

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

# Speech

Many slides courtesy of Dan Klein, Stuart Russell, or Andrew Moore

CS 5300 / CS 6300  
Artificial Intelligence  
Spring 2009

Hal Daumé III  
hal@cs.utah.edu

www.cs.utah.edu/~hal/courses/2009S\_AI

Slide 1 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

# Announcements

- Homeworks and project 5 pushed back a bit
- Project 3 grades out
  - Sorry for the delay, y'all broke the autograder :)
  - Some still have zeros due to a.g. issues
    - Don't freak out: talk to Scott!
- Initial extra credit assignments made
  - [www.cs.utah.edu/~hal/courses/2009S\\_AI/ec.html](http://www.cs.utah.edu/~hal/courses/2009S_AI/ec.html)
- Don't forget to register your team for the contest!

Slide 2 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

# Hidden Markov Models

- An HMM is
- Initial distribution: [redacted]
- Transitions: [redacted]
- Emissions: [redacted]
- Query: most likely seq: [redacted]

Slide 3 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

# State Path Trellis

- State trellis: graph of states and transitions over time

- Each arc represents some transition
- Each arc has weight [redacted]
- Each path is a sequence of states
- The product of weights on a path is the seq's probability
- Can think of the Forward (and now Viterbi) algorithms as computing sums of all paths (best paths) in this graph

Slide 4 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

# Digitizing Speech

Slide 5 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

# Speech in an Hour

- Speech input is an acoustic wave form

"I" to "a" transition:

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

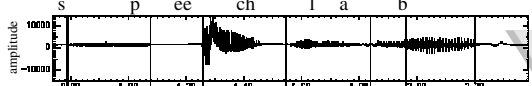
Slide 6 CS 5300: Speech

UNIVERSITY OF UTAH

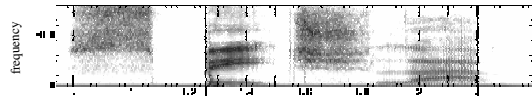
Hal Daumé III (hal@cs.utah.edu)

## 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

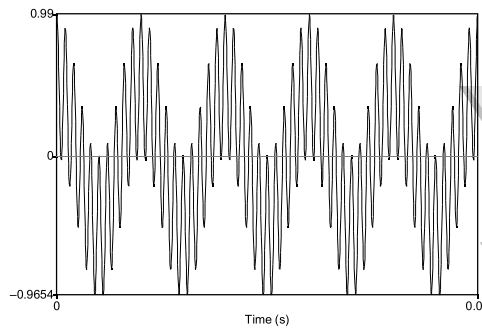


Slide 7 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

## Adding 100 Hz + 1000 Hz Waves



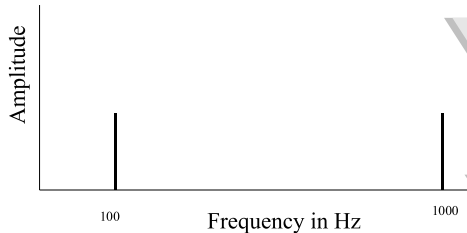
Slide 8 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

## Spectrum

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




Slide 9 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

## Part of [ae] from "lab"



- 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


Slide 10 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

## 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.

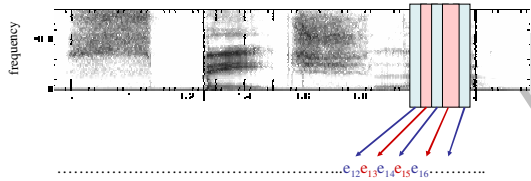
Slide 11 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

## Acoustic Feature Sequence

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



- These are the observations, now we need the hidden states X

Slide 14 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

## 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  $X$

Slide 15 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

## HMMs for Speech

Word Model

Observation Sequence (spectral feature vectors)

Slide 16 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

## Markov Process with Bigrams

Figure from Huang et al page 618

Slide 17 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

## 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  $e_{1:T}$ :

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

Slide 18 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

## Training (aka "preview of ML")

- > Two key components of a speech HMM:
  - > Acoustic model:  $p(E | X)$
  - > Language model:  $p(X | X')$
- > Where do these come from?
- > Can we estimate these models from data:
  - >  $p(E | X)$  might be estimated from transcribed speech
  - >  $p(X | X')$  might be estimated from large amounts of raw text

Slide 19 CS 5300: Speech

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

## n-gram Language Models

- > Assign a probability to a sequences of words

$$p(w_1, w_2, \dots, w_T) = \prod_{i=1}^T p(w_i | w_1, \dots, w_{i-1})$$

$$\approx \prod_{i=1}^T p(w_i | w_{i-k}, \dots, w_{i-1})$$

- > If I gave you a copy of the web, how would you estimate these probabilities?

Need to "smooth" estimates intelligently to avoid zero probability  $n$ -grams.

Language modeling is the art of good smoothing.

See [Goodman 1998], [Teh 2007]

Slide 20 CS 5300: Speech

UNIVERSITY OF UTAH Hal Daumé III (hal@cs.utah.edu)

## Acoustic models

- What if I gave you data like:

.....sp ee ch l ac b.....

- How would you estimate  $p(E|X)$ ?
- What's wrong with this approach?

Slide 21 CS 5300: Speech

UNIVERSITY OF UTAH Hal Daumé III (hal@cs.utah.edu)

## Acoustic models II

- What does our data really look like:

Acc:   
 W: yesterday I went to visit the speech lab

- We'd like to know *alignments* between transcript and waveform
- Suppose someone gave us a good speech recognizer.... could we figure out alignments from that?

Slide 22 CS 5300: Speech

UNIVERSITY OF UTAH Hal Daumé III (hal@cs.utah.edu)

## Expectation Maximization

- A general framework to do parameter estimation in the presence of hidden variables
- Repeat ad infinitum:
  - E-step: make probabilistic guesses at latent variables
  - M-step: fit parameters according to these guesses

W: I LIKE A I

Slide 23 CS 5300: Speech

UNIVERSITY OF UTAH Hal Daumé III (hal@cs.utah.edu)

## Expectation Maximization

e	$p(e   \text{"I"})$	$p(e   \text{"LIKE"})$	$p(e   \text{"A"})$
	0.33 → 2	0.33 → 1	0.33 → 1
	0.33 → 1	0.33 → 1	0.33 → 1
	0.33 → 1	0.33 → 1	0.33 → 1

W: I LIKE A I

Slide 24 CS 5300: Speech

UNIVERSITY OF UTAH Hal Daumé III (hal@cs.utah.edu)

## Expectation Maximization

e	$p(e   \text{"I"})$	$p(e   \text{"LIKE"})$	$p(e   \text{"A"})$
	0.5 → 4	0.33 → 1	0.33 → 1
	0.25 → 1	0.33 → 2	0.33 → 2
	0.25 → 1	0.33 → 2	0.33 → 2

W: I LIKE A I

Slide 25 CS 5300: Speech

UNIVERSITY OF UTAH Hal Daumé III (hal@cs.utah.edu)

## State of the Art DBNs for Speech

Slide 26 CS 5300: Speech

## Summary

- HMMs allow us to “separate” two models:
  - acoustic model (how does what I want to say sound?)
  - language model (what do I want to say)
- Speech recognition is “just” decoding in an HMM/DBN
  - Plus a heck of a lot of engineering
- Expectation maximization lets us estimate parameters in models with hidden variables
- Most research today focuses on language modeling

Shameless plug:  
take CS5350 and  
CS5964