Data Mining and Machine Learning

HMMs for Automatic Speech Recognition: Word and Sub-Word Level HMMs

Peter Jančovič

UNIVERSITY OF BIRMINGHAM

Content

- Word level HMMs
- Sub-word HMMs
  - Phoneme-level HMMs
- Context-sensitive sub-word HMMs
  - Biphone HMMs
  - Triphone HMMs
- Triphone HMM training issues
- Phoneme Decision Trees (PDTs)

Word Level HMMs

- Early systems (1980s) used <u>word</u> level HMMs
- I.e. each word modelled by a single, dedicated HMM (c.f. "zero" picture)
  - Advantages:
  - Good performance due to explicit modelling of word-dependent variability

6 state HMM of the digit ‘zero’

UNIVERSITY OF BIRMINGHAM

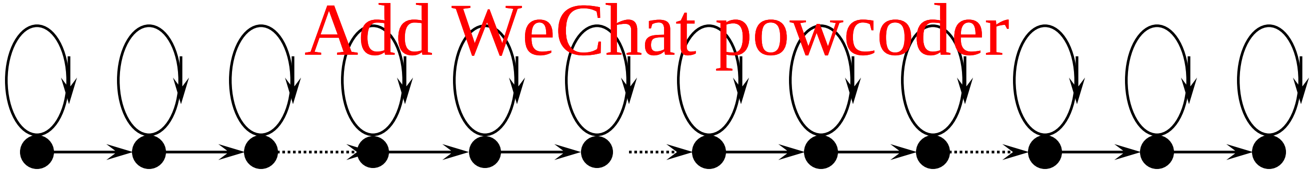
Word Level HMMs

- Disadvantages:
  – Many examples of each word required for training
  – Fails to exploit regularities in spoken language
- Word-level systems typically restricted to well-defined, demanding, small vocabulary applications

UNIVERSITY OF BIRMINGHAM

Sub-Word Level HMMs

- Build HMMs for a complete set of sub-word ‘building blocks’
- Construct word-level HMMs by concatenation of sub-word HMMs
- E.g. slide = / s l aI d /



/ s / **/ l /** **/ aI /** / d /

Sub-Word Level HMMs

- Advantages
  - Able to exploit regularities in speech patterns
  - More efficient use of training data - e.g. in phoneme-based system “five” (/ f aI v /) and “nine” (/n aI n /) both contribute to /aI/ model.
  - Flexibility - acoustic models can be built **immediately** for words which did not occur in the training data

UNIVERSITY OF BIRMINGHAM

Phoneme-Level HMMs

- Why choose phonemes rather than any other sub-word unit?
- Disadvantages
  - Phonemes are defined in terms of the contrastive properties of speech sounds within a language - not their consistency with HMM assumptions!

Advantages of Phoneme-HMMs

- Completeness & compactness – approx. 50 phonemes required to describe English

- Well studied – potential for exploitation of ‘speech knowledge’ (e.g. pronunciation differences due to accent…)

- Availability of extensive phoneme-based pronunciation dictionaries

Context-Sensitivity

- Problem
  - Acoustic realization of a phoneme depends on the context in which it occurs
  - Think of your lip shape for the “k” sound in the words “book shop” and “thick”

Biphones and Triphones

- Solution
  - **Context-sensitive** phoneme-level HMMs
  - E.g.
    - ‘biphones’ : (k:_S) in “book shop”
    - ‘triphones’ : (k:u_S) in “book shop”
- Almost all systems use triphone HMMs

UNIVERSITY OF BIRMINGHAM

Triphones - problems

- Increased number of model parameters
  - Need more (well-chosen) training data

- Which triphone?

  - If a word in the application contains a triphone which was not in the training set, which triphone HMM should we use?

UNIVERSITY OF BIRMINGHAM

Number of parameters

- If there are 50 phones, the maximum number of triphone HMMs is 50^3=125,000
- Most ruled out by **phonological** constraints – most phone triples never occur in speech
- But many are legal

Example: Model Parameters

- Each model has 3 emitting states
- Each state modelled as, say, a 10 component Gaussian mixture

- Each feature vector is 40 dimensional

- Hence number of parameters per model is:

$$3\times(10\times(40+40+1)+9)=2,457$$

Number of states | Number of mixture components | Mean vector | Variance vector | Mixture weight | Transition probs

Acoustic model parameters

- So, even if we only have 1,000 acoustic models (instead of 125,000), total acoustic model parameters will be 2,457,000
- Too many to estimate with practical quantity of data
- Most common solution is HMM **parameter tying**
- **Different** HMMs share **same** parameters

