

Statistical NLP Spring 2008



Lecture 9: Speech Signal

Dan Klein – UC Berkeley

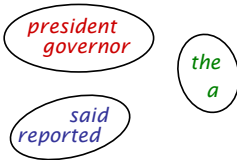
Unsupervised Tagging?

- 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

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 __ ♦



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

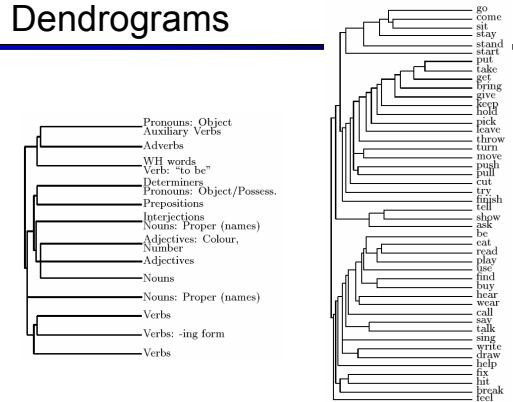
Distributional Clustering

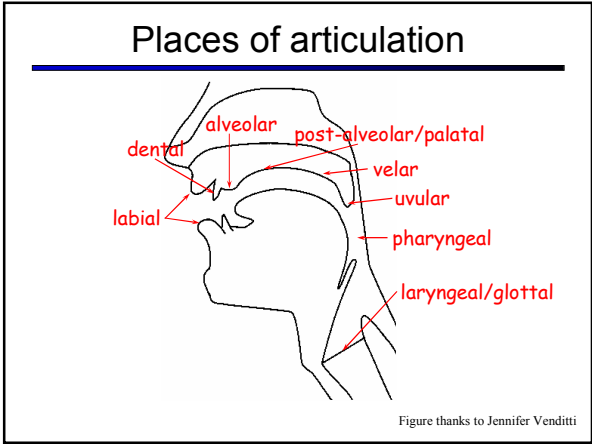
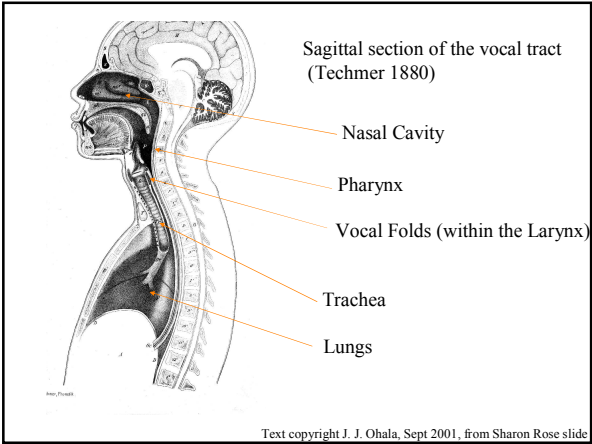
