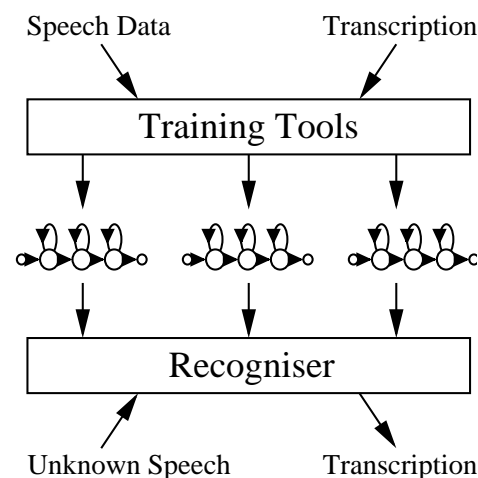


HMM tutorial 5

by Dr Philip Jackson

- Simple example
- Use of HMMs
 - Introduction to HTK
- Practical issues
 - Training and testing
- Application to speech
 - Isolated words
 - Continuous
- Summary



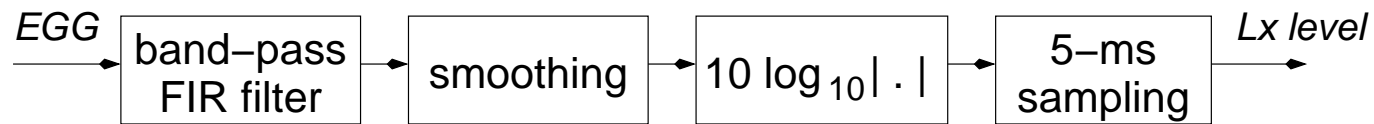
* Figures taken from Young et al. (1997)



Illustrative example

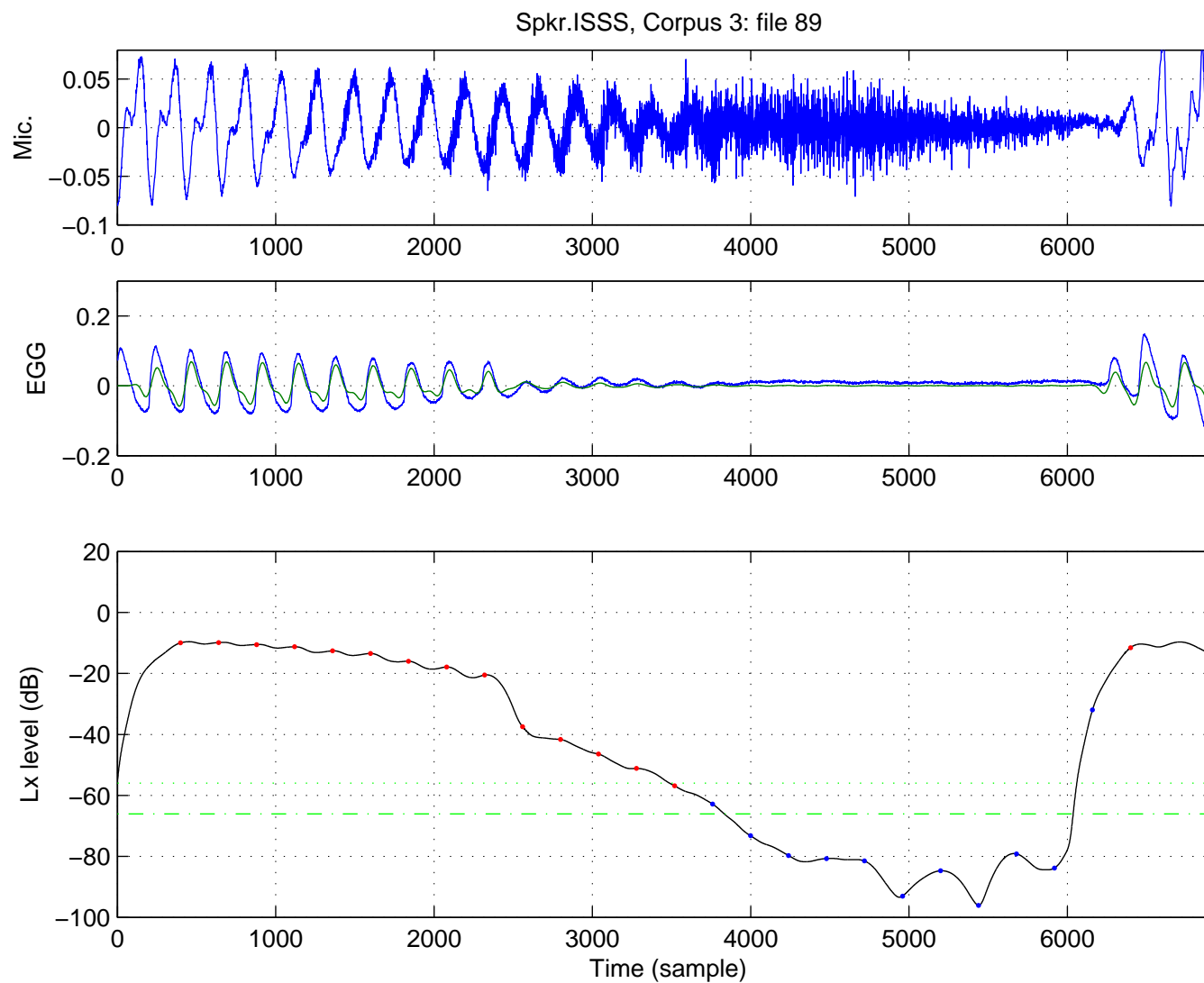
Aim:

