Lecture Notes in Speech Production, Speech Coding, and Speech Recognition

Mark Hasegawa-Johnson
University of Illinois at Urbana-Champaign

February 17, 2000
Contents

1 Basics of Digital Signal Processing ........................................ 13
   1.1 LTI Systems .................................................................. 13
       1.1.1 What is an LTI System? ........................................ 13
       1.1.2 Impulse Response ............................................... 13
       1.1.3 Eigenfunctions ................................................... 13
   1.2 Transforms .................................................................. 14
       1.2.1 Fourier Series: Dirichlet's Conditions .................... 14
       1.2.2 Z Transform and DFT: Region of Convergence ....... 14
   1.3 Transform Properties .................................................. 15
   1.4 Sampling Theorem ..................................................... 15
   1.5 Downsampling .......................................................... 16
   1.6 Upsampling .............................................................. 16

2 Speech Production .............................................................. 19
   2.1 Phonetic Classification of Speech Sounds ......................... 19
       2.1.1 Consonants ...................................................... 19
       2.1.2 Vowels ........................................................ 20
   2.2 A Source-Filter Model of Speech Production ..................... 20
       2.2.1 Source Function ............................................... 21
       2.2.2 Transfer Function ............................................ 22
       2.2.3 Radiation Characteristic .................................... 25
       2.2.4 Complete Synthesis Models ................................. 26
   2.3 One-Dimensional Acoustic Waves .................................... 26
       2.3.1 One-Dimensional Constitutive Equations ................. 26
       2.3.2 One-Dimensional Acoustic Wave Equations ............ 26
       2.3.3 General Form of the Solution: Time Domain ........... 26
       2.3.4 General Form of the Solution: Frequency Domain .... 26
   2.4 Parallel Admittance and Series Impedance Methods ............ 26
       2.4.1 Series Impedance Method ................................... 26
       2.4.2 Parallel Admittance Method ............................... 27
       2.4.3 Equivalence of the Parallel Admittance and Series Impedance Methods ............................................... 27
       2.4.4 Summary ....................................................... 28
       2.4.5 The Four Basic Lossless Impedances ...................... 28
   2.5 Concatenated Tubes Method for Finding Resonances .......... 31
       2.5.1 The Four Basic One-Dimensional Resonators ............ 31
       2.5.2 A Two-Tube Example: /aa/ .................................. 36
       2.5.3 A Three-Tube Example: /iy/ ................................. 37
   2.6 Summary ................................................................. 37
   2.7 Reflection and Transmission ........................................... 38
       2.7.1 Continuity of Flow and Pressure at a Boundary ....... 38
       2.7.2 Termination ................................................. 39
   2.8 Transfer Functions ..................................................... 39
2.8.1 Uniform Tube .................................................. 39
2.8.2 General Case .................................................. 40
2.9 How to Read Spectrograms ........................................ 41
  2.9.1 Broad Class Transitions .................................... 41
  2.9.2 How Many Segments? ....................................... 44
  2.9.3 Voiced or Unvoiced? ....................................... 45
  2.9.4 Measure the Formant Frequencies ......................... 50
  2.9.5 Context Effects .......................................... 55
2.10 Exercises ....................................................... 56

3 Short-Time Signal Processing ....................................... 57
  3.1 Short-Time Analysis ........................................... 57
    3.1.1 From 1D signal to Multidimensional “Frames” .......... 57
    3.1.2 Time-Domain Analysis .................................. 57
  3.2 Short-Time Fourier Transform ................................ 58
    3.2.1 Fourier Transform Interpretation ...................... 58
    3.2.2 Filterbank Interpretation .............................. 59
    3.2.3 Sampling in time ....................................... 59
    3.2.4 Sampling in frequency ................................ 59
    3.2.5 Filterbank Reconstruction ............................. 60
    3.2.6 Implementing non-uniform filterbanks using the STFT .. 61
  3.3 Window Characteristics ...................................... 61
    3.3.1 Rectangular Window .................................... 61
    3.3.2 Hamming and Hanning Windows ......................... 62
    3.3.3 Window Length ......................................... 62
3.4 Exercises ....................................................... 63

