HMM Theory and Practice

Multimodal Interaction Lab

EE4R Automatic Spoken Language Processing

UNIVERSITY OF BIRMINGHAM

Objectives

- To understand how to apply HMMs
- To understand the basic HMM algorithms
- *Notes: pp 46 - 55*

Multimodal Interaction Lab

EE4R Automatic Spoken Language Processing

UNIVERSITY OF BIRMINGHAM

Training & Recognition

- Major advantage of HMMs is the availability of a ‘toolkit’ of powerful, well-founded mathematical methods for HMM manipulation
- The **Baum-Welch** algorithm is used to train the parameters of a set of HMMs given a set of training data
- **Viterbi Decoding** is used to classify an unknown speech pattern in terms of the sequence of HMMs which is most likely to have produced it

Multimodal Interaction Lab

EE4R Automatic Spoken Language Processing

UNIVERSITY OF BIRMINGHAM

The Recognition Problem

- Given a sequence of acoustic feature vectors
  $Y = \{y_1,...,y_T\}$
  we want to find the sequence of words
  $W = \{w_1,...,w_L\}$
  such that the probability
  $P(W \mid Y)$
  is maximized.
- If $M = \{M_1,...,M_K\}$ is the sequence of HMMs which represents W, then $P(W \mid Y) = P(M \mid Y)$

Multimodal Interaction Lab

b ɜ: m ɪ ŋ ə m

UNIVERSITY OF BIRMINGHAM

Bayes' Theorem

- Computation of the probability $P(M \mid Y)$ is made possible using **Bayes' Theorem**

$$P(W \mid Y) = \frac{p(Y \mid W)P(W)}{p(Y)}$$

- *P(W)* is the "language model probability"
- $p(Y \mid W)$ is the "acoustic model probability"
- Bayes Theorem has been referred to as the "fundamental theorem of speech recognition"!

Multimodal Interaction Lab

b ɜ: m ɪ ŋ ə m

UNIVERSITY OF BIRMINGHAM

EE4R Automatic Spoken Language Processing

The Baum-Welch Algorithm

- The Baum-Welch algorithm is the method which is normally used for HMM parameter estimation
- Given a set of HMMs M_0 and a set of speech patterns Y, Baum's theorem defines how to produce a new model set M_1 such that

$$P(Y \mid M_1) \geq P(Y \mid M_0)$$

- Baum-Welch algorithm applies this method repeatedly until a HMM M_n is found which (locally) maximizes $P(Y \mid M)$
- Baum's theorem only valid for particular classes of state output PDF

Multimodal Interaction Lab

b ɜ: m ɪ ŋ ə m

UNIVERSITY OF BIRMINGHAM

EE4R Automatic Spoken Language Processing

Notes on B-W Reestimation

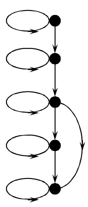
- The Baum-Welch algorithm is only guaranteed to find a **locally** optimal HMM set - hence choice of M_0 can be important
- Baum-Welch is a **supervised** training algorithm which requires labelled speech data
- The labelling need **not** be at the same level as the HMM set - phoneme level HMMs can be trained using data labelled orthographically at the phrase or sentence level
- For large applications B-W reestimation can be **very** computationally expensive

Viterbi Decoding

- Viterbi Decoding is the algorithm which is used to find the sequence of HMMs which is most likely to have generated a given speech pattern
- Based on **Dynamic Programming**
- Viterbi Decoding illustrates the type of computation typically done with HMMs

Viterbi Decoding (1)

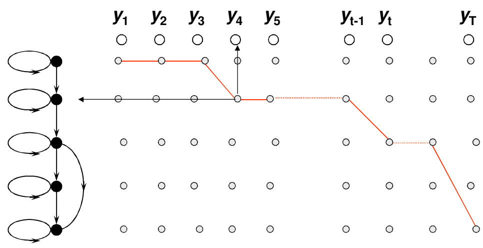
$y_1 \quad y_2 \quad y_3 \quad y_4 \quad y_5 \quad\quad y_{t-1} \quad y_t \quad\quad y_T$



Q: How can *M* have generated *Y*?

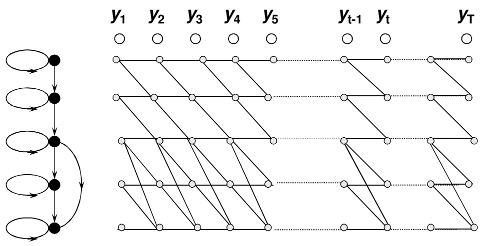
A: Via a state sequence of length T

Function of State Sequence



Viterbi Decoding (2)

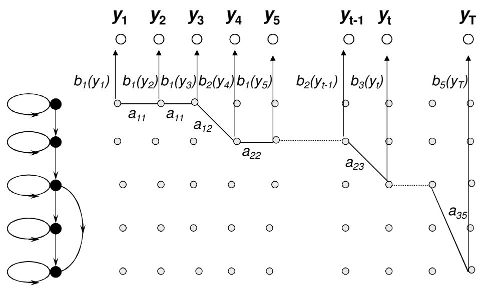
- **Construction of 'state-time trellis'**



Constructing the State-Time Trellis

- Simple Rule:
  - Connect node (i,t) of the trellis to node $(j,t+1)$ if and only if there is a transition between state i and state j in the HMM with probability a_{ij} greater than zero

