CS 188: Artificial Intelligence Spring 2010

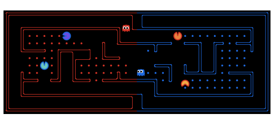
Lecture 21: DBNs, Viterbi, Speech Recognition

4/8/2010

Pieter Abbeel – UC Berkeley

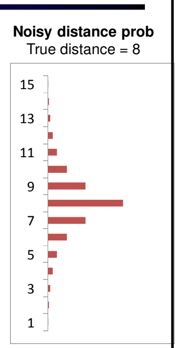
Announcements

- Written 6 due on tonight
- Project 4 up!
  - Due 4/15 – start early!
- Course contest update
  - Planning to post by Friday night



P4: Ghostbusters 2.0

- **Plot:** Pacman's grandfather, Grandpac, learned to hunt ghosts for sport.
- He was blinded by his power, but could hear the ghosts' banging and clanging.
- **Transition Model:** All ghosts move randomly, but are sometimes biased
- **Emission Model:** Pacman knows a "noisy" distance to each ghost

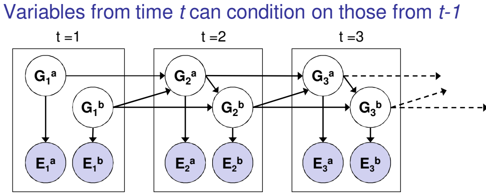


Today

- Dynamic Bayes Nets (DBNs)
  - [sometimes called temporal Bayes nets]
- HMMs: Most likely explanation queries
- Speech recognition
  - A massive HMM!
  - Details of this section not required
- Start machine learning

Dynamic Bayes Nets (DBNs)

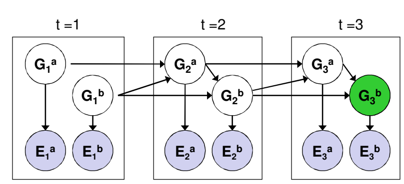
- We want to track multiple variables over time, using multiple sources of evidence
- Idea: Repeat a fixed Bayes net structure at each time
- Variables from time *t* can condition on those from *t-1*



- Discrete valued dynamic Bayes nets are also HMMs

Exact Inference in DBNs

- Variable elimination applies to dynamic Bayes nets
- Procedure: "unroll" the network for T time steps, then eliminate variables until $P(X_T|e_{1:T})$ is computed



- Online belief updates: Eliminate all variables from the previous time step; store factors for current time only

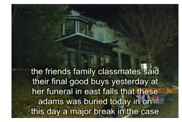
DBN Particle Filters

- A particle is a complete sample for a time step
- **Initialize**: Generate prior samples for the t=1 Bayes net
  - Example particle: $\mathbf{G_1^a}$ = (3,3) $\mathbf{G_1^b}$ = (5,3)
- **Elapse time**: Sample a successor for each particle
  - Example successor: $\mathbf{G_2^a}$ = (2,3) $\mathbf{G_2^b}$ = (6,3)
- **Observe**: Weight each entire sample by the likelihood of the evidence conditioned on the sample
  - Likelihood: $P(\mathbf{E_1^a}|\mathbf{G_1^a}) * P(\mathbf{E_1^b}|\mathbf{G_1^b})$
- **Resample:** Select prior samples (tuples of values) in proportion to their likelihood

8

Speech and Language

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



- Language processing technologies
  - Machine translation

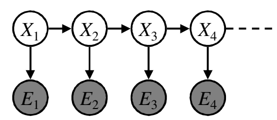
  - Information extraction
  - Web search, question answering
  - Text classification, spam filtering, etc…

SLAM

- SLAM = Simultaneous Localization And Mapping
  - We do not know the map or our location
  - Our belief state is over maps and positions!
  - Main techniques: Kalman filtering (Gaussian HMMs) and particle methods
- [DEMOS]
  - [intel-lab-raw-odo.wmv, intel-lab-scan-matching.wmv, visionSlam_heliOffice.wmv]

HMMs: MLE Queries

- HMMs defined by
  - States X
  - Observations E
  - Initial distr: $P(X_1)$
  - Transitions: $P(X|X_{-1})$
  - Emissions: $P(E|X)$



