Machine Learning for Speaker Recognition

Understand fundamental and advanced statistical models and deep learning models for robust speaker recognition and domain adaptation. This useful toolkit enables you to apply machine learning techniques to address practical issues, such as robustness under adverse acoustic environments and domain mismatch, when deploying speaker recognition systems. Presenting state-of-the-art machine learning techniques for speaker recognition and featuring a range of probabilistic models, learning algorithms, case studies, and new trends and directions for speaker recognition based on modern machine learning and deep learning, this is the perfect resource for graduates, researchers, practitioners, and engineers in electrical engineering, computer science, and applied mathematics.

Man-Wai Mak is an associate professor in the Department of Electronic and Information Engineering at The Hong Kong Polytechnic University.

Jen-Tzung Chien is the chair professor at the Department of Electrical and Computer Engineering at the National Chiao Tung University. He has published extensively, including the book *Bayesian Speech and Language Processing* (Cambridge University Press). He is currently serving as an elected member of the IEEE Machine Learning for Signal Processing (MLSP) Technical Committee.

Machine Learning for Speaker Recognition

MAN-WAI MAK
The Hong Kong Polytechnic University

JEN-TZUNG CHIEN
National Chiao Tung University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108428125
DOI: 10.1017/9781108552332

First published 2020

Printed in the United Kingdom by TJ International Ltd, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Mak, M. W., author. | Chien, Jen-Tzung, author.
Title: Machine learning for speaker recognition / Man-Wai Mak, Hong Kong Polytechnic University, Jen-Tzung Chien, National Chiao Tung University.
Description: New York : Cambridge University Press, 2020. | Includes bibliographical references and index.
Identifiers: LCCN 2019051262 (print) | LCCN 2019051263 (ebook) | ISBN 9781108428125 (hardback) | ISBN 9781108552332 (epub)
Subjects: LCSH: Automatic speech recognition. | Biometric identification. | Machine learning.
Classification: LCC TK7882.S65 M35 2020 (print) | LCC TK7882.S65 (ebook) | DDC 006.4/54–dc23
LC record available at https://lccn.loc.gov/2019051262
LC ebook record available at https://lccn.loc.gov/2019051263

ISBN 978-1-108-42812-5 Hardback

Contents

Cambridge University Press
978-1-108-42812-5 — Machine Learning for Speaker Recognition
Man-Wai Mak , Jen-Tzung Chien
Frontmatter
More Information

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 nonlinear 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 Xiaomin 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 Professor Helen Meng, Professor Brian Mak, Professor Tan Lee, Professor Koichi Shinoda, Professor Hsin-min Wang, Professor Sadaoki Furui, Professor Lin-shan Lee, Professor Sun-Yuan Kung, and Professor 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, MOST 108-2634-F-009-003 and MOST 109-2634-F-009-024).

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.

Abbreviations

AA-PLDA	adversarial augmentation PLDA
AAE	adversarial autoencoder
AC-GAN	auxiliary classifier GAN
AEVB	autoencoding variational Bayes
AfV	audio from video
AM-PLDA	adversarial manifold PLDA
CD	contrastive divergence
CNN	convolutional neural network
CTS	conversational telephone speech
DA	domain adaptation
DAE	denoising autoencoder
DBM	deep Boltzmann machine
DBN	deep belief network
DCF	decision cost function
DET	detection error tradeoff
DICN	dataset-invariant covariance normalization
DNN	deep neural network
EER	equal error rate
ELBO	evidence lower bound
EM	expectation-maximization
FA	factor analysis
FAR	false acceptance rate
FFT	fast Fourier transform
FRR	false rejection rate
GAN	generative adversarial network
GMM	Gaussian mixture model
HMM	hidden Markov model
IDVC	inter-dataset variability compensation
JFA	joint factor analysis
JS	Jensen–Shannon
KL	Kullback–Leibler
LSTM	long short-term memory
MAP	maximum *a posteriori*

