

Contents

Abbreviations and Acronyms xvii

1 Introduction 1
Reinhold Haeb-Umbach, Dorothea Kolossa

1.1 Overview of the Book 2

References 5

Part I Theoretical Foundations

2 Uncertainty Decoding and Conditional Bayesian Estimation 9
Reinhold Haeb-Umbach

2.1 Introduction 9

2.2 Speech Recognition Using HMMs 11

2.2.1 Bayesian Classification Rule 11

2.2.2 Presence of Corrupted Features 13

2.2.3 Variants of Uncertainty Decoding 17

2.2.4 Missing-Feature Approaches 18

2.3 Feature Space vs. Acoustic Model-Based Approaches 20

2.4 Posterior Estimation 21

2.4.1 A Priori Model of Clean Speech 24

2.4.2 Observation Model 25

2.4.3 Inference 26

2.5 From Theory to Practice 28

2.6 Conclusions 30

References 31

3 Uncertainty Propagation 35
Ramón Fernandez Astudillo, Dorothea Kolossa

3.1 Uncertainty Propagation 35

3.2 Modeling Uncertainty in the STFT Domain 37

3.3 Piecewise Uncertainty Propagation 39

3.3.1 Uncertainty Propagation Through Linear Transformations 40

3.3.2	Uncertainty Propagation with the Unscented Transform .	41
3.4	Uncertainty Propagation for the Mel Frequency Cepstral Features .	42
3.4.1	Propagation Through the STSA Computation	43
3.4.2	Propagation Through the Mel Filter Bank	44
3.4.3	Propagation Through the Logarithm of the Mel-STSA Analysis	45
3.4.4	Propagation Through the Discrete Cosine Transform	46
3.5	Uncertainty Propagation for the RASTA-PLP Cepstral Features . .	46
3.5.1	Propagation Through the PSD Computation	47
3.5.2	Propagation Through the Bark Filterbank	48
3.5.3	Propagation Through the Logarithm of the Bark-PSD . . .	48
3.5.4	Propagation Through RASTA Filter	49
3.5.5	Propagation Through Equal Loudness Pre-emphasis, Power Law of Hearing and Exponential	51
3.5.6	Propagation Through the All-Pole Model and LPCC Computation	51
3.6	Uncertainty Propagation for the ETSI ES 202 050 Standard	52
3.6.1	Propagation of the Complex Gaussian Model into the Time Domain	53
3.6.2	Propagation Through Log-Energy Transformation	55
3.6.3	Propagation Through Blind Equalization	56
3.7	Taking Full Covariances into Consideration	56
3.7.1	Covariances Between Features of the Same Frame	56
3.7.2	Covariance Between Features of Different Frames	57
3.8	Tests and Results	57
3.8.1	Monte Carlo Simulation of Uncertainty Propagation	57
3.8.2	Improving the Robustness of ASR with Uncertainty Propagation	61
	References	63

Part II Applications: Noise Robustness

4	Front-End, Back-End, and Hybrid Techniques for Noise-Robust Speech Recognition	67
	Li Deng	
4.1	Introduction	67
4.2	Bayesian Decision Rule with Unreliable Features	69
4.3	Use of Algonquin Model to Compute Feature Uncertainty	71
4.3.1	Algonquin Model of Speech Distortion	71
4.3.2	Step I in Computing Uncertainty: Means	72
4.3.3	Step II in Computing Uncertainty: Variances	74
4.3.4	Discussions	75
4.4	Use of a Phase-Sensitive Model for Feature Compensation	76
4.4.1	Phase-Sensitive Modeling of Acoustic Distortion — Deterministic Version	77

