Data Mining and Machine Learning

HMMs for Automatic Speech Recognition

Peter Jančovič

UNIVERSITY OF BIRMINGHAM

Objectives

To understand

- Application of HMMs for automatic speech recognition
- HMM assumptions

Pattern Recognition

- Suppose we have a finite number of classes, $w_1, \ldots, w_C$ and the goal is to decide which class has given rise to the measurement x

- The probability of the class w **given that the measurement x has been observed** is called **posterior probability of the class w** – denoted by $P(w|x)$

Bayes' Theorem

- The form of **Bayes' Theorem** which we need for pattern recognition is:

Class-conditional density

Prior probability

$$P(w \mid x) = \frac{p(x \mid w)P(w)}{p(x)}$$

Posterior probability

Automatic Speech Recognition

- Given a sequence of acoustic feature vectors

  $Y = \{y_1,\ldots,y_T\}$

  we want to find the sequence of words

  $W = \{w_1,\ldots,w_L\}$

  such that the probability

  $P(W|Y)$

  is maximized.
- If $M = \{M_1,\ldots,M_K\}$ is the sequence of HMMs which represents W, then $P(W|Y) = P(M|Y)$

Bayes' Theorem

- Computation of the probability *P(M | 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 | W)* is the "acoustic model probability"

Mathematical Modelling

- Mathematical modelling for speech recognition
- Two conflicting requirements:
  - Faithful model of human speech production/perception
  - Mathematically tractable & Computationally Useful
- HMMs are one of the best compromise at the moment

‘Divide and Conquer’

- One possible approach to ASR is sequential ‘divide and conquer’, e.g.
  - classify speech vectors as ‘acoustic features’
  - classify sequences of acoustic features as phonemes
  - classify sequences of phonemes as words
  - classify sequences of words ...

DISASTER!!

Delayed Decision Making

- Another name for this might be non-recoverable error propagation!

- Better to avoid all classification decisions until all sources of information are available. Then perform classification as a single, integrated process - delayed decision making

- Delayed Decision Making underlies HMM success

The ‘HMM Compromise’

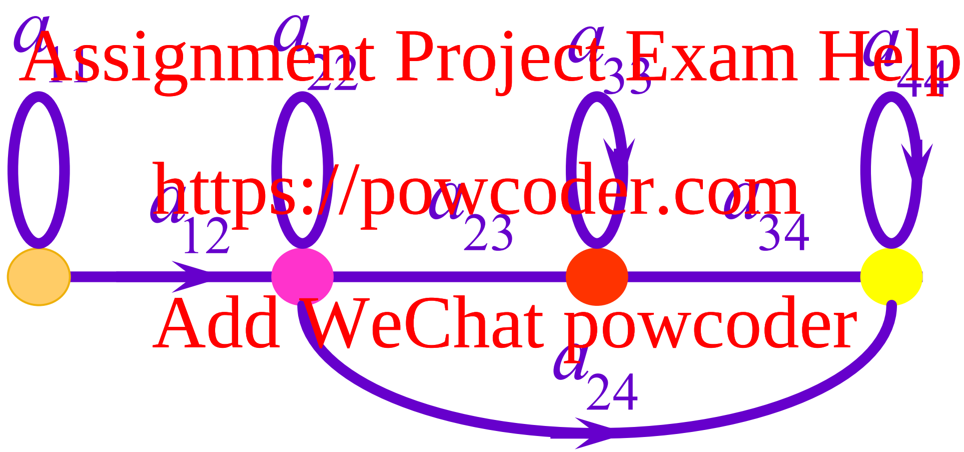
Assume that :

- A spoken utterance is a time-varying sequence which moves through a sequence of ‘segments’ – (yes)
- Underlying structure of the segments is constant w.r.t time – (no!)
- Durations of segments vary – (yes)
- All variations between different realizations of the segments are random – (no!)

UNIVERSITY OF BIRMINGHAM

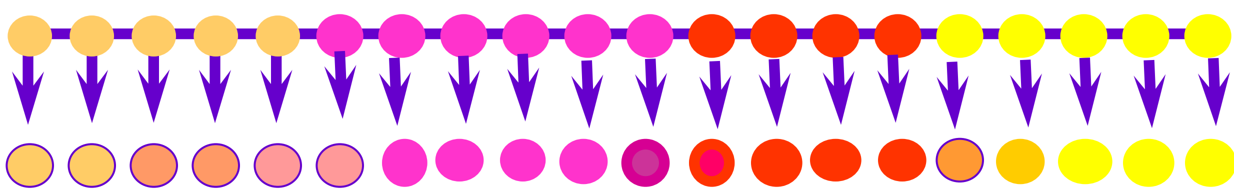
Hidden Markov Model

- In a **hidden** Markov model, the relationship between the observation sequence and the state sequence is ambiguous.



$X=\{x_t\}$

$Y=\{y_t\}$



UNIVERSITY OF BIRMINGHAM

Hidden Markov Models

A HMM consists of

- A set of states $S = \{s_1,\ldots,s_N\}$
- A state transition probability matrix $A = [a_{ij}]_{i,j=1,\ldots N}$,

  where $a_{ij} = \mathrm{Prob}(s_j \text{ at time } t \mid s_i \text{ at time } t\text{-}1)$
- For each state s_i, a PDF b_i defined on the set of possible observations O s.t.

  $b_i(o) = \mathrm{Prob}(y_t=o \mid x_t=s_i)$
- b_i is called the **state output PDF** for state i (or the i^{th} **state output PDF**)

10 state HMM of the digit ‘zero’

6 state HMM of the digit ‘zero’

HMM Assumptions

- **Temporal Independence** - the observation y_t depends on the state x_t but is otherwise independent of the rest of the observation sequence $Y = \{y_t\}$!

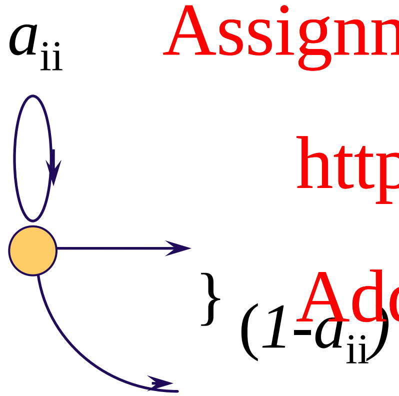
... so, the position of the vocal tract at time *t* is independent of its position at time *t-1*!

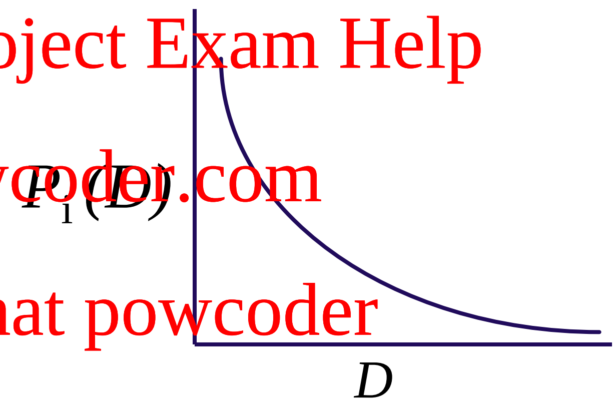
- **Piecewise stationarity** - the underlying structure of speech is a sequence of stationary segments

- **Random variability** - variations from this underlying structure are random

HMM State Duration Model

- Constant segments correspond to the HMM states





- Probability of state duration D is given by

$$P_i\,(D) = (1 - a_{ii})a_{ii}^{(D\,-\,1)}$$

Summary

- Introduction to application of HMMs for speech recognition – HMM assumptions