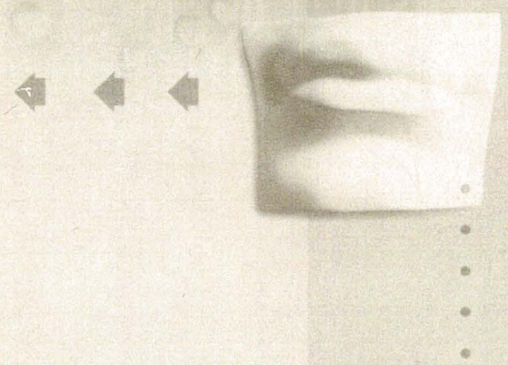
XUEDONG HUANG | ALEX ACERO | HSIAO-WUEN HON

SPOKEN LANGUAGE PROCESSING

A Guide to Theory, Algorithm, and System Development



Foreword by Dr. Raj Reddy
Carnegie Mellon University

Amazon/VB Assets
Exhibit 1011
Page 1

Spoken Language Processing



Amazon/VB Assets
Exhibit 1011
Page 3

Spoken Language Processing

A Guide to Theory, Algorithm, and System Development

Xuedong Huang
Alex Acero
Hsiao-Wuen Hon

Microsoft Research



Prentice Hall PTR
Upper Saddle River, New Jersey 07458
www.phptr.com

Amazon/VB Assets
Exhibit 1011
Page 4

Library of Congress Cataloging-in-Publication Data

Huang, Xuedong.
Spoken language processing: a guide to theory, algorithm, and system development/
Xuedong Huang, Alex Acero, Hsiao-Wuen Hon.
p. cm.
Includes bibliographical references and index.
ISBN 0-13-022616-5
1. Natural language processing (Computer science) I. Acero, Alex. II. Hon,
Hsiao-Wuen. III. Title.

QA76.9.N38 H83 2001
006.3'5—dc21 00-050196

Editorial/production supervision: *Jane Bonnell*
Cover design director: *Jerry Votta*
Cover design: *Anthony Gemmellaro*
Manufacturing buyer: *Maura Zaldivar*
Development editor: *Russ Hall*
Acquisitions editor: *Tim Moore*
Editorial assistant: *Allyson Kloss*
Marketing manager: *Debby van Dijk*



© 2001 by Prentice Hall PTR
Prentice-Hall, Inc.
Upper Saddle River, New Jersey 07458

Prentice Hall books are widely used by corporations and government agencies for training, marketing, and resale.
The publisher offers discounts on this book when ordered in bulk quantities. For more information, contact Corporate Sales Department, Phone: 800-382-3419; FAX: 201-236-7141;
E-mail: corpsales@prenhall.com
Or write: Prentice Hall PTR, Corporate Sales Dept., One Lake Street, Upper Saddle River, NJ 07458.

Company and product names mentioned herein are the trademarks or registered trademarks of their respective owners.

All rights reserved. No part of this book may be
reproduced, in any form or by any means,
without permission in writing from the publisher.

Printed in the United States of America

ISBN 0-13-022616-5

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Pearson Education Asia Pte. Ltd.
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

Amazon/VB Assets
Exhibit 1011
Page 5

To Yingzhi, Angela, Christina, and Derek

To Donna and Nicolas

To Phen, Stephanie, and Jacqueline

Contents

FOREWORD xxi

PREFACE xxv

1. INTRODUCTION 1

1.1. MOTIVATIONS 2
1.1.1. Spoken Language Interface 2
1.1.2. Speech-to-Speech Translation 3
1.1.3. Knowledge Partners 3
1.2. SPOKEN LANGUAGE SYSTEM ARCHITECTURE 4
1.2.1. Automatic Speech Recognition 4
1.2.2. Text-to-Speech Conversion 6
1.2.3. Spoken Language Understanding 7
1.3. BOOK ORGANIZATION 8
1.3.1. Part I: Fundamental Theory 9
1.3.2. Part II: Speech Processing 9
1.3.3. Part III: Speech Recognition 9
1.3.4. Part IV: Text-to-Speech Systems 10
1.3.5. Part V: Spoken Language Systems 10
1.4. TARGET AUDIENCES 10
1.5. HISTORICAL PERSPECTIVE AND FURTHER READING 11

PART I: FUNDAMENTAL THEORY

2. SPOKEN LANGUAGE STRUCTURE 19

2.1. SOUND AND HUMAN SPEECH SYSTEMS 21
2.1.1. Sound 21
2.1.2. Speech Production 24
2.1.3. Speech Perception 29

Amazon/VB Assets
Exhibit 1011
Page 7

2.2. PHONETICS AND PHONOLOGY 36
2.2.1. *Phonemes* 36
2.2.2. *The Allophone: Sound and Context* 47
2.2.3. *Speech Rate and Coarticulation* 49
2.3. SYLLABLES AND WORDS 51
2.3.1. *Syllables* 51
2.3.2. *Words* 53
2.4. SYNTAX AND SEMANTICS 58
2.4.1. *Syntactic Constituents* 58
2.4.2. *Semantic Roles* 63
2.4.3. *Lexical Semantics* 64
2.4.4. *Logical Form* 67
2.5. HISTORICAL PERSPECTIVE AND FURTHER READING 68

3. PROBABILITY, STATISTICS, AND INFORMATION THEORY 73
3.1. PROBABILITY THEORY 74
3.1.1. *Conditional Probability and Bayes' Rule* 75
3.1.2. *Random Variables* 77
3.1.3. *Mean and Variance* 79
3.1.4. *Covariance and Correlation* 82
3.1.5. *Random Vectors and Multivariate Distributions* 83
3.1.6. *Some Useful Distributions* 85
3.1.7. *Gaussian Distributions* 92
3.2. ESTIMATION THEORY 98
3.2.1. *Minimum/Least Mean Squared Error Estimation* 99
3.2.2. *Maximum Likelihood Estimation* 104
3.2.3. *Bayesian Estimation and MAP Estimation* 107
3.3. SIGNIFICANCE TESTING 113
3.3.1. *Level of Significance* 114
3.3.2. *Normal Test (Z-Test)* 115
3.3.3. *χ^2 Goodness-of-Fit Test* 116
3.3.4. *Matched-Pairs Test* 118
3.4. INFORMATION THEORY 120
3.4.1. *Entropy* 120
3.4.2. *Conditional Entropy* 123
3.4.3. *The Source Coding Theorem* 124
3.4.4. *Mutual Information and Channel Coding* 126
3.5. HISTORICAL PERSPECTIVE AND FURTHER READING 128