4 Linear Predictive Coding .......................................... 65
  4.1 All-Pole Model of the Speech Transfer Function .......... 65
  4.2 Normal Equations ............................................ 65
    4.2.1 Autocorrelation Method ................................ 66
    4.2.2 Covariance Method .................................... 67
    4.2.3 Choosing the LPC Order ................................. 68
    4.2.4 Choosing the LPC Gain ................................ 69
  4.3 Frequency-Domain Interpretation of LPC .................... 69
  4.4 Lattice Filtering ............................................. 69
    4.4.1 How to Calculate Reflection Coefficients ............ 70
    4.4.2 Why Use Lattice Filter Instead of Direct-Form LPC? .. 70
    4.4.3 Equivalence of Lattice and Concatenated-Tube Models .. 70
  4.5 Stability of the LPC Filter ................................ 70
    4.5.1 Stability of the Unquantized Filter \( H(z) \) ........ 70
    4.5.2 Stability of the Quantized Filter \( \hat{H}(z) \) ....... 71
    4.5.3 Quantizing Direct-Form Coefficients Leads to Unstable Filters 71
    4.5.4 THE SOLUTION .......................................... 71
  4.6 Log Area Ratios .............................................. 71
    4.6.1 The Problem ........................................... 71
    4.6.2 The Solution: Companded Quantization ................ 71
    4.6.3 Interpretation: Log Area Ratios ....................... 73
  4.7 Line Spectral Frequencies .................................. 73
    4.7.1 The Impedance Interpretation of the LSFs ............ 73
    4.7.2 Derivation of the LSF Polynomials .................... 74
    4.7.3 The Augmented-Tube Interpretation of Line Spectral Frequencies 76
  4.8 LPC Distance Measures ...................................... 77
    4.8.1 Itakura-Saito Distortion ............................... 77
4.8.2 Likelihood-Ratio Distortion, Itakura Distortion .......... 78
4.9 Exercises .................................................. 80

5 Spectral and Cepstral Distance Measures  .................... 83
5.1 Homomorphic Analysis ........................................ 83
5.2 Definitions .................................................. 83
5.2.1 Complex Cepstrum ........................................ 83
5.2.2 Cepstrum .................................................. 84
5.2.3 Example ................................................... 84
5.3 Minimum and Maximum Phase Sequences ........................ 84
5.4 Recursive Formula for the Cepstral Coefficients ............. 85
5.5 LPC Cepstrum ................................................ 85
5.5.1 Complex Cepstrum ........................................ 85
5.5.2 LPC Power Cepstrum ...................................... 85
5.5.3 How is the LPC Cepstrum Usually Used? .................. 87
5.6 Review ....................................................... 87
5.6.1 Complex Cepstrum ........................................ 87
5.6.2 Cepstrum .................................................. 87
5.6.3 Signals with Rational Z Transforms ........................ 87
5.7 Computational Considerations ................................ 88
5.8 Source-Filter Analysis ....................................... 88
5.8.1 Cepstrum of a Periodic Signal ............................ 89
5.8.2 Cepstrum of the Transfer Function ....................... 89
5.8.3 "Lifting" to separate source and filter ................... 89
5.9 Pole-Zero Analysis .......................................... 89
5.9.1 Estimating the Poles ....................................... 90
5.9.2 Estimating the Zeros ....................................... 90
5.10 Log Spectral Distance ...................................... 90
5.10.1 Power Spectrum .......................................... 90
5.10.2 Log Spectral Distance .................................... 91
5.11 Cepstral Distances .......................................... 91
5.11.1 Complex Cepstrum and Power Cepstrum .................. 91
5.11.2 Cepstral $L_2$ Norm ...................................... 91
5.11.3 LPC Cepstrum .......................................... 92
5.11.4 Cepstral Representation of Spectral Energy, Slope, and Finer Detail ................................................. 92
5.12 Cepstral Lifting ............................................. 93
5.12.1 Window in Time = Conolve in Frequency ................. 93
5.12.2 Weighted/Lifted Cepstral Distances ....................... 93
5.12.3 Symmetric Equivalent Window ............................ 93
5.12.4 Example: Rectangular Window ............................ 94
5.13 Exercises .................................................. 95

