Statistical NLP
Spring 2007

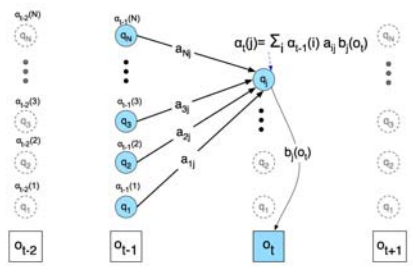
Lecture 8: Speech Signal

Dan Klein – UC Berkeley

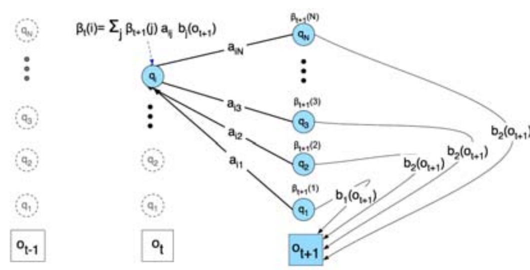
Unsupervised Tagging?

- AKA part-of-speech induction
- Task:
  - Raw sentences in
  - Tagged sentences out
- Obvious thing to do:
  - Start with a (mostly) uniform HMM
  - Run EM
  - Inspect results

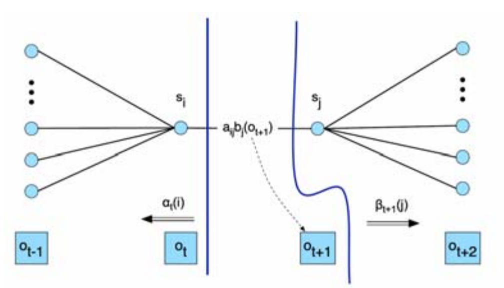
Forward Recurrence



Backward Recurrence



Fractional Transitions



EM for HMMs: Quantities

- Cache total path values:

$$\alpha_i(s) = P(w_0 \dots w_i, s_i) = \sum_{s_{i-1}} P(s_i|s_{i-1}) P(w_i|s_i) \alpha_{i-1}(s_{i-1})$$

$$\beta_i(s) = P(w_i + 1 \dots w_n|s_i) = \sum_{s_{i+1}} P(s_{i+1}|s_i) P(w_{i+1}|s_{i+1}) \beta_{i+1}(s_{i+1})$$

- Can calculate in $O(s^2 n)$ time (why?)

EM for HMMs: Process

- From these quantities, we can re-estimate transitions:

$$\text{count}(s \rightarrow s') = \frac{\sum_i \alpha_i(s) P(s'|s) P(w_i|s) \beta_{i+1}(s')}{P(\mathbf{w})}$$

- And emissions:

$$\text{count}(w, s) = \frac{\sum_{i: w_i = w} \alpha_i(s) \beta_{i+1}(s)}{P(\mathbf{w})}$$

- If you don't get these formulas immediately, just think about hard EM instead, where were re-estimate from the Viterbi sequences

Merialdo: Setup

- Some (discouraging) experiments [Merialdo 94]
- Setup:
  - You know the set of allowable tags for each word
  - Fix k training examples to their true labels
    - Learn P(w|t) on these examples
    - Learn P(t|t_{-1},t_{-2}) on these examples
  - On n examples, re-estimate with EM
- Note: we know allowed tags but not frequencies

Merialdo: Results

	Number of tagged sentences used for the initial model						
	0	100	2000	5000	10000	20000	all
Iter	Correct tags (% words) after ML on 1M words						
0	77.0	90.0	95.4	96.2	96.6	96.9	97.0
1	80.5	92.6	95.8	96.3	96.6	96.7	96.8
2	81.8	93.0	95.7	96.1	96.3	96.4	96.4
3	83.0	93.1	95.4	95.8	96.1	96.2	96.2
4	84.0	93.0	95.2	95.5	95.8	96.0	96.0
5	84.8	92.9	95.1	95.4	95.6	95.8	95.8
6	85.3	92.8	94.9	95.2	95.5	95.6	95.7
7	85.8	92.8	94.7	95.1	95.3	95.5	95.5
8	86.1	92.7	94.6	95.0	95.2	95.4	95.4
9	86.3	92.6	94.5	94.9	95.1	95.3	95.3
10	86.6	92.6	94.4	94.8	95.0	95.2	95.2

