

DISTANT SPEECH RECOGNITION

Matthias Wölfel

Universität Karlsruhe (TH), Germany

and

John McDonough

Universität des Saarlandes, Germany



A John Wiley and Sons, Ltd., Publication

Contents

Foreword	xiii
Preface	xvii
1 Introduction	1
1.1 Research and Applications in Academia and Industry	1
1.1.1 <i>Intelligent Home and Office Environments</i>	2
1.1.2 <i>Humanoid Robots</i>	3
1.1.3 <i>Automobiles</i>	4
1.1.4 <i>Speech-to-Speech Translation</i>	6
1.2 Challenges in Distant Speech Recognition	7
1.3 System Evaluation	9
1.4 Fields of Speech Recognition	10
1.5 Robust Perception	12
1.5.1 <i>A Priori Knowledge</i>	12
1.5.2 <i>Phonemic Restoration and Reliability</i>	12
1.5.3 <i>Binaural Masking Level Difference</i>	14
1.5.4 <i>Multi-Microphone Processing</i>	14
1.5.5 <i>Multiple Sources by Different Modalities</i>	15
1.6 Organizations, Conferences and Journals	16
1.7 Useful Tools, Data Resources and Evaluation Campaigns	18
1.8 Organization of this Book	18
1.9 Principal Symbols used Throughout the Book	23
1.10 Units used Throughout the Book	25
2 Acoustics	27
2.1 Physical Aspect of Sound	27
2.1.1 <i>Propagation of Sound in Air</i>	28
2.1.2 <i>The Speed of Sound</i>	29
2.1.3 <i>Wave Equation and Velocity Potential</i>	29
2.1.4 <i>Sound Intensity and Acoustic Power</i>	31
2.1.5 <i>Reflections of Plane Waves</i>	32
2.1.6 <i>Reflections of Spherical Waves</i>	33

2.2	Speech Signals	34
	2.2.1 <i>Production of Speech Signals</i>	34
	2.2.2 <i>Units of Speech Signals</i>	36
	2.2.3 <i>Categories of Speech Signals</i>	39
	2.2.4 <i>Statistics of Speech Signals</i>	39
2.3	Human Perception of Sound	41
	2.3.1 <i>Phase Insensitivity</i>	42
	2.3.2 <i>Frequency Range and Spectral Resolution</i>	42
	2.3.3 <i>Hearing Level and Speech Intensity</i>	42
	2.3.4 <i>Masking</i>	44
	2.3.5 <i>Binaural Hearing</i>	45
	2.3.6 <i>Weighting Curves</i>	45
	2.3.7 <i>Virtual Pitch</i>	46
2.4	The Acoustic Environment	47
	2.4.1 <i>Ambient Noise</i>	47
	2.4.2 <i>Echo and Reverberation</i>	48
	2.4.3 <i>Signal-to-Noise and Signal-to-Reverberation Ratio</i>	51
	2.4.4 <i>An Illustrative Comparison between Close and Distant Recordings</i>	52
	2.4.5 <i>The Influence of the Acoustic Environment on Speech Production</i>	53
	2.4.6 <i>Coloration</i>	54
	2.4.7 <i>Head Orientation and Sound Radiation</i>	55
	2.4.8 <i>Expected Distances between the Speaker and the Microphone</i>	57
2.5	Recording Techniques and Sensor Configuration	58
	2.5.1 <i>Mechanical Classification of Microphones</i>	58
	2.5.2 <i>Electrical Classification of Microphones</i>	59
	2.5.3 <i>Characteristics of Microphones</i>	60
	2.5.4 <i>Microphone Placement</i>	60
	2.5.5 <i>Microphone Amplification</i>	62
2.6	Summary and Further Reading	62
2.7	Principal Symbols	63
3	Signal Processing and Filtering Techniques	65
3.1	Linear Time-Invariant Systems	65
	3.1.1 <i>Time Domain Analysis</i>	66
	3.1.2 <i>Frequency Domain Analysis</i>	69
	3.1.3 <i>z-Transform Analysis</i>	72
	3.1.4 <i>Sampling Continuous-Time Signals</i>	79
3.2	The Discrete Fourier Transform	82
	3.2.1 <i>Realizing LTI Systems with the DFT</i>	85
	3.2.2 <i>Overlap-Add Method</i>	86
	3.2.3 <i>Overlap-Save Method</i>	87
3.3	Short-Time Fourier Transform	87
3.4	Summary and Further Reading	90
3.5	Principal Symbols	91

