CS 188: Artificial Intelligence

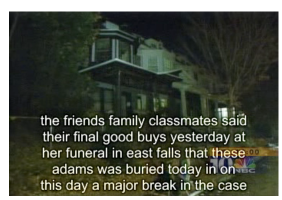
Lecture 18: Speech

Pieter Abbeel --- UC Berkeley

Many slides over this course adapted from Dan Klein, Stuart Russell, Andrew Moore

Speech and Language

- Speech technologies
  - Automatic speech recognition (ASR)
  - Text-to-speech synthesis (TTS)
  - Dialog systems



- Language processing technologies
  - Machine translation


"Il est impossible aux journalistes de rentrer dans les régions tibétaines"

Bruno Philip, correspondant du "Monde" en Chine, estime que les journalistes de l'AFP qui ont été expulsés de la province tibétaine du Qinghai "n'étaient pas dans l'illégalité".

Les faits Le dalaï-lama dénonce l'"enfer" imposé au Tibet depuis sa fuite, en 1959

Vidéo Anniversaire de la rébellion


"It is impossible for journalists to enter Tibetan areas"

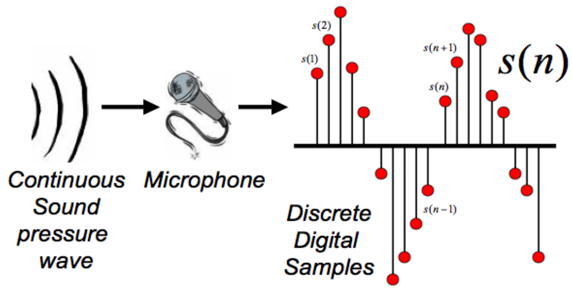
Philip Bruno, correspondent for "World" in China, said that journalists of the AFP who have been deported from the Tibetan province of Qinghai "were not illegal."

Facts The Dalai Lama denounces the "hell" imposed since he fled Tibet in 1959

Video Anniversary of the Tibetan rebellion: China on guard

  - Information extraction
  - Web search, question answering
  - Text classification, spam filtering, etc…

Digitizing Speech

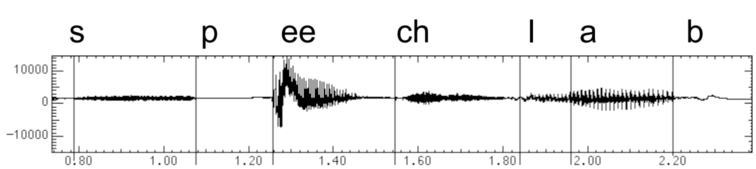


Thanks to Bryan Pellom for this slide!

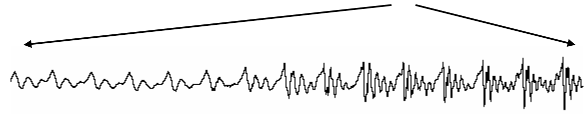
3

Speech in an Hour

- Speech input is an acoustic wave form



"l" to "a" transition:

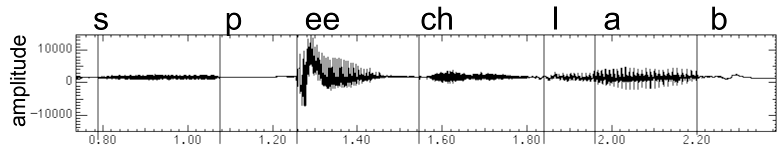


Graphs from Simon Arnfield's web tutorial on speech, Sheffield:
http://www.psyc.leeds.ac.uk/research/cogn/speech/tutorial/

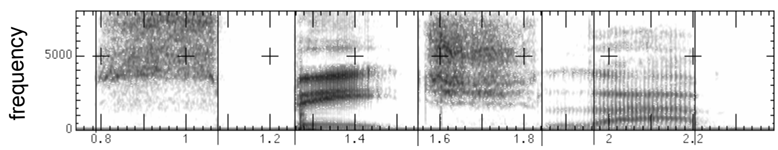
4

Spectral Analysis

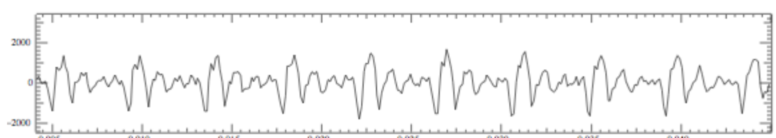
- Frequency gives pitch; amplitude gives volume
  - sampling at ~8 kHz phone, ~16 kHz mic (kHz=1000 cycles/sec)