Distributional Clustering

♦ *the president said that the downturn was over* ♦

president	the __ of
president	the __ said
governor	the __ of
governor	the __ appointed
said	sources __ ♦
said	president __ that
reported	sources __ ♦

president governor

the a

said reported

[Finch and Chater 92, Shuetze 93, many others]

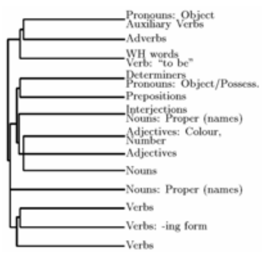
Distributional Clustering

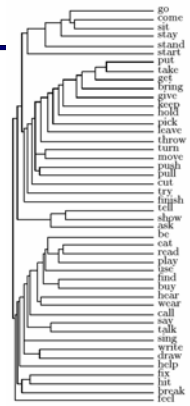
- Three main variants on the same idea:
  - Pairwise similarities and heuristic clustering
    - E.g. [Finch and Chater 92]
    - Produces dendrograms
  - Vector space methods
    - E.g. [Shuetze 93]
    - Models of ambiguity
  - Probabilistic methods
    - Various formulations, e.g. [Lee and Pereira 99]

Nearest Neighbors

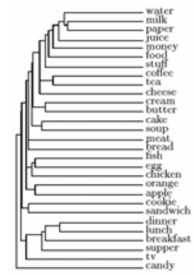
word	nearest neighbors
accompanied	submitted banned financed developed authorized headed canceled awarded barred
almost	virtually merely formally fully quite officially just nearly only less
causing	reflecting forcing providing creating producing becoming carrying particularly
classes	elections courses payments losses computers performances violations levels pictures
directors	professionals investigations materials competitors agreements papers transactions
goal	mood roof eye image tool song pool scene gap voice
japanese	chinese iraqi american western arab foreign european federal soviet indian
represent	reveal attend deliver reflect choose contain impose manage establish retain
think	believe wish know realize wonder assume feel say mean bet
york	angeles francisco sox rouge kong diego zone vegas inning layer
on	through in at over into with from for by across
must	might would could cannot will should can may does helps
they	we you i he she nobody who it everybody there

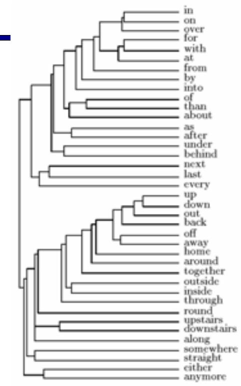
Dendrograms





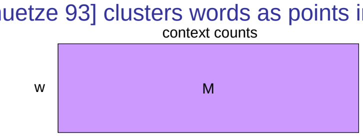
Dendrograms



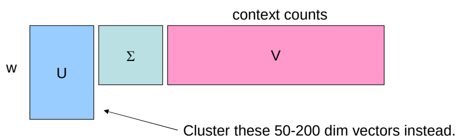


Vector Space Version

- [Shuetze 93] clusters words as points in R^n

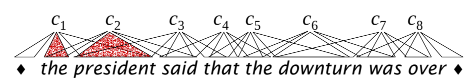


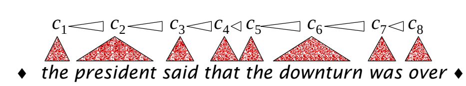
- Vectors too sparse, use SVD to reduce



A Probabilistic Version?