Basic Probability Calculation



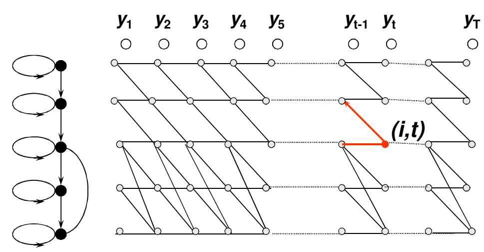
Viterbi Decoding (3)

- Let $X = \{x_1,...,x_T\}$ be a state sequence of length T
- The joint probability of Y and X is given by:

$$p(Y,X) = b_{x_1}(y_1)\prod_{t=2}^{T} a_{x_{t-1}x_t} b_{x_t}(y_t)$$

- i.e. the product of the state-output and state transition probabilities along the state sequence
- The optimal state sequence is the sequence X such that $p(Y,X)$ is maximized
- $p(Y)$ is the sum of $P(Y,X)$ over all sequences X

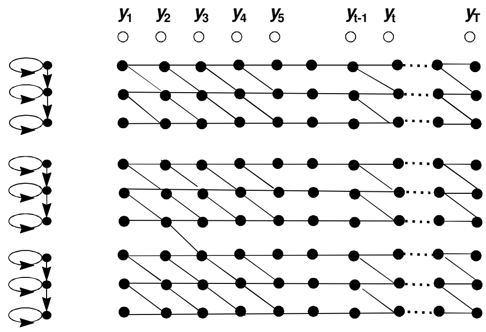
Viterbi Decoding (4)



$p_t(i)$ = Prob($y_1,...,y_t$, opt sequence to (i,t))

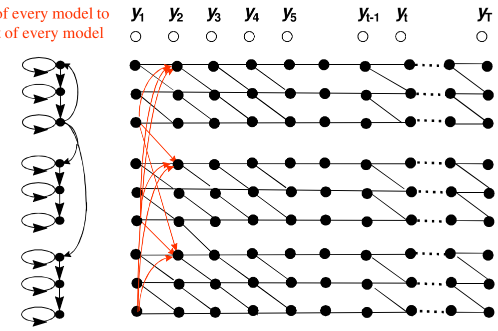
$p_t(i) = \max\{p_{t-1}(i-1)a_{i-1,i} ,\ p_{t-1}(i)a_{i,i}\}b_i(y_t)$

Isolated Speech Recognition

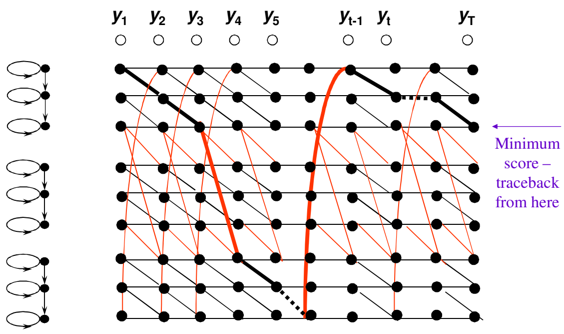


Connected Speech Recognition

New transitions connect end of every model to start of every model



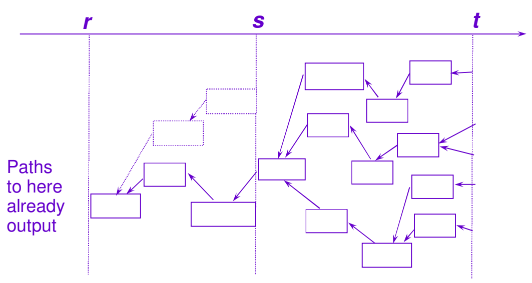
Connected Speech Recognition



Partial Traceback (1)

- In continuous speech recognition, cannot trace-back from the end of the utterance (there is no end!)
- Instead **partial traceback** operates as follows:
  - For each time t and state i a *word link record* describes the sequence of words on the best path to (t,i).
  - At regular intervals all active paths are traced back to see if they converge at some time s in the past
  - If so, the best path up to time s cannot change, and the sequence of words up to s can be output

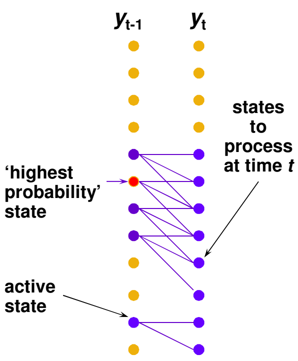
Partial Traceback (2)



Multimodal Interaction Lab
EE4R Automatic Spoken Language Processing
UNIVERSITY OF BIRMINGHAM

Beam Pruning

- Choose threshold T
- The active states at time $t-1$ are those whose partial path scores are within T of the best path score at time $t-1$
- At time t, only process those states which link back to an active state



Multimodal Interaction Lab
EE4R Automatic Spoken Language Processing
UNIVERSITY OF BIRMINGHAM

Partial Traceback, Beam Pruning & Recognition 'Speed'

- Partial traceback introduces a 'lag' into recognition process - **not** due to inadequate processor speed
- Lag worse when models are poor
- Beam Pruning less effective for ambiguous input
- Severe Beam Pruning will degrade performance
- Proper management of Partial Traceback and Beam Pruning is essential for optimal performance

Summary

- HMM Theory and Practice
- HMM Training
  - The Baum-Welch algorithm
- Recognition
  - Viterbi decoding
  - Beam pruning
  - Partial traceback