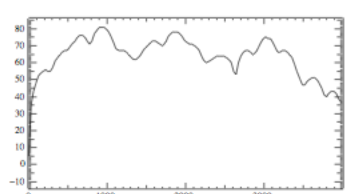
- Fourier transform of wave displayed as a spectrogram
  - darkness indicates energy at each frequency



Part of [ae] from "lab"



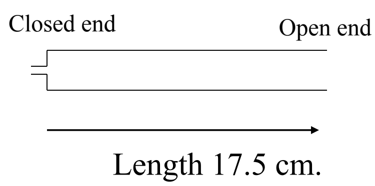
- Complex wave repeating nine times
  - Plus smaller wave that repeats 4x for every large cycle
  - Large wave: freq of 250 Hz (9 times in .036 seconds)
  - Small wave roughly 4 times this, or roughly 1000 Hz

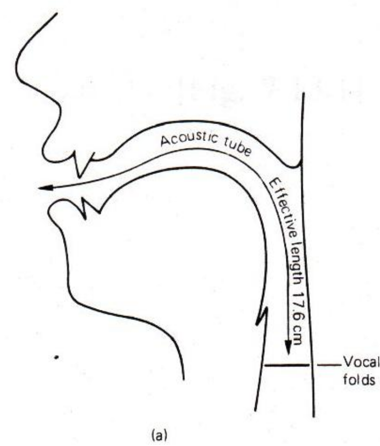


[demo]

Resonances of the vocal tract

- The human vocal tract as an open tube

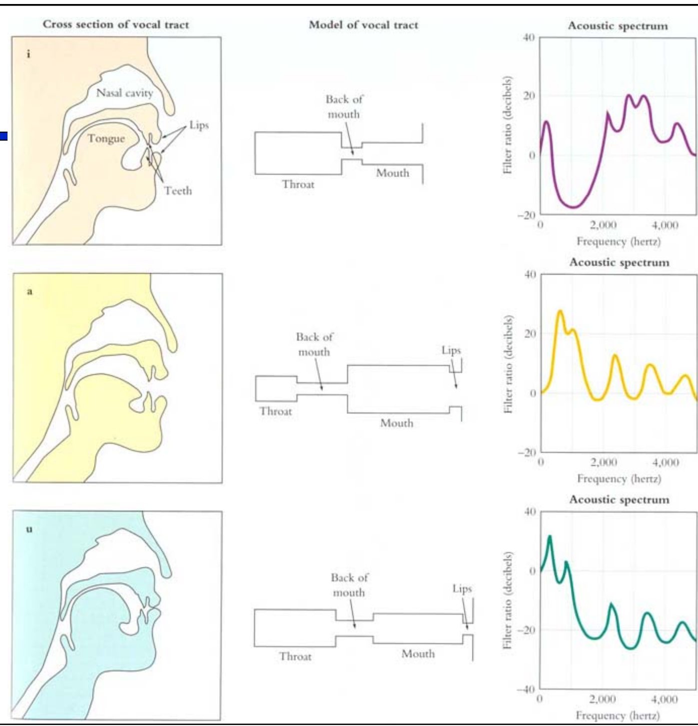




- Air in a tube of a given length will tend to vibrate at resonance frequency of tube.
- Constraint: Pressure differential should be maximal at (closed) glottal end and minimal at (open) lip end.

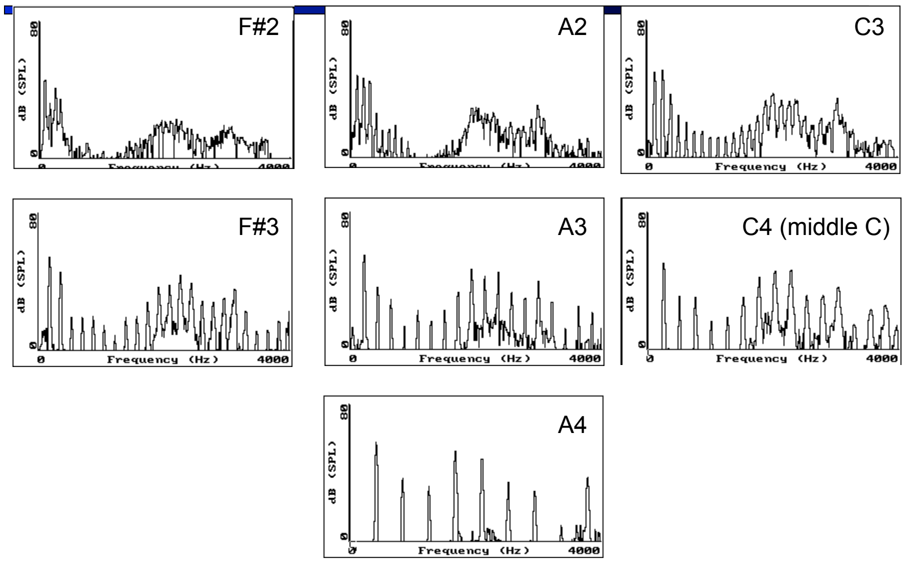
Figure from W. Barry Speech Science slides