- Query: most likely explanation:

$$\arg\max_{x_{1:t}} P(x_{1:t}|e_{1:t})$$

13

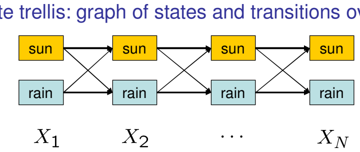
Today

- Dynamic Bayes Nets (DBNs)
  - [sometimes called temporal Bayes nets]
- *HMMs: Most likely explanation queries*
- Speech recognition
  - A massive HMM!
  - Details of this section not required
- Start machine learning

11

State Path Trellis

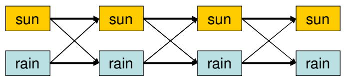
- State trellis: graph of states and transitions over time



- Each arc represents some transition $x_{t-1} \rightarrow x_t$
- Each arc has weight $P(x_t|x_{t-1})P(e_t|x_t)$
- 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

14

Viterbi Algorithm



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

$$m_t[x_t] = \max_{x_{1:t-1}} P(x_{1:t-1}, x_t, e_{1:t})$$

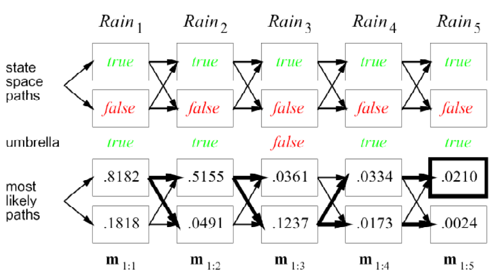
$$= \max_{x_{1:t-1}} P(x_{1:t-1}, e_{1:t-1}) P(x_t|x_{t-1}) P(e_t|x_t)$$

$$= P(e_t|x_t) \max_{x_{t-1}} P(x_t|x_{t-1}) \max_{x_{1:t-2}} P(x_{1:t-1}, e_{1:t-1})$$

$$= P(e_t|x_t) \max_{x_{t-1}} P(x_t|x_{t-1}) m_{t-1}[x_{t-1}]$$

15

Example



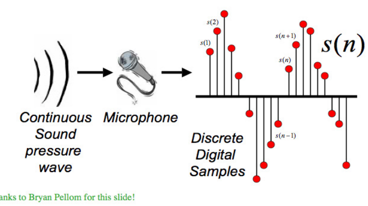
16

Today

- Dynamic Bayes Nets (DBNs)
  - [sometimes called temporal Bayes nets]
- HMMs: Most likely explanation queries
- *Speech recognition*
  - A massive HMM!
  - Details of this section not required
- Start machine learning

17

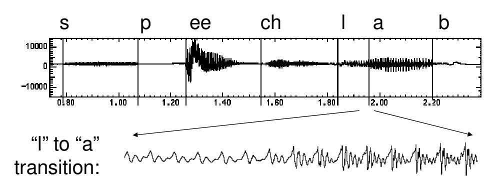
Digitizing Speech



18

Speech in an Hour

- Speech input is an acoustic wave form

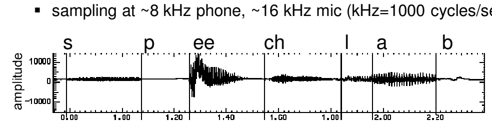


Graphs from Simon Arnfield's web tutorial on speech, Sheffield:
http://www.psyc.leeds.ac.uk/research/cogn/speech/tutorial/

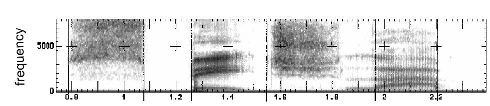
19

Spectral Analysis

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

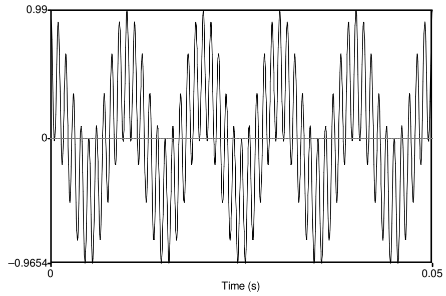


- Fourier transform of wave displayed as a spectrogram
  - darkness indicates energy at each frequency



