

# FUNDAMENTALS OF SPEECH RECOGNITION

Lawrence Rabiner  
Biing-Hwang Juang



PTR Prentice Hall  
Englewood Cliffs, New Jersey 07632

# CONTENTS

<b>LIST OF FIGURES</b>	<b>xiii</b>
<b>LIST OF TABLES</b>	<b>xxix</b>
<b>PREFACE</b>	<b>xxxix</b>
<b>1 FUNDAMENTALS OF SPEECH RECOGNITION</b>	<b>1</b>
1.1 Introduction	1
1.2 The Paradigm for Speech Recognition	3
1.3 Outline	3
1.4 A Brief History of Speech-Recognition Research	6
<b>2 THE SPEECH SIGNAL: PRODUCTION, PERCEPTION, AND ACOUSTIC-PHONETIC CHARACTERIZATION</b>	<b>11</b>
2.1 Introduction	11
2.1.1 The Process of Speech Production and Perception in Human Beings	11
2.2 The Speech-Production Process	14
2.3 Representing Speech in the Time and Frequency Domains	17
2.4 Speech Sounds and Features	20
	<b>vii</b>

2.4.1	The Vowels	21
2.4.2	Diphthongs	28
2.4.3	Semivowels	29
2.4.4	Nasal Consonants	30
2.4.5	Unvoiced Fricatives	31
2.4.6	Voiced Fricatives	32
2.4.7	Voiced and Unvoiced Stops	33
2.4.8	Review Exercises	37
2.5	<b>Approaches to Automatic Speech Recognition by Machine</b>	42
2.5.1	Acoustic-Phonetic Approach to Speech Recognition	45
2.5.2	Statistical Pattern-Recognition Approach to Speech Recognition	51
2.5.3	Artificial Intelligence (AI) Approaches to Speech Recognition	52
2.5.4	Neural Networks and Their Application to Speech Recognition	54
2.6	<b>Summary</b>	65

### **3 SIGNAL PROCESSING AND ANALYSIS METHODS FOR SPEECH RECOGNITION** **69**

3.1	<b>Introduction</b>	69
3.1.1	Spectral Analysis Models	70
3.2	<b>The Bank-of-Filters Front-End Processor</b>	73
3.2.1	Types of Filter Bank Used for Speech Recognition	77
3.2.2	Implementations of Filter Banks	80
3.2.3	Summary of Considerations for Speech-Recognition Filter Banks	92
3.2.4	Practical Examples of Speech-Recognition Filter Banks	93
3.2.5	Generalizations of Filter-Bank Analyzer	95
3.3	<b>Linear Predictive Coding Model for Speech Recognition</b>	97
3.3.1	The LPC Model	100
3.3.2	LPC Analysis Equations	101
3.3.3	The Autocorrelation Method	103
3.3.4	The Covariance Method	106
3.3.5	Review Exercise	107
3.3.6	Examples of LPC Analysis	108
3.3.7	LPC Processor for Speech Recognition	112
3.3.8	Review Exercises	117
3.3.9	Typical LPC Analysis Parameters	121
3.4	<b>Vector Quantization</b>	122
3.4.1	Elements of a Vector Quantization Implementation	123
3.4.2	The VQ Training Set	124
3.4.3	The Similarity or Distance Measure	125
3.4.4	Clustering the Training Vectors	125
3.4.5	Vector Classification Procedure	128
3.4.6	Comparison of Vector and Scalar Quantizers	129

3.4.7	Extensions of Vector Quantization	129
3.4.8	Summary of the VQ Method	131
3.5	<b>Auditory-Based Spectral Analysis Models</b>	132
3.5.1	The EIH Model	134
3.6	<b>Summary</b>	139