UNIVERSITY OF BIRMINGHAM

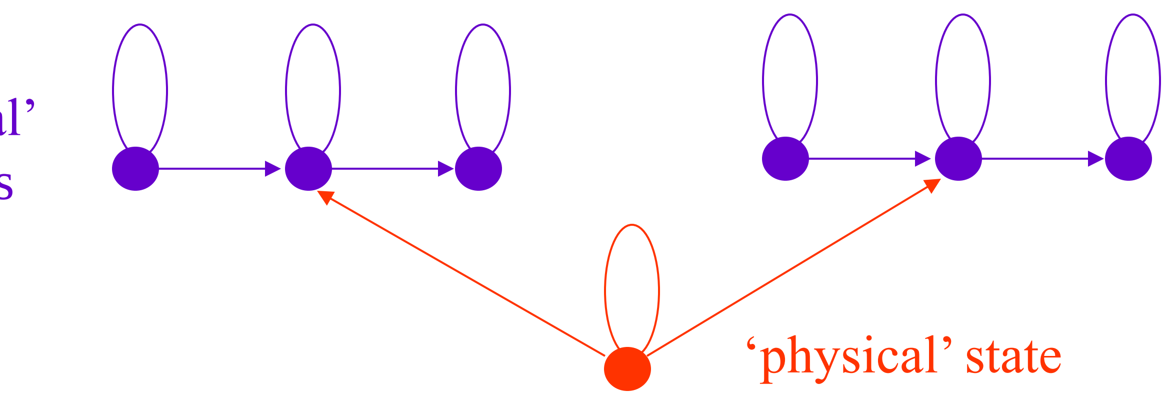
Tied variance

- Variances are more costly to estimate than means
- Simple solution – divide set of all HMMs into classes, so that within a class all HMM state PDFs have same variance
- This is **tied variance**
- If **all** HMM state PDFs share the same variance, the variance is referred to as **grand variance**

UNIVERSITY OF BIRMINGHAM

Phone decision trees

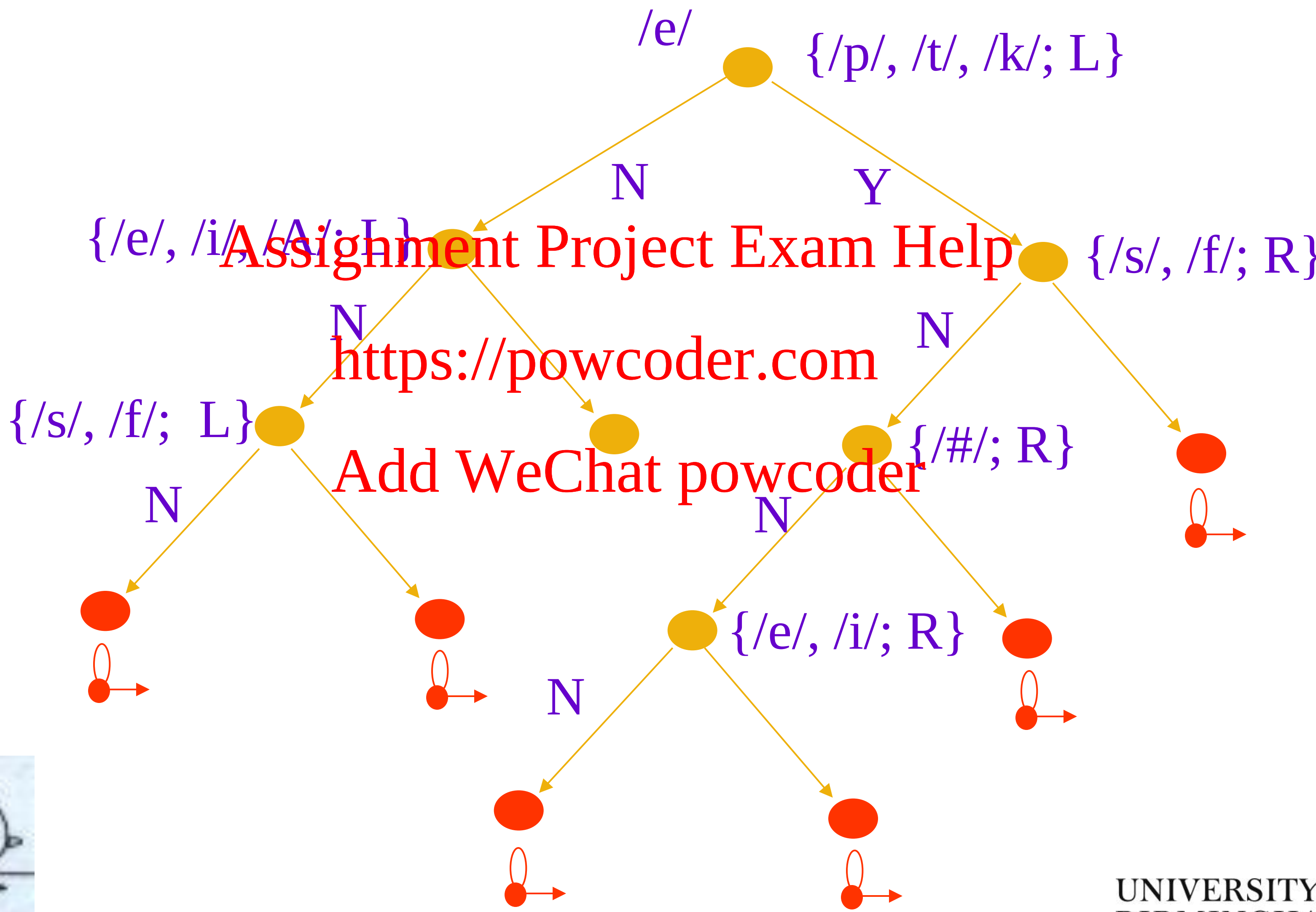
- Most common approach to general HMM tying is **decision tree clustering**
- Decision tree clustering can be applied to individual states or to whole HMMs – we'll consider states
- Basic idea is to use **knowledge** about which phones are likely to induce similar contextual effects



Phonetic knowledge

- For example, we know that /*f*/ and /s/ are both unvoiced fricatives, produced in a similar manner
- Therefore we might **hypothesise** that, for example, an utterance of the vowel /*e*/ preceded by /*f*/ might be similar to one preceded by /*s*/
- This is the basic idea behind decision tree clustering

Phone Decision Tree



Summary

- Word-level and Sub-Word HMMs
- Phoneme-level HMMs
- Context-sensitivity
  - Biphones & Triphones
- Triphone decision trees