4. PATTERN RECOGNITION 133
4.1. BAYES' DECISION THEORY 134
4.1.1. *Minimum-Error-Rate Decision Rules* 135

Amazon/VB Assets
Exhibit 1011
Page 8

4.1.2. *Discriminant Functions* 138
4.2. HOW TO CONSTRUCT CLASSIFIERS 140
4.2.1. *Gaussian Classifiers* 142
4.2.2. *The Curse of Dimensionality* 144
4.2.3. *Estimating the Error Rate* 146
4.2.4. *Comparing Classifiers* 148
4.3. DISCRIMINATIVE TRAINING 150
4.3.1. *Maximum Mutual Information Estimation* 150
4.3.2. *Minimum-Error-Rate Estimation* 156
4.3.3. *Neural Networks* 158
4.4. UNSUPERVISED ESTIMATION METHODS 163
4.4.1. *Vector Quantization* 163
4.4.2. *The EM Algorithm* 170
4.4.3. *Multivariate Gaussian Mixture Density Estimation* 172
4.5. CLASSIFICATION AND REGRESSION TREES 175
4.5.1. *Choice of Question Set* 177
4.5.2. *Splitting Criteria* 178
4.5.3. *Growing the Tree* 181
4.5.4. *Missing Values and Conflict Resolution* 182
4.5.5. *Complex Questions* 182
4.5.6. *The Right-Sized Tree* 184
4.6. HISTORICAL PERSPECTIVE AND FURTHER READING 190

PART II: SPEECH PROCESSING

5. DIGITAL SIGNAL PROCESSING 201
5.1. DIGITAL SIGNALS AND SYSTEMS 202
5.1.1. *Sinusoidal Signals* 203
5.1.2. *Other Digital Signals* 206
5.1.3. *Digital Systems* 206
5.2. CONTINUOUS-FREQUENCY TRANSFORMS 208
5.2.1. *The Fourier Transform* 208
5.2.2. *Z-Transform* 211
5.2.3. *Z-Transforms of Elementary Functions* 212
5.2.4. *Properties of the Z- and Fourier Transforms* 215
5.3. DISCRETE-FREQUENCY TRANSFORMS 216
5.3.1. *The Discrete Fourier Transform (DFT)* 218
5.3.2. *Fourier Transforms of Periodic Signals* 219
5.3.3. *The Fast Fourier Transform (FFT)* 222
5.3.4. *Circular Convolution* 227
5.3.5. *The Discrete Cosine Transform (DCT)* 228

Amazon/VB Assets
Exhibit 1011
Page 9

5.4. DIGITAL FILTERS AND WINDOWS 229
5.4.1. *The Ideal Low-Pass Filter* 229
5.4.2. *Window Functions* 230
5.4.3. *FIR Filters* 232
5.4.4. *IIR Filters* 238
5.5. DIGITAL PROCESSING OF ANALOG SIGNALS 242
5.5.1. *Fourier Transform of Analog Signals* 243
5.5.2. *The Sampling Theorem* 243
5.5.3. *Analog-to-Digital Conversion* 245
5.5.4. *Digital-to-Analog Conversion* 246
5.6. MULTIRATE SIGNAL PROCESSING 248
5.6.1. *Decimation* 248
5.6.2. *Interpolation* 249
5.6.3. *Resampling* 250
5.7. FILTERBANKS 251
5.7.1. *Two-Band Conjugate Quadrature Filters* 251
5.7.2. *Multiresolution Filterbanks* 254
5.7.3. *The DFT as a Filterbank* 255
5.7.4. *Modulated Lapped Transforms* 258
5.8. STOCHASTIC PROCESSES 260
5.8.1. *Statistics of Stochastic Processes* 261
5.8.2. *Stationary Processes* 264
5.8.3. *LTI Systems with Stochastic Inputs* 267
5.8.4. *Power Spectral Density* 268
5.8.5. *Noise* 269
5.9. HISTORICAL PERSPECTIVE AND FURTHER READING 270

6. SPEECH SIGNAL REPRESENTATIONS 275
6.1. SHORT-TIME FOURIER ANALYSIS 276
6.1.1. *Spectrograms* 281
6.1.2. *Pitch-Synchronous Analysis* 283
6.2. ACOUSTICAL MODEL OF SPEECH PRODUCTION 283
6.2.1. *Glottal Excitation* 284
6.2.2. *Lossless Tube Concatenation* 284
6.2.3. *Source-Filter Models of Speech Production* 288
6.3. LINEAR PREDICTIVE CODING 290
6.3.1. *The Orthogonality Principle* 291
6.3.2. *Solution of the LPC Equations* 292
6.3.3. *Spectral Analysis via LPC* 300
6.3.4. *The Prediction Error* 301
6.3.5. *Equivalent Representations* 303

Amazon/VB Assets
Exhibit 1011
Page 10

6.4. CEPSTRAL PROCESSING 306
6.4.1. The Real and Complex Cepstrum 307
6.4.2. Cepstrum of Pole-Zero Filters 308
6.4.3. Cepstrum of Periodic Signals 311
6.4.4. Cepstrum of Speech Signals 312
6.4.5. Source-Filter Separation via the Cepstrum 314
6.5. PERCEPTUALLY MOTIVATED REPRESENTATIONS 315
6.5.1. The Bilinear Transform 315
6.5.2. Mel-Frequency Cepstrum 316
6.5.3. Perceptual Linear Prediction (PLP) 318
6.6. FORMANT FREQUENCIES 319
6.6.1. Statistical Formant Tracking 320
6.7. THE ROLE OF PITCH 324
6.7.1. Autocorrelation Method 324
6.7.2. Normalized Cross-Correlation Method 327
6.7.3. Signal Conditioning 329
6.7.4. Pitch Tracking 330
6.8. HISTORICAL PERSPECTIVE AND FURTHER READING 332

