Algorithms for NLP

Acoustic Models

Taylor Berg-Kirkpatrick – CMU

Slides: Dan Klein – UC Berkeley
Speech Signals
Speech in a Slide

- Frequency gives pitch; amplitude gives volume

- Frequencies at each time slice processed into observation vectors

---

![Speech in a Slide](image)

---

$x_{12} x_{13} x_{12} x_{14} x_{14}$
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></th>
<th>LABIAL</th>
<th>CORONAL</th>
<th>DORSAL</th>
<th>RADICAL</th>
<th>LARYNGEAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bilabial</td>
<td>Labiodental</td>
<td>Dental</td>
<td>Alveolar</td>
<td>Palato-alveolar</td>
</tr>
<tr>
<td>Nasal</td>
<td>m</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>Plosive</td>
<td>p</td>
<td>b</td>
<td>t</td>
<td>d</td>
<td>t</td>
</tr>
<tr>
<td>Fricative</td>
<td>f</td>
<td>v</td>
<td>θ</td>
<td>δ</td>
<td>s</td>
</tr>
<tr>
<td>Approximant</td>
<td>u</td>
<td>u</td>
<td>u</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>
</tr>
<tr>
<td>Tap, Flap</td>
<td></td>
<td>f</td>
<td>r</td>
<td>r</td>
<td>R</td>
</tr>
<tr>
<td>Lateral fricative</td>
<td>h</td>
<td>l</td>
<td>t</td>
<td>k</td>
<td>k</td>
</tr>
<tr>
<td>Lateral approximant</td>
<td>l</td>
<td>l</td>
<td>θ</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>Lateral flap</td>
<td></td>
<td></td>
<td>θ</td>
<td>θ</td>
<td>θ</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

velar

uvular

pharyngeal

Figure thanks to Jennifer Venditti
Space of Phonemes

<table>
<thead>
<tr>
<th>Nasal</th>
<th>Plosive</th>
<th>Fricative</th>
<th>Approximant</th>
<th>Trill</th>
<th>Tap, Flap</th>
<th>Lateral fricative</th>
<th>Lateral approximant</th>
<th>Lateral flap</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>p b</td>
<td>φ β f v</td>
<td>θ δ s z</td>
<td>v</td>
<td>v</td>
<td>ḍ b</td>
<td>l</td>
<td>l</td>
</tr>
<tr>
<td>n</td>
<td>t d</td>
<td>θ δ s z</td>
<td>s z</td>
<td>j</td>
<td>r</td>
<td>ḍ b ḍ b</td>
<td>l</td>
<td>l</td>
</tr>
<tr>
<td>η</td>
<td>t q</td>
<td>s z</td>
<td>c j x γ</td>
<td>u</td>
<td>r</td>
<td>ḍ b ḍ b</td>
<td>l</td>
<td>l</td>
</tr>
<tr>
<td>j</td>
<td>c j</td>
<td>x γ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>η</td>
<td>k g</td>
<td>x γ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>η</td>
<td>q g</td>
<td>x γ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>η</td>
<td>q g</td>
<td>x γ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>?</td>
<td>h ꝏ h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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>LABIAL</th>
<th>CORONAL</th>
<th>DORSAL</th>
<th>RADICAL</th>
<th>LARYNGEAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilabial</td>
<td>Labiodental</td>
<td>Dental</td>
<td>Alveolar</td>
<td>Palato-alveolar</td>
</tr>
<tr>
<td>Nasal</td>
<td>m</td>
<td>n</td>
<td>η</td>
<td>η</td>
</tr>
<tr>
<td>Plosive</td>
<td>p</td>
<td>b</td>
<td>t</td>
<td>d</td>
</tr>
<tr>
<td>Fricative</td>
<td>φ β</td>
<td>f v</td>
<td>θ δ</td>
<td>s z</td>
</tr>
<tr>
<td>Approximant</td>
<td>v</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>
</tr>
<tr>
<td>Tap, Flap</td>
<td>v</td>
<td>r</td>
<td>r</td>
<td>r</td>
</tr>
<tr>
<td>Lateral fricative</td>
<td>ɾ</td>
<td>ɾ</td>
<td>r</td>
<td>r</td>
</tr>
<tr>
<td>Lateral approximant</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
</tr>
<tr>
<td>Lateral flap</td>
<td>j</td>
<td>j</td>
<td>j</td>
<td>j</td>
</tr>
</tbody>
</table>

