Speech and Language

- Speech technologies
  - Automatic speech recognition (ASR)
  - Text-to-speech synthesis (TTS)
  - Dialog systems

- Language processing technologies
  - Machine translation
  - Information extraction
  - Web search, question answering
  - Text classification, spam filtering, etc...

Digitizing Speech

- Speech input is an acoustic wave form

Spectral Analysis

- Frequency gives pitch; amplitude gives volume
  - sampling at ~8 kHz phone, ~16 kHz mic (kHz=1000 cycles/sec)

- Fourier transform of wave displayed as a spectrogram
  - darkness indicates energy at each frequency

Part of [ae] from “lab”

- Complex wave repeating nine times
  - Plus smaller wave that repeats 4x for every large cycle
  - Large wave: freq of 250 Hz (9 times in .036 seconds)
  - Small wave roughly 4 times this, or roughly 1000 Hz
Resonances of the vocal tract

- The human vocal tract as an open tube
  - Air in a tube of a given length will tend to vibrate at resonance frequency of tube.
  - Constraint: Pressure differential should be maximal at (closed) glottal end and minimal at (open) lip end.

Vowel [i] sung at successively higher pitches

Acoustic Feature Sequence

- Time slices are translated into acoustic feature vectors (~39 real numbers per slice)
  - These are the observations, now we need the hidden states X

State Space

- P(E|X) encodes which acoustic vectors are appropriate for each phoneme (each kind of sound)
- P(X|X') encodes how sounds can be strung together
  - We will have one state for each sound in each word
  - From some state x, can only:
    - Stay in the same state (e.g. speaking slowly)
    - Move to the next position in the word
    - At the end of the word, move to the start of the next word
  - We build a little state graph for each word and chain them together to form our state space X

HMMs for Speech
Decoding

- While there are some practical issues, finding the words given the acoustics is an HMM inference problem.

- We want to know which state sequence $x_{1:T}$ is most likely given the evidence $e_{1:T}$:

$$x^*_{1:T} = \arg \max_{x_{1:T}} P(x_{1:T}|e_{1:T})$$

- From the sequence $x$, we can simply read off the words.