4.4.2	The Phase-Sensitive Model of Acoustic Distortion — Probabilistic Version	80
4.4.3	Feature Compensation Experiments and Lessons Learned	82
4.4.4	Discussions	83
4.5	Bayesian Decision Rule with Unreliable Model Parameters	84
4.5.1	Bayesian Predictive Classification Rule	84
4.5.2	Model Compensation Viewed from the Perspective of the BPC Rule	85
4.6	Model and Feature Compensation — A Taxonomy-Oriented Overview	86
4.6.1	Model-Domain Compensation	86
4.6.2	Feature-Domain Compensation	88
4.6.3	Hybrid Compensation Techniques	89
4.7	Noise Adaptive Training	90
4.7.1	The Basic NAT Scheme and its Performance	90
4.7.2	NAT and Related Work — A Brief Overview	91
4.8	Summary and Conclusions	93
	References	95
5	Model-Based Approaches to Handling Uncertainty	101
	M. J. F. Gales	
5.1	Introduction	101
5.2	General Acoustic Model Adaptation	102
5.3	Impact of Noise on Speech	104
5.3.1	Static Parameter Mismatch Function	104
5.3.2	Dynamic Parameter Mismatch Functions	106
5.3.3	Corrupted Speech Distributions	106
5.4	Feature Enhancement Approaches	107
5.5	Model-Based Noise Compensation	110
5.5.1	Vector Taylor Series Compensation	111
5.5.2	Sampling-Based Approximations	112
5.6	Efficient Model Compensation and Likelihood Calculation	113
5.6.1	Compensation Parameter Estimation	114
5.6.2	Compensating the Model Parameters	115
5.7	Adaptive Training and Noise Estimation	116
5.7.1	EM-Based Approaches	117
5.7.2	Second-Order Approaches	119
5.8	Conclusions and Future Research Directions	120
	References	123
6	Reconstructing Noise-Corrupted Spectrographic Components for Robust Speech Recognition	127
	Bhiksha Raj and Rita Singh	
6.1	Introduction	127
6.2	Spectrographic Representation	130
6.3	Notation	131

6.4	A Note on Estimating Spectrographic Masks	132
6.5	Classifier Compensation Methods	133
6.5.1	State-Based Imputation	134
6.5.2	Marginalization	134
6.5.3	Marginalization with Soft Masks	135
6.6	Feature Compensation Methods	135
6.6.1	Correlation-Based Reconstruction	136
6.6.2	Cluster-Based Reconstruction	138
6.6.3	Minimum Mean Squared Error Estimation of Spectral Components from Soft Masks	141
6.7	Experimental Evaluation	144
6.7.1	Experimental Setup	144
6.7.2	Recognition Performance with Knowledge of True SNR ..	145
6.7.3	Recognition with Log Spectra	145
6.7.4	Effect of Preprocessing	146
6.7.5	Recognition with Cepstra	147
6.7.6	Effect of Errors in Identifying Unreliable Components ..	148
6.7.7	Experiments with MMSE Estimation from Soft Masks ..	151
6.8	Conclusion and Discussion	153
	References	155
7	Automatic Speech Recognition Using Missing Data Techniques:	
	Handling of Real-World Data	157
	Jort F. Gemmeke, Maarten Van Segbroeck, Yujun Wang, Bert Cranen, Hugo Van hamme	
7.1	Introduction	157
7.2	MDT ASR	160
7.2.1	Missing Data Techniques	160
7.2.2	Gaussian-Dependent Imputation	161
7.3	Real-World Data: The SPEECON and SpeechDat-Car Databases ..	162
7.3.1	Isolated Word Test Set	162
7.3.2	Training Sets	163
7.4	Experimental Setup	163
7.4.1	Mask Estimation	163
7.4.2	Recognizer Setup	165
7.5	MDT and Multi-Condition Training	167
7.5.1	Recognition Results	167
7.5.2	Discussion	170
7.6	MDT and Dereverberation	171
7.6.1	Experimental Setup	172
7.6.2	Results and Discussion	173
7.7	MDT and Feature Enhancement	176
7.7.1	Experimental Setup	177
7.7.2	Results and Discussion	180
7.8	General Discussion and Conclusions	182

7.9	Acknowledgements	183
	References	184
8	Conditional Bayesian Estimation Employing a Phase-Sensitive Observation Model for Noise Robust Speech Recognition	187
	Volker Leutnant and Reinhold Haeb-Umbach	
8.1	Introduction	188
8.2	Uncertainty Decoding for ASR	189
8.2.1	Bayesian Framework of Speech Recognition	190
8.2.2	Corrupted Features	190
8.3	Conditional Bayesian Estimation of the Feature Posterior	193
8.3.1	The A Priori Model	194
8.3.2	The Observation Model	197
8.3.3	Moments of the Phase Factor Distribution	202
8.3.4	Inference	205
8.4	Experiments	210
8.4.1	Aurora 2 Database	210
8.4.2	Aurora 4 Database	211
8.4.3	Training of the HMMs	211
8.4.4	Training of the A Priori Models	212
8.4.5	Experimental Setup	212
8.4.6	Experimental Results and Discussion	214
8.5	Conclusions	218
	References	220

