

# Speech Technology

## A Theoretical and Practical Introduction

**Michael Hammond** University of Arizona



**CAMBRIDGE**  
UNIVERSITY PRESS

# Contents

|                                         |         |
|-----------------------------------------|---------|
| List of Figures                         | page xi |
| List of Tables                          | xv      |
| Preface                                 | xvii    |
| <b>1 Overview</b>                       | 1       |
| 1.1 What Is Speech Technology?          | 1       |
| 1.2 Automatic Speech Recognition        | 1       |
| 1.3 Speech Synthesis                    | 2       |
| 1.4 Success                             | 2       |
| 1.5 Scale                               | 3       |
| 1.6 Linguistic Theory                   | 4       |
| 1.7 Structure                           | 5       |
| 1.7.1 Sounds                            | 5       |
| 1.7.2 Morphemes                         | 6       |
| 1.7.3 Sentences                         | 9       |
| 1.8 Spelling                            | 11      |
| <b>2 Speech</b>                         | 15      |
| 2.1 Articulation                        | 15      |
| 2.2 Phonology and Prosody               | 21      |
| 2.3 Acoustics                           | 24      |
| 2.4 Decomposing a Complex Wave          | 31      |
| 2.5 Acoustics and Computers             | 33      |
| 2.6 The Acoustics of Different Sounds   | 38      |
| 2.7 Representations                     | 48      |
| 2.8 Summary                             | 61      |
| <b>3 Finite-State Language Modeling</b> | 63      |
| 3.1 Formal Language Theory              | 63      |
| 3.2 Chomsky Hierarchy                   | 64      |
| 3.3 Finite Languages                    | 65      |
| 3.4 Regular Languages                   | 65      |
| 3.5 Modeling with Regular Languages     | 81      |
| 3.6 Regular Relations and Transducers   | 82      |

|          |                                          |            |
|----------|------------------------------------------|------------|
| 3.7      | Modeling with Regular Relations          | 90         |
| 3.8      | PyFoma                                   | 93         |
| 3.9      | Higher-Level Languages                   | 95         |
| 3.10     | Summary                                  | 101        |
| <b>4</b> | <b>Statistical Language Models</b>       | <b>102</b> |
| 4.1      | Weighted Finite-State Models             | 102        |
| 4.2      | Weights                                  | 106        |
| 4.3      | PyFoma Again                             | 107        |
| 4.4      | Transducer Operations                    | 108        |
| 4.5      | Examples of WFSTs                        | 112        |
| 4.6      | Hidden Markov Models                     | 114        |
| 4.7      | <i>N</i> -gram Models                    | 126        |
| 4.8      | Summary                                  | 134        |
| <b>5</b> | <b>Non-neural Synthesis</b>              | <b>135</b> |
| 5.1      | A Baseline                               | 135        |
| 5.2      | The Vocoder                              | 137        |
| 5.3      | Rule-Based Synthesis                     | 138        |
| 5.4      | Espeak                                   | 140        |
| 5.5      | Concatenative Synthesis                  | 142        |
| 5.6      | HMM-Based Synthesis                      | 151        |
| 5.6.1    | Gaussian Mixture Models                  | 151        |
| 5.6.2    | Duration                                 | 157        |
| 5.6.3    | State Transitions                        | 159        |
| 5.6.4    | Decision Trees                           | 159        |
| 5.7      | Festival                                 | 162        |
| 5.8      | Summary                                  | 167        |
| <b>6</b> | <b>Non-neural Recognition</b>            | <b>168</b> |
| 6.1      | The Logical Problem of Recognition       | 168        |
| 6.2      | Audrey and Shoebox                       | 169        |
| 6.3      | Dynamic Representations                  | 170        |
| 6.4      | Dynamic Time Warping                     | 171        |
| 6.5      | Speech Commands with DTW                 | 175        |
| 6.6      | HMMs and Vector Quantization             | 176        |
| 6.7      | Speech Commands with Vector Quantization | 179        |
| 6.8      | GMM-HMMs                                 | 180        |
| 6.9      | Kaldi                                    | 181        |
| 6.9.1    | Cepstral Mean and Variance Normalization | 182        |
| 6.9.2    | HCLG Transducer                          | 183        |