4	Bayesian Filters	93
4.1	Sequential Bayesian Estimation	95
4.2	Wiener Filter	98
	4.2.1 <i>Time Domain Solution</i>	98
	4.2.2 <i>Frequency Domain Solution</i>	99
4.3	Kalman Filter and Variations	101
	4.3.1 <i>Kalman Filter</i>	101
	4.3.2 <i>Extended Kalman Filter</i>	106
	4.3.3 <i>Iterated Extended Kalman Filter</i>	107
	4.3.4 <i>Numerical Stability</i>	108
	4.3.5 <i>Probabilistic Data Association Filter</i>	110
	4.3.6 <i>Joint Probabilistic Data Association Filter</i>	115
4.4	Particle Filters	121
	4.4.1 <i>Approximation of Probabilistic Expectations</i>	121
	4.4.2 <i>Sequential Monte Carlo Methods</i>	125
4.5	Summary and Further Reading	132
4.6	Principal Symbols	133
5	Speech Feature Extraction	135
5.1	Short-Time Spectral Analysis	136
	5.1.1 <i>Speech Windowing and Segmentation</i>	136
	5.1.2 <i>The Spectrogram</i>	137
5.2	Perceptually Motivated Representation	138
	5.2.1 <i>Spectral Shaping</i>	138
	5.2.2 <i>Bark and Mel Filter Banks</i>	139
	5.2.3 <i>Warping by Bilinear Transform – Time vs Frequency Domain</i>	142
5.3	Spectral Estimation and Analysis	145
	5.3.1 <i>Power Spectrum</i>	145
	5.3.2 <i>Spectral Envelopes</i>	146
	5.3.3 <i>LP Envelope</i>	147
	5.3.4 <i>MVDR Envelope</i>	150
	5.3.5 <i>Perceptual LP Envelope</i>	153
	5.3.6 <i>Warped LP Envelope</i>	153
	5.3.7 <i>Warped MVDR Envelope</i>	156
	5.3.8 <i>Warped-Twice MVDR Envelope</i>	157
	5.3.9 <i>Comparison of Spectral Estimates</i>	159
	5.3.10 <i>Scaling of Envelopes</i>	160
5.4	Cepstral Processing	163
	5.4.1 <i>Definition and Characteristics of Cepstral Sequences</i>	163
	5.4.2 <i>Homomorphic Deconvolution</i>	166
	5.4.3 <i>Calculating Cepstral Coefficients</i>	167
5.5	Comparison between Mel Frequency, Perceptual LP and warped MVDR Cepstral Coefficient Front-Ends	168
5.6	Feature Augmentation	169
	5.6.1 <i>Static and Dynamic Parameter Augmentation</i>	169

5.6.2	<i>Feature Augmentation by Temporal Patterns</i>	171
5.7	Feature Reduction	171
5.7.1	<i>Class Separability Measures</i>	172
5.7.2	<i>Linear Discriminant Analysis</i>	173
5.7.3	<i>Heteroscedastic Linear Discriminant Analysis</i>	176
5.8	Feature-Space Minimum Phone Error	178
5.9	Summary and Further Reading	178
5.10	Principal Symbols	179
6	Speech Feature Enhancement	181
6.1	Noise and Reverberation in Various Domains	183
6.1.1	<i>Frequency Domain</i>	183
6.1.2	<i>Power Spectral Domain</i>	185
6.1.3	<i>Logarithmic Spectral Domain</i>	186
6.1.4	<i>Cepstral Domain</i>	187
6.2	Two Principal Approaches	188
6.3	Direct Speech Feature Enhancement	189
6.3.1	<i>Wiener Filter</i>	189
6.3.2	<i>Gaussian and Super-Gaussian MMSE Estimation</i>	191
6.3.3	<i>RASTA Processing</i>	191
6.3.4	<i>Stereo-Based Piecewise Linear Compensation for Environments</i>	192
6.4	Schematics of Indirect Speech Feature Enhancement	193
6.5	Estimating Additive Distortion	194
6.5.1	<i>Voice Activity Detection-Based Noise Estimation</i>	194
6.5.2	<i>Minimum Statistics Noise Estimation</i>	195
6.5.3	<i>Histogram- and Quantile-Based Methods</i>	196
6.5.4	<i>Estimation of the a Posteriori and a Priori Signal-to-Noise Ratio</i>	197
6.6	Estimating Convolutional Distortion	198
6.6.1	<i>Estimating Channel Effects</i>	199
6.6.2	<i>Measuring the Impulse Response</i>	200
6.6.3	<i>Harmful Effects of Room Acoustics</i>	201
6.6.4	<i>Problem in Speech Dereverberation</i>	201
6.6.5	<i>Estimating Late Reflections</i>	202
6.7	Distortion Evolution	204
6.7.1	<i>Random Walk</i>	204
6.7.2	<i>Semi-random Walk by Polyak Averaging and Feedback</i>	205
6.7.3	<i>Predicted Walk by Static Autoregressive Processes</i>	206
6.7.4	<i>Predicted Walk by Dynamic Autoregressive Processes</i>	207
6.7.5	<i>Predicted Walk by Extended Kalman Filters</i>	209
6.7.6	<i>Correlated Prediction Error Covariance Matrix</i>	210
6.8	Distortion Evaluation	211
6.8.1	<i>Likelihood Evaluation</i>	212
6.8.2	<i>Likelihood Evaluation by a Switching Model</i>	213
6.8.3	<i>Incorporating the Phase</i>	214
6.9	Distortion Compensation	215