## **4 PATTERN-COMPARISON TECHNIQUES** **141**

4.1	<b>Introduction</b>	141
4.2	<b>Speech (Endpoint) Detection</b>	143
4.3	<b>Distortion Measures—Mathematical Considerations</b>	149
4.4	<b>Distortion Measures—Perceptual Considerations</b>	150
4.5	<b>Spectral-Distortion Measures</b>	154
4.5.1	Log Spectral Distance	158
4.5.2	Cepstral Distances	163
4.5.3	Weighted Cepstral Distances and Liftering	166
4.5.4	Likelihood Distortions	171
4.5.5	Variations of Likelihood Distortions	177
4.5.6	Spectral Distortion Using a Warped Frequency Scale	183
4.5.7	Alternative Spectral Representations and Distortion Measures	190
4.5.8	Summary of Distortion Measures—Computational Considerations	193
4.6	<b>Incorporation of Spectral Dynamic Features into the Distortion Measure</b>	194
4.7	<b>Time Alignment and Normalization</b>	200
4.7.1	Dynamic Programming—Basic Considerations	204
4.7.2	Time-Normalization Constraints	208
4.7.3	Dynamic Time-Warping Solution	221
4.7.4	Other Considerations in Dynamic Time Warping	229
4.7.5	Multiple Time-Alignment Paths	232
4.8	<b>Summary</b>	238

## **5 SPEECH RECOGNITION SYSTEM DESIGN AND IMPLEMENTATION ISSUES** **242**

5.1	<b>Introduction</b>	242
5.2	<b>Application of Source-Coding Techniques to Recognition</b>	244
5.2.1	Vector Quantization and Pattern Comparison Without Time Alignment	244
5.2.2	Centroid Computation for VQ Codebook Design	246
5.2.3	Vector Quantizers with Memory	254
5.2.4	Segmental Vector Quantization	256
5.2.5	Use of a Vector Quantizer as a Recognition Preprocessor	257
5.2.6	Vector Quantization for Efficient Pattern Matching	263
5.3	<b>Template Training Methods</b>	264
5.3.1	Casual Training	265

5.3.2	Robust Training	266
5.3.3	Clustering	267
<b>5.4</b>	<b>Performance Analysis and Recognition Enhancements</b>	<b>274</b>
5.4.1	Choice of Distortion Measures	274
5.4.2	Choice of Clustering Methods and kNN Decision Rule	277
5.4.3	Incorporation of Energy Information	280
5.4.4	Effects of Signal Analysis Parameters	282
5.4.5	Performance of Isolated Word-Recognition Systems	284
<b>5.5</b>	<b>Template Adaptation to New Talkers</b>	<b>285</b>
5.5.1	Spectral Transformation	286
5.5.2	Hierarchical Spectral Clustering	288
<b>5.6</b>	<b>Discriminative Methods in Speech Recognition</b>	<b>291</b>
5.6.1	Determination of Word Equivalence Classes	294
5.6.2	Discriminative Weighting Functions	297
5.6.3	Discriminative Training for Minimum Recognition Error	302
<b>5.7</b>	<b>Speech Recognition in Adverse Environments</b>	<b>305</b>
5.7.1	Adverse Conditions in Speech Recognition	306
5.7.2	Dealing with Adverse Conditions	309
<b>5.8</b>	<b>Summary</b>	<b>317</b>

## **6 THEORY AND IMPLEMENTATION OF HIDDEN MARKOV MODELS 321**

<b>6.1</b>	<b>Introduction</b>	<b>321</b>
<b>6.2</b>	<b>Discrete-Time Markov Processes</b>	<b>322</b>
<b>6.3</b>	<b>Extensions to Hidden Markov Models</b>	<b>325</b>
6.3.1	Coin-Toss Models	326
6.3.2	The Urn-and-Ball Model	328
6.3.3	Elements of an HMM	329
6.3.4	HMM Generator of Observations	330
<b>6.4</b>	<b>The Three Basic Problems for HMMs</b>	<b>333</b>
6.4.1	Solution to Problem 1—Probability Evaluation	334
6.4.2	Solution to Problem 2—"Optimal" State Sequence	337
6.4.3	Solution to Problem 3—Parameter Estimation	342
6.4.4	Notes on the Reestimation Procedure	347
<b>6.5</b>	<b>Types of HMMs</b>	<b>348</b>
<b>6.6</b>	<b>Continuous Observation Densities in HMMs</b>	<b>350</b>
<b>6.7</b>	<b>Autoregressive HMMs</b>	<b>352</b>
<b>6.8</b>	<b>Variants on HMM Structures—Null Transitions and Tied States</b>	<b>356</b>
<b>6.9</b>	<b>Inclusion of Explicit State Duration Density in HMMs</b>	<b>358</b>
<b>6.10</b>	<b>Optimization Criterion—ML, MMI, and MDI</b>	<b>362</b>
<b>6.11</b>	<b>Comparisons of HMMs</b>	<b>364</b>
<b>6.12</b>	<b>Implementation Issues for HMMs</b>	<b>365</b>
6.12.1	Scaling	365
6.12.2	Multiple Observation Sequences	369
6.12.3	Initial Estimates of HMM Parameters	370