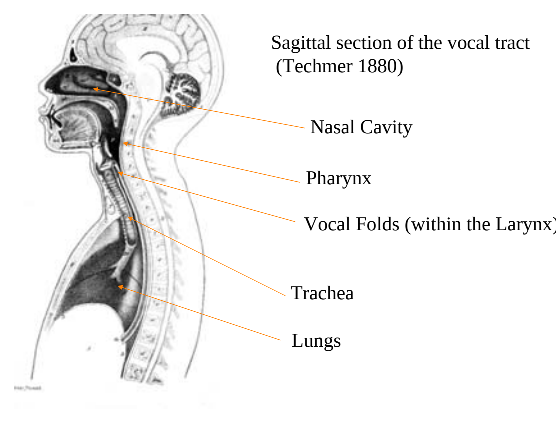
$$P(S,C)=\prod_i P(c_i)P(w_i \mid c_i)P(w_{i-1},w_{i+1} \mid c_i)$$





What Else?

- Various newer ideas:
  - Context distributional clustering [Clark 00]
  - Morphology-driven models [Clark 03]
  - Contrastive estimation [Smith and Eisner 05]
- Also:
  - What about ambiguous words?
  - Using wider context signatures has been used for learning synonyms (what's wrong with this approach?)
  - Can extend these ideas for grammar induction (later)



Places of articulation

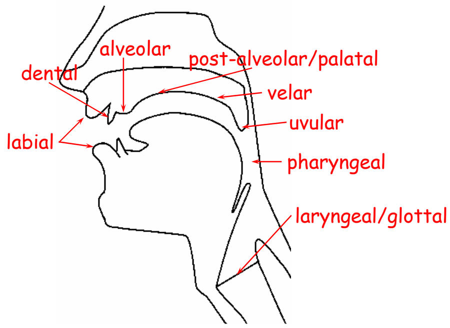
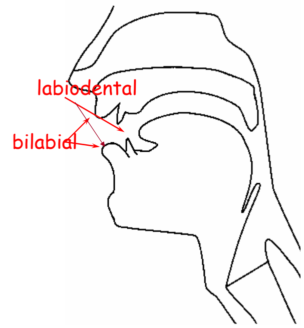


Figure thanks to Jennifer Venditti

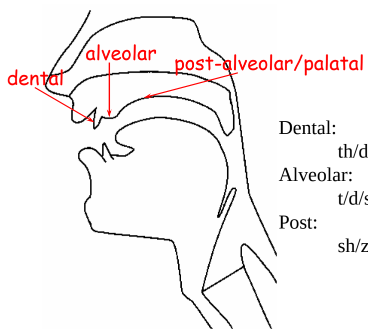
Labial place



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

Figure thanks to Jennifer Venditti

Coronal place



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

Figure thanks to Jennifer Venditti

Dorsal Place

Velar:
k/g/ng

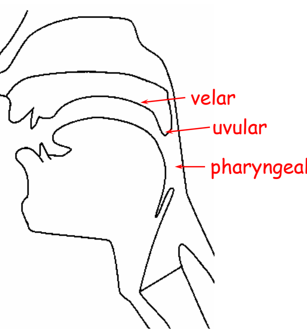
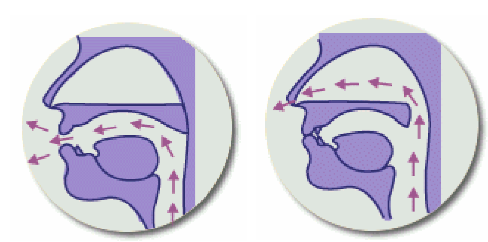


Figure thanks to Jennifer Venditti

Manner of Articulation

- Stop: complete closure of articulators, so no air escapes through mouth
- Oral stop: palate is raised, no air escapes through nose. Air pressure builds up behind closure, explodes when released
  - p, t, k, b, d, g
- Nasal stop: oral closure, but palate is lowered, air escapes through nose.
  - m, n, ng

Oral vs. Nasal Sounds



Thanks to Jong-bok Kim for this figure!