20

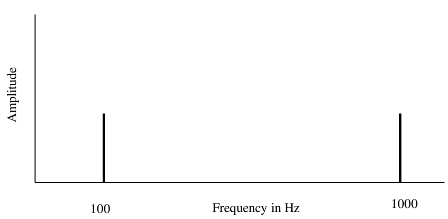
Adding 100 Hz + 1000 Hz Waves



21

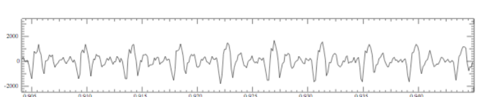
Spectrum

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



22

Part of [ae] from “lab”

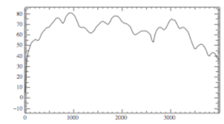


- Note complex wave repeating nine times in figure
- Plus smaller waves which repeats 4 times for every large pattern
- Large wave has frequency of 250 Hz (9 times in .036 seconds)
- Small wave roughly 4 times this, or roughly 1000 Hz
- Two little tiny waves on top of peak of 1000 Hz waves

23

Back to Spectra

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

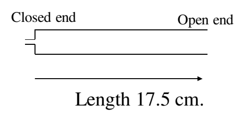


- x-axis shows frequency, y-axis shows magnitude (in decibels, a log measure of amplitude)
- Peaks at 930 Hz, 1860 Hz, and 3020 Hz.

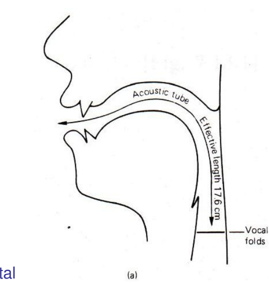
25

Resonances of the vocal tract

- The human vocal tract as an open tube

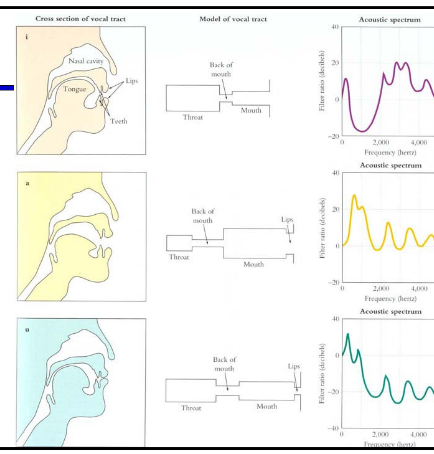


- Air in a tube of a given length will tend to vibrate at resonance frequency of tube.
- Constraint: Pressure differential should be maximal at (closed) glottal end and minimal at (open) lip end.



26

Figure from W. Barry Speech Science slides

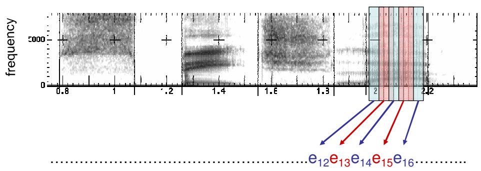


From Mark Liberman’s website

28

Acoustic Feature Sequence

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

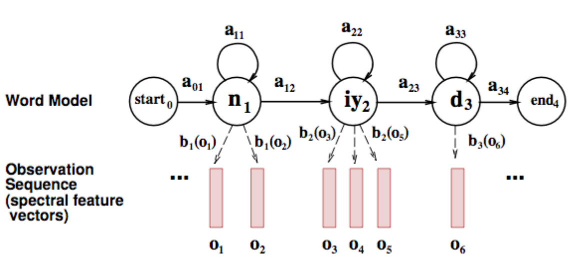


- These are the observations, now we need the hidden states X

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



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

End of Part II!

- Now we're done with our unit on probabilistic reasoning
- Last part of class: machine learning

Parameter Estimation

- Estimating the distribution of a random variable
- *Elicitation:* ask a human!
  - Usually need domain experts, and sophisticated ways of eliciting probabilities (e.g. betting games)
  - Trouble calibrating
- *Empirically:* use training data
  - For each outcome x, look at the *empirical rate* of that value:

$$P_{\text{ML}}(x) = \frac{\text{count}(x)}{\text{total samples}}$$

r g g

$$P_{\text{ML}}(\text{r}) = 1/3$$

  - This is the estimate that maximizes the *likelihood of the data*

$$L(x, \theta) = \prod_i P_\theta(x_i)$$