6 Engineering Models of Audition ............................... 97
6.1 Models of Hearing ........................................... 97
6.2 Perceptual Frequency Scales .................................. 97
6.2.1 Critical-Band Scale ....................................... 98
6.2.2 Mel-Frequency Scale ...................................... 98
6.2.3 Place on the Basilar Membrane ........................... 98
6.2.4 Frequency Just-Noticeable-Difference ...................... 98
6.3 Loudness ................................................... 99
6.3.1 Loudness vs. Frequency ................................... 99
6.3.2 Intensity JND ............................................. 99
6.4 Masking ................................................... 100
6.4.1 Noise-Masker Ratio ....................................... 101
7 Speech Coding

7.1 What is Speech Coding? .................................................... 107
7.2 Engineering Tradeoffs ..................................................... 107
  7.2.1 Applications and Standards ....................................... 109
7.3 Subjective Quality Metrics ................................................ 109
  7.3.1 Measures of Speech Quality: Mean Opinion Scores ............ 109
7.4 Objective Measures of Speech Quality ................................. 110
  7.4.1 Signal to Noise Ratio .............................................. 110
  7.4.2 Segmental SNR ..................................................... 111
  7.4.3 Perceptually-Weighted SEGSNR ................................. 111
  7.4.4 Spectral Amplitude Distortion ................................... 111
  7.4.5 Noise-to-Masker Ratio ........................................... 112
7.5 Memoryless Quantization (“Pulse Code Modulation” PCM) ........ 112
  7.5.1 Uniform Quantization ............................................. 113
  7.5.2 Compressed PCM .................................................. 113
7.6 Quantization: Basic Principles .......................................... 115
  7.6.1 Minimum Distortion Rule ......................................... 115
  7.6.2 Mean-Squared Error, SEGNSR .................................... 115
7.7 Scalar Quantization ($L = 1$) ............................................ 116
  7.7.1 Linear PCM ...................................................... 117
  7.7.2 Minimum-MSE Scalar Quantizer ................................... 117
  7.7.3 Semilog Companded Quantization ................................ 118
7.8 Vector Quantization ($L > 1$) ............................................ 119
  7.8.1 Minimum-MSE Vector Quantizer ................................... 119
  7.8.2 Product Coding: VQ with the Complexity of Scalar Quantization .......................................................... 120
7.9 Adaptive Step Size Quantization (APCM) ................................ 121
7.10 Differential PCM (DPCM) ................................................. 122
  7.10.1 Fixed Differential PCM (DPCM) .................................. 122
  7.10.2 Adaptive Differential PCM (ADPCM) ............................ 122
7.11 Perceptual Error Weighting .............................................. 123
  7.11.1 Error Spectrum is Nearly White .................................. 123
  7.11.2 Using the Signal to Mask the Noise .............................. 123
7.12 DPCM with Noise Feedback ............................................. 124
  7.12.1 An Alternative Representation of DPCM ......................... 124
  7.12.2 DPCM with Noise Shaping ....................................... 125
7.13 LPC Vocoder ............................................................... 125
  7.13.1 A Simple Model of Speech Production ......................... 125
  7.13.2 Vocoder Parameter Calculations ................................ 127
7.14 Pitch Prediction Vocoder ................................................. 129
  7.14.1 Purpose .......................................................... 129
  7.14.2 Excitation ....................................................... 129
7.15 LPC-Based Analysis-by-Synthesis Coding ............................ 130
  7.15.1 What is Analysis-by-Synthesis? ................................ 130
  7.15.2 LPC-based Analysis-by-Synthesis ............................... 131
  7.15.3 Frame-Based Analysis ........................................... 131
  7.15.4 Self-Excited LPC ................................................ 134
  7.15.5 Multi-Vector LPC-AS ............................................ 135
  7.15.6 Perceptual Error Weighting ..................................... 137
7.16 Exercises ............................................................... 138
8 Speech Recognition

