

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Automatic Speech Recognition: A Bridge for Better Communication	1
1.1.1	Human–Human Communication	2
1.1.2	Human–Machine Communication	2
1.2	Basic Architecture of ASR Systems	4
1.3	Book Organization	5
1.3.1	Part I: Conventional Acoustic Models	6
1.3.2	Part II: Deep Neural Networks	6
1.3.3	Part III: DNN-HMM Hybrid Systems for ASR	7
1.3.4	Part IV: Representation Learning in Deep Neural Networks	7
1.3.5	Part V: Advanced Deep Models	7
	References	8
 <b>Part I Conventional Acoustic Models</b>		
<b>2</b>	<b>Gaussian Mixture Models</b>	<b>13</b>
2.1	Random Variables	13
2.2	Gaussian and Gaussian-Mixture Random Variables	14
2.3	Parameter Estimation	17
2.4	Mixture of Gaussians as a Model for the Distribution of Speech Features	18
	References	20
<b>3</b>	<b>Hidden Markov Models and the Variants</b>	<b>23</b>
3.1	Introduction	23
3.2	Markov Chains	25

3.3	Hidden Markov Sequences and Models . . . . .	26
3.3.1	Characterization of a Hidden Markov Model . . . . .	27
3.3.2	Simulation of a Hidden Markov Model . . . . .	29
3.3.3	Likelihood Evaluation of a Hidden Markov Model . . . . .	29
3.3.4	An Algorithm for Efficient Likelihood Evaluation . . . . .	30
3.3.5	Proofs of the Forward and Backward Recursions . . . . .	32
3.4	EM Algorithm and Its Application to Learning	
	HMM Parameters . . . . .	33
3.4.1	Introduction to EM Algorithm . . . . .	33
3.4.2	Applying EM to Learning the HMM—Baum-Welch Algorithm . . . . .	35
3.5	Viterbi Algorithm for Decoding HMM State Sequences. . . . .	39
3.5.1	Dynamic Programming and Viterbi Algorithm . . . . .	39
3.5.2	Dynamic Programming for Decoding HMM States . . . . .	40
3.6	The HMM and Variants for Generative Speech Modeling and Recognition . . . . .	42
3.6.1	GMM-HMMs for Speech Modeling and Recognition . . . . .	43
3.6.2	Trajectory and Hidden Dynamic Models for Speech Modeling and Recognition. . . . .	44
3.6.3	The Speech Recognition Problem Using Generative Models of HMM and Its Variants. . . . .	46
	References. . . . .	48

## Part II Deep Neural Networks

<b>4</b>	<b>Deep Neural Networks . . . . .</b>	<b>57</b>
4.1	The Deep Neural Network Architecture . . . . .	57
4.2	Parameter Estimation with Error Backpropagation. . . . .	59
4.2.1	Training Criteria. . . . .	60
4.2.2	Training Algorithms . . . . .	61
4.3	Practical Considerations . . . . .	65
4.3.1	Data Preprocessing . . . . .	65
4.3.2	Model Initialization. . . . .	67
4.3.3	Weight Decay . . . . .	68
4.3.4	Dropout. . . . .	69
4.3.5	Batch Size Selection . . . . .	70
4.3.6	Sample Randomization . . . . .	72
4.3.7	Momentum . . . . .	73
4.3.8	Learning Rate and Stopping Criterion . . . . .	73
4.3.9	Network Architecture . . . . .	75
4.3.10	Reproducibility and Restartability . . . . .	75
	References. . . . .	76

<b>5</b>	<b>Advanced Model Initialization Techniques</b>	79
5.1	Restricted Boltzmann Machines	79
5.1.1	Properties of RBMs	81
5.1.2	RBM Parameter Learning	83
5.2	Deep Belief Network Pretraining	86
5.3	Pretraining with Denoising Autoencoder	89
5.4	Discriminative Pretraining	91
5.5	Hybrid Pretraining	92
5.6	Dropout Pretraining	93
	References	94

### Part III Deep Neural Network-Hidden Markov Model Hybrid Systems for Automatic Speech Recognition

<b>6</b>	<b>Deep Neural Network-Hidden Markov Model Hybrid Systems</b>	99
6.1	DNN-HMM Hybrid Systems	99
6.1.1	Architecture	99
6.1.2	Decoding with CD-DNN-HMM	101
6.1.3	Training Procedure for CD-DNN-HMMs	102
6.1.4	Effects of Contextual Window	104
6.2	Key Components in the CD-DNN-HMM and Their Analysis	106
6.2.1	Datasets and Baselines for Comparisons and Analysis	106
6.2.2	Modeling Monophone States or Senones	108
6.2.3	Deeper Is Better	109
6.2.4	Exploit Neighboring Frames	111
6.2.5	Pretraining	111
6.2.6	Better Alignment Helps	112
6.2.7	Tuning Transition Probability	113
6.3	Kullback-Leibler Divergence-Based HMM	113
	References	114
<b>7</b>	<b>Training and Decoding Speedup</b>	117
7.1	Training Speedup	117
7.1.1	Pipelined Backpropagation Using Multiple GPUs	118
7.1.2	Asynchronous SGD	121
7.1.3	Augmented Lagrangian Methods and Alternating Directions Method of Multipliers	124
7.1.4	Reduce Model Size	126
7.1.5	Other Approaches	127