to provide an objective labelling of voiced and unvoiced regions during fricative speech.



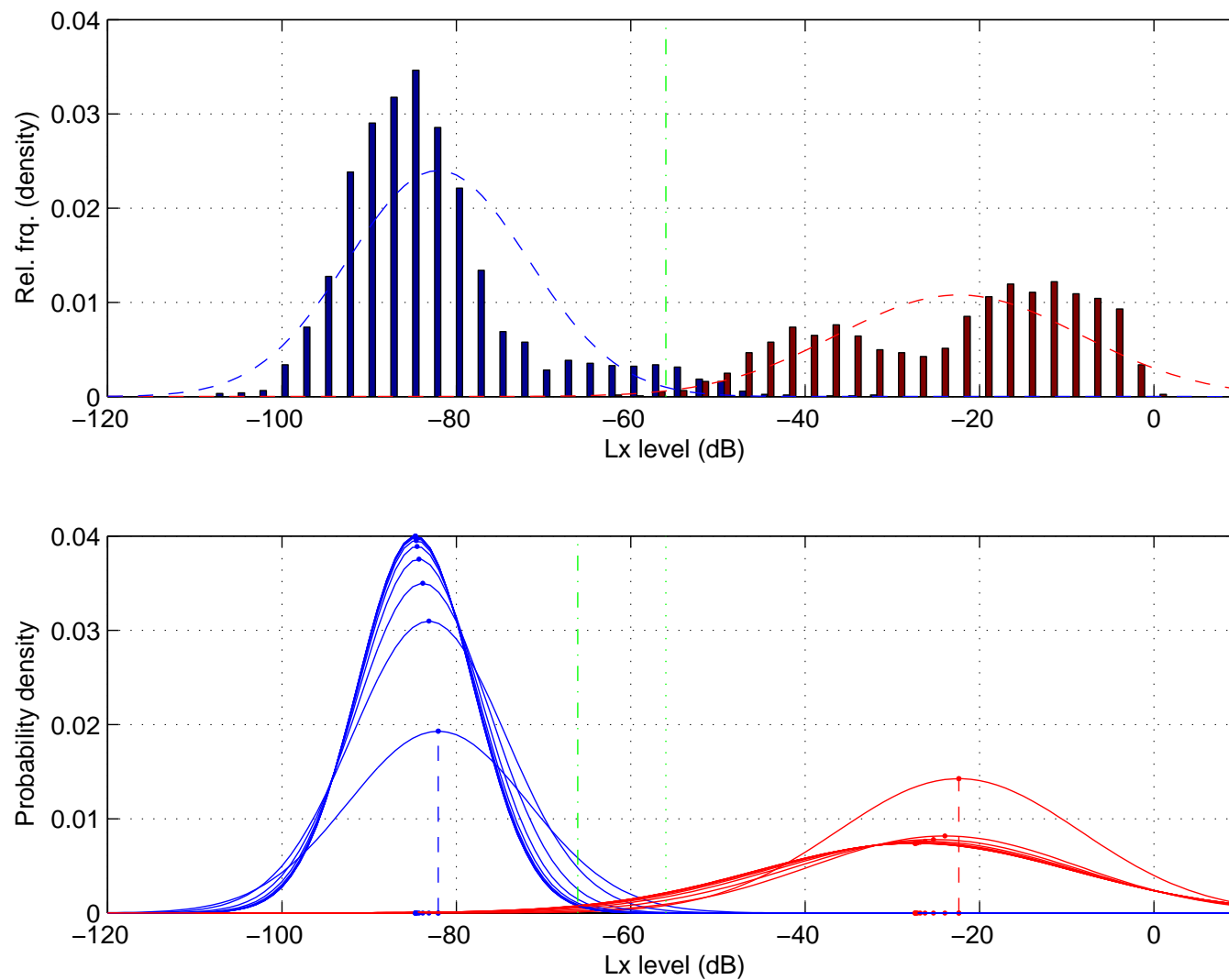
Processing of EGG signal to give Lx-level features.

