© 2008 AGI-Information Management Consultants
May be used for personal purporses only or by
libraries associated to dandelon.com network.

Statistical Methods for Speech Recognition

Frederick Jelinek

The MIT Press
Cambridge, Massachusetts
London, England

Contents

Preface xix

Chapter 1

The Speech Recognition Problem 1

1.1 Introduction 1

1.2 A Mathematical Formulation 4

1.3 Components of a Speech Recognizer 5

1.3.1 Acoustic Processing 5
1.3.2 Acoustic Modeling 7
1.3.3 Language Modeling 8
1.3.4 Hypothesis Search 8
7.5.5 *The Source-Channel Model of Speech Recognition* 9

1.4 About This Book 9

1.5 Vector Quantization 10

1.6 Additional Reading 12

References 12

Chapter 2

Hidden Markov Models IS

2.1 About Markov Chains 15

2.2 The Hidden Markov Model Concept 17

2.3 The Trellis 19

2.4 Search for the Likeliest State Transition Sequence 21

2.5 Presence of Null Transitions 23

2.6 Dealing with an HMM That Has Null Transitions That Do Not Form a Loop 25

2.7 Estimation of Statistical Parameters of HMMs 27

2.8 Practical Need for Normalization 33

2.9 Alternative Definitions of HMMs 35

2.9.1 HMMs Outputting Real Numbers 35

2.9.2 HMM Outputs Attached to States 35

2.10 Additional Reading 36

References 37

Chapter 3

The Acoustic Model 39

3.1 Introduction 39

3.2 Phonetic Acoustic Models 40

3.3 More on Acoustic Model Training 43

3.4 The Effect of Context 44

3.5 Viterbi Alignment 45

3.6 Singleton Fenonic Base Forms 45

3.7 A Needed Generalization 47

3.8 Generation of Synthetic Base Forms 48

3.9 A Further Refinement 51

3.10 Singleton Base Forms for Words Outside the Vocabulary 52

3.11 Additional Reading 52

References 54

Chapter 4

Basic Language Modeling 57

4.1 Introduction 57

4.2 Equivalence Classification of History 59

4.3 The Trigram Language Model 60

4.4 Optimal Linear Smoothing 62

4.5 An Example of a Trigram Language Model 66

4.6 Practical Aspects of Deleted Interpolation 66

4.7 Backing-Off 69

4.8 HMM Tagging 70

4.9 Use of Tag Equivalence Classification in a Language Model 72

4.10 Vocabulary Selection and Personalization from Text Databases 73

4.11 Additional Reading 75

References 76

Chapter 5

The **Viterbi** Search 79

5.1 Introduction 79

5.2 Finding the Most Likely Word Sequence 79

5.3 The Beam Search 81

5.4 Successive Language Model Refinement Search 84

5.5 Search versus Language Model State Spaces 86

5.6 Af-Best Search 86

5.7 A Maximum Probability Lattice 89

5.8 Additional Reading 90

References 90

Chapter 6

Hypothesis Search on a Tree and the Fast Match 93

6.1 Introduction 93

6.2 Tree Search versus Trellis (Viterbi) Search 95

6.3 A* Search 95

6.4 Stack Algorithm for Speech Recognition 97

6.5 Modifications of the Tree Search 99

6.6 Multiple-Stack Search 99

6.6.1 First Algorithm 100
6.6.2 A Multistack Algorithm 101
6.6.3 Actual Multistack Algorithm 102

6.7 Fast Match 103

6.8 The Cost of Search Shortcuts 109

6.9 Additional Reading 110

References 110

Chapter 7
Elements of Information Theory 113

7.1 Introduction 113

7.2 Functional Form of the Basic Information Measure 114

7.3 Some Mathematical Properties of Entropy 119

7.4 An Alternative Point of View and Notation 123

7.5 A Source-Coding Theorem 126

7.6 A Brief Digression 132

7.7 Mutual Information 132

7.8 Additional Reading 135

References 135

Chapter 8

The Complexity of Tasks—The Quality of Language **Models** 137

8.1 The Problem with Estimation of Recognition Task Complexity 137

8.2 The Shannon Game 139

8.3 Perplexity 141

8.4 The Conditional Entropy of the System 142

8.5 Additional Reading 144

References 145

Chapter 9

The Expectation-Maximization **Algorithm and Its Consequences** 147

9.1 Introduction 147

9.2 The EM Theorem 147

9.3 The Baum-Welch Algorithm 149

9.4 Real Vector Outputs of the Acoustic Processor 152

9.4.1 Development for Two Dimensions 152

9.4.2 The Generalization to k Dimensions 158