7. SPEECH CODING 337

7.1. SPEECH CODERS ATTRIBUTES 338
7.2. SCALAR WAVEFORM CODERS 340
7.2.1. Linear Pulse Code Modulation (PCM) 340
7.2.2. μ-law and A-law PCM 342
7.2.3. Adaptive PCM 344
7.2.4. Differential Quantization 345
7.3. SCALAR FREQUENCY DOMAIN CODERS 348
7.3.1. Benefits of Masking 349
7.3.2. Transform Coders 350
7.3.3. Consumer Audio 351
7.3.4. Digital Audio Broadcasting (DAB) 352
7.4. CODE EXCITED LINEAR PREDICTION (CELP) 353
7.4.1. LPC Vocoder 353
7.4.2. Analysis by Synthesis 353
7.4.3. Pitch Prediction: Adaptive Codebook 356
7.4.4. Perceptual Weighting and Postfiltering 357
7.4.5. Parameter Quantization 358
7.4.6. CELP Standards 359
7.5. LOW-BIT RATE SPEECH CODERS 361
7.5.1. Mixed-Excitation LPC Vocoder 362
7.5.2. Harmonic Coding 363
7.5.3. Waveform Interpolation 367
7.6. HISTORICAL PERSPECTIVE AND FURTHER READING 371

Amazon/VB Assets
Exhibit 1011
Page 11

PART III: SPEECH RECOGNITION

8. HIDDEN MARKOV MODELS 377

8.1. THE MARKOV CHAIN 378
8.2. DEFINITION OF THE HIDDEN MARKOV MODEL 380
8.2.1. Dynamic Programming and DTW 383
8.2.2. How to Evaluate an HMM—The Forward Algorithm 385
8.2.3. How to Decode an HMM—The Viterbi Algorithm 387
8.2.4. How to Estimate HMM Parameters—Baum-Welch Algorithm 389
8.3. CONTINUOUS AND SEMICONTINUOUS HMMS 394
8.3.1. Continuous Mixture Density HMMs 394
8.3.2. Semicontinuous HMMs 396
8.4. PRACTICAL ISSUES IN USING HMMS 398
8.4.1. Initial Estimates 398
8.4.2. Model Topology 399
8.4.3. Training Criteria 401
8.4.4. Deleted Interpolation 401
8.4.5. Parameter Smoothing 403
8.4.6. Probability Representations 404
8.5. HMM LIMITATIONS 405
8.5.1. Duration Modeling 406
8.5.2. First-Order Assumption 408
8.5.3. Conditional Independence Assumption 409
8.6. HISTORICAL PERSPECTIVE AND FURTHER READING 409

9. ACOUSTIC MODELING 415

9.1. VARIABILITY IN THE SPEECH SIGNAL 416
9.1.1. Context Variability 417
9.1.2. Style Variability 418
9.1.3. Speaker Variability 418
9.1.4. Environment Variability 419
9.2. HOW TO MEASURE SPEECH RECOGNITION ERRORS 419
9.3. SIGNAL PROCESSING—EXTRACTING FEATURES 421
9.3.1. Signal Acquisition 422
9.3.2. End-Point Detection 422
9.3.3. MFCC and Its Dynamic Features 424
9.3.4. Feature Transformation 426
9.4. PHONETIC MODELING—SELECTING APPROPRIATE UNITS 428
9.4.1. Comparison of Different Units 429
9.4.2. Context Dependency 430
9.4.3. Clustered Acoustic-Phonetic Units 432
9.4.4. Lexical Baseforms 436

Amazon/VB Assets
Exhibit 1011
Page 12

9.5. ACOUSTIC MODELING—SCORING ACOUSTIC FEATURES 439
9.5.1. Choice of HMM Output Distributions 439
9.5.2. Isolated vs. Continuous Speech Training 441
9.6. ADAPTIVE TECHNIQUES—MINIMIZING MISMATCHES 444
9.6.1. Maximum a Posteriori (MAP) 445
9.6.2. Maximum Likelihood Linear Regression (MLLR) 447
9.6.3. MLLR and MAP Comparison 450
9.6.4. Clustered Models 452
9.7. CONFIDENCE MEASURES: MEASURING THE RELIABILITY 453
9.7.1. Filler Models 453
9.7.2. Transformation Models 454
9.7.3. Combination Models 456
9.8. OTHER TECHNIQUES 457
9.8.1. Neural Networks 457
9.8.2. Segment Models 459
9.9. CASE STUDY: WHISPER 464
9.10. HISTORICAL PERSPECTIVE AND FURTHER READING 465

10. ENVIRONMENTAL ROBUSTNESS 477

10.1. THE ACOUSTICAL ENVIRONMENT 478
10.1.1. Additive Noise 478
10.1.2. Reverberation 480
10.1.3. A Model of the Environment 482
10.2. ACOUSTICAL TRANSDUCERS 486
10.2.1. The Condenser Microphone 486
10.2.2. Directionality Patterns 489
10.2.3. Other Transduction Categories 496
10.3. ADAPTIVE ECHO CANCELLATION (AEC) 497
10.3.1. The LMS Algorithm 499
10.3.2. Convergence Properties of the LMS Algorithm 500
10.3.3. Normalized LMS Algorithm 501
10.3.4. Transform-Domain LMS Algorithm 502
10.3.5. The RLS Algorithm 503
10.4. MULTIMICROPHONE SPEECH ENHANCEMENT 504
10.4.1. Microphone Arrays 505
10.4.2. Blind Source Separation 510
10.5. ENVIRONMENT COMPENSATION PREPROCESSING 515
10.5.1. Spectral Subtraction 516
10.5.2. Frequency-Domain MMSE from Stereo Data 519
10.5.3. Wiener Filtering 520
10.5.4. Cepstral Mean Normalization (CMN) 522
10.5.5. Real-Time Cepstral Normalization 525
10.5.6. The Use of Gaussian Mixture Models 525

