Text-to-Speech Synthesis

PAUL TAYLOR
University of Cambridge

CAMBRIDGE
UNIVERSITY PRESS

Contents

Foreword *page* xxiii
Preface xxvii

1 Introduction 1

1.1 What are text-to-speech systems for? 2
1.2 What should the goals of text-to-speech system development be? 3
1.3 The engineering approach 4
1.4 Overview of the book 5
1.4.1 Viewpoints within the book 5
1.4.2 Readers' backgrounds 6
1.4.3 Background and specialist sections 7

2 Communication and language 8

2.1 Types of communication 8
2.1.1 Affective communication 8
2.1.2 Iconic communication 9
2.1.3 Symbolic communication 10
2.1.4 Combinations of symbols 11
2.1.5 Meaning, form and signal 12
2.2 Human communication 13
2.2.1 Verbal communication 14
2.2.2 Linguistic levels 16
2.2.3 Affective prosody 17
2.2.4 Augmentative prosody 18
2.3 Communication processes 18
2.3.1 Communication factors 19
2.3.2 Generation 20
2.3.3 Encoding 21
2.3.4 Decoding 22
2.3.5 Understanding 22
2.4 Discussion 23
2.5 Summary 24

3 The text-to-speech problem 26

3.1 Speech and writing 26
3.1.1 Physical nature 27
3.1.2 Spoken form and written form 28
3.1.3 Use 29
3.1.4 Prosodic and verbal content 30
3.1.5 Component balance 31
3.1.6 Non-linguistic content 32
3.1.7 Semiotic systems 33
3.1.8 Writing systems 34
3.2 Reading aloud 35
3.2.1 Reading silently and reading aloud 35
3.2.2 Prosody in reading aloud 36
3.2.3 Verbal content and style in reading aloud 37
3.3 Text-to-speech system organisation 37
3.3.1 The common-form model 38
3.3.2 Other models 39
3.3.3 Comparison 40
3.4 Systems 41
3.4.1 A simple text-to-speech system 41
3.4.2 Concept to speech 42
3.4.3 Canned speech and limited-domain synthesis 43
3.5 Key problems in text-to-speech 44
3.5.1 Text classification with respect to semiotic systems 44
3.5.2 Decoding natural-language text 46
3.5.3 Naturalness 47
3.5.4 Intelligibility: encoding the message in signal 48
3.5.5 Auxiliary generation for prosody 49
3.5.6 Adapting the system to the situation 50
3.6 Summary 50

4 Text segmentation and organisation 52

4.1 Overview of the problem 52
4.2 Words and sentences 53
4.2.1 What is a word? 53
4.2.2 Defining words in text-to-speech 55
4.2.3 Scope and morphology 59
4.2.4 Contractions and clitics 60
4.2.5 Slang forms 61
4.2.6 Hyphenated forms 61
4.2.7 What is a sentence? 62
4.2.8 The lexicon 63
4.3 Text segmentation 63

4.3.1 Tokenisation 64
4.3.2 Tokenisation and punctuation 65
4.3.3 Tokenisation algorithms 66
4.3.4 Sentence splitting 67
4.4 Processing documents 68
4.4.1 Markup languages 68
4.4.2 Interpreting characters 70
4.5 Text-to-speech architectures 71
4.6 Discussion 75
4.6.1 Further reading 75
4.6.2 Summary 76

5 Text decoding: finding the words from the text 78

5.1 Overview of text decoding 78
5.2 Text-classification algorithms 79
5.2.1 Features and algorithms 79
5.2.2 Tagging and word-sense disambiguation 82
5.2.3 Ad-hoc approaches 83
5.2.4 Deterministic rule approaches 83
5.2.5 Decision lists 85
5.2.6 Naive Bayes classifier 86
5.2.7 Decision trees 87
5.2.8 Part-of-speech tagging 88
5.3 Non-natural-language text 92
5.3.1 Semiotic classification 92
5.3.2 Semiotic decoding 95
5.3.3 Verbalisation 95
5.4 Natural-language text 97
5.4.1 Acronyms and letter sequences 99
5.4.2 Homograph disambiguation 99
5.4.3 Non-homographs 101
5.5 Natural-language parsing 102
5.5.1 Context-free grammars 102
5.5.2 Statistical parsing 105
5.6 Discussion 105
5.6.1 Further reading 108
5.6.2 Summary 109