8.1 Introduction to Recognition .................................................. 145
  8.1.1 Applications of Speech Recognition .................................. 145
  8.1.2 Front-End Processor and Pattern-Matcher ......................... 146
  8.1.3 Types of Front-End Processing ......................................... 146
  8.1.4 Statistical Speech Recognition: A Type of Pattern Matcher .... 147

8.2 Classification of a Single Spectrum ....................................... 147
  8.2.1 Gaussian Probability Models ......................................... 148
  8.2.2 Contour Plots of a Gaussian Distribution ........................ 148
  8.2.3 Mixture Gaussian Models ............................................. 148

8.3 Classification of a Sequence of Spectra .................................. 149
  8.3.1 Symbol-Timed Markov Process ....................................... 149
  8.3.2 Clock-Timed Markov Process: Self-Loops .......................... 151
  8.3.3 Left-to-Right Models ................................................ 151
  8.3.4 Hidden Markov Models ............................................... 151
  8.3.5 Example: Hidden Coin Toss ......................................... 152
  8.3.6 Continuous-Distribution HMMs ...................................... 153
  8.3.7 Example: Automatic Language Identification ...................... 153

8.4 Recognition Using a Hidden Markov Model ............................. 154
  8.4.1 Maximum Likelihood Recognition: The Forward Algorithm ....... 155
  8.4.2 Approximate Recognition: The Viterbi Algorithm .............. 156

8.5 Training a Hidden Markov Model .......................................... 157
  8.5.1 Initializing the Observation Densities: Segmental K-Means .... 158
  8.5.2 Refining the Model: Baum-Welch Algorithm ....................... 158
  8.5.3 Multiple Observation Sequences .................................... 159

8.6 Explicit State Duration Models ............................................ 160
  8.6.1 Duration Probabilities and Transition Probabilities ........... 160
  8.6.2 Recognition Using Explicit Probability Densities ............. 160
  8.6.3 Approximate Duration Modeling using Viterbi and Forward Algorithms 161

8.7 Continuous Observation Probability Densities ........................ 161
  8.7.1 Multivariate Gaussian Densities .................................... 161
  8.7.2 Mixture Gaussian Models ............................................ 163
  8.7.3 Feedforward Neural Networks ...................................... 164
  8.7.4 Initializing the Observation Densities: Segmental K-Means .... 167
  8.7.5 Tied Mixtures, Continuous Density Codebook ................... 167

8.8 Spectral Dynamics .......................................................... 168
  8.8.1 Spectral and Cepstral Derivatives .................................. 168
  8.8.2 Cepstral Differences ................................................ 169
  8.8.3 Cepstral Derivative Estimates .................................... 169
  8.8.4 RASTA ............................................................... 169

8.9 Probability Scaling in the Forward-Backward Algorithm ........... 171
  8.9.1 What's the Problem? .................................................. 171
  8.9.2 The Scaled Forward Algorithm ...................................... 171
  8.9.3 Recognition Using the Scaled Forward Algorithm ............... 171
  8.9.4 The Scaled Backward Algorithm .................................... 172
  8.9.5 Re-Estimation Using Scaled Parameters .......................... 172

8.10 Phone Models, Lexicon, Phonological Rules .......................... 173
  8.10.1 Recognition ......................................................... 173
  8.10.2 Training ............................................................ 173

8.11 Context-Dependent Phone Models ........................................ 174

8.12 Deleted Interpolation ..................................................... 174

8.13 Connected Word Recognition ............................................. 175
  8.13.1 The One-Pass Algorithm ........................................... 175

8.14 Language Modeling ......................................................... 177
8.14.1 Maximum A Posteriori Recognition .............................................. 178
8.14.2 N-Gram Language Models .......................................................... 178
8.14.3 Perplexity ...................................................................................... 179
8.14.4 Class N-Gram .............................................................................. 179
8.14.5 Hierarchical Language Models ..................................................... 179
8.15 User Interface Design ....................................................................... 181
  8.15.1 Dialog Modeling ........................................................................... 181
  8.15.2 Guidelines for User Interface Design ........................................... 181
8.16 Exercises ......................................................................................... 183
8.17 Final Project ..................................................................................... 184
  8.17.1 Recognizer Specifications ............................................................ 184
  8.17.2 The Design Process ..................................................................... 185
List of Figures