Amazon/VB Assets
Exhibit 1011
Page 13

10.6. ENVIRONMENTAL MODEL ADAPTATION 528
10.6.1. Retraining on Corrupted Speech 528
10.6.2. Model Adaptation 530
10.6.3. Parallel Model Combination 531
10.6.4. Vector Taylor Series 535
10.6.5. Retraining on Compensated Features 537
10.7. MODELING NONSTATIONARY NOISE 538
10.8. HISTORICAL PERSPECTIVE AND FURTHER READING 540

11. LANGUAGE MODELING 545
11.1. FORMAL LANGUAGE THEORY 546
11.1.1. Chomsky Hierarchy 547
11.1.2. Chart Parsing for Context-Free Grammars 549
11.2. STOCHASTIC LANGUAGE MODELS 554
11.2.1. Probabilistic Context-Free Grammars 554
11.2.2. N-gram Language Models 558
11.3. COMPLEXITY MEASURE OF LANGUAGE MODELS 560
11.4. N-GRAM SMOOTHING 562
11.4.1. Deleted Interpolation Smoothing 564
11.4.2. Backoff Smoothing 565
11.4.3. Class N-grams 570
11.4.4. Performance of N-gram Smoothing 573
11.5. ADAPTIVE LANGUAGE MODELS 574
11.5.1. Cache Language Models 574
11.5.2. Topic-Adaptive Models 575
11.5.3. Maximum Entropy Models 576
11.6. PRACTICAL ISSUES 578
11.6.1. Vocabulary Selection 578
11.6.2. N-gram Pruning 580
11.6.3. CFG vs. N-gram Models 581
11.7. HISTORICAL PERSPECTIVE AND FURTHER READING 584

12. BASIC SEARCH ALGORITHMS 591
12.1. BASIC SEARCH ALGORITHMS 592
12.1.1. General Graph Searching Procedures 593
12.1.2. Blind Graph Search Algorithms 597
12.1.3. Heuristic Graph Search 601
12.2. SEARCH ALGORITHMS FOR SPEECH RECOGNITION 608
12.2.1. Decoder Basics 609
12.2.2. Combining Acoustic and Language Models 610
12.2.3. Isolated Word Recognition 610
12.2.4. Continuous Speech Recognition 611

12.3. LANGUAGE MODEL STATES 613
12.3.1. Search Space with FSM and CFG 613
12.3.2. Search Space with the Unigram 616
12.3.3. Search Space with Bigrams 617
12.3.4. Search Space with Trigrams 619
12.3.5. How to Handle Silences Between Words 621
12.4. TIME-SYNCHRONOUS VITERBI BEAM SEARCH 622
12.4.1. The Use of Beam 624
12.4.2. Viterbi Beam Search 625
12.5. STACK DECODING (A* SEARCH) 626
12.5.1. Admissible Heuristics for Remaining Path 630
12.5.2. When to Extend New Words 631
12.5.3. Fast Match 634
12.5.4. Stack Pruning 638
12.5.5. Multistack Search 639
12.6. HISTORICAL PERSPECTIVE AND FURTHER READING 640

13. LARGE-VOCABULARY SEARCH ALGORITHMS 645
13.1. EFFICIENT MANIPULATION OF A TREE LEXICON 646
13.1.1. Lexical Tree 646
13.1.2. Multiple Copies of Pronunciation Trees 648
13.1.3. Factored Language Probabilities 650
13.1.4. Optimization of Lexical Trees 653
13.1.5. Exploiting Subtree Polymorphism 656
13.1.6. Context-Dependent Units and Inter-Word Triphones 658
13.2. OTHER EFFICIENT SEARCH TECHNIQUES 659
13.2.1. Using Entire HMM as a State in Search 659
13.2.2. Different Layers of Beams 660
13.2.3. Fast Match 661
13.3. N-BEST AND MULTIPASS SEARCH STRATEGIES 663
13.3.1. N-best Lists and Word Lattices 664
13.3.2. The Exact N-best Algorithm 666
13.3.3. Word-Dependent N-best and Word-Lattice Algorithm 667
13.3.4. The Forward-Backward Search Algorithm 670
13.3.5. One-Pass vs. Multipass Search 673
13.4. SEARCH-ALGORITHM EVALUATION 674
13.5. CASE STUDY—MICROSOFT WHISPER 676
13.5.1. The CFG Search Architecture 676
13.5.2. The N-gram Search Architecture 677
13.6. HISTORICAL PERSPECTIVE AND FURTHER READING 681

Amazon/VB Assets
Exhibit 1011
Page 15

PART IV: TEXT-TO-SPEECH SYSTEMS

14. TEXT AND PHONETIC ANALYSIS 689
14.1. MODULES AND DATA FLOW 690
14.1.1. Modules 692
14.1.2. Data Flows 694
14.1.3. Localization Issues 696
14.2. LEXICON 697
14.3. DOCUMENT STRUCTURE DETECTION 699
14.3.1. Chapter and Section Headers 700
14.3.2. Lists 701
14.3.3. Paragraphs 702
14.3.4. Sentences 702
14.3.5. Email 704
14.3.6. Web Pages 705
14.3.7. Dialog Turns and Speech Acts 705
14.4. TEXT NORMALIZATION 706
14.4.1. Abbreviations and Acronyms 709
14.4.2. Number Formats 712
14.4.3. Domain-Specific Tags 718
14.4.4. Miscellaneous Formats 719
14.5. LINGUISTIC ANALYSIS 720
14.6. HOMOGRAPH DISAMBIGUATION 724
14.7. MORPHOLOGICAL ANALYSIS 725
14.8. LETTER-TO-SOUND CONVERSION 728
14.9. EVALUATION 730
14.10. CASE STUDY: FESTIVAL 732
14.10.1. Lexicon 733
14.10.2. Text Analysis 733
14.10.3. Phonetic Analysis 735
14.11. HISTORICAL PERSPECTIVE AND FURTHER READING 735