9.5 Constant and Tied Parameters 158

9.5.1 Keeping Some Parameters Constant 158

9.5.2 Tying of Parameter Sets 159

9.6 Tied Mixtures 161

9.7 Additional Reading 163

References 163

Chapter 10

Decision Trees and Tree Language Models 165

10.1 Introduction 165

10.2 Application of Decision Trees to Language Modeling 166

10.3 Decision Tree Example 166

10.4 What Questions? 168

10.5 The Entropy Goodness Criterion for the Selection of Questions, and a Stopping Rule 170

10.6 A Restricted Set of Questions 172

10.7 Selection of Questions by Chou's Method 173

10.8 Selection of the Initial Split of a Set *Sf* into Complementary Subsets 176

10.9 The Two-ing Theorem 177

10.10 Practical Considerations of Chou's Method 179

10.10.1 Problem ofOs in the q-Distribution: The Gini Index 179

10.10.2 Equivalence Classification Induced by Decision Trees 182

10.10.3 Computationally Feasible Specification of Decision Tree Equivalence Classes 183

10.11 Construction of Decision Trees Based on Word Encoding 184

10.12 A Hierarchical Classification of Vocabulary Words 186

10.13 More on Decision Trees Based on Word Encoding 188

10.13.1 Implementing Hierarchical Word Classification 188
10.13.2 Predicting Encoded Words One Bit at a Time 189
10.13.3 Treatment of Unseen Training Data 190
10.13.4 Problems and Advantages of Word Encoding 190

10.14 Final Remarks on the Decision Tree Method 191

10.14.1 Smoothing 191
10.14.2 Fragmentation of Data 192

10.15 Additional Reading 193

References 194

Chapter 11

Phonetics from Orthography: Spelling-to-Base Form Mappings 197

11.1 Overview of Base Form Generation from Spelling 197

11.2 Generating Alignment Data 199

11.3 Decision Tree Classification of Phonetic Environments 201

11.4 Finding the Base Forms 204

11.5 Additional reading 204

References 205

Chapter 12

Triphones and Allophones 207

12.1 Introduction 207

12.2 Triphones 208

12.3 The General Method 210

12.4 Collecting Realizations of Particular Phones 210

12.5 A Direct Method 211

12.6 The Consequences 214

12.7 Back to Triphones 215

12.8 Additional Reading 217

References 218

Chapter 13

Maximum Entropy Probability Estimation and Language Models 219

13.1 Outline of the Maximum Entropy Approach 219

13.2 The Main Idea 220

13.3 The General Solution 221

13.4 The Practical Problem 222

13.5 An Example 224

13.6 A Trigram Language Model 227

13.7 Limiting Computation 228

13.8 Iterative Scaling 231

13.9 The Problem of Finding Appropriate Constraints 233

13.10 Weighting of Diverse Evidence: Voting 234

13.11 Limiting Data Fragmentation: Multiple Decision Trees 236

13.11.1 Combining Different Knowledge Sources 236

13.11.2 Spontaneous Multiple Tree Development 238

13.12 Remaining Unsolved Problems 240

13.13 Additional Reading 241

References 242

Chapter 14

Three Applications of Maximum Entropy Estimation to Language Modeling 245

14.1 About the Applications 245

14.2 Simple Language Model Adaptation to a New Domain ' 246

14.3 A More Complex Adaptation 248

14.4 A Dynamic Language Model: Triggers 251

14.5 The Cache Language Model 253

14.6 Additional Reading 255

References 255

Chapter 15

Estimation of Probabilities from Counts and the Back-Off Method 257

15.1 Inadequacy of Relative Frequency Estimates 257

15.2 Estimation of Probabilities from Counts Using Held-Out Data 258

15.2.1 The Basic Idea 258
15.2.2 The Estimation 259
15.2.3 Deciding the Value ofM 261

15.3 Universality of the Held-Out Estimate 262

15.4 The Good-Turing Estimate 263

15.5 Applicability of the Held-Out and Good-Turing Estimates 265

15.6 Enhancing Estimation Methods 268

15.6.1 Frequency Enhancement of Held-Out Estimation of Bigrams 268
15.6.2 Frequency Enhancement of Good-Turing Estimation of Bigrams 269
15.6.3 Other Enhancements 270

15.7 The Back-Off Language Model 271

15.8 Additional Reading 273

References 274

Name Index 275

Subject Index 279