6 Prosody prediction from text 111

6.1 Prosodic form 111
6.2 Phrasing 112
6.2.1 Phrasing phenomena 112
6.2.2 Models of phrasing 113

6.3 Prominence 115
6.3.1 Syntactic prominence patterns 116
6.3.2 Discourse prominence patterns 118
6.3.3 Prominence systems, data and labelling 119
6.4 Intonation and tune 121
6.5 Prosodic meaning and function 122
6.5.1 Affective prosody 123
6.5.2 Suprasegmentality 124
6.5.3 Augmentative prosody 125
6.5.4 Symbolic communication and prosodic style 126
6.6 Determining prosody from the text 127
6.6.1 Prosody and human reading 127
6.6.2 Controlling the degree of augmentative prosody 128
6.6.3 Prosody and synthesis techniques 128
6.7 Phrasing prediction 129
6.7.1 Experimental formulation 129
6.7.2 Deterministic approaches 130
6.7.3 Classifier approaches 132
6.7.4 HMM approaches 133
6.7.5 Hybrid approaches 135
6.8 Prominence prediction 136
6.8.1 Compound-noun phrases 136
6.8.2 Function-word prominence 138
6.8.3 Data-driven approaches 138
6.9 Intonational-tune prediction 139
6.10 Discussion 139
6.10.1 Labelling schemes and labelling accuracy 139
6.10.2 Linguistic theories and prosody 141
6.10.3 Synthesising suprasegmental and true prosody 142
6.10.4 Prosody in real dialogues 143
6.10.5 Conclusion 144
6.10.6 Summary 144

7 Phonetics and phonology 146

7.1 Articulatory phonetics and speech production 146
7.1.1 The vocal organs 147
7.1.2 Sound sources 147
7.1.3 Sound output 150
7.1.4 The vocal-tract filter 150
7.1.5 Vowels 151
7.1.6 Consonants 153
7.1.7 Examining speech production 155
7.2 Acoustics, phonetics and speech perception 156

7.2.1 Acoustic representations 156
7.2.2 Acoustic characteristics 159
7.3 The communicative use of speech 160
7.3.1 Communicating discrete information with a continuous channel 161
7.3.2 Phonemes, phones and allophones 162
7.3.3 Allophonic variation and phonetic context 166
7.3.4 Coarticulation, targets and transients 168
7.3.5 The continuous nature of speech 169
7.3.6 Transcription 170
7.3.7 The distinctiveness of speech in communication 171
7.4 Phonology: the linguistic organisation of speech 172
7.4.1 Phonotactics 172
7.4.2 Word formation 179
7.4.3 Distinctive features and phonological theories 181
7.4.4 Syllables 184
7.4.5 Lexical stress 186
7.5 Discussion 189
7.5.1 Further reading 189
7.5.2 Summary 190

8 Pronunciation 192

8.1 Pronunciation representations 192
8.1.1 Why bother? 192
8.1.2 Phonemic and phonetic input 193
8.1.3 Difficulties in deriving phonetic input 194
8.1.4 A structured approach to pronunciation 195
8.1.5 Abstract phonological representations 196
8.2 Formulating a phonological representation system 197
8.2.1 Simple consonants and vowels 197
8.2.2 Difficult consonants 199
8.2.3 Diphthongs and affricates 201
8.2.4 Approximant–vowel combinations 201
8.2.5 Defining the full inventory 203
8.2.6 Phoneme names 204
8.2.7 Syllabic issues 206
8.3 The lexicon 207
8.3.1 Lexicon and rules 208
8.3.2 Lexicon formats 210
8.3.3 The offline lexicon 213
8.3.4 The system lexicon 214
8.3.5 Lexicon quality 215
8.3.6 Determining the pronunciations of unknown words 216
8.4 Grapheme-to-phoneme conversion 218

8.4.1 Rule-based techniques 218
8.4.2 Grapheme-to-phoneme alignment 219
8.4.3 Neural networks 219
8.4.4 Pronunciation by analogy 220
8.4.5 Other data-driven techniques 221
8.4.6 Statistical techniques 221
8.5 Further issues 222
8.5.1 Morphology 222
8.5.2 Language origin and names 223
8.5.3 Post-lexical processing 223
8.6 Summary 224