2.1 Several sources are active at the release of a stop. A period of silence is followed by a fricative burst lasting 5-10 ms. If the stop is unvoiced and syllable-initial, fricative is followed by 30-80 ms of aspiration. When the vocal folds are brought close enough together to begin vibrating, voicing begins. .......................................................... 22
2.2 Standing wave patterns of a hard-walled acoustic resonator of uniform area, closed at $x = 0$. The amplitude of the pressure standing wave, $|P(x, s)|$, is shown using a solid line, normalized to unity magnitude. The amplitude of the flow standing wave, $|Z_0(x)|U(x, s)|$, normalized to unity magnitude, is shown using a dashed line. ................................................................. 23
2.3 Tube open at opposite end. .......................................................... 29
2.4 Tube closed at opposite end. ........................................................ 30
2.5 Quarter-wave resonator ............................................................... 31
2.6 Modes of a quarter-wave resonator ............................................. 32
2.7 Open-ended half-wave resonator ................................................. 34
2.8 Closed-ended half-wave resonator .............................................. 34
2.9 Helmholtz resonator ................................................................. 35
2.10 Two-tube model of the vowel /aa/. ............................................. 36
2.11 Three-tube model of the vowel /iy/. ............................................ 37
2.12 Spectrogram of the phrase, “When the sunlight strikes mindrops in the air...” A spectrogram is a plot of signal energy as a function of both time and frequency. ......................................................... 42
2.13 “His ship.” The /z/ has become an /sh/ through processes of coarticulation, so that there are now two identical /sh/ phonemes in a row. The fricative noise section in the middle of the phrase looks just like a single /sh/; the only acoustic cue suggesting the presence of two fricatives is the duration, which is a bit more than 1.5 times the expected duration of a strident fricative in this context. ............................................................ 44
2.14 “Why do I owe you a letter?” Without knowing the phrase, we can guess the number of phonemes by counting the number of formant transitions, i.e. the number of significant up or down movements of either F1 or F2 or both. .......................................................... 44
2.15 Strident fricative consonants, showing differences in place of articulation (e.g. “supper” versus “zuh”) and voicing (e.g. “fuss” versus “futter”). Non-strident fricatives are formed by directing a turbulent jet of air against the teeth or some other obstacle, creating loud frication noise. ................................................................. 46
2.16 Non-strident fricative consonants, showing differences in place of articulation (e.g. “fuss” versus “thug”) and voicing (e.g. “fuss” versus “vanilla.”) Non-strident fricatives are similar to strident fricatives, but the turbulent jet of air usually does not hit an obstacle such as the teeth, so frication noise is not as loud. .......................................................... 47
2.17 Stop consonants, showing differences in place of articulation (e.g. “put” versus “tug”) and voicing (e.g. “put” versus “bug.”) .......................................................... 48
2.18 Syllable-initial unvoiced stops in English are always aspirated. These two examples show two different /p/ releases which are not strictly syllable-initial, and are therefore not aspirated. The middle /p/ in “paper” is both syllable-final and syllable-initial, so it has about half the normal amount of aspiration. The /p/ in spin is not syllable-initial at all, so it is unaspirated. 49
2.19 “Bet.” In this example, the formant frequencies are relatively constant throughout the vowel. 50
2.20 Formant frequencies of the vowels of English, as measured from a large database of male speakers [17].

2.21 The words “about,” “buy,” and “boy,” showing the three diphthongs of English: /ÆW/, /ÆI/, and /ɔW/, respectively (ARPAEFT notation). Each diphthong starts with a relatively static set of formant targets, then moves toward a briefly marked second set of formant targets.

2.22 Nasal consonants. The nasal consonants in English have the same three places of articulation (lips, tongue blade, and tongue body) as the stop consonants.

