Table of Contents

Preface xiii

Chapter 1. Speech Analysis 1
Christophe D'ALESSANDRO

1.1. Introduction 1
1.1.1. Source-filter model 1
1.1.2. Speech sounds 2
1.1.3. Sources 6
1.1.4. Vocal tract 12
1.1.5. Lip-radiation 18
1.2. Linear prediction 18
1.2.1. Source-filter model and linear prediction 18
1.2.2. Autocorrelation method: algorithm 21
1.2.3. Lattice filter 28
1.2.4. Models of the excitation 31
1.3. Short-term Fourier transform 35
1.3.1. Spectrogram 35
1.3.2. Interpretation in terms of filter bank 36
1.3.3. Block-wise interpretation 37
1.3.4. Modification and reconstruction 38
1.4. A few other representations 39
1.4.1. Bilinear time-frequency representations 39
1.4.2. Wavelets 41
1.4.3. Cepstrum 43
1.4.4. Sinusoidal and harmonic representations 46
1.5. Conclusion 49
1.6. References 50

Chapter 2. Principles of Speech Coding 55
Gang FENG and Laurent GIRIN

2.1. Introduction. 55
2.1.1. Main characteristics of a speech coder 57
2.1.2. Key components of a speech coder 59
2.2. Telephone-bandwidth speech coders 63
2.2.1. From predictive coding to CELP. 65
2.2.2. Improved CELP coders 69
2.2.3. Other coders for telephone speech 77
2.3. Wideband speech coding 79
2.3.1. Transform coding. 81
2.3.2. Predictive transform coding. 85
2.4. Audiovisual speech coding. 86
2.4.1. A transmission channel for audiovisual speech 86
2.4.2. Joint coding of audio and video parameters 88
2.4.3. Prospects 93
2.5. References 93

Chapter 3. Speech Synthesis. 99
Olivier BOËFFARD and Christophe D'ALESSANDRO

3.1. Introduction. 99
3.2. Key goal: speaking for communicating 100
3.2.1. What acoustic content? 101
3.2.2. What melody? 102
3.2.3. Beyond the strict minimum 103
3.3 Synoptic presentation of the elementary modules in speech synthesis systems 104
3.3.1. Linguistic processing. 105
3.3.2. Acoustic processing 105
3.3.3. Training models automatically 106
3.3.4. Operational constraints 107
3.4. Description of linguistic processing 107
3.4.1. Text pre-processing. 107
3.4.2. Grapheme-to-phoneme conversion 108
3.4.3. Syntactic-prosodic analysis 110
3.4.4. Prosodic analysis 112
3.5. Acoustic processing methodology 114
3.5.1. Rule-based synthesis 114
3.5.2. Unit-based concatenative synthesis 115
3.6. Speech signal modeling. 117
3.6.1. The source-filter assumption 118
3.6.2. Articulatory model 119
3.6.3. Formant-based modeling 119

3.6.4. Auto-regressive modeling 120
3.6.5. Harmonic plus noise model 120
3.7. Control of prosodic parameters: the PSOLA technique 122
3.7.1. Methodology background 124
3.7.2. The ancestors of the method 125
3.7.3. Descendants of the method 128
3.7.4. Evaluation 131
3.8. Towards variable-size acoustic units 131
3.8.1. Constitution of the acoustic database 134
3.8.2. Selection of sequences of units 138
3.9. Applications and standardization 142
3.10. Evaluation of speech synthesis. 144
3.10.1. Introduction 144
3.10.2. Global evaluation 146
3.10.3. Analytical evaluation 151
3.10.4. Summary for speech synthesis evaluation. 153
3.11. Conclusions 154
3.12. References. 154

Chapter 4. Facial Animation for Visual Speech 169
Thierry GUIARD-MARIGNY

4.1. Introduction. 169
4.2. Applications of facial animation for visual speech. 170
4.2.1. Animation movies 170
4.2.2. Telecommunications 170
4.2.3. Human-machine interfaces 170
4.2.4. A tool for speech research. 171
4.3. Speech as a bimodal process. 171
4.3.1. The intelligibility of visible speech 172
4.3.2. Visemes for facial animation 174
4.3.3. Synchronization issues. 175
4.3.4. Source consistency 176
4.3.5. Key constraints for the synthesis of visual speech. 177
4.4. Synthesis of visual speech 178
4.4.1. The structure of an artificial talking head. 178
4.4.2. Generating expressions 178
4.5. Animation. 180
4.5.1. Analysis of the image of a face. 180
4.5.2. The puppeteer 181
4.5.3. Automatic analysis of the speech signal 181
4.5.4. From the text to the phonetic string 181
4.6. Conclusion 182
4.7. References 182