9 Synthesis of prosody 225

9.1 Intonation overview 225
9.1.1 F0 and pitch 226
9.1.2 Intonational form 226
9.1.3 Models of F0 contours 227
9.1.4 Micro-prosody 229
9.2 Intonational behaviour 229
9.2.1 Intonational tune 229
9.2.2 Downdrift 230
9.2.3 Pitch range 233
9.2.4 Pitch accents and boundary tones 234
9.3 Intonation theories and models 236
9.3.1 Traditional models and the British school 236
9.3.2 The Dutch school 237
9.3.3 Autosegmental–metrical and ToBI models 237
9.3.4 The INTSINT model 239
9.3.5 The Fujisaki model and superimpositional models 239
9.3.6 The Tilt model 242
9.3.7 Comparison 244
9.4 Intonation synthesis with AM models 245
9.4.1 Prediction of AM labels from text 246
9.4.2 Deterministic synthesis methods 246
9.4.3 Data-driven synthesis methods 247
9.4.4 Analysis with autosegmental models 248
9.5 Intonation synthesis with deterministic acoustic models 248
9.5.1 Synthesis with superimpositional models 249
9.5.2 Synthesis with the Tilt model 249
9.5.3 Analysis with Fujisaki and Tilt models 250
9.6 Data-driven intonation models 250
9.6.1 Unit-selection-style approaches 251
9.6.2 Dynamic-system models 252

9.6.3 Hidden Markov models 253
9.6.4 Functional models 254
9.7 Timing 254
9.7.1 Formulation of the timing problem 255
9.7.2 The nature of timing 255
9.7.3 Klatt rules 256
9.7.4 The sums-of-products model 257
9.7.5 The Campbell model 258
9.7.6 Other regression techniques 259
9.8 Discussion 259
9.8.1 Further reading 260
9.8.2 Summary 260

10 Signals and filters 262

10.1 Analogue signals 262
10.1.1 Simple periodic signals: sinusoids 263
10.1.2 General periodic signals 265
10.1.3 Sinusoids as complex exponentials 266
10.1.4 Fourier analysis 269
10.1.5 The frequency domain 270
10.1.6 The Fourier transform 275
10.2 Digital signals 278
10.2.1 Digital waveforms 279
10.2.2 Digital representations 280
10.2.3 The discrete-time Fourier transform 280
10.2.4 The discrete Fourier transform 281
10.2.5 The z-transform 282
10.2.6 The frequency domain for digital signals 283
10.3 Properties of transforms 284
10.3.1 Linearity 284
10.3.2 Time and frequency duality 284
10.3.3 Scaling 285
10.3.4 Impulse properties 285
10.3.5 Time delay 286
10.3.6 Frequency shift 286
10.3.7 Convolution 287
10.3.8 Analytical and numerical analysis 287
10.3.9 Stochastic signals 288
10.4 Digital filters 288
10.4.1 Difference equations 289
10.4.2 The impulse response 289
10.4.3 The filter convolution sum 292
10.4.4 The filter transfer function 293

10.4.5 The transfer function and the impulse response 293
10.5 Digital filter analysis and design 294
10.5.1 Polynomial analysis: poles and zeros 294
10.5.2 Frequency interpretation of the z-domain transfer function 297
10.5.3 Filter characteristics 298
10.5.4 Putting it all together 304
10.6 Summary 305

11 Acoustic models of speech production 309

11.1 The acoustic theory of speech production 309
11.1.1 Components in the model 309
11.2 The physics of sound 311
11.2.1 Resonant systems 311
11.2.2 Travelling waves 313
11.2.3 Acoustic waves 315
11.2.4 Acoustic reflection 317
11.3 The vowel-tube model 318
11.3.1 Discrete time and distance 319
11.3.2 Junction of two tubes 320
11.3.3 Special cases of junctions 322
11.3.4 The two-tube vocal-tract model 323
11.3.5 The single-tube model 325
11.3.6 The multi-tube vocal-tract model 327
11.3.7 The all-pole resonator model 329
11.4 Source and radiation models 330
11.4.1 Radiation 330
11.4.2 The glottal source 330
11.5 Model refinements 333
11.5.1 Modelling the nasal cavity 333
11.5.2 Source positions in the oral cavity 334
11.5.3 Models with vocal-tract losses 335
11.5.4 Source and radiation effects 336
11.6 Discussion 336
11.6.1 Further reading 339
11.6.2 Summary 339