15. PROSODY 739
15.1. THE ROLE OF UNDERSTANDING 740
15.2. PROSODY GENERATION SCHEMATIC 743
15.3. SPEAKING STYLE 744
15.3.1. Character 744
15.3.2. Emotion 744
15.4. SYMBOLIC PROSODY 744
15.4.1. Pauses 745
15.4.2. Prosodic Phrases 747
749

Amazon/VB Assets
Exhibit 1011
Page 16

15.4.3. Accent 751
15.4.4. Tone 753
15.4.5. Tune 757
15.4.6. Prosodic Transcription Systems 759
15.5. DURATION ASSIGNMENT 761
15.5.1. Rule-Based Methods 762
15.5.2. CART-Based Durations 763
15.6. PITCH GENERATION 763
15.6.1. Attributes of Pitch Contours 764
15.6.2. Baseline F0 Contour Generation 768
15.6.3. Parametric F0 Generation 774
15.6.4. Corpus-Based F0 Generation 778
15.7. PROSODY MARKUP LANGUAGES 783
15.8. PROSODY EVALUATION 784
15.9. HISTORICAL PERSPECTIVE AND FURTHER READING 785

16. SPEECH SYNTHESIS 793

16.1. ATTRIBUTES OF SPEECH SYNTHESIS 794
16.2. FORMANT SPEECH SYNTHESIS 796
16.2.1. Waveform Generation from Formant Values 797
16.2.2. Formant Generation by Rule 800
16.2.3. Data-Driven Formant Generation 803
16.2.4. Articulatory Synthesis 803
16.3. CONCATENATIVE SPEECH SYNTHESIS 804
16.3.1. Choice of Unit 805
16.3.2. Optimal Unit String: The Decoding Process 810
16.3.3. Unit Inventory Design 817
16.4. PROSODIC MODIFICATION OF SPEECH 818
16.4.1. Synchronous Overlap and Add (SOLA) 818
16.4.2. Pitch Synchronous Overlap and Add (PSOLA) 820
16.4.3. Spectral Behavior of PSOLA 822
16.4.4. Synthesis Epoch Calculation 823
16.4.5. Pitch-Scale Modification Epoch Calculation 825
16.4.6. Time-Scale Modification Epoch Calculation 826
16.4.7. Pitch-Scale Time-Scale Epoch Calculation 827
16.4.8. Waveform Mapping 827
16.4.9. Epoch Detection 828
16.4.10. Problems with PSOLA 829
16.5. SOURCE-FILTER MODELS FOR PROSODY MODIFICATION 831
16.5.1. Prosody Modification of the LPC Residual 832
16.5.2. Mixed Excitation Models 832
16.5.3. Voice Effects 834

16.6. EVALUATION OF TTS SYSTEMS 834
16.6.1. *Intelligibility Tests* 837
16.6.2. *Overall Quality Tests* 840
16.6.3. *Preference Tests* 842
16.6.4. *Functional Tests* 842
16.6.5. *Automated Tests* 843
16.7. HISTORICAL PERSPECTIVE AND FURTHER READING 844

PART V: SPOKEN LANGUAGE SYSTEMS

17. SPOKEN LANGUAGE UNDERSTANDING 853

17.1. WRITTEN VS. SPOKEN LANGUAGES 855
17.1.1. *Style* 856
17.1.2. *Disfluency* 857
17.1.3. *Communicative Prosody* 858
17.2. DIALOG STRUCTURE 859
17.2.1. *Units of Dialog* 860
17.2.2. *Dialog (Speech) Acts* 861
17.2.3. *Dialog Control* 866
17.3. SEMANTIC REPRESENTATION 867
17.3.1. *Semantic Frames* 867
17.3.2. *Conceptual Graphs* 872
17.4. SENTENCE INTERPRETATION 873
17.4.1. *Robust Parsing* 874
17.4.2. *Statistical Pattern Matching* 878
17.5. DISCOURSE ANALYSIS 881
17.5.1. *Resolution of Relative Expression* 882
17.5.2. *Automatic Inference and Inconsistency Detection* 885
17.6. DIALOG MANAGEMENT 886
17.6.1. *Dialog Grammars* 887
17.6.2. *Plan-Based Systems* 888
17.6.3. *Dialog Behavior* 892
17.7. RESPONSE GENERATION AND RENDITION 894
17.7.1. *Response Content Generation* 895
17.7.2. *Concept-to-Speech Rendition* 899
17.7.3. *Other Renditions* 901
17.8. EVALUATION 901
17.8.1. *Evaluation in the ATIS Task* 901
17.8.2. *PARADISE Framework* 903
17.9. CASE STUDY—DR. WHO 906
17.9.1. *Semantic Representation* 906
17.9.2. *Semantic Parser (Sentence Interpretation)* 908

Amazon/VB Assets
Exhibit 1011
Page 18

17.9.3. Discourse Analysis 909
17.9.4. Dialog Manager 910
17.10. HISTORICAL PERSPECTIVE AND FURTHER READING 913

18. APPLICATIONS AND USER INTERFACES 919
18.1. APPLICATION ARCHITECTURE 920
18.2. TYPICAL APPLICATIONS 921
18.2.1. Computer Command and Control 921
18.2.2. Telephony Applications 924
18.2.3. Dictation 926
18.2.4. Accessibility 929
18.2.5. Handheld Devices 930
18.2.6. Automobile Applications 930
18.2.7. Speaker Recognition 931
18.3. SPEECH INTERFACE DESIGN 931
18.3.1. General Principles 931
18.3.2. Handling Errors 937
18.3.3. Other Considerations 941
18.3.4. Dialog Flow 942
18.4. INTERNATIONALIZATION 943
18.5. CASE STUDY—MIPAD 945
18.5.1. Specifying the Application 946
18.5.2. Rapid Prototyping 948
18.5.3. Evaluation 949
18.5.4. Iterations 951
18.6. HISTORICAL PERSPECTIVE AND FURTHER READING 952

INDEX 957

Amazon/VB Assets
Exhibit 1011
Page 19

Foreword

Recognition and understanding of spontaneous unrehearsed speech remains an elusive goal. To understand speech, a human considers not only the specific information conveyed to the ear, but also the context in which the information is being discussed. For this reason, people can understand spoken language even when the speech signal is corrupted by noise. However, understanding the context of speech is, in turn, based on a broad knowledge of the world. And this has been the source of the difficulty and over forty years of research.