6.9.1	<i>Spectral Subtraction</i>	215
6.9.2	<i>Compensating for Channel Effects</i>	217
6.9.3	<i>Distortion Compensation for Distributions</i>	218
6.10	Joint Estimation of Additive and Convolutional Distortions	222
6.11	Observation Uncertainty	227
6.12	Summary and Further Reading	228
6.13	Principal Symbols	229
7	Search: Finding the Best Word Hypothesis	231
7.1	Fundamentals of Search	233
7.1.1	<i>Hidden Markov Model: Definition</i>	233
7.1.2	<i>Viterbi Algorithm</i>	235
7.1.3	<i>Word Lattice Generation</i>	238
7.1.4	<i>Word Trace Decoding</i>	240
7.2	Weighted Finite-State Transducers	241
7.2.1	<i>Definitions</i>	241
7.2.2	<i>Weighted Composition</i>	244
7.2.3	<i>Weighted Determinization</i>	246
7.2.4	<i>Weight Pushing</i>	249
7.2.5	<i>Weighted Minimization</i>	251
7.2.6	<i>Epsilon Removal</i>	253
7.3	Knowledge Sources	255
7.3.1	<i>Grammar</i>	256
7.3.2	<i>Pronunciation Lexicon</i>	263
7.3.3	<i>Hidden Markov Model</i>	264
7.3.4	<i>Context Dependency Decision Tree</i>	264
7.3.5	<i>Combination of Knowledge Sources</i>	273
7.3.6	<i>Reducing Search Graph Size</i>	274
7.4	Fast On-the-Fly Composition	275
7.5	Word and Lattice Combination	278
7.6	Summary and Further Reading	279
7.7	Principal Symbols	281
8	Hidden Markov Model Parameter Estimation	283
8.1	Maximum Likelihood Parameter Estimation	284
8.1.1	<i>Gaussian Mixture Model Parameter Estimation</i>	286
8.1.2	<i>Forward-Backward Estimation</i>	290
8.1.3	<i>Speaker-Adapted Training</i>	296
8.1.4	<i>Optimal Regression Class Estimation</i>	300
8.1.5	<i>Viterbi and Label Training</i>	301
8.2	Discriminative Parameter Estimation	302
8.2.1	<i>Conventional Maximum Mutual Information Estimation Formulae</i>	302
8.2.2	<i>Maximum Mutual Information Training on Word Lattices</i>	306
8.2.3	<i>Minimum Word and Phone Error Training</i>	308

8.2.4	<i>Maximum Mutual Information Speaker-Adapted Training</i>	310
8.3	Summary and Further Reading	313
8.4	Principal Symbols	315
9	Feature and Model Transformation	317
9.1	Feature Transformation Techniques	318
9.1.1	<i>Vocal Tract Length Normalization</i>	318
9.1.2	<i>Constrained Maximum Likelihood Linear Regression</i>	319
9.2	Model Transformation Techniques	320
9.2.1	<i>Maximum Likelihood Linear Regression</i>	321
9.2.2	<i>All-Pass Transform Adaptation</i>	322
9.3	Acoustic Model Combination	332
9.3.1	<i>Combination of Gaussians in the Logarithmic Domain</i>	333
9.4	Summary and Further Reading	334
9.5	Principal Symbols	336
10	Speaker Localization and Tracking	337
10.1	Conventional Techniques	338
10.1.1	<i>Spherical Intersection Estimator</i>	339
10.1.2	<i>Spherical Interpolation Estimator</i>	341
10.1.3	<i>Linear Intersection Estimator</i>	342
10.2	Speaker Tracking with the Kalman Filter	345
10.2.1	<i>Implementation Based on the Cholesky Decomposition</i>	348
10.3	Tracking Multiple Simultaneous Speakers	351
10.4	Audio-Visual Speaker Tracking	352
10.5	Speaker Tracking with the Particle Filter	354
10.5.1	<i>Localization Based on Time Delays of Arrival</i>	356
10.5.2	<i>Localization Based on Steered Beamformer Response Power</i>	356
10.6	Summary and Further Reading	357
10.7	Principal Symbols	358
11	Digital Filter Banks	359
11.1	Uniform Discrete Fourier Transform Filter Banks	360
11.2	Polyphase Implementation	364
11.3	Decimation and Expansion	365
11.4	Noble Identities	368
11.5	Nyquist(M) Filters	369
11.6	Filter Bank Design of De Haan <i>et al.</i>	371
11.6.1	<i>Analysis Prototype Design</i>	372
11.6.2	<i>Synthesis Prototype Design</i>	375
11.7	Filter Bank Design with the Nyquist(M) Criterion	376
11.7.1	<i>Analysis Prototype Design</i>	376
11.7.2	<i>Synthesis Prototype Design</i>	377
11.7.3	<i>Alternative Design</i>	378