Vowels

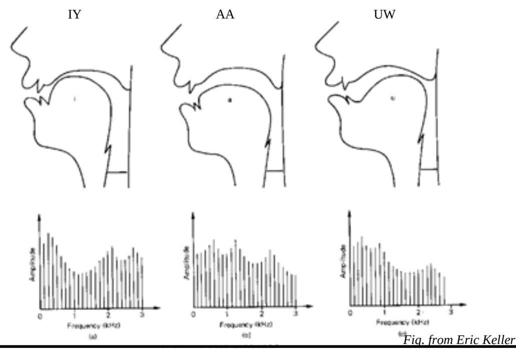
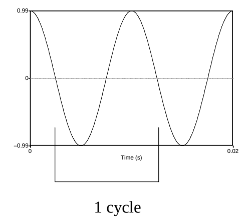


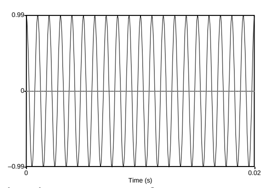
Fig. from Eric Keller

Simple Periodic Waves

- Characterized by:
  - period: T
  - amplitude A
  - phase ϕ
- Fundamental frequency in cycles per second, or Hz
  - $F_0 = 1/T$

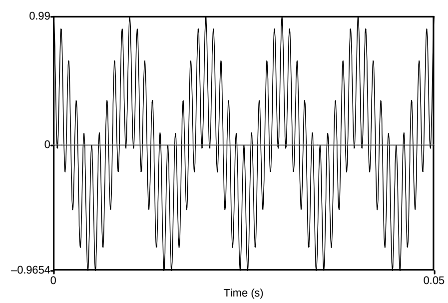


Simple periodic waves of sound



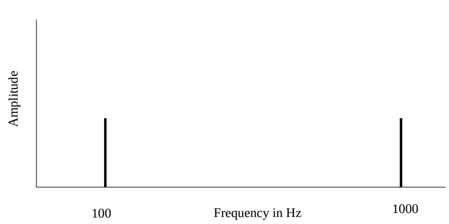
- **Y axis: Amplitude = amount of air pressure at that point in time**
  - **Zero is normal air pressure, negative is rarefaction**
- **X axis: time. Frequency = number of cycles per second.**
- **Frequency = 1/Period**
- **20 cycles in .02 seconds = 1000 cycles/second = 1000 Hz**

Complex waves: 100Hz+1000Hz

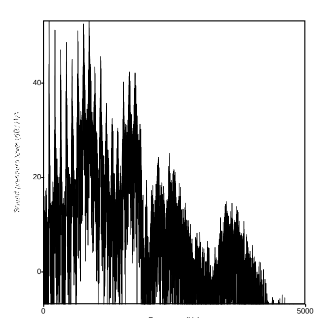


Spectrum

Frequency components (100 and 1000 Hz) on x-axis

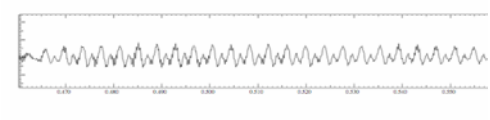


Spectrum of an actual soundwave



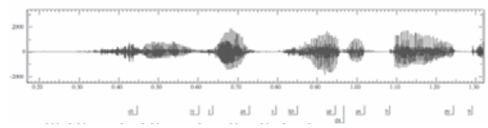
Waveforms for speech

- Waveform of the vowel [iy]



