



Speech Processing 15-492/18-492

Speech Recognition

Intro

Acoustic modelling

HMMs

Speech Recognition

- ◆ *From acoustics to text*
- ◆ *Acoustic modeling*
 - *Recognizing all forms of all phonemes*
- ◆ *Language modeling*
 - *Expectation of what might be said*
- ◆ *We need both to do recognition*

Acoustics are not enough

- ◆ *Last Saturday in Hawaii, numerous Waipouli vacationers were shocked to find their beach cordoned off for a UC Berkeley Drama enactment of "Personal office space". The play features exclusively topless men and women in an everyday office environment. Richard Carlson, one of the annoyed tourists and a regular swimmer at Waipouli beach, complained that they really knew how to wreck a nice beach with the nudist play. Many of the tourists appeared ruffled by the content and fled the scene to avoid compromising photos.*
- ◆ *In yesterday's press release, AT&T unveiled SpeechKit, its new speech recognition toolkit. According to Michael Armstrong, the COO of the company, the most innovative feature of the system is its revolutionary three-dimensional interface, which opens a new universe of possibilities for the speech recognition community. During the official software release, Jonathan Blues, a senior researcher at AT&T Labs, explained how to recognize speech with the new display, and how the toolkit has already played a crucial role in his research.*

Acoustics are not enough

- ◆ *Last Saturday in Hawaii, numerous Waipouli vacationers were shocked to find their beach cordoned off for a UC Berkeley Drama enactment of "Personal office space". The play features exclusively topless men and women in an everyday office environment. Richard Carlson, one of the annoyed tourists and a regular swimmer at Waipouli beach, complained that they really knew **how to wreck a nice beach with this nudist play**. Many of the tourists appeared ruffled by the content and fled the scene to avoid compromising photos.*
- ◆ *In yesterday's press release, AT&T unveiled SpeechKit, its new speech recognition toolkit. According to Michael Armstrong, the COO of the company, the most innovative feature of the system is its revolutionary three-dimensional interface, which opens a new universe of possibilities for the speech recognition community. During the official software release, Jonathan Blues, a senior researcher at AT&T Labs, explained **how to recognize speech with this new display**, and how the toolkit has already played a crucial role in his research.*

Split the task

- ◆ *Build Acoustic models*
 - *Probability of phones given acoustics*
- ◆ *Build Language models*
 - *Probability of word string*

Acoustic models

- ◆ *Represent all ways to say each phoneme*
 - *Like “templates” for each phoneme*
 - *Averages over multiple examples*
 - *Different phonetic contexts*
 - ⊗ *“sow” vs “see” etc*
 - *Different people speaking*
 - *Different acoustic environment*
 - *Different channels*
 - ⊗ *(assume channel is similar)*

Better Acoustic Models

◆ *DTW Template*

- *Could be averages over multiple examples*
- *Need to be time normalized*
 - ⊗ *Linear interpolate or try to match*
- *Matching probabilistically*
 - ⊗ *What is the probability that example matches*
 - ⊗ *Test each frame*

Hidden Markov Models

- Markov Process
 - Future can be predicted from the past

$$P(X_{t+1} | X_t, X_{t-1}, \dots, X_{t-m})$$

- Hidden Markov Models:
 - When the state is unknown
 - A probability is given for each states

Hidden Markov Model

Set of states

$$S = \{s_1, \dots, s_N\}$$

Output alphabet

$$K = \{k_1, \dots, k_M\}$$

Initial state probabilities

$$\Pi = \{\pi_i\}, i \in S$$

State transition probabilities

$$A = \{a_{ij}\}, i, j \in S$$

State emission probabilities

$$B = \{b_{ijk}\}, i, j \in S, k \in K$$

A model $\mu = (A, B, \Pi)$

Key Requirements

1. Given a model $\mu = (A, B, \Pi)$, how do we efficiently compute how likely an observation is, $P(O | \mu)$.
 - which model is most probable
2. Given observation O and model μ , which state sequence best explains the observations
 - in a model what states are most likely
3. Given O and a space of models, how do we find the best model to explain O
 - how do we training the thing

Find Probability of Observation

- ◆ *Given observation O and model M*
 - *Efficiently find $P(O|M)$*
 - *Called **decoding***
- ◆ *Find sum of all paths probabilities*
- ◆ *Each path prob is product of each transition in state sequence*
- ◆ *Use dynamic programming (generalized DTW)*
 - *Also used in Chart Parsers, Theorem Provers*

Finding the Best Path

- ◆ *What is the most probable state sequence*
- ◆ *Use **Viterbi** algorithm*
 - *Maximize best sequence*
 - *At each point hold list possible states*
 - *Hold back-pointer to best previous state*
 - *Cumulate values along path*
- ◆ *Because we are looking for **BEST***
 - *Can ignore other back-pointers*
- ◆ *(When looking for N-best need more complex structure)*