2.23 Examples of phoneme deletion and feature assimilation caused by context effects.

4.1 LPC synthesis using a lattice filter structure.

4.2 Spectral sensitivity to changes in the reflection coefficients.

4.3 LAR companding.

4.4 Acoustic resonator and lattice model with a matched impedance termination at the glottis.

4.5 Acoustic resonator and lattice filter model with a zero-admittance termination at the glottis.

4.6 Acoustic resonator and lattice filter model with a zero-impedance termination at the glottis.

6.1 An example of an equal-loudness curve as a function of frequency.

7.1 The three major engineering applications of speech signal processing.

7.2 An approximate comparison of the speech qualities, bit rates, and computational complexities achieved by various speech and audio coding algorithms.

7.3 Mean opinion scores for various types of speech coders.

7.4 Pulse-code modulation (PCM) quantizes each sample of a signal by rounding it to the nearest of a set of fixed quantization levels.

7.5 Mu-law companding function.

7.6 The nth Voronoi region is defined as the set of points in N-space which are closer to the nth codebook vector than to any other codebook vector.

7.7 Scalar quantization involves quantizing each sample independently of the previous and following samples.

7.8 The mu-law companding function.

7.9 Schematic of a DPCM coder.

7.10 The minimum-energy quantization noise is usually white noise.

7.11 Shaped quantization noise may be less audible than white quantization noise, even at slightly higher SNR.

7.12 Bandwidths of the LPC poles are expanded by moving the poles away from the unit circle (poles at r are shown as circles, poles at e as triangles).

7.13 An alternative implementation of DPCM.

7.14 DPCM with noise filtering.

7.15 A model of speech production which might be used in a speech synthesis program.

7.16 A simplified model of speech production, whose parameters can be transmitted efficiently across a digital channel.

7.17 A pitch-predictive vocoder.

7.18 Transfer function of the pitch prediction filter for several values of the prediction coefficient.

7.19 LPC analysis by synthesis coder.

7.20 The frame/sub-frame structure of most LPC analysis by synthesis coders.

7.21 An LPC analysis by synthesis coder with two codebooks: an “adaptive” codebook, which represents the pitch periodicity, and a “stochastic” codebook, which represents the unpredictable innovations in each speech frame.

8.1 Most speech recognizers include a front-end processor, which converts the signal into some sort of spectral vectors, and a pattern-matching unit, which tries to match the spectral vectors to a set of pre-defined models.

8.2 Contour plots of Gaussian and mixture-Gaussian probability densities.

8.3 A model which generates a random sequence of ones and twos.

8.4 A model of a process which speaks the words “one” and “two” in random order.
8.5 A left-to-right Markov process ........................................ 151
8.6 A hidden Markov model generates spectral vectors based on some internal state; the internal state of the model can never be known with certainty ........................................ 152
8.7 Simple Markov models of the words “hai” (/ai/, if we ignore the /h/) and “ja” (/ia/, if we pretend that /j/ and /i/ are the same). Transition probabilities are designed so that the /i/ states last an average of 1.5 frames, and the /a/ states last an average of 5 frames ........................................ 154
8.8 Contour plots of Gaussian and mixture-Gaussian probability densities ........................................ 162
8.9 Flow-chart and classification space of a single-neuron neural network ........................................ 165
8.10 Flow-chart, and classification space, of a two-level neural network ........................................ 165
8.11 Classification space showing a non-convex region $S_i$ ........................................ 166
8.12 Flow-chart of a three-level neural network ........................................ 166
8.13 In the RASTA method, frame-to-frame variations in a spectral estimate are smoothed using a filter like the one shown here ........................................ 170
8.14 A network of triphone models representing the phrase “one cat.” Phones are written in the ARPABET transcription system ........................................ 174
8.15 Deleted interpolation linearly combines the trained model parameters of monophone, diphone, and triphone models ........................................ 175
8.16 A detailed model of word transition probabilities can be created by parsing words into phrases, and phrases into complete sentences ........................................ 180