - Universal background model, 59
  - Generative adversarial network, 170, 181
  - Auxiliary classifier GAN, 283
- Generative model, 171
- Gibbs sampling, 50, 141
- GMM–SVM, 64
  - GMM–SVM scoring, 75
  - GMM-supervector, 74
- Nuisance attribute projection, 77
- Objective function, 79
Index

Relation with WCCN, 81
GMM–UBM
  GMM–UBM scoring, 62
  MAP adaptation, 60
Gradient vanishing, 178
Greedy training, 153, 161
Heavy-tailed PLDA, 96
  Compared with GPLDA, 103
Generative model, 97
Model parameter estimation, 100
  Posterior calculations, 97
  Scoring, 101
I-vector whitening, 254
I-vectors, 104
  Cosine-Distance Scoring, 117
  Generative model, 104
  I-vector extraction, 108
  I-vector extractor, 108
PLDA scoring, 117
  Relation with MAP, 110
  Senone I-vectors, 123
Session variability suppression, 111
Total factor, 106
Importance reweighting, 186
Importance sampling, 48
Importance weight, 49
Important sampling, 186
Incomplete data, 34
Isotropic Gaussian distribution, 160, 169
Jensen’s inequality, 33
Jensen-Shannon divergence, 175, 287
JFA
  Generative model, 125
  JFA scoring, 130
  Latent posterior, 126
  likelihood ratio score, 132
  Linear scoring, 132
  Parameter estimation, 127
  Point estimate, 130
  JFA: Estimating eigenspace matrix, 128
  JFA: Estimating speaker loading matrix, 130
Joint Factor Analysis, 124
Kullback-Leibler divergence, 36, 279
Linear discriminant analysis, 111
Local gradient, 148
Logistic sigmoid function, 146
Markov chain, 172
Markov chain Monte Carlo, 49
Markov switching, 155
Markov-chain Monte Carlo sampling, 141
Maximum a posteriori, 38
Maximum likelihood, 34, 171
Mean-field inference, 157, 164
Minimax optimization, 173, 285
Mixture of PLDA, 215
  DNN-driven mixture of PLDA, 227
  SNR-dependent mixture of PLDA, 223
  SNR-independent mixture of PLDA, 216
Mode collapse, 178
Model regularization, 52, 288
Monte Carlo estimate, 168
Multi-task DNN, 226, 240
Multiobjective learning, 285
Nonparametric discriminant analysis, 113
Nuisance-attribute autoencoder, 261
Partition function, 139
PLDA
  EM algorithm, 90
  Enhancement of PLDA, 96
  Generative model, 89
  PLDA Scoring, 92
PLDA adaptation, 255
Probabilistic linear discriminant analysis, 89
Probabilistic model, 30
Proposal distribution, 48
Regression DNN, 237
Reparameterization trick, 167, 168
Restricted Boltzmann machine, 137, 151
  Bernoulli-Bernoulli RBM, 138
  Gaussian-Bernoulli RBM, 139
Robust PLDA, 200
  Duration-invariant PLDA, 202
SNR- and duration-invariant PLDA, 210
SNR-invariant PLDA, 200
Saddle point, 177
Sampling method, 47
Score calibration, 228, 232
Score matching, 160
Semi-supervised learning, 153, 184, 193
Softmax function, 146
Source-normalized LDA, 250
Speaker embedding, 196, 297
Meta-embedding, 199
x-vectors, 196
Stack-wise training procedure, 152
Stacking autoencoder, 158
Stochastic backpropagation, 166
Stochastic gradient descent, 142, 147, 159, 176
Stochastic gradient variational Bayes, 167
Stochastic neural network, 164
Student’s t-distribution, 280
Sum-of-squares error function, 146
Supervised manifold learning, 281
Support vector discriminant analysis, 116
Support vector machines, 64
dual problem, 66
primal problem, 66
slack variables, 67
Wolfe dual, 68
Total-factor prior, 250
Triplet loss, 263
Two-player game, 173
Uncertainty modeling, 164
Uncertainty propagation, 118
Variational autoencoder, 164, 181
VAE for domain adaptation, 267
VAE scoring, 265
Variational Bayesian learning, 47, 164, 279
Variational lower bound, 43, 165
Variational manifold learning, 280
VB-EM algorithm, 42, 46, 158
Within-class covariance analysis, 111