It is difficult to develop computer programs that are sufficiently sophisticated to understand continuous speech by a random speaker. Only when programmers simplify the problem—by isolating words, limiting the vocabulary or number of speakers, or constraining the way in which sentences may be formed—is speech recognition by computer possible.

Since the early 1970s, researchers at AT&T, BBN, CMU, IBM, Lincoln Labs, MIT, and SRI have made major contributions in Spoken Language Understanding Research. In 1971, the Defense Advanced Research Projects Agency (DARPA) initiated an ambitious five-year, $15 million, multisite effort to develop speech understanding systems. The goals were to develop systems that would accept continuous speech from many speakers, with minimal speaker adaptation, and operate on a 1000-word vocabulary, artificial syntax, and a

Amazon/VB Assets
Exhibit 1011
Page 21

constrained task domain. Two of the systems, Harpy and Hearsay-II, both developed at Carnegie Mellon University, achieved the original goals and in some instances surpassed them.

During the last three decades I have been at Carnegie Mellon, I have been very fortunate to be able to work with many brilliant students and researchers. Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon were arguably among the outstanding researchers in the speech group at CMU. Since then, they have moved to Microsoft and have put together a world-class team at Microsoft Research. Over the years, they have contributed standards for building spoken language understanding systems with Microsoft's SAPI/SDK family of products and pushed the technologies forward with the rest of the community. Today, they continue to play a premier leadership role in both the research community and in industry.

This new book, *Spoken Language Processing*, represents a welcome addition to the technical literature on this increasingly important emerging area of Information Technology. As we move from desktop PCs to personal digital assistants (PDAs), wearable computers, and Internet cell phones, speech becomes a central, if not the only, means of communication between the human and machine! Huang, Acero, and Hon have undertaken a commendable task of creating a comprehensive reference that covers theoretical, algorithmic, and systems aspects of the spoken language tasks of recognition, synthesis, and understanding.

The task of spoken language communication requires a system to recognize, interpret, execute, and respond to a spoken query. This task is complicated by the fact that the speech signal is corrupted by many sources: noise in the background, characteristics of the microphone, vocal tract characteristics of the speakers, and differences in pronunciation. In addition, the system has to cope with non-grammaticality of spoken communication and ambiguity of language. An effective system must strive to utilize all the available sources of knowledge—acoustics, phonetics and phonology, lexical, syntactic, and semantic structure of language, and task-specific context-dependent information.

Speech is based on a sequence of discrete sound segments that are linked in time. These segments, called phonemes, are assumed to have unique articulatory and acoustic characteristics. While the human vocal apparatus can produce an almost infinite number of articulatory gestures, the number of phonemes is limited. English as spoken in the United States, for example, contains 16 vowel and 24 consonant sounds. Each phoneme has distinguishable acoustic characteristics and, in combination with other phonemes, forms larger units such as syllables and words. Knowledge about the acoustic differences among these sound units is essential to distinguish one word from another, say, *bit* from *pit*.

When speech sounds are connected to form larger linguistic units, the acoustic characteristics of a given phoneme will change as a function of its immediate phonetic environment because of the interaction among various anatomical structures (such as the tongue, lips, and vocal chords) and their different degrees of sluggishness. The result is an overlap of phonemic information in the acoustic signal from one segment to the other. For example, the same underlying phoneme *t* can have drastically different acoustic characteristics in different words, say, in *tea*, *tree*, *city*, *beaten*, and *steep*. This effect, known as coarticulation, can occur within a given word or across a word boundary. Thus, the word *this* will have very different acoustic properties in phrases such as *this car* and *this ship*.

Amazon/VB Assets
Exhibit 1011
Page 22

This book is self-contained for those who wish to familiarize themselves with the current state of spoken language systems technology. However, a researcher or a professional in the field will benefit from a thorough grounding in a number of disciplines, including:

- *Signal processing*: Fourier Transforms, DFT, and FFT
- *Acoustics*: physics of sounds and speech, models of vocal tract
- *Pattern recognition*: clustering and pattern matching techniques
- *Artificial intelligence*: knowledge representation and search, natural language processing
- *Computer science*: hardware, parallel systems, algorithm optimization
- *Statistics*: probability theory, hidden Markov models, dynamic programming
- *Linguistics*: acoustic phonetics, lexical representation, syntax, and semantics

A newcomer to this field, easily overwhelmed by the vast number of different algorithms scattered across many conference proceedings, can find in this book a set of techniques that Huang, Acero, and Hon have found to work well in practice. This book is unique in that it includes both the theory and implementation details necessary to build spoken language systems. If you were able to assemble all the individual material that is covered in the book and put it on a shelf, it would be several times larger than this volume and yet you would be missing vital information. You would not have the material that is in this book that threads it all into one story, one context. If you need additional resources, the authors include extensive references to get that additional detail. *Spoken Language Processing* is very appealing both as a textbook and as a reference book for practicing engineers. Some readers familiar with a specific topic may decide to skip a few chapters; others may want to focus in other chapters. This is not a book that you will pick up and read once from cover to cover, but one you will keep near you for reference as long as you work in this field.

Raj Reddy
Dean, School of Computer Science
Carnegie Mellon University

Amazon/VB Assets
Exhibit 1011
Page 23

Preface

Our primary motivation in writing this book is to share our working experience to bridge the gap between the knowledge of industry gurus and newcomers to the spoken language processing community. Many powerful techniques hide in conference proceedings and academic papers for years before becoming widely recognized by the research community or the industry. We spent many years pursuing spoken language technology research at Carnegie Mellon University before we started spoken language R&D at Microsoft. We fully understand that it is by no means a small undertaking to transfer a state-of-the-art spoken language research system into a commercially viable product that can truly help people improve their productivity. Our experience in both industry and academia is reflected in the context of this book, which presents a contemporary and comprehensive description of both theoretic and practical issues in spoken language processing. This book is intended for people of diverse academic and practical backgrounds. Speech scientists, computer scientists, linguists, engineers, physicists, and psychologists all have a unique perspective on spoken language processing. This book will be useful to all of these special interest groups.