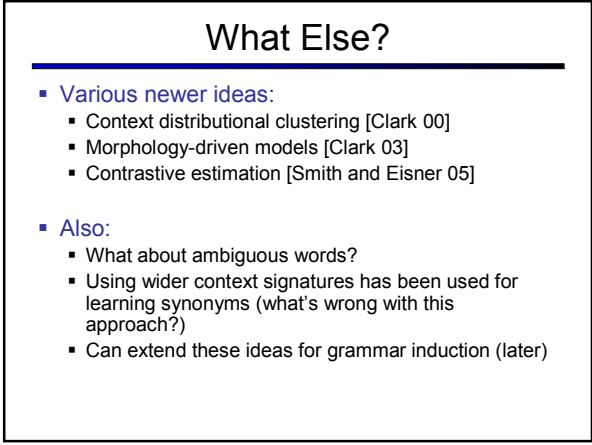
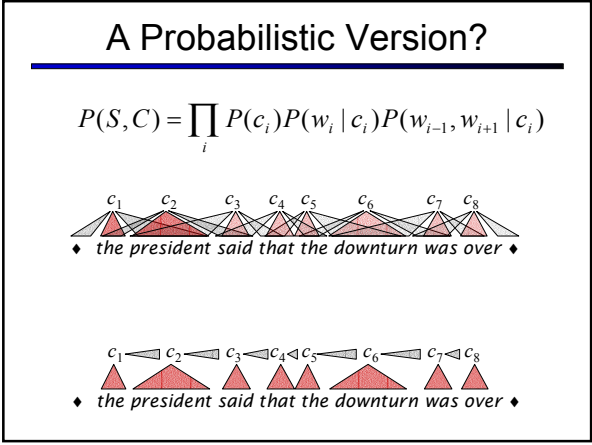
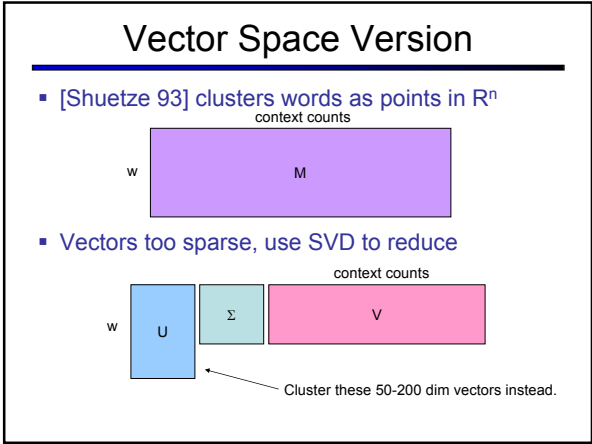
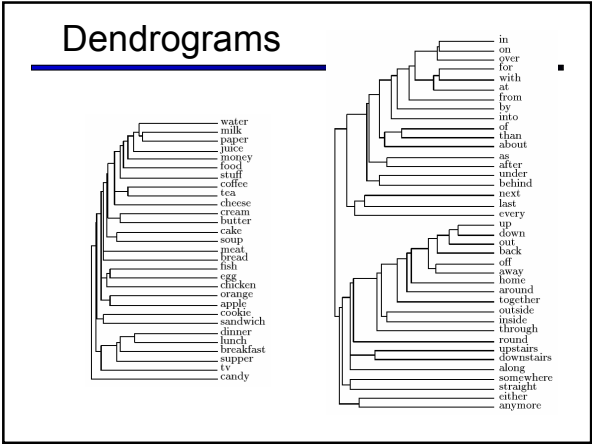
- Three main variants on the same idea:
 - Pairwise similarities and heuristic clustering
 - E.g. [Finch and Chater 92]
 - Produces dendrograms
 - Vector space methods
 - E.g. [Shuetze 93]
 - Models of ambiguity
 - Probabilistic methods
 - Various formulations, e.g. [Lee and Pereira 99]