Part III Applications: Reverberation Robustness

9	Variance Compensation for Recognition of Reverberant Speech with Dereverberation Preprocessing	225
	Marc Delcroix, Shinji Watanabe, Tomohiro Nakatani	
9.1	Introduction	226
9.2	Related Work	227
9.3	Overview of Recognition System	229
9.3.1	Dereverberation Based on Late Reverberation Energy Suppression	229
9.3.2	Speech Recognition Based on Dynamic Variance Compensation	231
9.3.3	Relation with a Conventional Uncertainty Decoding Approach	233
9.4	Proposed Method for Variance Compensation	235
9.4.1	Combination of <i>Static</i> and <i>Dynamic</i> Variance Adaptation	235
9.4.2	Adaptation of Variance Model Parameters	237
9.5	Digit Recognition Experiments	239
9.5.1	Experimental Settings	240
9.5.2	Baseline Results	240
9.5.3	Supervised Adaptation	242

9.5.4	Unsupervised Adaptation	244
9.5.5	Insights about the Proposed Algorithm	244
9.6	Experiment on Large Vocabulary Task	247
9.6.1	Experimental Settings	247
9.6.2	Large Vocabulary Results	248
9.7	Conclusion	249
	Appendix1	250
	Appendix2	251
	References	252
10	A Model-Based Approach to Joint Compensation of Noise and Reverberation for Speech Recognition	257
	Alexander Krueger and Reinhold Haeb-Umbach	
10.1	Introduction	258
10.2	Problem Formulation	261
10.3	Feature Extraction	262
10.4	Bayesian Feature Enhancement Approach	264
10.4.1	A Priori Model	266
10.4.2	Observation Model	268
10.5	Suboptimal Inference	276
10.5.1	Feature Enhancement Algorithm	277
10.6	Experiments	281
10.6.1	Baseline Recognition Results	281
10.6.2	Choice of Model Parameters	282
10.6.3	Recognition Results for Feature Enhancement	285
10.7	Conclusion	287
	References	288
	Part IV Applications: Multiple Speakers and Modalities	
11	Evidence Modeling for Missing Data Speech Recognition Using Small Microphone Arrays	293
	Marco Kühne, Roberto Togneri and Sven Nordholm	
11.1	Introduction	294
11.2	Robust HMM Likelihood Computation Using Evidence Models	296
11.3	Microphone Array-Based Evidence Model Parameter Estimation	301
11.3.1	BSS Front-End	301
11.3.2	Evidence PDF Estimation	306
11.4	Speech Recognition Experiments	308
11.4.1	Room Simulation	308
11.4.2	Model Training	309
11.4.3	Performance Measures	309
11.4.4	Results	310
11.4.5	General Discussion	312
11.5	Summary	314
	References	316

12 Recognition of Multiple Speech Sources Using ICA	319
Eugen Hoffmann, Dorothea Kolossa and Reinhold Orglmeister	
12.1 Introduction	319
12.2 Blind Source Separation	320
12.2.1 Problem Formulation	321
12.2.2 ICA	322
12.2.3 Permutation Correction	324
12.2.4 Postmasking	326
12.3 Uncertainty Estimation	332
12.3.1 Ideal Uncertainties	333
12.3.2 Masking Error Estimate	333
12.3.3 Uncertainty Propagation	334
12.4 Experiments and Results	334
12.4.1 Recording Conditions	334
12.4.2 Parameter Settings	335
12.4.3 Performance Measures	336
12.4.4 Separation Results	336
12.4.5 Model Training	337
12.4.6 Recognition of Uncertain Data	337
12.4.7 Recognition Results	338
12.5 Conclusions	340
References	341
13 Use of Missing and Unreliable Data for Audiovisual Speech Recognition	345
Alexander Vorwerk, Steffen Zeiler, Dorothea Kolossa, Ramón Fernández Astudillo, Dennis Lerch	
13.1 Introduction	346
13.2 Coupled HMMs in Audiovisual Speech Recognition	346
13.2.1 Two-Stream Coupled HHM for Word Models	347
13.2.2 Missing Features in Coupled HMMs	348
13.3 Audio Features	351
13.4 Video Features	351
13.4.1 Face Recognition	351
13.4.2 Feature Extraction	352
13.5 Audiovisual Speech Recognizer Implementation and Results	356
13.5.1 Face Detection	356
13.5.2 Video Features and Their Uncertainties	358
13.5.3 Audio Features and Their Uncertainties	362
13.5.4 Audiovisual Recognition System	363
13.6 Efficiency Considerations	364
13.6.1 Gaussian Density Computation	364
13.6.2 Conventional Likelihood	366
13.6.3 Uncertainty Decoding	367
13.6.4 Modified Imputation	367

13.6.5	Binary Uncertainties	368
13.6.6	Overview of Uncertain Recognition Techniques	368
13.6.7	Recognition Results	369
13.7	Conclusion.....	371
	References	373
Index		377