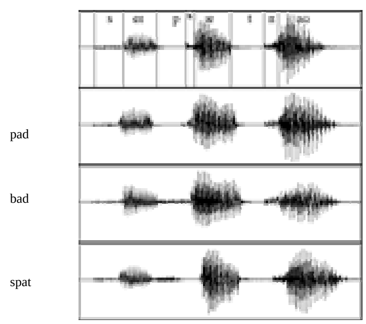
- Frequency: repetitions/second of a wave
- Above vowel has 28 reps in .11 secs
- So freq is 28/.11 = 255 Hz
- This is speed that vocal folds move, hence voicing
- Amplitude: y axis: amount of air pressure at that point in time
- Zero is normal air pressure, negative is rarefaction

She just had a baby



-
  - 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

Examples from Ladefoged

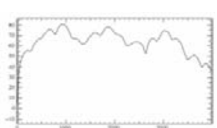


Part of [ae] waveform from "had"

- 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 .036 seconds)
- Small wave roughly 4 times this, or roughly 1000 Hz
- Two little tiny waves on top of peak of 1000 Hz waves

Back to Spectra

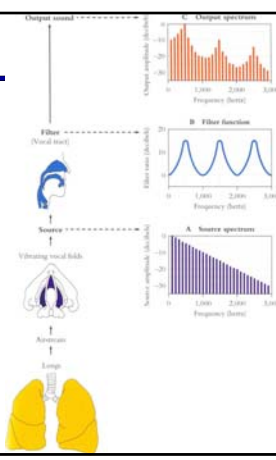
- Spectrum represents these freq components
- Computed by Fourier transform, algorithm which separates out each frequency component of wave.



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

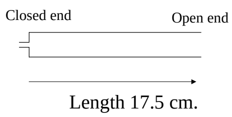
Why these Peaks?

- **Articulatory facts:**
  - **The vocal cord vibrations create** harmonics
  - **The mouth is an amplifier**
  - **Depending on shape of mouth, some harmonics are amplified more than others**



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.

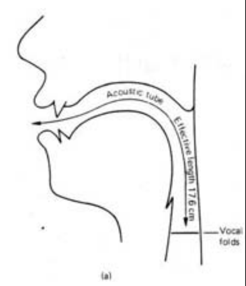
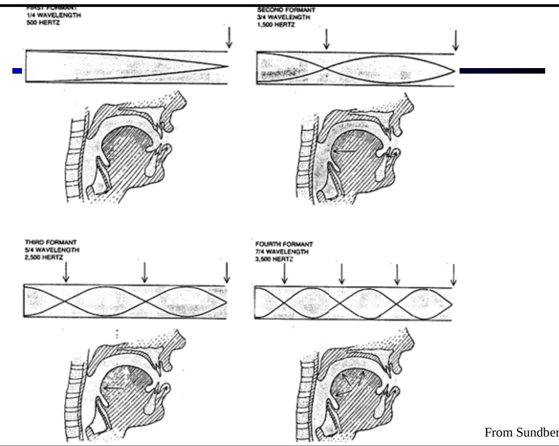


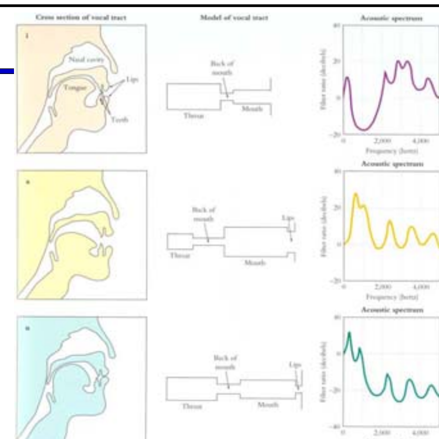
Figure from W. Barry Speech Science slides