Feature extraction front end



From top: acoustic waveform (top) of VFV utterance of devoiced /zh/; original and band-passed EGG waveforms; derived Lx level.

Effect of training



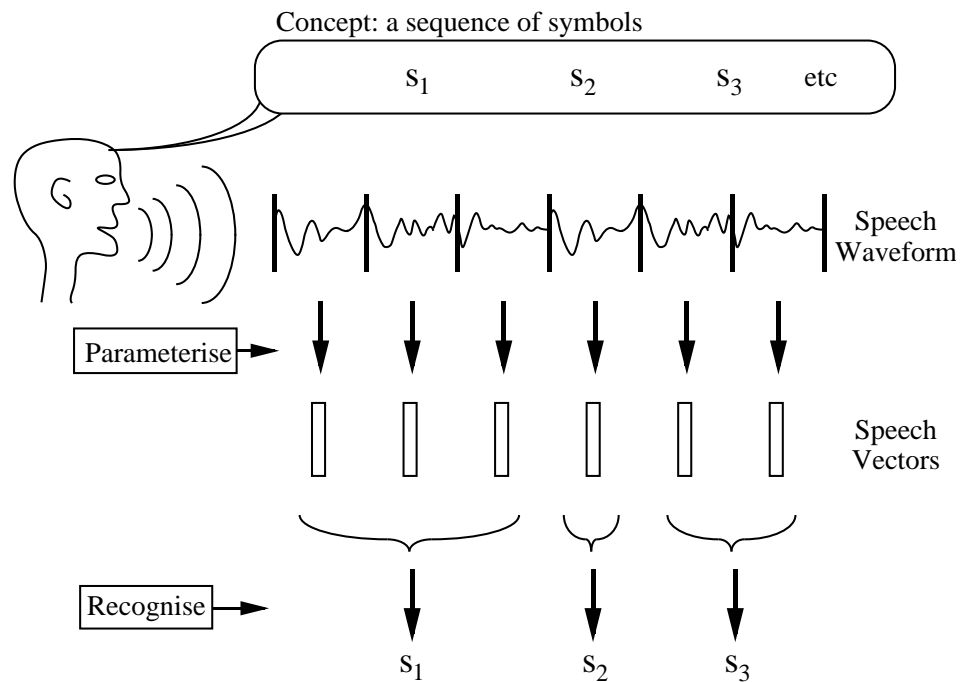
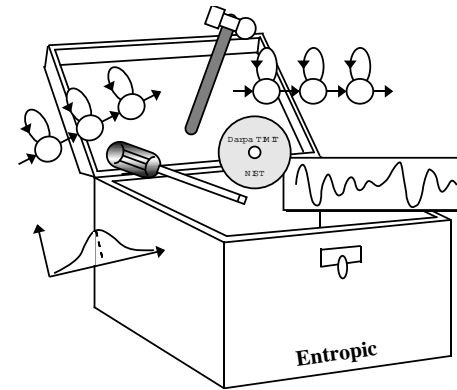
From top: histograms of Lx level classified manually with initial models; convergence of model re-estimation (from dashed start).

Example summary

- Hand labelled training data
- Initialised model
- Trained model
- Decoded with model to determine voiced/unvoiced boundaries
- Timing results used in analysis