[demo]



From Mark Liberman's website

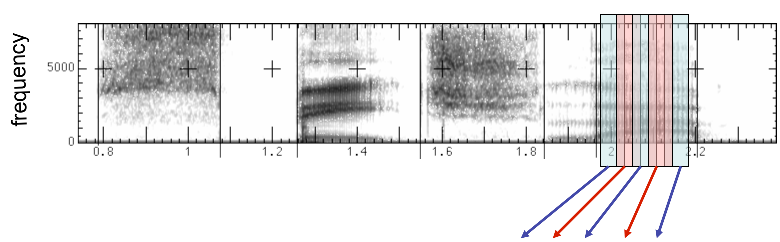
Vowel [i] sung at successively higher pitches



Figures from Ratree Wayland

Acoustic Feature Sequence

- Time slices are translated into acoustic feature vectors (~39 real numbers per slice)



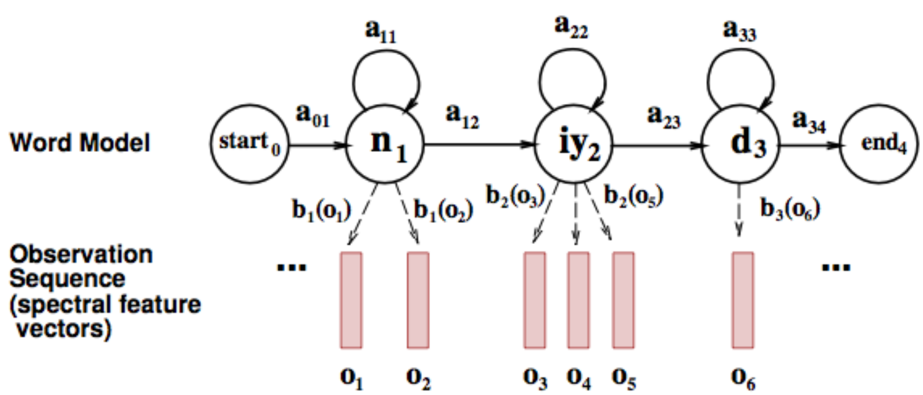
…………………………………………….$e_{12}e_{13}e_{14}e_{15}e_{16}$………..

- 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

HMMs for Speech



Transitions with Bigrams

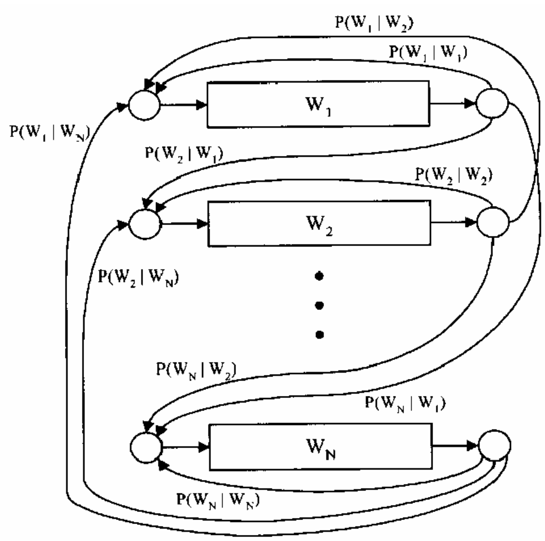


Figure from Huang et al page 618

Training Counts

198015222 the first
194623024 the same
168504105 the following
158562063 the world
…
14112454 the door

23135851162 the *

$$\hat{P}(\text{door}|\text{the}) = \frac{14112454}{23135851162}$$

$$= 0.0006$$

Decoding

- 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 $e_{1:T}$:

$$x^*_{1:T} = \arg\max_{x_{1:T}} P(x_{1:T}|e_{1:T})$$
$$= \arg\max_{x_{1:T}} P(x_{1:T}, e_{1:T})$$

- From the sequence x, we can simply read off the words

14