Spoken Language Processing I:
Speech Recognition

Instructor: Tim Bunnell
Office: 77 E. Del. Ave. (Greenhouse)
Lab: duPont Hospital for Kids, Room A/R 220
Phone: 831-3130 (UD) / 651-6835 (LAB)
email: bunnell@asel.udel.edu

Comments:
1) There are 26 scheduled class meetings. I must miss 3 or 4 of them because of a conference in October in China. I will schedule an examination for one of the missed classes and arrange coverage for a second (form to be determined). I’d like to fit the remaining two class meetings in before and/or after the gap. We will discuss how this can be done with minimum bother for everyone.

2) Each class meeting will be broken into two components: a lecture component; and a discussion component. The length of the two components will grow or shrink in a complementary way to adjust to topic needs. Some topics will get a long lecture and short discussion, others some brief comments followed by extended discussion.

3) We will be developing two speech recognition systems in the course of the semester. One will recognize isolated words, the other continuous speech. This will require an ongoing programming effort from the “paying customers” and any assistance we can wrangle from the listeners.

4) There will be 3 tests, one following each topic area. The final exam will cover only the third topic area. These will account for 60% of your grade. The remaining 40% will be based on programming projects and participation in class discussions.
Major Topics

• Structure of Speech
  – Physical acoustics
  – Speech acoustics
  – Acoustic Phonetics

• Digital/Numerical Representation
  – Digital signals
  – Analysis techniques
  – Data reduction

• Recognition Techniques
  – Dynamic Time Warping
  – Hidden Markov Modeling

Comments:

1) We are going to blaze as fast as possible through the first topic. There will be some readings and some supplementary material on the web that you should be familiar with for class. My descriptions will be pretty cryptic, touching mainly on what I see as the non-obvious and/or most interesting points to bring out. You will be responsible for raising questions during the discussion period regarding any material that needs clarification for you. We will try to cover this topic in two weeks.

2) Beginning with the second topic, we will be including programming activities as well as readings & web pages to supplement lectures. Discussion periods will include practical discussion of coding issues as well as discussion of the major and sub topics. Coding will be in C (rather than C++) and written to be highly portable. We will discuss in class what platform to develop on.

3) In the second topic area we will discuss the algorithms for various signal analysis techniques, but you will be supplied with functions that implement those algorithms. Your job will be to learn to use the functions in an effective way and incorporate them into a system capable of supporting real-time speech analysis and recognition.

3) The programming for the third topic area will build on code developed within the second topic area.
Structure: Physical Acoustics

- Time-domain representation
  - Simple waveforms
  - Complex waveforms
  - Periodic and aperiodic waveforms
- Frequency domain representation
  - Line spectra
  - Continuous spectra
- Sound shaping
  - Sources
  - Filters
  - Resonators

Comments:
1) Readings will be drawn from web pages and handouts
2) Class will be lecture biased.

Classes 1&2
Structure: Speech Acoustics

• Source-Filter Theory
• Speech Source characteristics
  – Frication
  – Phonation
• Speech Filter characteristics
  – Formants
  – Vocal tract area functions

Comments:
1) Readings will be drawn from web pages and handouts
2) Class will be lecture biased.
   Classes 2&3
Structure: Acoustic Phonetics

• Classification and Features
  – Manner / Place / Voicing
  – Other

• Segmental structure
  – Vowels & F1/F2 vowel space
  – Consonants

• Coarticulation

• Suprasegmentals
  – Intonation
  – Timing
  – Amplitude

Comments:
1) Readings will be drawn from web pages and handouts
2) Class will be lecture biased.

Class 4

Test on Topic - Class 5
Representation: Digital Signals

- Analog to digital conversion
  - Aliasing
  - Clipping
- Waveform Display and editing
  - Introduction to WEDW for Windows
  - Introduction to Waves+ for Unix (?)

Comments:
1) Readings drawn from web pages and user manuals for Wedw & Waves+
2) Everyone will be assigned five sentences to practice segmentation and labeling of digitized speech using a waveform editor.
3) We will begin construction of our speech recognition system by implementing its data acquisition module.