7.2	Decoding Speedup . . . . .	127
7.2.1	Parallel Computation. . . . .	128
7.2.2	Sparse Network . . . . .	130
7.2.3	Low-Rank Approximation . . . . .	132
7.2.4	Teach Small DNN with Large DNN . . . . .	133
7.2.5	Multiframe DNN . . . . .	134
	References. . . . .	135
<b>8</b>	<b>Deep Neural Network Sequence-Discriminative Training . . . . .</b>	<b>137</b>
8.1	Sequence-Discriminative Training Criteria . . . . .	137
8.1.1	Maximum Mutual Information . . . . .	137
8.1.2	Boosted MMI . . . . .	139
8.1.3	MPE/sMBR . . . . .	140
8.1.4	A Uniformed Formulation . . . . .	141
8.2	Practical Considerations. . . . .	142
8.2.1	Lattice Generation . . . . .	142
8.2.2	Lattice Compensation . . . . .	143
8.2.3	Frame Smoothing . . . . .	145
8.2.4	Learning Rate Adjustment . . . . .	146
8.2.5	Training Criterion Selection . . . . .	146
8.2.6	Other Considerations. . . . .	147
8.3	Noise Contrastive Estimation . . . . .	147
8.3.1	Casting Probability Density Estimation Problem as a Classifier Design Problem. . . . .	148
8.3.2	Extension to Unnormalized Models. . . . .	150
8.3.3	Apply NCE in DNN Training . . . . .	151
	References. . . . .	153

## Part IV Representation Learning in Deep Neural Networks

<b>9</b>	<b>Feature Representation Learning in Deep Neural Networks . . . . .</b>	<b>157</b>
9.1	Joint Learning of Feature Representation and Classifier . . . . .	157
9.2	Feature Hierarchy . . . . .	159
9.3	Flexibility in Using Arbitrary Input Features . . . . .	162
9.4	Robustness of Features . . . . .	163
9.4.1	Robust to Speaker Variations . . . . .	163
9.4.2	Robust to Environment Variations . . . . .	165
9.5	Robustness Across All Conditions . . . . .	167
9.5.1	Robustness Across Noise Levels. . . . .	167
9.5.2	Robustness Across Speaking Rates . . . . .	169
9.6	Lack of Generalization Over Large Distortions . . . . .	170
	References. . . . .	173

<b>10 Fuse Deep Neural Network and Gaussian Mixture</b>	
<b>Model Systems</b> . . . . .	177
10.1 Use DNN-Derived Features in GMM-HMM Systems . . . . .	177
10.1.1 GMM-HMM with Tandem and Bottleneck	
Features . . . . .	177
10.1.2 DNN-HMM Hybrid System Versus GMM-HMM	
System with DNN-Derived Features . . . . .	180
10.2 Fuse Recognition Results . . . . .	182
10.2.1 ROVER . . . . .	183
10.2.2 SCARF . . . . .	184
10.2.3 MBR Lattice Combination . . . . .	185
10.3 Fuse Frame-Level Acoustic Scores . . . . .	186
10.4 Multistream Speech Recognition . . . . .	187
References . . . . .	189
<b>11 Adaptation of Deep Neural Networks</b> . . . . .	193
11.1 The Adaptation Problem for Deep Neural Networks . . . . .	193
11.2 Linear Transformations . . . . .	194
11.2.1 Linear Input Networks . . . . .	195
11.2.2 Linear Output Networks . . . . .	196
11.3 Linear Hidden Networks . . . . .	198
11.4 Conservative Training . . . . .	199
11.4.1 $L_2$ Regularization . . . . .	199
11.4.2 KL-Divergence Regularization . . . . .	200
11.4.3 Reducing Per-Speaker Footprint . . . . .	202
11.5 Subspace Methods . . . . .	204
11.5.1 Subspace Construction Through Principal	
Component Analysis . . . . .	204
11.5.2 Noise-Aware, Speaker-Aware,	
and Device-Aware Training . . . . .	205
11.5.3 Tensor . . . . .	209
11.6 Effectiveness of DNN Speaker Adaptation . . . . .	210
11.6.1 KL-Divergence Regularization Approach . . . . .	210
11.6.2 Speaker-Aware Training . . . . .	212
References . . . . .	213