6.12.4	Effects of Insufficient Training Data	370
6.12.5	Choice of Model	371
<b>6.13</b>	<b>Improving the Effectiveness of Model Estimates</b>	<b>372</b>
6.13.1	Deleted Interpolation	372
6.13.2	Bayesian Adaptation	373
6.13.3	Corrective Training	376
<b>6.14</b>	<b>Model Clustering and Splitting</b>	<b>377</b>
<b>6.15</b>	<b>HMM System for Isolated Word Recognition</b>	<b>378</b>
6.15.1	Choice of Model Parameters	379
6.15.2	Segmental K-Means Segmentation into States	382
6.15.3	Incorporation of State Duration into the HMM	384
6.15.4	HMM Isolated-Digit Performance	385
<b>6.16</b>	<b>Summary</b>	<b>386</b>

## **7 SPEECH RECOGNITION BASED ON CONNECTED WORD MODELS 390**

7.1	Introduction	390
7.2	General Notation for the Connected Word-Recognition Problem	393
7.3	The Two-Level Dynamic Programming (Two-Level DP) Algorithm	395
7.3.1	Computation of the Two-Level DP Algorithm	399
7.4	The Level Building (LB) Algorithm	400
7.4.1	Mathematics of the Level Building Algorithm	401
7.4.2	Multiple Level Considerations	405
7.4.3	Computation of the Level Building Algorithm	407
7.4.4	Implementation Aspects of Level Building	410
7.4.5	Integration of a Grammar Network	414
7.4.6	Examples of LB Computation of Digit Strings	416
7.5	The One-Pass (One-State) Algorithm	416
7.6	Multiple Candidate Strings	420
7.7	Summary of Connected Word Recognition Algorithms	423
7.8	Grammar Networks for Connected Digit Recognition	425
7.9	Segmental K-Means Training Procedure	427
7.10	Connected Digit Recognition Implementation	428
7.10.1	HMM-Based System for Connected Digit Recognition	429
7.10.2	Performance Evaluation on Connected Digit Strings	430
7.11	Summary	432

## **8 LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION 434**

8.1	Introduction	434
8.2	Subword Speech Units	435
8.3	Subword Unit Models Based on HMMs	439
8.4	Training of Subword Units	441

8.5	Language Models for Large Vocabulary Speech Recognition	447
8.6	Statistical Language Modeling	448
8.7	Perplexity of the Language Model	449
8.8	Overall Recognition System Based on Subword Units	450
8.8.1	Control of Word Insertion/Word Deletion Rate	454
8.8.2	Task Semantics	454
8.8.3	System Performance on the Resource Management Task	454
8.9	Context-Dependent Subword Units	458
8.9.1	Creation of Context-Dependent Diphones and Triphones	460
8.9.2	Using Interword Training to Create CD Units	461
8.9.3	Smoothing and Interpolation of CD PLU Models	462
8.9.4	Smoothing and Interpolation of Continuous Densities	464
8.9.5	Implementation Issues Using CD Units	464
8.9.6	Recognition Results Using CD Units	467
8.9.7	Position Dependent Units	469
8.9.8	Unit Splitting and Clustering	470
8.9.9	Other Factors for Creating Additional Subword Units	475
8.9.10	Acoustic Segment Units	476
8.10	Creation of Vocabulary-Independent Units	477
8.11	Semantic Postprocessor for Recognition	478
8.12	Summary	478

## **9 TASK ORIENTED APPLICATIONS OF AUTOMATIC SPEECH RECOGNITION** **482**

9.1	Introduction	482
9.2	Speech-Recognizer Performance Scores	484
9.3	Characteristics of Speech-Recognition Applications	485
9.3.1	Methods of Handling Recognition Errors	486
9.4	Broad Classes of Speech-Recognition Applications	487
9.5	Command-and-Control Applications	488
9.5.1	Voice Repertory Dialer	489
9.5.2	Automated Call-Type Recognition	490
9.5.3	Call Distribution by Voice Commands	491
9.5.4	Directory Listing Retrieval	491
9.5.5	Credit Card Sales Validation	492
9.6	Projections for Speech Recognition	493

## **INDEX** **497**