Classes 6 & 7
Representation: Analysis Tech.

- The general “representation” problem
- Preliminaries (frames & windows)
- Fundamental analysis methods
  - Fast Fourier Transform
  - Cepstrum
  - LPC analysis
  - Principal Components Factor analysis
  - Analysis by Synthesis
- Current best practice
  - Perception-based transformations (Bark and Mel cepstra, PLP, and RASTA)

Comments:
1) For readings, we will go over several journal articles & book chapters describing these techniques. Further readings will be selected papers related to the relationship between the perceptual properties of speech and the output of acoustic analyses.
2) Handouts/web pages will provide documentation on using functions which implement some of these techniques. We will not actually try to implement all of them for our program.
2) Coding issues to be discussed are (1) data structures that allow multiple/alternative analysis techniques to be applied within the same program, and (2) pipelined data flow for efficient realtime analysis.
3) Analysis module will be “connected to” the data acquisition module previously developed.

Classes 8, 9, 10, 11
Representation: Data Reduction

- Vector Quantization
  - Training algorithms
  - Classifying
- Artificial Neural Networks
  - Structures
  - Training
  - Classifying
- Phonetic feature classifiers
- Additional topics (Gaussian mixtures)

Comments:

1) Readings will be articles related to each technique. These techniques differ from “pure” analysis techniques in that they impose some form of classification upon the purely parametric analysis vectors. Such classifiers are based on the properties of real data and thus require training.

2) We will implement vector quantization (training and classification). The classifier will be incorporated into our recognition program, connecting it to the output of the analysis stage.

3) We will also introduce the TIMIT database at this time because it is an excellent source of speech material for training and testing all sorts of classifiers and recognizers.

4) We will develop, implement, and test one method for perceptual evaluation of classifiers. This will be a “vocoder” (i.e., an analysis/resynthesis program) that lets us hear how well a given acoustic feature classifier captures the perceptually important aspects of the speech signal.

Classes 12, 13, 14, 15

Test on Topic - Class 16
Recognition: Template Matching

- Simplifying Assumptions & Limitations
  - Speaker dependence
  - Isolated words
- Basic algorithm for DTW
  - Distance matrix generation
  - Path search
- Issues & Problems
  - Segmentation
  - The distance metric

Comments:

1) Readings for this topic will cover: the general notion of template matching for speech recognition; path search methods; speech segmentation; and how distances should be computed.

2) Training and testing modules will be developed and connected to the modules for acquisition, analysis, and reduction. The class will test the recognizer.

Classes 17,18,19,20
Recognition: HMM

• Basic structure
• Discrete versus Continuous HMM
• HMM decoding (recognition)
  – Viterbi algorithm
• HMM training

Comments:
1) Readings will be (1) a fairly long but well-written book chapter, and (2) going over code implementing aspects of the training and decoding.
2) Programming for this topic will implement training and decoding modules for a discrete HMM that will recognize continuous speech drawn from a small vocabulary and limited grammar.
3) Time permitting, we will discuss (and possibly implement) additional search constraints that would allow larger vocabularies and more complex grammars by using bigram phonotactic and/or language models.

Classes 21, 22, 23, 24, 25, 26

In-class final per schedule
The system will consist of a pipeline of four processes. The first (Data Acquisition) acquires an unbroken data stream of speech samples from an audio device or file and breaks it into frames of data for subsequent processes. The frames are evenly spaced in time (typically one frame every 10 ms) and normally overlap one another (i.e. each frame contains more than 10 ms of data). The second process converts the time-series in a frame to an alternative representation that is more useful for representing speech acoustics. The third stage further reduces the dimensionality of the representation to a manageable number of features. The final stage performs pattern matching on a sequence of frames to identify the speech.

Each processes will run as a separate thread with an input queue so that frames can be passed from faster processes (e.g., Data Acquisition) to slower processes (e.g., Signal Processing) without blocking.