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

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
# Time-Domain Information

<table>
<thead>
<tr>
<th></th>
<th>s</th>
<th>eɪ</th>
<th>pʰ</th>
<th>æ</th>
<th>t</th>
<th>n</th>
<th>aʊ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pat</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>pad</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>bad</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>spat</strong></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
Frequency components (100 and 1000 Hz) on x-axis
- 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 .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 Speech
Spectrograms
Spectrograms

[Image of spectrograms with frequency and time axes]
Spectrograms
Types of Graphs
Back to Spectra

- 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.
Source / Filter
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:

<table>
<thead>
<tr>
<th>Closed end</th>
<th>Open end</th>
</tr>
</thead>
</table>

  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
FIRST FORMANT
1/4 WAVELENGTH
500 HERTZ

SECOND FORMANT
3/4 WAVELENGTH
1,500 HERTZ

THIRD FORMANT
5/4 WAVELENGTH
2,500 HERTZ

FOURTH FORMANT
7/4 WAVELENGTH
3,500 HERTZ

From Sundberg
Computing the 3 Formants of Schwa

Let the length of the tube be L

\[ F_1 = \frac{c}{\lambda_1} = \frac{c}{4L} = \frac{35,000}{4 \times 17.5} = 500 \text{Hz} \]
\[ F_2 = \frac{c}{\lambda_2} = \frac{c}{4/3L} = \frac{3c}{4L} = \frac{3 \times 35,000}{4 \times 17.5} = 1500 \text{Hz} \]
\[ F_3 = \frac{c}{\lambda_3} = \frac{c}{4/5L} = \frac{5c}{4L} = \frac{5 \times 35,000}{4 \times 17.5} = 2500 \text{Hz} \]

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

Front  Near front  Central  Near back  Back

Close  i  y — i  ü  u
Near close  I  Y  I  ü  u
Close mid  e  ø — ø  œ  o
Mid  œ  œ  œ  o
Open mid  ø  æ — œ  œ  œ
Near open  æ  œ  œ  œ
Open

Vowels at right & left of bullets are rounded & unrounded.
Seeing Formants: the Spectrogram

[Diagram showing spectrograms for various phonemes: [i], [ɪ], [ɛ], [æ], [ɑ], [ɔ], [ʊ], [u].]
American English Vowel Space

Figures from Jennifer Venditti, H. T. Bunnell
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
2. closure for k
3. burst of aspiration for k
4. ey vowel; faint 1100 Hz formant is nasalization
5. bilabial nasal
6. short b closure, voicing barely visible.
7. ae; note upward transitions after bilabial stop at beginning
8. note F2 and F3 coming together for "k"

From Ladefoged “A Course in Phonetics”
Dialect Issues

- Speech varies from dialect to dialect (examples are American vs. British English)
  - Syntactic (“I could” vs. “I could do”)
  - Lexical (“elevator” vs. “lift”)
  - Phonological
  - Phonetic

- Mismatch between training and testing dialects can cause a large increase in error rate
Speech Recognition
The Noisy Channel Model

**Acoustic model:** HMMs over word positions with mixtures of Gaussians as emissions

**Language model:** Distributions over sequences of words (sentences)

\[
\begin{align*}
  w^* &= \arg \max_w P(w|a) \\
  &\propto \arg \max_w P(a|w) P(w)
\end{align*}
\]
Speech Model

![Diagram of a speech model with nodes and arrows indicating the flow from words to sound types and acoustic observations.](image)
Acoustic Model

Sound types:
- $s_1$
- $s_2$
- $s_3$
- $s_4$
- $s_5$
- $s_6$
- $s_7$

Acoustic observations:
- $a_1$
- $a_2$
- $a_3$
- $a_4$
- $a_5$
- $a_6$
- $a_7$

Acoustic model
A frame (25 ms wide) extracted every 10 ms

25 ms

10ms

$a_1$, $a_2$, $a_3$

Preview of feature extraction for each frame:
1) DFT (Spectrum)
2) Log (Calibrate?)
3) another DFT (!!???)

