Algorithms for NLP

Speech Signals

Taylor Berg-Kirkpatrick – CMU

Slides: Dan Klein – UC Berkeley
Log-linear Parameterization

- **Model form:**

\[
P(y|x; w) = \frac{\exp(w^\top f(x, y))}{\sum_{y'} \exp(w^\top f(x, y'))}
\]

- **Learn by following gradient of training LL:**

\[
\frac{\partial L(w)}{\partial w} = \sum_i f(x_i, y_i^*) - \sum_i \left( \mathbb{E}_{P(y|x_i; w)} [f(x_i, y)] \right)
\]
Mixed Interpolation

- But can’t we just interpolate:
  - $P(w|\text{most recent words})$
  - $P(w|\text{skip contexts})$
  - $P(w|\text{caching})$
  - ...

- Yes, and people do (well, did)
  - But additive combination tends to flatten distributions, not zero out candidates
Neural LMs
Neural LMs

\[ i\text{-th output} = P(w_t = i | context) \]

Image: (Bengio et al, 03)
Neural vs Maxent

- **Maxent LM**
  \[ P(y|x; w) \propto \exp(w^\top f(x, y)) \]

- **Simple Neural LM**
  \[ P(y|x; A, B, C) \propto \exp \left( A_y^\top \sigma(B^\top C x) \right) \]

\( \sigma \) nonlinear, e.g. tanh
Neural LM Example

\[ P(y|x) \propto e^{(A^\top h)} \]

\[ h = \sigma \left( B^\top C_x \right) \]

\[ C_{closing} = 1.2 \quad 7.4 \quad \ldots \]

\[ C_{the} = -3.3 \quad 1.1 \quad \ldots \]

\[ x_{-2} = closing \]

\[ x_{-1} = the \]
Neural LMs

\[ i\text{-th output} = P(w_t = i \mid \text{context}) \]

Image: (Bengio et al, 03)
Speech Signals
- Frequency gives pitch; amplitude gives volume

- Frequencies at each time slice processed into observation vectors
Articulation
Articulatory System

Nasal cavity
Oral cavity
Pharynx
Vocal folds (in the larynx)
Trachea
Lungs

Sagittal section of the vocal tract (Techmer 1880)
Text from Ohala, Sept 2001, from Sharon Rose slide
Space of Phonemes

<table>
<thead>
<tr>
<th>LARYRAL</th>
<th>Labiodental</th>
<th>CORONAL</th>
<th>DORSAL</th>
<th>RADICAL</th>
<th>LARYNGEAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilabial</td>
<td>m</td>
<td>n</td>
<td>η</td>
<td>η</td>
<td>η</td>
</tr>
<tr>
<td>Plosive</td>
<td>p b φ ϕ</td>
<td>t d</td>
<td>t d</td>
<td>c j</td>
<td>k g q g</td>
</tr>
<tr>
<td>Fricative</td>
<td>φ β f v θ ð</td>
<td>s z j z</td>
<td>s z j z</td>
<td>x γ x k</td>
<td>h h H H</td>
</tr>
<tr>
<td>Approximant</td>
<td>u j w</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trill</td>
<td>b r</td>
<td></td>
<td></td>
<td></td>
<td>R R</td>
</tr>
<tr>
<td>Tap, Flap</td>
<td>v r</td>
<td></td>
<td></td>
<td></td>
<td>R R</td>
</tr>
<tr>
<td>Lateral fricative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral approximant</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral flap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Standard international phonetic alphabet (IPA) chart of consonants
Place
Places of Articulation

- Labial
- Dental
- Alveolar
- Post-alveolar/palatal
- Velar
- Uvular
- Pharyngeal
- Laryngeal/glottal

Figure thanks to Jennifer Venditti
Labial place

Bilabial: p, b, m
Labiodental: f, v

Figure thanks to Jennifer Venditti
Coronal place

Dental: th/dh
Alveolar: t/d/s/z/l/n
Post: sh/zh/y

Figure thanks to Jennifer Venditti
Dorsal Place

Velar:

k/g/ng

Figure thanks to Jennifer Venditti
**Space of Phonemes**

<table>
<thead>
<tr>
<th>Naval</th>
<th>Labiodental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Palatoalveolar</th>
<th>Retroflex</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Pharyngeal</th>
<th>Epiglottal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasal</td>
<td>m</td>
<td>m</td>
<td>n</td>
<td>η</td>
<td>j</td>
<td>η</td>
<td>η</td>
<td>η</td>
<td>N</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Plosive</td>
<td>p b φ ϕ</td>
<td>t d</td>
<td>t d</td>
<td>c η</td>
<td>k g</td>
<td>q g</td>
<td>h</td>
<td>η</td>
<td>H</td>
<td>η</td>
<td>h h</td>
</tr>
<tr>
<td>Fricative</td>
<td>θ θ s z</td>
<td>θ s z</td>
<td>θ s z</td>
<td>θ s z</td>
<td>θ s z</td>
<td>θ s z</td>
<td>θ s z</td>
<td>θ s z</td>
<td>θ s z</td>
<td>θ s z</td>
<td>θ s z</td>
</tr>
<tr>
<td>Approximant</td>
<td>v</td>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
</tr>
<tr>
<td>Trill</td>
<td>b</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
</tr>
<tr>
<td>Tap, Flap</td>
<td>v</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
<td>r</td>
</tr>
<tr>
<td>Lateral fricative</td>
<td>l</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>d</td>
</tr>
<tr>
<td>Lateral approximant</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
</tr>
<tr>
<td>Lateral flap</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
</tr>
</tbody>
</table>

- **Standard international phonetic alphabet (IPA) chart of consonants**
Manner
Manner of Articulation