Nearest Neighbors

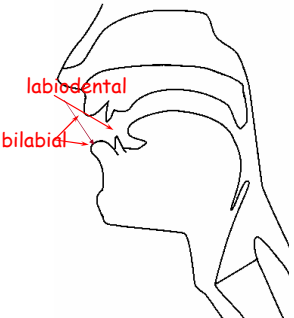
word	nearest neighbors
accompanied	submitted banned financed developed authorized headed canceled awarded barred
almost	virtually merely formally fully quite officially just nearly only less
causing	reflecting forcing providing creating producing becoming carrying particularly
classes	elections courses 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
think	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





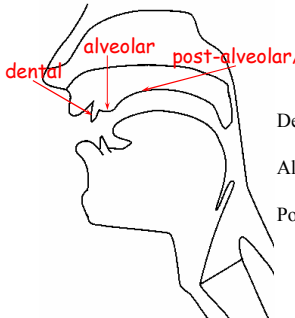
Labial place



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

Figure thanks to Jennifer Venditti

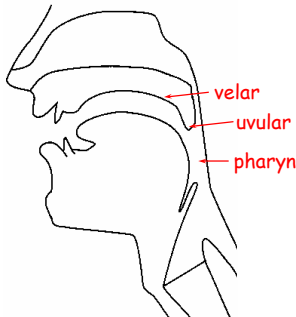
Coronal place



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

Figure thanks to Jennifer Venditti

Dorsal Place



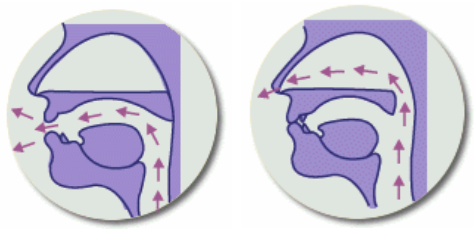
Velar:
k/g/ng

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

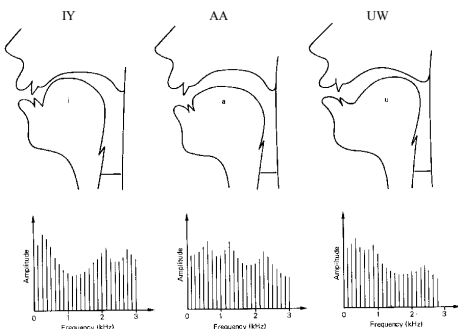
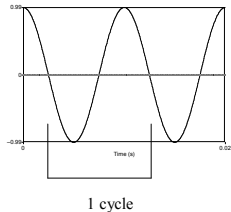


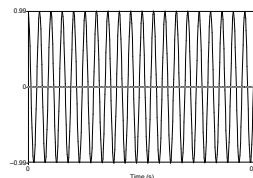
Fig. from Eric Keller

Simple Periodic Waves

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

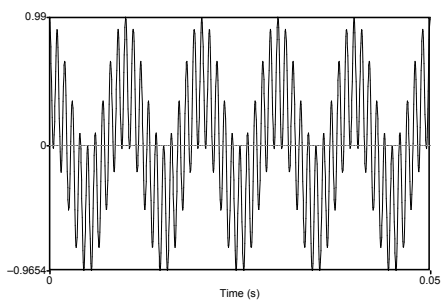


Simple periodic waves of sound



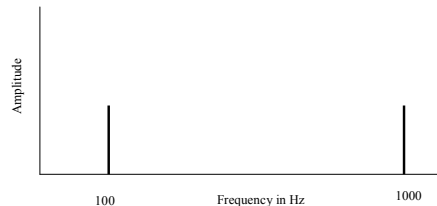
- Y axis: Amplitude = amount of air pressure at that point in time
 - Zero is normal air pressure, negative is rarefaction
- X axis: time. Frequency = number of cycles per second.
- Frequency = $1/\text{Period}$
- 20 cycles in .02 seconds = 1000 cycles/second = 1000 Hz

Complex waves: 100Hz+1000Hz

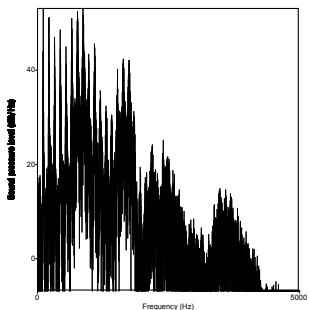


Spectrum

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

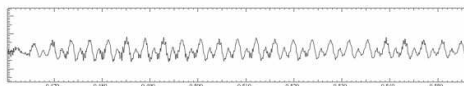


Spectrum of an actual soundwave



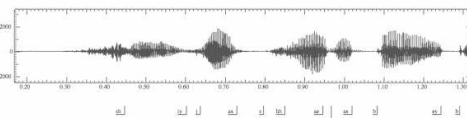
Waveforms for speech

- Waveform of the vowel [iy]