From Sundberg

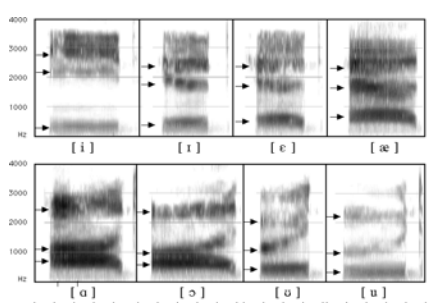
Computing the 3 Formants of Schwa

- Let the length of the tube be L
  - $F_1 = c/\lambda_1 = c/(4L) = 35{,}000/4*17.5 =$ 500Hz
  - $F_2 = c/\lambda_2 = c/(4/3L) = 3c/4L = 3*35{,}000/4*17.5 =$ 1500Hz
  - $F_3 = c/\lambda_3 = c/(4/5L) = 5c/4L = 5*35{,}000/4*17.5 =$ 2500Hz
- So we expect a neutral vowel to have 3 resonances at 500, 1500, and 2500 Hz
- These vowel resonances are called formants



From Mark Liberman's Web site

Seeing formants: the spectrogram



American English Vowel Space

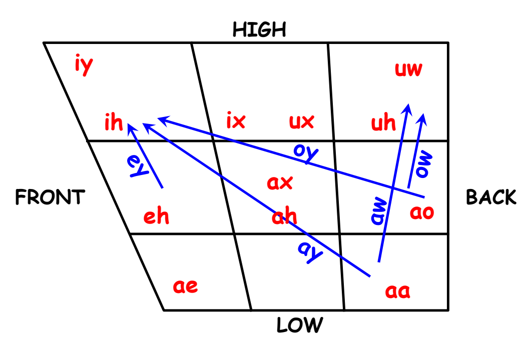
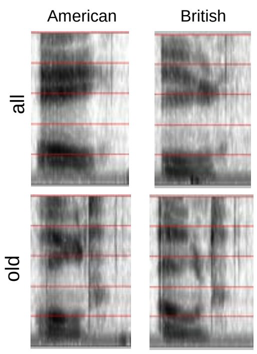


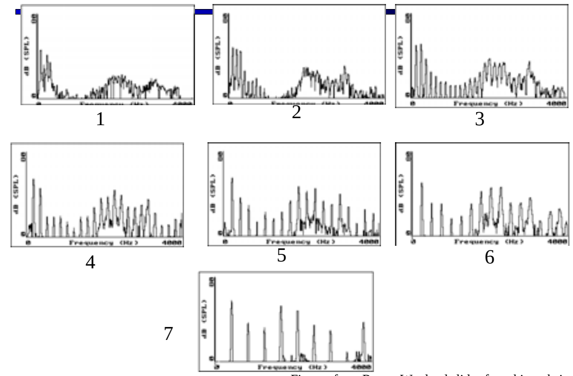
Figure from Jennifer Venditti

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 (butter: [♌✍☼❒] vs. [♌✍♦☜])
  - Phonetic
- Mismatch between training and testing dialects can cause a large increase in error rate

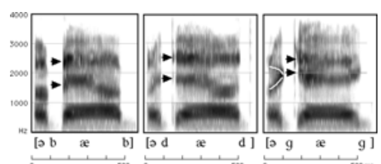


Vowel [i] sung at successively higher pitch.



Figures from Ratree Wayland slides from his website

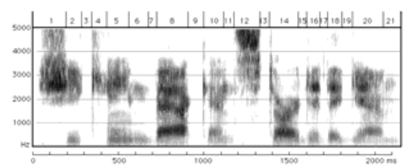
How to read spectrograms



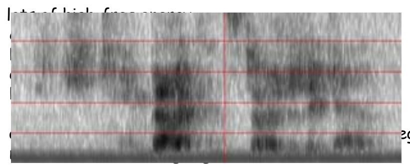
- bab**: closure of lips lowers all formants: so rapid increase in all formants at beginning of "bab"**
- dad**: first formant increases, but F2 and F3 slight fall**
- gag**: F2 and F3 come together: this is a characteristic of velars. Formant transitions take longer in velars than in alveolars or labials**

From Ladefoged "A Course in Phonetics"

She came back and started again



1. l
3.
4.
5.
6.
7.
8. eginning
9.



From Ladefoged "A Course in Phonetics"