## Part V Advanced Deep Models

<b>12 Representation Sharing and Transfer in Deep</b>	
<b>Neural Networks</b> . . . . .	219
12.1 Multitask and Transfer Learning . . . . .	219
12.1.1 Multitask Learning . . . . .	219
12.1.2 Transfer Learning . . . . .	220

12.2	Multilingual and Crosslingual Speech Recognition . . . . .	221
12.2.1	Tandem/Bottleneck-Based Crosslingual Speech Recognition . . . . .	222
12.2.2	Shared-Hidden-Layer Multilingual DNN . . . . .	223
12.2.3	Crosslingual Model Transfer . . . . .	226
12.3	Multiobjective Training of Deep Neural Networks for Speech Recognition . . . . .	230
12.3.1	Robust Speech Recognition with Multitask Learning . . . . .	230
12.3.2	Improved Phone Recognition with Multitask Learning . . . . .	230
12.3.3	Recognizing both Phonemes and Graphemes . . . . .	231
12.4	Robust Speech Recognition Exploiting Audio-Visual Information . . . . .	232
	References. . . . .	233
<b>13</b>	<b>Recurrent Neural Networks and Related Models . . . . .</b>	<b>237</b>
13.1	Introduction . . . . .	237
13.2	State-Space Formulation of the Basic Recurrent Neural Network . . . . .	239
13.3	The Backpropagation-Through-Time Learning Algorithm. . . . .	240
13.3.1	Objective Function for Minimization. . . . .	241
13.3.2	Recursive Computation of Error Terms . . . . .	241
13.3.3	Update of RNN Weights . . . . .	242
13.4	A Primal-Dual Technique for Learning Recurrent Neural Networks. . . . .	244
13.4.1	Difficulties in Learning RNNs . . . . .	244
13.4.2	Echo-State Property and Its Sufficient Condition . . . . .	245
13.4.3	Learning RNNs as a Constrained Optimization Problem . . . . .	245
13.4.4	A Primal-Dual Method for Learning RNNs . . . . .	246
13.5	Recurrent Neural Networks Incorporating LSTM Cells . . . . .	249
13.5.1	Motivations and Applications. . . . .	249
13.5.2	The Architecture of LSTM Cells . . . . .	250
13.5.3	Training the LSTM-RNN . . . . .	250
13.6	Analyzing Recurrent Neural Networks—A Contrastive Approach. . . . .	251
13.6.1	Direction of Information Flow: Top-Down versus Bottom-Up . . . . .	251
13.6.2	The Nature of Representations: Localist or Distributed . . . . .	254
13.6.3	Interpretability: Inferring Latent Layers versus End-to-End Learning. . . . .	255

13.6.4	Parameterization: Parsimonious Conditionals versus Massive Weight Matrices. . . . .	256
13.6.5	Methods of Model Learning: Variational Inference versus Gradient Descent . . . . .	258
13.6.6	Recognition Accuracy Comparisons . . . . .	258
13.7	Discussions . . . . .	259
	References. . . . .	261
<b>14</b>	<b>Computational Network . . . . .</b>	<b>267</b>
14.1	Computational Network. . . . .	267
14.2	Forward Computation . . . . .	269
14.3	Model Training . . . . .	271
14.4	Typical Computation Nodes. . . . .	275
14.4.1	Computation Node Types with No Operand. . . . .	276
14.4.2	Computation Node Types with One Operand. . . . .	276
14.4.3	Computation Node Types with Two Operands . . . . .	281
14.4.4	Computation Node Types for Computing Statistics . . . . .	287
14.5	Convolutional Neural Network . . . . .	288
14.6	Recurrent Connections. . . . .	291
14.6.1	Sample by Sample Processing Only Within Loops . . . . .	292
14.6.2	Processing Multiple Utterances Simultaneously . . . . .	293
14.6.3	Building Arbitrary Recurrent Neural Networks. . . . .	293
	References. . . . .	297
<b>15</b>	<b>Summary and Future Directions . . . . .</b>	<b>299</b>
15.1	Road Map . . . . .	299
15.1.1	Debut of DNNs for ASR. . . . .	299
15.1.2	Speedup of DNN Training and Decoding . . . . .	302
15.1.3	Sequence Discriminative Training. . . . .	302
15.1.4	Feature Processing . . . . .	303
15.1.5	Adaptation. . . . .	304
15.1.6	Multitask and Transfer Learning. . . . .	305
15.1.7	Convolution Neural Networks . . . . .	305
15.1.8	Recurrent Neural Networks and LSTM . . . . .	306
15.1.9	Other Deep Models. . . . .	306
15.2	State of the Art and Future Directions . . . . .	307
15.2.1	State of the Art—A Brief Analysis . . . . .	307
15.2.2	Future Directions . . . . .	308
	References. . . . .	309
	<b>Index . . . . .</b>	<b>317</b>

Automatic Speech Recognition

A Deep Learning Approach

Yu, D.; Deng, L.

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

ISBN: 978-1-4471-5778-6