MCMC	Markov chain Monte Carlo
MFCC	mel-frequency cepstral coefficient
ML	maximum likelihood
MLP	multilayer perceptron
MMD	maximum mean discrepancy
NAP	nuisance attribute projection
NDA	nonparametric discriminant analysis
NIST	National Institute of Standards and Technology
PCA	principal component analysis
PLDA	probabilistic linear discriminant analysis
RBF	radial basis function
RBM	restricted Boltzmann machine
RKHS	reproducing kernel Hilbert space
ReLU	rectified linear unit
SD-mPLDA	SNR-dependent mixture of PLDA
SDI-PLDA	SNR- and duration-invariant PLDA
SI-mPLDA	SNR-independent mixture of PLDA
SGD	stochastic gradient descent
SGVB	stochastic gradient variational Bayes
SNE	stochastic neighbor embedding
SRE	speaker recognition evaluation
SVDA	support vector discriminant analysis
SVM	support vector machine
t-SNE	t-distributed stochastic neighbor embedding
UBM	universal background model
VDANN	variational domain adversarial neural network
VAE	variational autoencoder
VB	variational Bayesian
VB-EM	variational Bayesian expectation-maximization
VM-PLDA	variational manifold PLDA
WCC	within-class covariance correction
WCCN	within-class covariance normalization

Notations

$\mathbf{o}$	Acoustic vectors (observations)
$\mathcal{O}$	Set of acoustic vectors
π_c	Prior probability of the cth mixture component
$\boldsymbol{\mu}_c$	The mean vector of the cth mixture component
$\boldsymbol{\Sigma}_c$	The covariance matrix of the cth mixture component
C	The number of mixture components in GMMs
ℓ_c	Indicator variable for the cth mixture in a GMM
$\gamma(\ell_c)$ and $\gamma_c(\cdot)$	Posterior probability of mixture c
$\mathbf{x}$	I-vector
$\mathcal{X}$	A set of i-vectors
$\mathbf{V}$	Speaker loading matrix of PLDA and JFA
$\mathbf{V}^{\mathsf{T}}$	Transpose of matrix $\boldsymbol{V}$
$\mathbf{U}$	SNR loading matrix in SNR-invariant PLDA and channel loading matrix in JFA
$\mathbf{G}$	Channel loading matrix in PLDA model
$\boldsymbol{\epsilon}$	Residue term of the PLDA model or the factor analysis model in i-vector systems
$\boldsymbol{\Sigma}$	Covariance matrix of $\boldsymbol{\epsilon}$ in PLDA
$\mathbf{z}$	Latent factor in PLDA
$\mathcal{Z}$	Set of latent factors $\mathbf{z}$
$\mathbf{m}$	Global mean of i-vectors
$\mathbb{E}\{\mathbf{x}\}$	Expectation of $\mathbf{x}$
$\langle\mathbf{x}\rangle$	Expectation of $\mathbf{x}$
$\boldsymbol{\Lambda}$	A set of model parameters
$\mathbf{0}$	Vector with all elements equal to 0
$\mathbf{1}$	Vector with all elements equal to 1
$\mathbf{I}$	Identity matrix
$\langle\mathbf{z}_i\|\mathbf{x}_i\rangle$	Conditional expectation of $\mathbf{z}_i$ given $\mathbf{x}_i$
$Q(\cdot)$	Auxiliary function of EM algorithms
$\mathbf{T}$	Total variability matrix in i-vector systems
$\mathbf{w}_i$	Latent factor of the factor analysis (FA) model in i-vector systems
$\boldsymbol{\mu}^{(b)}$	Supervector of the UBM in the FA model of i-vector systems
$\boldsymbol{\mu}_i$	Utterance-dependent supervector in i-vector systems