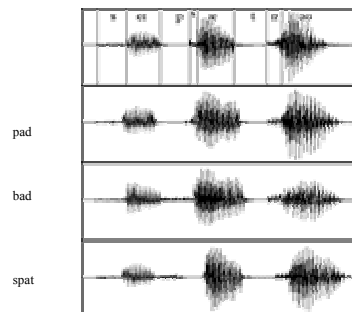
- Frequency: repetitions/second of a wave
- Above vowel has 28 reps in .11 secs
- So freq is $28/.11 = 255$ Hz
- This is speed that vocal folds move, hence voicing
- Amplitude: y axis: amount of air pressure at that point in time
- Zero is normal air pressure, negative is rarefaction

She just had a baby

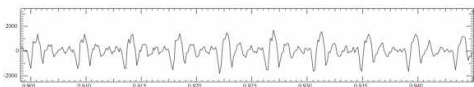


- What can we learn from a wavefile?
 - Vowels are voiced, long, loud
 - Length in time = length in space in waveform picture
 - Voicing: regular peaks in amplitude
 - When stops closed: no peaks: silence.
 - Peaks = voicing: .46 to .58 (vowel [y], from second .65 to .74 (vowel [ax]) and so on
 - Silence 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



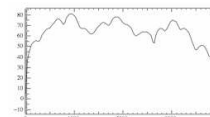
Part of [ae] waveform from "had"



- 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

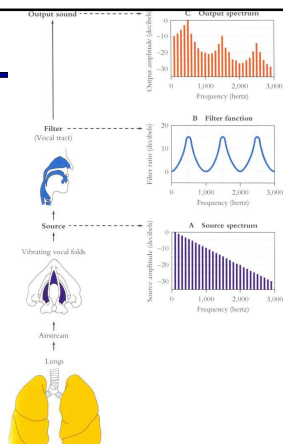
- Spectrum represents these freq components
- Computed by Fourier transform, algorithm which separates out each frequency component of wave.



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

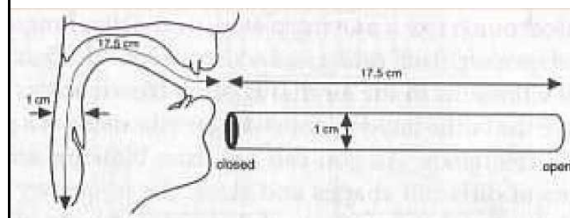
Why these Peaks?

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



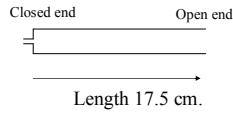
Deriving Schwa

- Reminder of basic facts about sound waves
 - $f = c/\lambda$.
 - c = speed of sound (approx 35,000 cm/sec)
 - A sound with $\lambda=10$ meters: $f = 35$ Hz (35,000/1000)
 - A sound with $\lambda=2$ centimeters: $f = 17,500$ Hz (35,000/2)



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.

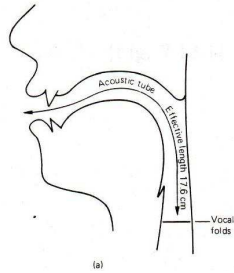
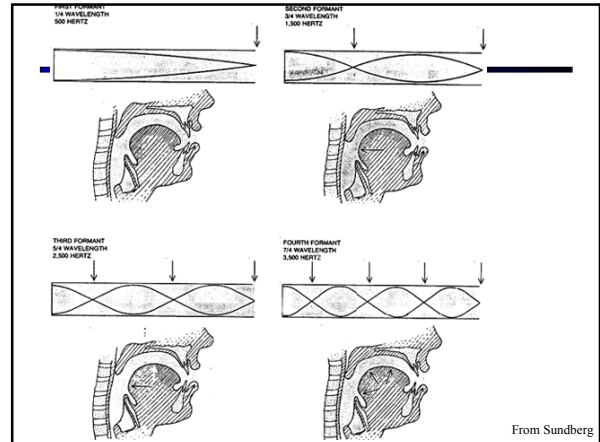
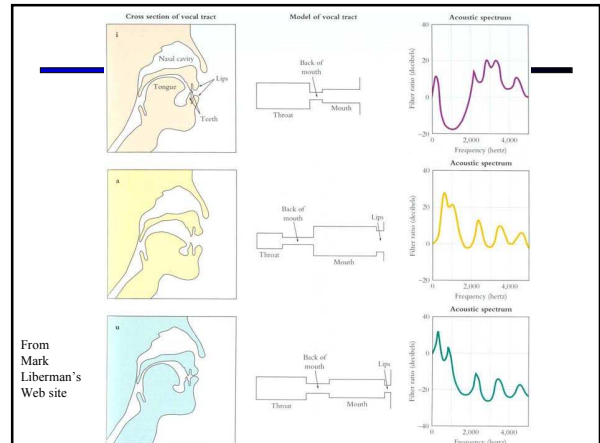