12 Analysis of speech signals 341

12.1 Short-term speech analysis 341
12.1.1 Windowing 342
12.1.2 Short-term spectral representations 343
12.1.3 Frame lengths and shifts 345
12.1.4 The spectrogram 348
12.1.5 Auditory scales 351

12.2 Filter-bank analysis 352
12.3 The cepstrum 353
12.3.1 Cepstrum definition 353
12.3.2 Treating the magnitude spectrum as a signal 353
12.3.3 Cepstral analysis as deconvolution 355
12.3.4 Cepstral analysis discussion 356
12.4 Linear-prediction analysis 357
12.4.1 Finding the coefficients: the covariance method 358
12.4.2 The autocorrelation method 360
12.4.3 Levinson–Durbin recursion 361
12.5 Spectral-envelope and vocal-tract representations 362
12.5.1 Linear-prediction spectra 362
12.5.2 Transfer-function poles 364
12.5.3 Reflection coefficients 366
12.5.4 Log area ratios 367
12.5.5 Line-spectrum frequencies 367
12.5.6 Linear-prediction cepstra 369
12.5.7 Mel-scaled cepstra 370
12.5.8 Perceptual linear prediction 370
12.5.9 Formant tracking 370
12.6 Source representations 372
12.6.1 Residual signals 372
12.6.2 Closed-phase analysis 374
12.6.3 Open-phase analysis 377
12.6.4 Impulse/noise models 378
12.6.5 Parameterisation of glottal-flow signals 379
12.7 Pitch and epoch detection 379
12.7.1 Pitch detection 379
12.7.2 Epoch detection: finding the instant of glottal closure 381
12.8 Discussion 384
12.8.1 Further reading 385
12.8.2 Summary 386

13 Synthesis techniques based on vocal-tract models 387

13.1 Synthesis specification: the input to the synthesiser 387
13.2 Formant synthesis 388
13.2.1 Sound sources 389
13.2.2 Synthesising a single formant 390
13.2.3 Resonators in series and parallel 391
13.2.4 Synthesising consonants 392
13.2.5 A complete synthesiser 394
13.2.6 The phonetic input to the synthesiser 394
13.2.7 Formant-synthesis quality 397

13.3 Classical linear-prediction synthesis 399
13.3.1 Comparison with formant synthesis 399
13.3.2 The impulse/noise source model 400
13.3.3 Linear-prediction diphone-concatenative synthesis 401
13.3.4 A complete synthesiser 403
13.3.5 Problems with the source 404
13.4 Articulatory synthesis 405
13.5 Discussion 407
13.5.1 Further reading 409
13.5.2 Summary 410

14 Synthesis by concatenation and signal-processing modification 412

14.1 Speech units in second-generation systems 413
14.1.1 Creating a diphone inventory 414
14.1.2 Obtaining diphones from speech 414
14.2 Pitch-synchronous overlap and add (PSOLA) 415
14.2.1 Time-domain PSOLA 416
14.2.2 Epoch manipulation 417
14.2.3 How does PSOLA work? 421
14.3 Residual-excited linear prediction (RELP) 421
14.3.1 Residual manipulation 423
14.3.2 Linear-prediction PSOLA 423
14.4 Sinusoidal models 424
14.4.1 Pure sinusoidal models 425
14.4.2 Harmonic/noise models 426
14.5 MBROLA 429
14.6 Synthesis from cepstral coefficients 429
14.7 Concatenation issues 431
14.8 Discussion 433
14.8.1 Further reading 433
14.8.2 Summary 433

15 Hidden-Markov-model synthesis 435

15.1 The HMM formalism 435
15.1.1 Observation probabilities 436
15.1.2 Delta coefficients 438
15.1.3 Acoustic representations and covariance 439
15.1.4 States and transitions 440
15.1.5 Recognising with HMMs 440
15.1.6 Language models 443
15.1.7 The Viterbi algorithm 444
15.1.8 Training HMMS 447
15.1.9 Context-sensitive modelling 451