Figure: Simon Arnfield
Feature Extraction
Digitizing Speech

Figure: Bryan Pellom
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
Problem with Raw Spectrum

Figures from Ratree Wayland
Deconvolution / Liftering
Deconvolution / Lifting
Deconvolution / Lifting

\[ \log(s) = \log(e) + \log(f) \]
Deconvolution / Lifting

\[ s = e \circ f \]

\[ \log(s) = \log(e) + \log(f) \]

\[ \text{IDFT}(\log(s)) \]
- Do FFT to get spectral information
  - Like the spectrogram we saw earlier

- Apply Mel scaling (New)
  - Models human ear; more sensitivity in lower freqs
  - Approx linear below 1kHz, log above, equal samples above and below 1kHz

- Take Log
- Do discrete cosine transform
Final Feature Vector

- 39 (real) features per 10 ms frame:
  - 12 MFCC features
  - 12 delta MFCC features
  - 12 delta-delta MFCC features
  - 1 (log) frame energy
  - 1 delta (log) frame energy
  - 1 delta-delta (log frame energy)

- So each frame is represented by a 39D vector
Emission Model
HMMs for Continuous Observations

- Before: discrete set of observations
- Now: feature vectors are real-valued
- Solution 1: discretization
- Solution 2: continuous emissions
  - Gaussians
  - Multivariate Gaussians
  - Mixtures of multivariate Gaussians
- A state is progressively
  - Context independent subphone (~3 per phone)
  - Context dependent phone (triphones)
  - State tying of CD phone
Vector Quantization

- **Idea:** discretization
  - Map MFCC vectors onto discrete symbols
  - Compute probabilities just by counting

- This is called vector quantization or VQ

- Not used for ASR any more

- But: useful to consider as a starting point
VQ is insufficient for top-quality ASR
- Hard to cover high-dimensional space with codebook
- Moves ambiguity from the model to the preprocessing

Instead: assume the possible values of the observation vectors are normally distributed.
- Represent the observation likelihood function as a Gaussian?
A Gaussian is parameterized by a mean and a variance:

\[
P(x | \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right)
\]

- \(P(x)\):
  - \(P(x)\) is highest here at mean
  - \(P(x)\) is low here, far from mean
Multivariate Gaussians

- Instead of a single mean $\mu$ and variance $\sigma^2$:

  $$P(x|\mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right)$$

- Vector of means $\mu$ and covariance matrix $\Sigma$

  $$P(x|\mu, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right)$$

- Usually assume diagonal covariance (!!)
  - This isn’t very true for FFT features, but is less bad for MFCC features
Gaussians: Size of $\Sigma$

- $\mu = [0 \ 0]$  
- $\Sigma = I$  
- $\Sigma = 0.6I$  
- $\Sigma = 2I$

As $\Sigma$ becomes larger, Gaussian becomes more spread out; as $\Sigma$ becomes smaller, Gaussian more compressed

Text and figures from Andrew Ng
Gaussians: Shape of $\Sigma$

$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$; $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$; $\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$

- As we increase the off diagonal entries, more correlation between value of $x$ and value of $y$
But we’re not there yet

- Single Gaussians may do a bad job of modeling a complex distribution in any dimension

- Even worse for diagonal covariances

- Solution: mixtures of Gaussians

From openlearn.open.ac.uk
Mixtures of Gaussians:

\[ P(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{k/2}|\Sigma_i|^{1/2}} \exp \left( -\frac{1}{2}(x - \mu_i)^\top \Sigma_i^{-1}(x - \mu_i) \right) \]

\[ P(x|\mu, \Sigma, c) = \sum_i c_i P(x|\mu_i, \Sigma_i) \]
Summary: each state has an emission distribution \( P(x|s) \) (likelihood function) parameterized by:

- \( M \) mixture weights
- \( M \) mean vectors of dimensionality \( D \)
- Either \( M \) covariance matrices of \( D \times D \) or \( M \) \( D \times 1 \) diagonal variance vectors

Like soft vector quantization after all

- Think of the mixture means as being learned codebook entries
- Think of the Gaussian densities as a learned codebook distance function
- Think of the mixture of Gaussians like a multinomial over codes
- (Even more true given shared Gaussian inventories, cf next week)
State Model
State Transition Diagrams

- Bayes Net: HMM as a Graphical Model

- State Transition Diagram: Markov Model as a Weighted FSA
Lexical State Structure

figure: J & M
Adding an LM

Figure from Huang et al page 618
State Space

- State space must include
  - Current word ($|V|$ on order of 20K+)
  - Index within current word ($|L|$ on order of 5)
  - E.g. (lec[t]ure) (though not in orthography!)

