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

https://doi.org/10.1017/9781108552332.001 Published online by Cambridge University Press

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.