- In addition to varying by place, sounds vary by manner

- **Stop:** complete closure of articulators, no air escapes via mouth
  - Oral stop: palate is raised \((p, t, k, b, d, g)\)
  - Nasal stop: oral closure, but palate is lowered \((m, n, ng)\)

- **Fricatives:** substantial closure, turbulent: \((f, v, s, z)\)

- **Approximants:** slight closure, sonorant: \((l, r, w)\)

- **Vowels:** no closure, sonorant: \((i, e, a)\)
Space of Phonemes

<table>
<thead>
<tr>
<th>Position</th>
<th>Labial</th>
<th>Labio-dental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Palato-alveolar</th>
<th>Retroflex</th>
<th>Palatal</th>
<th>Velar</th>
<th>Uvular</th>
<th>Pharyngeal</th>
<th>Epiglottal</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasal</td>
<td>m</td>
<td>n</td>
<td>η</td>
<td>n</td>
<td>η</td>
<td>η</td>
<td>η</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plosive</td>
<td>p b</td>
<td>t d</td>
<td>t̪ ̬ d</td>
<td>c j</td>
<td>k g</td>
<td>q g</td>
<td>?</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fricative</td>
<td>f v</td>
<td>θ ð s z</td>
<td>s z</td>
<td>c j</td>
<td>x y</td>
<td>x k</td>
<td>h</td>
<td>H</td>
<td>ḥ h</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximant</td>
<td>v</td>
<td>j</td>
<td>j</td>
<td>w</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trill</td>
<td>b r</td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tap, Flap</td>
<td>v r</td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral fricative</td>
<td>j ̓ ̓</td>
<td>̺ ̺ ̺ ̺ ̺ ̺</td>
<td>̺ ̺ ̺ ̺</td>
<td>̺ ̺ ̺ ̺</td>
<td>̺ ̺ ̺ ̺</td>
<td>̺ ̺ ̺ ̺</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral approximant</td>
<td>l l ̥</td>
<td>̥ ̥ ̥ ̥ ̥ ̥</td>
<td>̥ ̥ ̥ ̥</td>
<td>̥ ̥ ̥ ̥</td>
<td>̥ ̥ ̥ ̥</td>
<td>̥ ̥ ̥ ̥</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lateral flap</td>
<td>l</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Standard international phonetic alphabet (IPA) chart of consonants
Vowels
Vowel Space

Front  Near front  Central  Near back  Back

Close  i <---- i <----- u <----- u
Near close
Close mid  e <---- ø <----- ø <---- ø
Mid
Open mid  e <---- ø <----- ø <---- ø
Near open
Open

Vowels at right & left of bullets are rounded & unrounded.
Acoustics
“She just had a baby”

- What can we learn from a wavefile?
  - No gaps between words (!)
  - 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 [iy], from second .65 to .74 (vowel [ax]) 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 [sh]: intense irregular pattern; see .33 to .46
<table>
<thead>
<tr>
<th>Time-Domain Information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Example from Ladefoged</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>word</th>
<th>s</th>
<th>ə</th>
<th>p</th>
<th>ʰ</th>
<th>ə</th>
<th>t</th>
<th>ʰ</th>
<th>ə</th>
<th>o</th>
</tr>
</thead>
<tbody>
<tr>
<td>pat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pad</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bad</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>spat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• Y axis: Amplitude = amount of air pressure at that point in time
  • Zero is normal air pressure, negative is rarefaction
• X axis: Time.
• Frequency = number of cycles per second.
• 20 cycles in .02 seconds = 1000 cycles/second = 1000 Hz
Complex Waves: 100Hz+1000Hz
Spectrum

Frequency components (100 and 1000 Hz) on x-axis
Part of [ae] waveform from “had”

- Note complex wave repeating nine times in figure
- Plus smaller waves which repeats 4 times for every large pattern
- Large wave has frequency of 250 Hz (9 times in 0.036 seconds)
- Small wave roughly 4 times this, or roughly 1000 Hz
- Two little tiny waves on top of peak of 1000 Hz waves
Spectrum of an Actual Soundwave
Source / Channel
Why these Peaks?

- **Articulation process:**
  - The vocal cord vibrations create harmonics
  - The mouth is an amplifier
  - Depending on shape of mouth, some harmonics are amplified more than others
Vowel [i] at increasing pitches

Figures from Ratree Wayland
Resonances of the Vocal Tract

- The human vocal tract as an open tube:

  
  ![Diagram of vocal tract]

  - Closed end
  - Open end
  
  Length 17.5 cm.

- 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.

Figure from W. Barry
Computing the 3 Formants of Schwa

- Let the length of the tube be $L$
  - $F_1 = \frac{c}{\lambda_1} = \frac{c}{4L} = \frac{35,000}{4*17.5} = 500\text{Hz}$
  - $F_2 = \frac{c}{\lambda_2} = \frac{c}{4/3L} = \frac{3c}{4L} = \frac{3*35,000}{4*17.5} = 1500\text{Hz}$
  - $F_3 = \frac{c}{\lambda_3} = \frac{c}{4/5L} = \frac{5c}{4L} = \frac{5*35,000}{4*17.5} = 2500\text{Hz}$

- So we expect a neutral vowel to have 3 resonances at 500, 1500, and 2500 Hz

- These vowel resonances are called formants
Seeing Formants: the Spectrogram
Vowel Space

Vowels at right & left of bullets are rounded & unrounded.
Spectrograms
**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

From Ladefoged “A Course in Phonetics”
“She came back and started again”

1. lots of high-freq energy
3. vowel; faint 1100 Hz formant is nasalized
5. ey
7. short b closure, voicing barely visible.
9. note F2 and F3 coming together for “k”

From Ladefoged “A Course in Phonetics”