15.1.10 Are HMMs a good model of speech? 454
15.2 Synthesis from hidden Markov models 456
15.2.1 Finding the likeliest observations given the state sequence 457
15.2.2 Finding the likeliest observations and state sequence 459
15.2.3 Acoustic representations 459
15.2.4 Context-sensitive synthesis models 463
15.2.5 Duration modelling 464
15.2.6 Signal Processing in HMM synthesis 464
15.2.7 HMM synthesis systems 465
15.3 Labelling databases with HMMs 467
15.3.1 Determining the word sequence 467
15.3.2 Determining the phone sequence 468
15.3.3 Determining the phone boundaries 468
15.3.4 Measuring the quality of the alignments 470
15.4 Other data-driven synthesis techniques 471
15.5 Discussion 471
15.5.1 Further reading 471
15.5.2 Summary 472

16 Unit-selection synthesis 474

16.1 From concatenative synthesis to unit selection 474
16.1.1 Extending concatenative synthesis 475
16.1.2 The Hunt and Black algorithm 477
16.2 Features 479
16.2.1 Base types 479
16.2.2 Linguistic and acoustic features 480
16.2.3 Choice of features 481
16.2.4 Types of features 482
16.3 The independent-feature target-function formulation 484
16.3.1 The purpose of the target function 484
16.3.2 Defining a perceptual space 485
16.3.3 Perceptual spaces defined by independent features 486
16.3.4 Setting the target weights using acoustic distances 488
16.3.5 Limitations of the independent-feature formulation 491
16.4 The acoustic-space target-function formulation 493
16.4.1 Decision-tree clustering 494
16.4.2 General partial-synthesis functions 496
16.5 Join functions 497
16.5.1 Basic issues in joining units 497
16.5.2 Phone-class join costs 498
16.5.3 Acoustic-distance join costs 499
16.5.4 Combining categorical and and acoustic join costs 500
16.5.5 Probabilistic and sequence join costs 501

16.5.6 Join classifiers 502
16.6 Searching 504
16.6.1 Base types and searching 505
16.6.2 Pruning 508
16.6.3 Pre-selection 508
16.6.4 Beam pruning 509
16.6.5 Multi-pass searching 509
16.7 Discussion 510
16.7.1 Unit selection and signal processing 511
16.7.2 Features, costs and perception 511
16.7.3 Example unit-selection systems 512
16.7.4 Further reading 514
16.7.5 Summary 515

17 Further issues 517

17.1 Databases 517
17.1.1 Unit-selection databases 517
17.1.2 Text materials 518
17.1.3 Prosody databases 518
17.1.4 Labelling 519
17.1.5 What exactly is hand labelling? 519
17.1.6 Automatic labelling 521
17.1.7 Avoiding explicit labels 521
17.2 Evaluation 522
17.2.1 System testing: intelligibility and naturalness 523
17.2.2 Word-recognition tests 523
17.2.3 Naturalness tests 524
17.2.4 Test data 525
17.2.5 Unit or component testing 525
17.2.6 Competitive evaluations 526
17.3 Audio-visual speech synthesis 527
17.3.1 Speech control 528
17.4 Synthesis of emotional and expressive speech 529
17.4.1 Describing emotion 529
17.4.2 Synthesising emotion with prosody control 529
17.4.3 Synthesising emotion with voice transformation 530
17.4.4 Unit selection and HMM techniques 531
17.5 Summary 531

18 Conclusion 533

18.1 Speech technology and linguistics 533
18.2 Future directions 536
18.3 Conclusion 539

Appendix A Probability 540

A.1 Discrete probabilities 540
A.1.1 Discrete random variables 540
A.1.2 Probability mass functions 541
A.1.3 Expected values 541
A.1.4 Moments of a PMF 542
A.2 Pairs of discrete random variables 542
A.2.1 Marginal distributions 543
A.2.2 Independence 543
A.2.3 Expected values 543
A.2.4 Moments of a joint distribution 544
A.2.5 Higher-order moments and covariance 544
A.2.6 Correlation 544
A.2.7 Conditional probability 545
A.2.8 Bayes' rule 545
A.2.9 Sum of random variables 545
A.2.10 The chain rule 546
A.2.11 Entropy 546
A.3 Continuous random variables 547
A.3.1 Continuous random variables 548
A.3.2 Expected values 548
A.3.3 The Gaussian distribution 549
A.3.4 The uniform distribution 549
A.3.5 Cumulative density functions 549
A.4 Pairs of continuous random variables 550
A.4.1 Independent versus uncorrelated 551
A.4.2 The sum of two random variables 551
A.4.3 Entropy 552
A.4.4 Kullback–Leibler distance 552

Appendix B Phone definitions 553

References 556
Index 583