Spoken language processing is a diverse subject that relies on knowledge of many levels, including acoustics, phonology, phonetics, linguistics, semantics, pragmatics, and discourse. The diverse nature of spoken language processing requires knowledge in computer science, electrical engineering, mathematics, syntax, and psychology. There are a number of excellent books on the subfields of spoken language processing, including speech recognition, text-to-speech conversion, and spoken language understanding, but there is no single book that covers both theoretical and practical aspects of these subfields and spoken language interface design. We devote many chapters systematically introducing fundamental

Amazon/VB Assets
Exhibit 1011
Page 25

theories needed to understand how speech recognition, text-to-speech synthesis, and spoken language understanding work. Even more important is the fact that the book highlights what works well in practice, which is invaluable if you want to build a practical speech recognizer, a practical text-to-speech synthesizer, or a practical spoken language system. Using numerous real examples in developing Microsoft's spoken language systems, we concentrate on showing how the fundamental theories can be applied to solve real problems in spoken language processing.

We would like to thank many people who helped us during our spoken language processing R&D careers. We are particularly indebted to Professor Raj Reddy at the School of Computer Science, Carnegie Mellon University. Under his leadership, Carnegie Mellon University has become a center of research excellence on spoken language processing. Today's computer industry and academia benefit tremendously from his leadership and contributions.

Special thanks are due to Microsoft for its encouragement of spoken language R&D. The management team at Microsoft has been extremely generous to the speech technology group. We are particularly grateful to Bill Gates, Nathan Myhrvold, Rick Rashid, Dan Ling, and Jack Breese for the great environment they have created for us at Microsoft Research. We would also like to thank Bob Muglia and Kai-Fu Lee for their leadership role in Microsoft's speech product development.

Scott Meredith helped us write a number of chapters in this book and deserves to be a co-author. His insight and experience in text-to-speech synthesis enriched this book a great deal. We also owe gratitude to many colleagues we worked with in the speech technology group of Microsoft Research. In alphabetic order, Jim Adcock, Bruno Alabiso, Fil Alleva, Eric Bidstrup, Antonio Bigazzi, Ciprian Chelba, Li Deng, James Droppo, Doug Duchene, Joshua Goodman, Mei-Yuh Hwang, Larry Israel, Derek Jacoby, Li Jiang, Yun-Cheng Ju, David Larson, Kevin Larson, Jingsong Liu, Ricky Loynd, Milind Mahajan, Peter Mau, John Merrill, Yunus Mohammed, Salman Mughal, Mike Plumpe, Scott Quinn, Bill Rockenbeck, Mike Rozak, Kevin Schofield, Roxana Teodorescu, Gina Venolia, Kuansan Wang, Ye-Yi Wang, and Shenzhi Zhang.

In addition, we want to thank Les Atlas, Jeff Bilmes, Alan Black, David Caulton, Eric Chang, Phil Chou, Dinei Florencio, Allen Gersho, Francisco Gimenez-Galanes, Hynek Hermansky, Henrique Malvar, Julian Odell, Mari Ostendorf, Joseph Pentheroudakis, Tandy Trower, and Charles Wayne. They provided us with many wonderful comments to refine this book. Tim Moore, Russ Hall, and Jane Bonnell at Prentice Hall helped us finish this book in a finite amount of time.

Finally, writing this book was a marathon that could not have been finished without the support of our spouses, Yingzhi, Donna, and Phen, during the many evenings and weekends we spent on this project.

Xuedong Huang
Alex Acero
Hsiao-Wuen Hon
Redmond, WA

Amazon/VB Assets
Exhibit 1011
Page 26

CHAPTER 1

Introduction

From human prehistory to the new media of the future, speech communication has been and will be the dominant mode of human social bonding and information exchange. The spoken word is now extended, through technological mediation such as telephony, movies, radio, television, and the Internet. This trend reflects the primacy of spoken communication in human psychology.

In addition to human-human interaction, this human preference for spoken language communication finds a reflection in human-machine interaction as well. Most computers currently utilize a *graphical user interface* (GUI), based on graphically represented interface objects and functions such as windows, icons, menus, and pointers. Most computer operating systems and applications also depend on a user's keyboard strokes and mouse clicks, with a display monitor for feedback. Today's computers lack the fundamental human abilities to speak, listen, understand, and learn. Speech, supported by other natural modalities, will be one of the primary means of interfacing with computers. And, even before speech-based interaction reaches full maturity, applications in home, mobile, and office segments are incorporating spoken language technology to change the way we live and work.

Amazon/VB Assets
Exhibit 1011
Page 27

A spoken language system needs to have both speech recognition and speech synthesis capabilities. However, those two components by themselves are not sufficient to build a useful spoken language system. An understanding and dialog component is required to manage interactions with the user; and domain knowledge must be provided to guide the system's interpretation of speech and allow it to determine the appropriate action. For all these components, significant challenges exist, including robustness, flexibility, ease of integration, and engineering efficiency. The goal of building commercially viable spoken language systems has long attracted the attention of scientists and engineers all over the world. The purpose of this book is to share our working experience in developing advanced spoken language processing systems with both our colleagues and newcomers. We devote many chapters to systematically introducing fundamental theories and to highlighting what works well based on numerous lessons we learned in developing Microsoft's spoken language systems.

1.1. MOTIVATIONS

What motivates the integration of spoken language as the primary interface modality? We present a number of scenarios, roughly in order of expected degree of technical challenges and expected time to full deployment.