11.8	Quality Assessment of Filter Bank Prototypes	379
11.9	Summary and Further Reading	384
11.10	Principal Symbols	384
12	Blind Source Separation	387
12.1	Channel Quality and Selection	388
12.2	Independent Component Analysis	390
	12.2.1 <i>Definition of ICA</i>	390
	12.2.2 <i>Statistical Independence and its Implications</i>	392
	12.2.3 <i>ICA Optimization Criteria</i>	396
	12.2.4 <i>Parameter Update Strategies</i>	403
12.3	BSS Algorithms based on Second-Order Statistics	404
12.4	Summary and Further Reading	407
12.5	Principal Symbols	408
13	Beamforming	409
13.1	Beamforming Fundamentals	411
	13.1.1 <i>Sound Propagation and Array Geometry</i>	411
	13.1.2 <i>Beam Patterns</i>	415
	13.1.3 <i>Delay-and-Sum Beamformer</i>	416
	13.1.4 <i>Beam Steering</i>	421
13.2	Beamforming Performance Measures	426
	13.2.1 <i>Directivity</i>	426
	13.2.2 <i>Array Gain</i>	428
13.3	Conventional Beamforming Algorithms	430
	13.3.1 <i>Minimum Variance Distortionless Response Beamformer</i>	430
	13.3.2 <i>Array Gain of the MVDR Beamformer</i>	433
	13.3.3 <i>MVDR Beamformer Performance with Plane Wave Interference</i>	433
	13.3.4 <i>Superdirective Beamformers</i>	437
	13.3.5 <i>Minimum Mean Square Error Beamformer</i>	439
	13.3.6 <i>Maximum Signal-to-Noise Ratio Beamformer</i>	441
	13.3.7 <i>Generalized Sidelobe Canceler</i>	441
	13.3.8 <i>Diagonal Loading</i>	445
13.4	Recursive Algorithms	447
	13.4.1 <i>Gradient Descent Algorithms</i>	448
	13.4.2 <i>Least Mean Square Error Estimation</i>	450
	13.4.3 <i>Recursive Least Squares Estimation</i>	455
	13.4.4 <i>Square-Root Implementation of the RLS Beamformer</i>	461
13.5	Nonconventional Beamforming Algorithms	465
	13.5.1 <i>Maximum Likelihood Beamforming</i>	466
	13.5.2 <i>Maximum Negentropy Beamforming</i>	471
	13.5.3 <i>Hidden Markov Model Maximum Negentropy Beamforming</i>	477
	13.5.4 <i>Minimum Mutual Information Beamforming</i>	480
	13.5.5 <i>Geometric Source Separation</i>	487

13.6	Array Shape Calibration	488
13.7	Summary and Further Reading	489
13.8	Principal Symbols	491
14	Hands On	493
14.1	Example Room Configurations	494
14.2	Automatic Speech Recognition Engines	496
14.3	Word Error Rate	498
14.4	Single-Channel Feature Enhancement Experiments	499
14.5	Acoustic Speaker-Tracking Experiments	501
14.6	Audio-Video Speaker-Tracking Experiments	503
14.7	Speaker-Tracking Performance vs Word Error Rate	504
14.8	Single-Speaker Beamforming Experiments	505
14.9	Speech Separation Experiments	507
14.10	Filter Bank Experiments	508
14.11	Summary and Further Reading	509
	Appendices	511
A	List of Abbreviations	513
B	Useful Background	517
B.1	Discrete Cosine Transform	517
B.2	Matrix Inversion Lemma	518
B.3	Cholesky Decomposition	519
B.4	Distance Measures	519
B.5	Super-Gaussian Probability Density Functions	521
	<i>B.5.1 Generalized Gaussian pdf</i>	521
	<i>B.5.2 Super-Gaussian pdfs with the Meier G-function</i>	523
B.6	Entropy	528
B.7	Relative Entropy	529
B.8	Transformation Law of Probabilities	529
B.9	Cascade of Warping Stages	530
B.10	Taylor Series	530
B.11	Correlation and Covariance	531
B.12	Bessel Functions	531
B.13	Proof of the Nyquist–Shannon Sampling Theorem	532
B.14	Proof of Equations (11.31–11.32)	532
B.15	Givens Rotations	534
B.16	Derivatives with Respect to Complex Vectors	537
B.17	Perpendicular Projection Operators	540
	Bibliography	541
	Index	561