- Acoustic probabilities only depend on phone type
  - E.g. $P(x|\text{lec[t]ure}) = P(x|t)$

- From a state sequence, can read a word sequence
State Refinement
Phones Aren’t Homogeneous
Need to Use Subphones

Figure: J & M
A Word with Subphones

Figure: J & M
Modeling phonetic context
“Need” with triphone models
Lots of Triphones

- Possible triphones: $50 \times 50 \times 50 = 125,000$

- How many triphone types actually occur?

- 20K word WSJ Task (from Bryan Pellom)
  - Word internal models: need 14,300 triphones
  - Cross word models: need 54,400 triphones

- Need to generalize models, tie triphones
State Tying / Clustering

- [Young, Odell, Woodland 1994]
- How do we decide which triphones to cluster together?
- Use phonetic features (or ‘broad phonetic classes’)
  - Stop
  - Nasal
  - Fricative
  - Sibilant
  - Vowel
  - lateral

Figure: J & M
State Space

- State space now includes
  - Current word: $|W|$ is order 20K
  - Index in current word: $|L|$ is order 5
  - Subphone position: 3
  - E.g. (lec[t-mid]ure)

- Acoustic model depends on clustered phone context
  - But this doesn’t grow the state space

- But, adding the LM context for trigram+ does
  - (after the, lec[t-mid]ure)
  - This is a real problem for decoding
Decoding
Inference Tasks

Most likely word sequence:

\[
d \quad - \quad ae \quad - \quad d
\]

Most likely state sequence:

\[
d_1-d_6-d_6-d_4-ae_5-ae_2-ae_3-ae_0-d_2-d_2-d_3-d_7-d_5
\]
Viterbi Decoding

\[
\phi_t(s_t, s_{t-1}) = P(x_t|s_t)P(s_t|s_{t-1})
\]

\[
v_t(s_t) = \max_{s_{t-1}} \phi_t(s_t, s_{t-1})v_{t-1}(s_{t-1})
\]

Figure: Enrique Benimeli
Viterbi Decoding

Figure: Enrique Benimeli
Emission Caching

- Problem: scoring all the $P(x|s)$ values is too slow
- Idea: many states share tied emission models, so cache them
Prefix Trie Encodings

- Problem: many partial-word states are indistinguishable
- Solution: encode word production as a prefix trie (with pushed weights)

- A specific instance of minimizing weighted FSAs [Mohri, 94]
Beam Search

- Problem: trellis is too big to compute $v(s)$ vectors
- Idea: most states are terrible, keep $v(s)$ only for top states at each time
- Important: still dynamic programming; collapse equiv states
Problem: Higher-order n-grams explode the state space

(One) Solution:
- Factor state space into (word index, lm history)
- Score unigram prefix costs while inside a word
- Subtract unigram cost and add trigram cost once word is complete
LM Reweighting

- Noisy channel suggests
  \[ P(x|w)P(w) \]

- In practice, want to boost LM
  \[ P(x|w)P(w)^\alpha \]

- Also, good to have a “word bonus” to offset LM costs
  \[ P(x|w)P(w)^\alpha|w|^\beta \]

- These are both consequences of broken independence assumptions in the model
Training
Training Mixture Models

- **Input:** wav files with unaligned transcriptions

- **Forced alignment**
  - Computing the “Viterbi path” over the training data (where the transcription is known) is called “forced alignment”
  - We know which word string to assign to each observation sequence.
  - We just don’t know the state sequence.
  - So we constrain the path to go through the correct words (by using a special example-specific language model)
  - And otherwise run the Viterbi algorithm

- **Result:** aligned state sequence
Creating CD phones:
- Start with monophone, do EM training
- Clone Gaussians into triphones
- Build decision tree and cluster Gaussians
- Clone and train mixtures (GMMs)

General idea:
- Introduce complexity gradually
- Interleave constraint with flexibility
Standard subphone/mixture HMM

Temporal Structure

Gaussian Mixtures

<table>
<thead>
<tr>
<th>Model</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM Baseline</td>
<td>25.1%</td>
</tr>
</tbody>
</table>
An Induced Model

Fully Connected

Single Gaussians

Standard Model

[Petrov, Pauls, and Klein, 07]
Hierarchical Split Training with EM

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM Baseline</td>
<td>25.1%</td>
</tr>
<tr>
<td>5 Split rounds</td>
<td>21.4%</td>
</tr>
</tbody>
</table>

- 32.1%
- 28.7%
- 25.6%
- 23.9%
Refinement of the /ih/-phone
Refinement of the /ih/-phone
Refinement of the /ih/-phone
HMM states per phone