Figure from W. Barry Speech Science slides

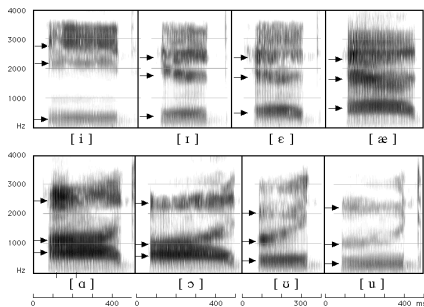


Computing the 3 Formants of Schwa

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



American English Vowel Space

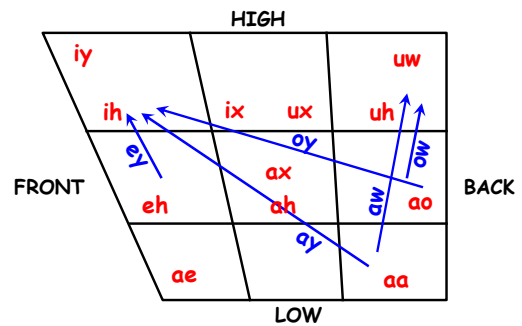
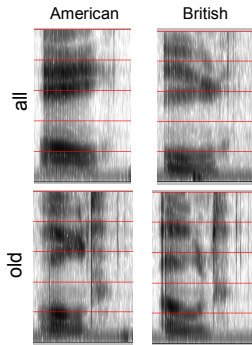


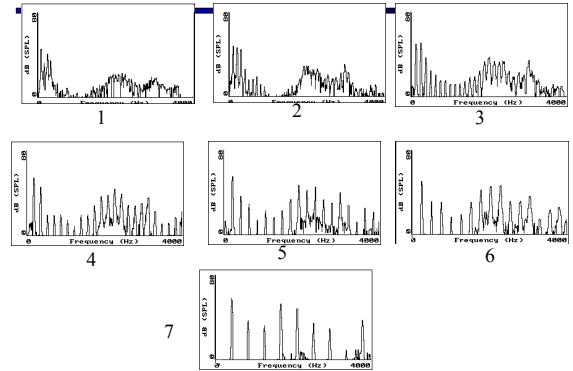
Figure from Jennifer Venditti

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
 - Phonetic
- Mismatch between training and testing dialects can cause a large increase in error rate

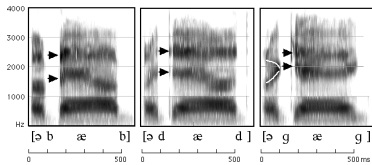


Vowel [i] sung at successively higher pitch.



Figures from Katreë Wayland slides from his website

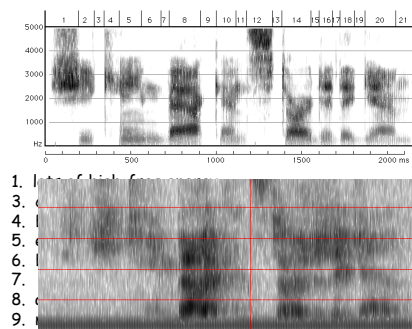
How to read spectrograms



- 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



From Ladefoged "A Course in Phonetics"

The Noisy Channel Model



- Search through space of all possible sentences.
- Pick the one that is most probable given the waveform.

Speech Recognition Architecture