1.1.1. Spoken Language Interface

There are generally two categories of users who can benefit from adoption of speech as a control modality in parallel with others, such as the mouse, keyboard, touch-screen, and joystick. For novice users, functions that are conceptually simple should be directly accessible. For example, raising the voice output volume under software control on the desktop speakers, a conceptually simple operation, in some GUI systems of today requires opening one or more windows or menus, and manipulating sliders, check-boxes, or other graphical elements. This requires some knowledge of the system's interface conventions and structures. For the novice user, to be able to say *raise the volume* would be more direct and natural. For expert users, the GUI paradigm is sometimes perceived as an obstacle or nuisance and shortcuts are sought. Frequently these shortcuts allow the power user's hands to remain on the keyboard or mouse while mixing content creation with system commands. For example, an operator of a graphic design system for CAD/CAM might wish to specify a text formatting command while keeping the pointer device in position over a selected screen element.

Speech has the potential to accomplish these functions more powerfully than keyboard and mouse clicks. Speech becomes more powerful when supplemented by information streams encoding other dynamic aspects of user and system status, which can be resolved by the semantic component of a complete multimodal interface. We expect such multimodal interactions to proceed based on more complete user modeling, including speech, visual orientation, natural and device-based gestures, and facial expression, and these will be coordinated with detailed system profiles of typical user tasks and activity patterns.

Amazon/VB Assets
Exhibit 1011
Page 28

In some situations you must rely on speech as an input or output medium. For example, with wearable computers, it may be impossible to incorporate a large keyboard. When driving, safety is compromised by any visual distraction, and hands are required for controlling the vehicle. The ultimate speech-only device, the telephone, is far more widespread than the PC. Certain manual tasks may also require full visual attention to the focus of the work. Finally, spoken language interfaces offer obvious benefits for individuals challenged with a variety of physical disabilities, such as loss of sight or limitations in physical motion and motor skills. Chapter 18 contains a detailed discussion on spoken language applications.

1.1.2. Speech-to-Speech Translation

Speech-to-speech translation has been depicted for decades in science fiction stories. Imagine questioning a Chinese-speaking conversational partner by speaking English into an unobtrusive device, and hearing real-time replies you can understand. This scenario, like the spoken language interface, requires both speech recognition and speech synthesis technology. In addition, sophisticated multilingual spoken language understanding is needed. This highlights the need for tightly coupled advances in speech recognition, synthesis, and understanding systems, a point emphasized throughout this book.

1.1.3. Knowledge Partners

The ability of computers to process spoken language as proficient as humans will be a landmark to signal the arrival of truly intelligent machines. Alan Turing [29] introduced his famous *Turing test*. He suggested a game, in which a computer's use of language would form the criterion for intelligence. If the machine could win the game, it would be judged intelligent. In Turing's game, you play the role of an interrogator. By asking a series of questions via a teletype, you must determine the identity of the other two participants: a machine and a person. The task of the machine is to fool you into believing it is a person by responding as a person to your questions. The task of the other person is to convince you the other participant is the machine. The critical issue for Turing was that using language as humans do is sufficient as an operational test for intelligence.

The ultimate use of spoken language is to pass the Turing test in allowing future extremely intelligent systems to interact with human beings as knowledge partners in all aspects of life. This has been a staple of science fiction, but its day will come. Such systems require reasoning capabilities and extensive world knowledge embedded in sophisticated search, communication, and inference tools that are beyond the scope of this book. We expect that spoken language technologies described in this book will form the essential enabling mechanism to pass the Turing test.

Amazon/VB Assets
Exhibit 1011
Page 29

1.2. SPOKEN LANGUAGE SYSTEM ARCHITECTURE

Spoken language processing refers to technologies related to speech recognition, text-to-speech, and spoken language understanding. A spoken language system has at least one of the following three subsystems: a speech recognition system that converts speech into words, a text-to-speech system that conveys spoken information, and a spoken language understanding system that maps words into actions and that plans system-initiated actions.

There is considerable overlap in the fundamental technologies for these three subareas. Manually created rules have been developed for spoken language systems with limited success. But, in recent decades, data-driven statistical approaches have achieved encouraging results, which are usually based on modeling the speech signal using well-defined statistical algorithms that can automatically extract knowledge from the data. The data-driven approach can be viewed fundamentally as a pattern recognition problem. In fact, speech recognition, text-to-speech conversion, and spoken language understanding can all be regarded as pattern recognition problems. The patterns are either recognized during the runtime operation of the system or identified during system construction to form the basis of runtime generative models such as prosodic templates needed for text-to-speech synthesis. While we use and advocate the statistical approach, we by no means exclude the knowledge engineering approach from consideration. If we have a good set of rules in a given problem area, there is no need to use the statistical approach at all. The problem is that, at time of this writing, we do not have enough knowledge to produce a complete set of high-quality rules. As scientific and theoretical generalizations are made from data collected to construct data-driven systems, better rules may be constructed. Therefore, the rule-based and statistical approaches are best viewed as complementary.

1.2.1. Automatic Speech Recognition

A source-channel mathematical model described in Chapter 3 is often used to formulate speech recognition problems. As illustrated in Figure 1.1, the speaker's mind decides the source word sequence $\mathbf{W}$ that is delivered through his/her text generator. The source is passed through a noisy communication channel that consists of the speaker's vocal apparatus to produce the speech waveform and the speech signal processing component of the speech recognizer. Finally, the speech decoder aims to decode the acoustic signal $\mathbf{X}$ into a word sequence $\hat{\mathbf{W}}$, which is hopefully close to the original word sequence $\mathbf{W}$.

A typical practical speech recognition system consists of basic components shown in the dotted box of Figure 1.2. Applications interface with the decoder to get recognition results that may be used to adapt other components in the system. *Acoustic models* include the representation of knowledge about acoustics, phonetics, microphone and environment variability, gender and dialect differences among speakers, etc. *Language models* refer to a system's knowledge of what constitutes a possible word, what words are likely to co-occur, and in what sequence. The semantics and functions related to an operation a user may wish to perform may also be necessary for the language model. Many uncertainties exist in these areas, associated with speaker characteristics, speech style and rate, recognition of basic

Amazon/VB Assets
Exhibit 1011
Page 30