Chapter 5. Computational Auditory Scene Analysis 189
Alain DE CHEVEIGNÉ

5.1. Introduction. 189
5.2. Principles of auditory scene analysis 191
5.2.1. Fusion versus segregation: choosing a representation 191
5.2.2. Features for simultaneous fusion 191
5.2.3. Features for sequential fusion. 192
5.2.4. Schemes 193
5.2.5. Illusion of continuity, phonemic restoration 193
5.3. CASA principles. 193
5.3.1. Design of a representation. 193
5.4. Critique of the CASA approach 200
5.4.1. Limitations of ASA. 201
5.4.2. The conceptual limits of “separable representation” 202
5.4.3. Neither a model, nor a method? 203
5.5. Perspectives 203
5.5.1. Missing feature theory 203
5.5.2. The cancellation principle 204
5.5.3. Multimodal integration 205
5.5.4. Auditory scene synthesis: transparency measure 205
5.6. References 206

Chapter 6. Principles of Speech Recognition 213
Renato DE MORI and Brigitte BIGI

6.1. Problem definition and approaches to the solution. 213
6.2. Hidden Markov models for acoustic modeling 216
6.2.1. Definition. 216
6.2.2. Observation probability and model parameters 217
6.2.3. HMM as probabilistic automata 218
6.2.4. Forward and backward coefficients 219
6.3. Observation probabilities. 222
6.4. Composition of speech unit models 223
6.5. The Viterbi algorithm. 226
6.6. Language models 228
6.6.1. Perplexity as an evaluation measure for language models 230
6.6.2. Probability estimation in the language model 232
6.6.3. Maximum likelihood estimation 234
6.6.4. Bayesian estimation 235
6.7. Conclusion 236
6.8. References 237

Chapter 7. Speech Recognition Systems 239
Jean-Luc GAUVAIN and Lori LAMEL

7.1. Introduction. 239
7.2. Linguistic model 241
7.3. Lexical representation. 244
7.4. Acoustic modeling. 247
7.4.1. Feature extraction. 247
7.4.2. Acoustic-phonetic models 249
7.4.3. Adaptation techniques 253
7.5. Decoder 256
7.6. Applicative aspects 257
7.6.1. Efficiency: speed and memory 257
7.6.2. Portability: languages and applications 259
7.6.3. Confidence measures. 260
7.6.4. Beyond words 261
7.7. Systems 261
7.7.1. Text dictation 262
7.7.2. Audio document indexing 263
7.7.3. Dialog systems 265
7.8. Perspectives 268
7.9. References 270

Chapter 8. Language Identification 279
Martine ADDA-DECKER

8.1. Introduction. 279
8.2. Language characteristics 281
8.3. Language identification by humans. 286
8.4. Language identification by machines. 287
8.4.1. LId tasks 288
8.4.2. Performance measures 288
8.4.3. Evaluation 289
8.5. LId resources 290
8.6. LId formulation 295
8.7. Lid modeling 298
8.7.1. Acoustic front-end 299
8.7.2. Acoustic language-specific modeling 300
8.7.3. Parallel phone recognition. 302
8.7.4. Phonotactic modeling 304
8.7.5. Back-end optimization. 309
8.8. Discussion 309
8.9. References 311

Chapter 9. Automatic Speaker Recognition 321
Frédéric BIMBOT.

9.1. Introduction 321
9.1.1. Voice variability and characterization 321
9.1.2. Speaker recognition 323
9.2. Typology and operation of speaker recognition systems 324
9.2.1. Speaker recognition tasks 324
9.2.2. Operation 325
9.2.3. Text-dependence 326
9.2.4. Types of errors 327
9.2.5. Influencing factors 328
9.3. Fundamentals 329
9.3.1. General structure of speaker recognition systems 329
9.3.2. Acoustic analysis 330
9.3.3. Probabilistic modeling 331
9.3.4. Identification and verification scores 335
9.3.5. Score compensation and decision 337
9.3.6. From theory to practice 342
9.4. Performance evaluation 343
9.4.1. Error rate 343
9.4.2. DET curve and EER 344
9.4.3. Cost function, weighted error rate and HTER 346
9.4.4. Distribution of errors 346
9.4.5. Orders of magnitude 347
9.5. Applications 348
9.5.1. Physical access control 348
9.5.2. Securing remote transactions 349
9.5.3. Audio information indexing 350
9.5.4. Education and entertainment 350
9.5.5. Forensic applications 351
9.5.6. Perspectives 352
9.6. Conclusions 352
9.7. Further reading 353

