

# Preface

*Automatic Speech Recognition* (ASR), which is aimed to enable natural human-machine interaction, has been an intensive research area for decades. Many core technologies, such as Gaussian mixture models (GMMs), hidden Markov models (HMMs), mel-frequency cepstral coefficients (MFCCs) and their derivatives, n-gram language models (LMs), discriminative training, and various adaptation techniques have been developed along the way, mostly prior to the new millenium. These techniques greatly advanced the state of the art in ASR and in its related fields. Compared to these earlier achievements, the advancement in the research and application of ASR in the decade before 2010 was relatively slow and less exciting, although important techniques such as GMM-HMM sequence discriminative training were made to work well in practical systems during this period.

In the past several years, however, we have observed a new surge of interest in ASR. In our opinion, this change was led by the increased demands on ASR in mobile devices and the success of new speech applications in the mobile world such as voice search (VS), short message dictation (SMD), and virtual speech assistants (e.g., Apple's Siri, Google Now, and Microsoft's Cortana). Equally important is the development of the deep learning techniques in large vocabulary continuous speech recognition (LVCSR) powered by big data and significantly increased computing ability. A combination of a set of deep learning techniques has led to more than 1/3 error rate reduction over the conventional state-of-the-art GMM-HMM framework on many real-world LVCSR tasks and helped to pass the adoption threshold for many real-world users. For example, the word accuracy in English or the character accuracy in Chinese in most SMD systems now exceeds 90 % and even 95 % on some systems.

Given the recent surge of interest in ASR in both industry and academia we, as researchers who have actively participated in and closely witnessed many of the recent exciting deep learning technology development, believe the time is ripe to write a book to summarize the advancements in the ASR field, especially those during the past several years.

Along with the development of the field over the past two decades or so, we have seen a number of useful books on ASR and on machine learning related to ASR, some of which are listed here:

- Deep Learning: Methods and Applications, by Li Deng and Dong Yu (June 2014)
- Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods, by Joseph Keshet, Samy Bengio (January 2009)
- Speech Recognition Over Digital Channels: Robustness and Standards, by Antonio Peinado and Jose Segura (September 2006)
- Pattern Recognition in Speech and Language Processing, by Wu Chou and Biing-Hwang Juang (February 2003)
- Speech Processing—A Dynamic and Optimization-Oriented Approach, by Li Deng and Doug O’Shaughnessy (June 2003)
- Spoken Language Processing: A Guide to Theory, Algorithm and System Development, by Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon (April 2001)
- Digital Speech Processing: Synthesis, and Recognition, Second Edition, by Sadaoki Furui (June 2001)
- Speech Communications: Human and Machine, Second Edition, by Douglas O’Shaughnessy (June 2000)
- Speech and Language Processing—An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, by Daniel Jurafsky and James Martin (April 2000)
- Speech and Audio Signal Processing, by Ben Gold and Nelson Morgan (April 2000)
- Statistical Methods for Speech Recognition, by Fred Jelinek (June 1997)
- Fundamentals of Speech Recognition, by Lawrence Rabiner and Biing-Hwang Juang (April 1993)
- Acoustical and Environmental Robustness in Automatic Speech Recognition, by Alex Acero (November 1992).

All these books, however, were either published before the rise of deep learning for ASR in 2009 or, as our 2014 overview book, were focused on less technical aspects of deep learning for ASR than is desired. These earlier books did not include the new deep learning techniques developed after 2010 with sufficient technical and mathematical detail as would be demanded by ASR or deep learning specialists. Different from the above books and in addition to some necessary background material, our current book is mainly a collation of research in most recent advances in deep learning or discriminative and hierarchical models, as applied specific to the field of ASR. Our new book presents insights and theoretical foundation of a series of deep learning models such as deep neural network (DNN), restricted Boltzmann machine (RBM), denoising autoencoder, deep belief network, recurrent neural network (RNN) and long short-term memory (LSTM) RNN, and their application in ASR through a variety of techniques including the DNN-HMM

hybrid system, the tandem and bottleneck systems, multi-task and transfer learning, sequence-discriminative training, and DNN adaptation. The book further discusses practical considerations, tricks, setups, and speedups on applying the deep learning models and related techniques in building real-world real-time ASR systems. To set the background, our book also includes two chapters that introduce GMMs and HMMs with their variants. However, we omit details of the GMM–HMM techniques that do not directly relate to the theme of the book—the hierarchical modeling or deep learning approach. Our book is thus complementary to, instead of replacement of, the published books listed above on many of similar topics. We believe this book will be of interest to advanced graduate students, researchers, practitioners, engineers, and scientists in speech processing and machine learning fields. We hope our book not only provides reference to many of the techniques used in the field but also ignites new ideas to further advance the field.

During the preparation of the book, we have received encouragement and help from Alex Acero, Geoffrey Zweig, Qiang Huo, Frank Seide, Jasha Droppo, Mike Seltzer, and Chin-Hui Lee. We also thank Springer editors, Agata Oelschlaeger and Kiruthika Poomalai, for their kind and timely help in polishing up the book and for handling its publication.

Seattle, USA, July 2014

Dong Yu  
Li Deng

Automatic Speech Recognition

A Deep Learning Approach

Yu, D.; Deng, L.

2015, XXVI, 321 p. 62 illus., Hardcover

ISBN: 978-1-4471-5778-6