|          |                                                     |            |
|----------|-----------------------------------------------------|------------|
| 6.9.3    | Scoring                                             | 183        |
| 6.9.4    | Hebrew <i>Yes</i> and <i>No</i>                     | 184        |
| 6.9.5    | Speech Commands Digits                              | 186        |
| 6.9.6    | Reduced mini-LibriSpeech                            | 188        |
| 6.10     | Summary                                             | 190        |
| <b>7</b> | <b>Neural Nets</b>                                  | <b>191</b> |
| 7.1      | Logic Gates                                         | 191        |
| 7.2      | Perceptrons                                         | 193        |
| 7.3      | Multi-layer Perceptrons                             | 199        |
| 7.4      | Backpropagation                                     | 201        |
| 7.5      | Pytorch                                             | 205        |
| 7.6      | Convolution                                         | 207        |
| 7.7      | Recurrent Neural Nets                               | 209        |
| 7.8      | Sequence-to-Sequence Mapping                        | 215        |
| 7.8.1    | Encoder-Decoder Architecture                        | 215        |
| 7.8.2    | Attention                                           | 219        |
| 7.8.3    | Transformer                                         | 221        |
| 7.9      | Summary                                             | 223        |
| <b>8</b> | <b>Neural Synthesis</b>                             | <b>224</b> |
| 8.1      | WaveNet                                             | 225        |
| 8.2      | Deep Voice                                          | 230        |
| 8.2.1    | Grapheme-to-Phoneme Model                           | 230        |
| 8.2.2    | Segmentation Model                                  | 230        |
| 8.2.3    | Connectionist Temporal Classification               | 230        |
| 8.2.4    | Phoneme Duration and Fundamental Frequency<br>Model | 239        |
| 8.2.5    | Audio Synthesis Model                               | 240        |
| 8.2.6    | Later Versions of Deep Voice                        | 240        |
| 8.3      | Tacotron                                            | 242        |
| 8.4      | VoiceLoop                                           | 245        |
| 8.5      | FastSpeech                                          | 248        |
| 8.6      | Glow-TTS                                            | 250        |
| 8.7      | HiFi-GAN                                            | 251        |
| 8.8      | VITS                                                | 254        |
| 8.9      | Coqui TTS                                           | 255        |
| 8.10     | Summary                                             | 259        |
| <b>9</b> | <b>Neural Recognition</b>                           | <b>260</b> |
| 9.1      | DNN-HMMs                                            | 260        |

|           |                                |            |
|-----------|--------------------------------|------------|
| 9.2       | LSTMs and CTC                  | 262        |
| 9.3       | RNN Transducers                | 263        |
| 9.4       | DeepSpeech                     | 267        |
| 9.5       | Listen, Attend, and Spell      | 268        |
| 9.6       | Wav2letter                     | 269        |
| 9.7       | Wav2vec                        | 270        |
| 9.8       | HuBERT                         | 273        |
| 9.9       | Conformer                      | 273        |
| 9.10      | QuartzNet                      | 275        |
| 9.11      | Whisper                        | 276        |
| 9.12      | Summary                        | 278        |
| <b>10</b> | <b>Other Technologies</b>      | <b>280</b> |
| 10.1      | Identification                 | 280        |
| 10.2      | Voice Activity Detection       | 285        |
| 10.3      | Diarization                    | 288        |
| 10.4      | Spoken Language Identification | 289        |
| 10.5      | Voice Cloning                  | 290        |
| 10.6      | Large Language Models          | 292        |
| 10.7      | Summary                        | 292        |
|           | Appendix: Docker               | 293        |
|           | Bibliography                   | 298        |
|           | Index                          | 310        |