CS 188: Artificial Intelligence
Fall 2006

Lecture 21: Speech / Viterbi
11/09/2006

Dan Klein – UC Berkeley

Announcements

- Optional midterm
  - On Tuesday 11/21 in class
  - Review session 11/19, 7-9pm, in 306 Soda

- Projects
  - 3.2 due 11/9
  - 3.3 due 11/15
  - 3.4 due 11/27

- Contest
  - Pacman contest details on web site this week
  - Entries due 12/3

Hidden Markov Models

- Hidden Markov models (HMMs)
  - Underlying Markov chain over states X
  - You observe outputs (effects) E at each time step
  - As a Bayes’ net:

  \[
  X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow \ldots
  \]

  \[
  E_1 \rightarrow E_2 \rightarrow E_3 \rightarrow E_4 \rightarrow \ldots
  \]

- Several questions you can answer for HMMs:
  - Last time: filtering to track belief about current X given evidence

Speech Recognition

- [demos]

Digitizing Speech

- Speech input is an acoustic wave form

Speech in an Hour

- Speech input is an acoustic wave form

Graphs from Simon Arnfield’s web tutorial on speech, Sheffield:
http://www.psyp.leeds.ac.uk/research/cognispeech/tutorial/
She just had a baby

- What can we learn from a wavefile?
  - Vowels are voiced, long, loud
  - Length in time = length in space in waveform picture
  - Voicing: regular peaks in amplitude
  - When stops closed: no peaks: silence.
  - Peaks = voicing: .46 to .58 (vowel [i]), from second .65 to .74 (vowel [ɪ]) and so on
  - Silence of stop closure (1.06 to 1.08 for first [b], or 1.26 to 1.28 for second [b])
  - Fricatives like [ʃ] intense irregular pattern; see .33 to .46

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

Adding 100 Hz + 1000 Hz Waves

- Spectrum represents these freq components
  - Computed by Fourier transform, algorithm which separates out each frequency component of wave.

  - x-axis shows frequency, y-axis shows magnitude (in decibels, a log measure of amplitude)
  - Peaks at 930 Hz, 1860 Hz, and 3020 Hz.

Vowel Formants

- Back to Spectra
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.

<table>
<thead>
<tr>
<th>Closed end</th>
<th>Open end</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length 17.5 cm.</td>
<td></td>
</tr>
</tbody>
</table>

Why these Peaks?

- Articulatory facts:
  - Vocal cord vibrations create harmonics
  - The mouth is a selective amplifier
  - Depending on shape of mouth, some harmonics are amplified more than others

How to read spectrograms

- bab: closure of lips lowers all formants: so rapid increase in all formants at beginning of "bab"
- dad: first formant increases, but F2 and F3 slight fall
- gag: F2 and F3 come together: this is a characteristic of velars. Formant transitions take longer in velars than in alveolars or labials

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 $X$

---

**ASR Lexicon: Markov Models**

- Markov Process with Bigrams

**HMMs for Speech**

- 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 $e_{1:T}$:

  $$ z_{1:T} = \arg \max_{x_{1:T}} P(x_{1:T}|e_{1:T}) $$

  $$ m_t[x_t] = \max_{e_{1:t-1}} P(x_{1:t-1}, x_t, e_{1:t}) $$

  $$ = \max_{x_{1:t-1}} P(x_{1:t-1}, e_{1:t-1})P(x_t|x_{t-1})P(e_t|x_t) $$

  $$ = P(e_t|x_t) \max_{x_{1:t-1}} P(x_{1:t-1})m_{t-1}[x_{t-1}] $$

**Viterbi Algorithm**

- Question: what is the most likely state sequence given the observations?
  - Slow answer: enumerate all possibilities
  - Better answer: cached incremental version

**Markov Process with Bigrams**

- Figure from Huang et al page 618
Next Class

- Final part of the course: machine learning
- We’ll start talking about how to learn model parameters (like probabilities) from data
- One of the most heavily used technologies in all of AI