$\mathbf{y}$	Indicator vector in PLDA mixture models
$\mathcal{Y}$	Set of indicator vectors, complete data or target labels
$y_{\cdot,\cdot,c}$	indicator variables for the cth mixture in i-vector FA model
$\mathcal{H}_{\cdot,c}$	Set comprising the frame indexes whose acoustic vectors are aligned to mixture c
$\mathbf{N}$	Matrix comprising the zeroth order sufficient statistics in its diagonal
n_c	Zeroth order sufficient statistics of mixture c
$\tilde{\mathbf{f}}$	Vector comprising the first order sufficient statistics
ξ_i	Slack variables in SVM
α_i	Lagrange multipliers in SVM
$\phi(\cdot)$	Feature map in SVM
$K(\cdot,\cdot)$	Kernel function in SVM
b	Bias
$\mathcal{L}(\cdot)$	Lower bound of a log likelihood function
$\mathbb{R}^D$	Real numbers in D-dimensional space
$\mathbf{v}$	Visible units $\{v_i\}$ in RBM
$\mathbf{h}$	Hidden units $\{h_j\}$ in RBM
$E(\mathbf{v},\mathbf{h})$	Energy function of visible units $\mathbf{v}$ and hidden units $\mathbf{h}$ in RBM
$E(\mathbf{w})$	Error function with DNN parameters $\mathbf{w}$
$\mathbf{X}_n$	Minibatch data with length T_n
E_n	Error function using minibatch data $\mathbf{X}_n$
$\mathbf{y}(\mathbf{x}_t,\mathbf{w})$	Regression outputs corresponding to inputs $\mathbf{x}_t$ in DNN
$\mathbf{r}_t$	Regression targets in DNN
$\mathbf{z}_t$	Hidden units in DNN
a_{tk}	Activation of unit k
$\mathbb{H}[\cdot]$	Entropy function
$q(\mathbf{h}\|\mathbf{v})$	Variational distribution of hidden units $\mathbf{h}$ given visible units $\mathbf{v}$ in RBM
$\widetilde{\mathbf{x}}$	Corrupted version of an original sample $\mathbf{x}$ in DAE
$\widehat{\mathbf{x}}$	Reconstructed data in DAE
$\mathbf{h}$	Deterministic latent code in DAE
$\boldsymbol{\theta}$	Model parameter in VAE
$\boldsymbol{\phi}$	Variational parameter in VAE
L	Total number of samples
$\mathbf{z}^{(l)}$	The lth latent variable sample
$\mathcal{D}_{\text{KL}}(q\,\\|\,p)$	Kullback–Leibler divergence between distributions q and p
$\mathcal{D}_{\text{JS}}(q\,\\|\,p)$	Jensen–Shannon divergence between distributions q and p
$p_{\text{data}}(\mathbf{x})$	Data distribution
$p_{\text{model}}(\mathbf{x})$	Model distribution
$p_g(\mathbf{x})$	Distribution of generator (or equivalently model distribution $p_{\text{model}}(\mathbf{x})$)
G	Generator with distribution $p_g(\mathbf{x})$ in GAN

D	Discriminator in GAN
$\boldsymbol{\theta}_g$	Parameter of generator in GAN
$\boldsymbol{\theta}_d$	Parameter of discriminator in GAN
$\boldsymbol{\theta}_e$	Parameter of encoder in an AAE
$\boldsymbol{\theta}_{\text{enc}}$	Parameter of encoder in manifold or adversarial learning
$\boldsymbol{\theta}_{\text{dec}}$	Parameter of decoder in manifold or adversarial learning
$\boldsymbol{\theta}_{\text{dis}}$	Parameter of discriminator in manifold or adversarial learning
$\boldsymbol{\theta}_{\text{gen}}$	Parameter of generator in manifold or adversarial learning
t_n	Target value of an i-vector $\mathbf{x}_n$
t_{nm}	Target value for indication if $\mathbf{x}_n$ and $\mathbf{x}_m$ belong to the same class
$\mathcal{T}$	A set of target values
θ_g	Parameter of decoder in VDANN
θ_e	Parameter of encoder in VDANN
θ_c	Parameter of classifier in VDANN
θ_d	Parameter of discriminator in VDANN
$\boldsymbol{\mu}_\phi$	Mean vector of encoder's output in a VAE
$\boldsymbol{\sigma}_\phi$	Standard deviation vector of encoder's output in a VAE