Chapter 10. Robust Recognition Methods 355
Jean-Paul HATON

10.1. Introduction 355
10.2. Signal pre-processing methods 357
10.2.1. Spectral subtraction 357
10.2.2. Adaptive noise cancellation 358
10.2.3. Space transformation 359
10.2.4. Channel equalization 359
10.2.5. Stochastic models 360
10.3. Robust parameters and distance measures 360

10.3.1. Spectral representations 361
10.3.2. Auditory models. 364
10.3.3 Distance measure 365
10.4. Adaptation methods 366
10.4.1 Model composition 366
10.4.2. Statistical adaptation 367
10.5. Compensation of the Lombard effect 368
10.6. Missing data scheme. 369
10.7. Conclusion 369
10.8. References. 370

Chapter 11. Multimodal Speech: Two or Three senses are Better than One 377
Jean-Luc SCHWARTZ, Pierre ESCUDIER and Pascal TEISSIER

11.1. Introduction 377
11.2. Speech is a multimodal process 379
11.2.1. Seeing without hearing 379
11.2.2. Seeing for hearing better in noise. 380
11.2.3. Seeing for better hearing… even in the absence of noise. 382
11.2.4. Bimodal integration imposes itself to perception 383
11.2.5. Lip reading as taking part to the ontogenesis of speech. 385
11.2.6. ...and to its phylogenesis ? 386
11.3. Architectures for audio-visual fusion in speech perception 388
11.3.1.Three paths for sensory interactions in cognitive psychology . . . 389
11.3.2. Three paths for sensor fusion in information processing 390
11.3.3. The four basic architectures for audiovisual fusion 391
11.3.4. Three questions for a taxonomy 392
11.3.5. Control of the fusion process 394
11.4. Audio-visual speech recognition systems 396
11.4.1. Architectural alternatives 397
11.4.2. Taking into account contextual information 401
11.4.3. Pre-processing 403
11.5. Conclusions 405
11.6. References. 406

Chapter 12. Speech and Human-Computer Communication 417
Wolfgang MINKER & Françoise NÉEL

12.1. Introduction 417
12.2. Context. 418
12.2.1. The development of micro-electronics. 419
12.2.2. The expansion of information and communication technologies and increasing interconnection of computer systems 420

12.2.3. The coordination of research efforts and the improvement of automatic speech processing systems 421
12.3. Specificities of speech. 424
12.3.1. Advantages of speech as a communication mode 424
12.3.2. Limitations of speech as a communication mode 425
12.3.3. Multidimensional analysis of commercial speech recognition products 427
12.4. Application domains with voice-only interaction. 430
12.4.1. Inspection, control and data acquisition 431
12.4.2. Home automation: electronic home assistant. 432
12.4.3. Office automation: dictation and speech-to-text systems 432
12.4.4. Training. 435
12.4.5. Automatic translation 438
12.5. Application domains with multimodal interaction 439
12.5.1. Interactive terminals 440
12.5.2. Computer-aided graphic design. 441
12.5.3. On-board applications 442
12.5.4. Human-human communication facilitation 444
12.5.5. Automatic indexing of audio-visual documents. 446
12.6. Conclusions 446
12.7. References. 447

Chapter 13. Voice Services in the Telecom Sector 455
Laurent COURTOIS, Patrick BRISARD and Christian GAGNOULET

13.1. Introduction 455
13.2. Automatic speech processing and telecommunications 456
13.3. Speech coding in the telecommunication sector 456
13.4. Voice command in telecom services 457
13.4.1. Advantages and limitations of voice command 457
13.4.2. Major trends 459
13.4.3. Major voice command services. 460
13.4.4. Call center automation (operator assistance) 460
13.4.5. Personal voice phonebook 462
13.4.6. Voice personal telephone assistants 463
13.4.7. Other services based on voice command 463
13.5. Speaker verification in telecom services 464
13.6. Text-to-speech synthesis in telecommunication systems 464
13.7. Conclusions 465
13.8. References. 466

List of Authors 467

Index 471