Parameter Estimation

- ◆ *Called **training***
- ◆ *Use **Maximum Likelihood Estimation***
 - *Baum-Welch (forward/backward algorithm)*
- ◆ *Special case of **EM (Expectation Maximization)***
 - *Run observation and find current probs (forward)*
 - *Modify probabilities to make observations best path (backward)*
 - *Repeat until convergences*
- ◆ *Not globally optimal*
 - *May find local maximum*

HMM recognition

- ◆ *A bunch of HMM*
 - *One for each phone type*
- ◆ *Each observation (e.g. 10ms frame)*
 - *Probability distribution of possible phone type*
- ◆ *Thus can find most probably sequence*
 - *Use Viterbi to find best path*

But that's not enough

- But not all phones are equi-probable
- Find word sequences that maximizes

$$P(W | O)$$

- Using Bayes' Law

$$\frac{P(W)P(O|W)}{P(O)}$$

- Combine models

- Use HMMs to provide

$$P(O | W)$$

- Use language model to provide

$$P(W)$$

How many HMM models

◆ *How many models*

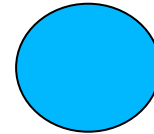
- *One for each thing you want to recognize:*

- ⊗ *One per phone*
- ⊗ *One per word*
- ⊗ *One per city name ...*

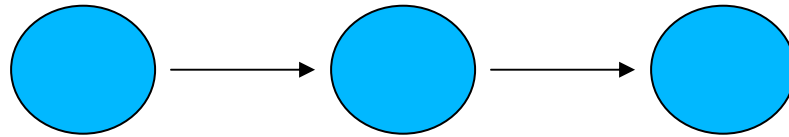
◆ *What is the size and shape of the model*

HMM Topology

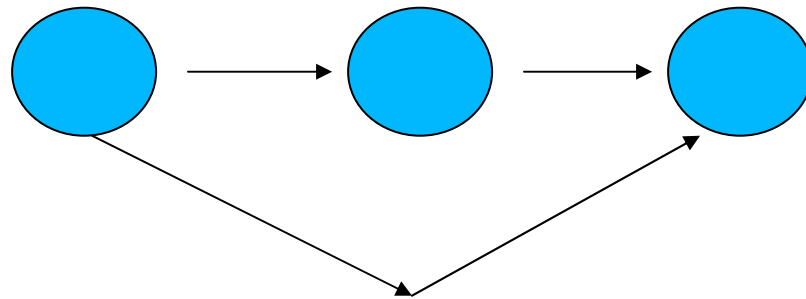
1 state



3 state



3 state with skips



How many models

- ◆ *Context Independent models:*
 - *One for each phoneme*
 - *One for silence, noises*
- ◆ *Triphone models*
 - *Context dependent*
 - *Phone before and after*
 - *Need lots of data to train this*
- ◆ *Tied states (semi-continuous)*
 - *Build full triphone models*
 - *Combine low frequency “similar” phones*
 - *Train again on smaller set*

But even that's not enough

◆ *HMM for words*

- *For common words or common in domain*
- *E.g. City, State (need more than 3 states)*

Search space is very large

- ◆ *Prune Viterbi search*
 - *Best number of paths*
 - *Some percentage of probability mass*
- ◆ *Prune lexical trees*
 - *Restrict vocabulary*
 - *Use language model*
 - *Or even grammar*

Some computational issues

- ◆ *Probabilities are multiplied along paths*
 - *They get **very** small*
- ◆ *Treat probabilities as logs*
 - *Thus add rather than multiple*
 - *Typically use negative log probabilities*

Training

- ◆ *How much data do you need*
 - *As much as you can get*
 - *More than 10Hrs (100Hrs, 1000Hrs)*
 - *Can take months to train*
- ◆ *The larger the models*
 - *The larger the number of parameters*
 - *More data needs to be used for training*
 - *Examples are equi-probably (find oy-oy examples is hard)*

The right type of data

- ◆ *Training data must match intended domain*
 - *Male/Female, Native/non-native, UK/US*
 - *As close to target domain as possible*
 - *Right channel (cell phone/land line)*

How to improve ASR

- ◆ *Get more data*
- ◆ *Fix bugs*

Summary

◆ *HMMs*

- *Find probability of observation (decoding)*
- *Find best path (Viterbi)*
- *Train the parameters (Baum-Welch)*

◆ *Bayes Law*

- *Acoustic model and Language model*

Reading

- ◆ *Section 8.2 Definition of Hidden Markov Model pp 380-393*
- ◆ *Section 8.4 Practical Issues in using HMMS pp 398-405*
- ◆ *In Huang et al.*
- ◆ *Two page description of the contents emailed to nbach@cs.cmu.edu before 3:30pm Monday (5th) or Wednesday (7th Oct)*