Tools for using HMMs

- Data preparation
- Training
- Testing
- Analysis



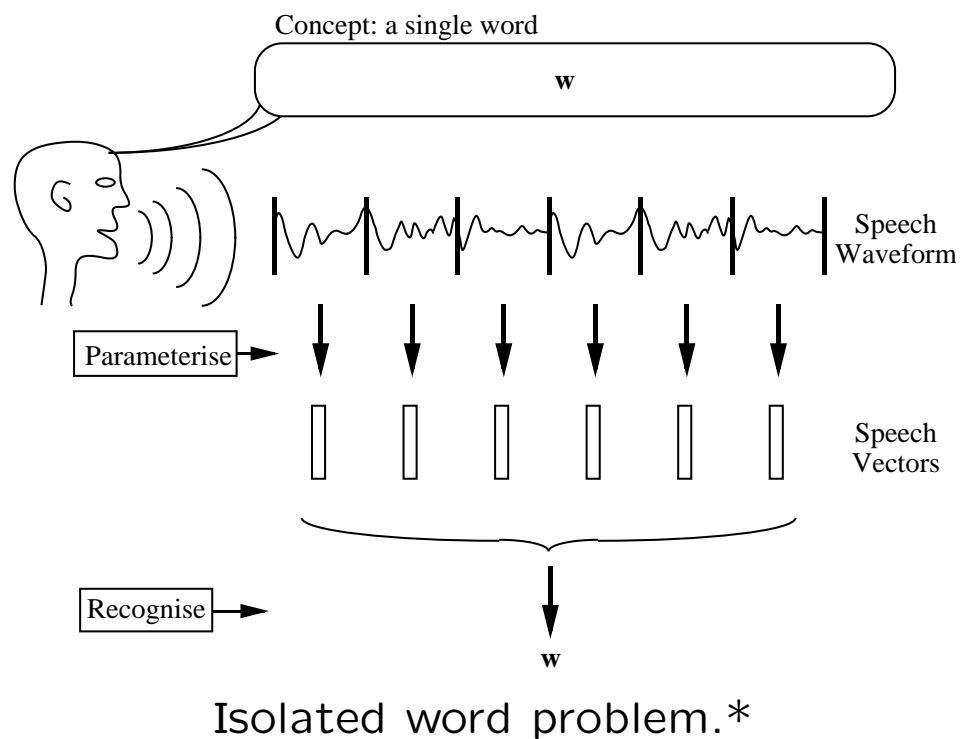
Message encoding and decoding.*

Isolated word recognition

The problem is to find

$$\arg \max_i \{P(w_i|\mathcal{O})\}, \quad (1)$$

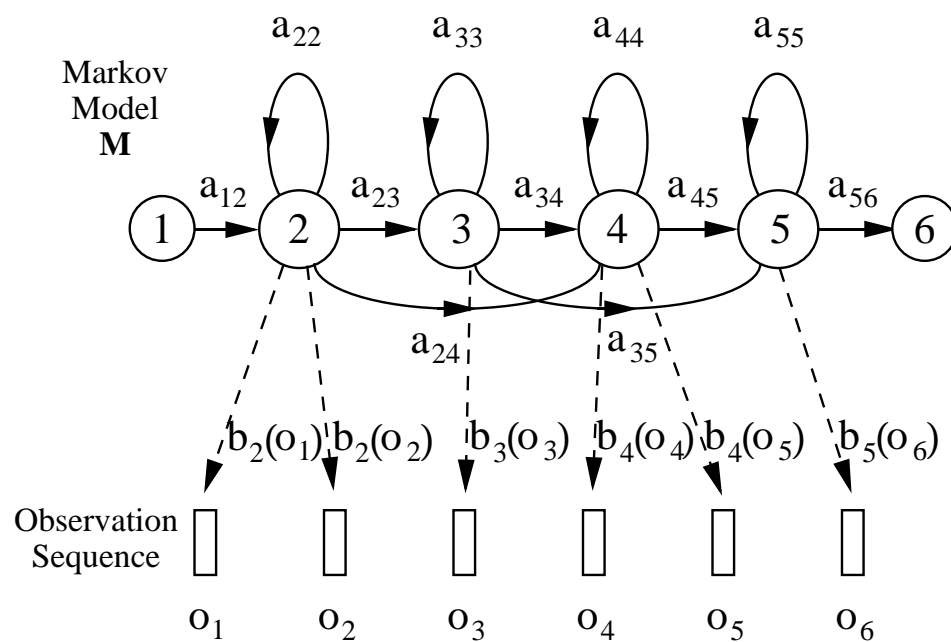
where, according to Bayes, $P(w_i|\mathcal{O}) = \frac{P(\mathcal{O}|w_i)P(w_i)}{P(\mathcal{O})}$.



The hidden Markov model

In this case, we assume

$$\begin{aligned}
 P(O|w_i) &= P(O|\lambda_i) & (2) \\
 &= \arg \max_X \pi_{x_1} b_{x_1}(o_1) \cdot \prod_{t=2}^T a_{x_{t-1}x_t} b_{x_t}(o_t).
 \end{aligned}$$



The Markov generation model.*

Building the grammar

Example utterances:

Dial three three two six five four

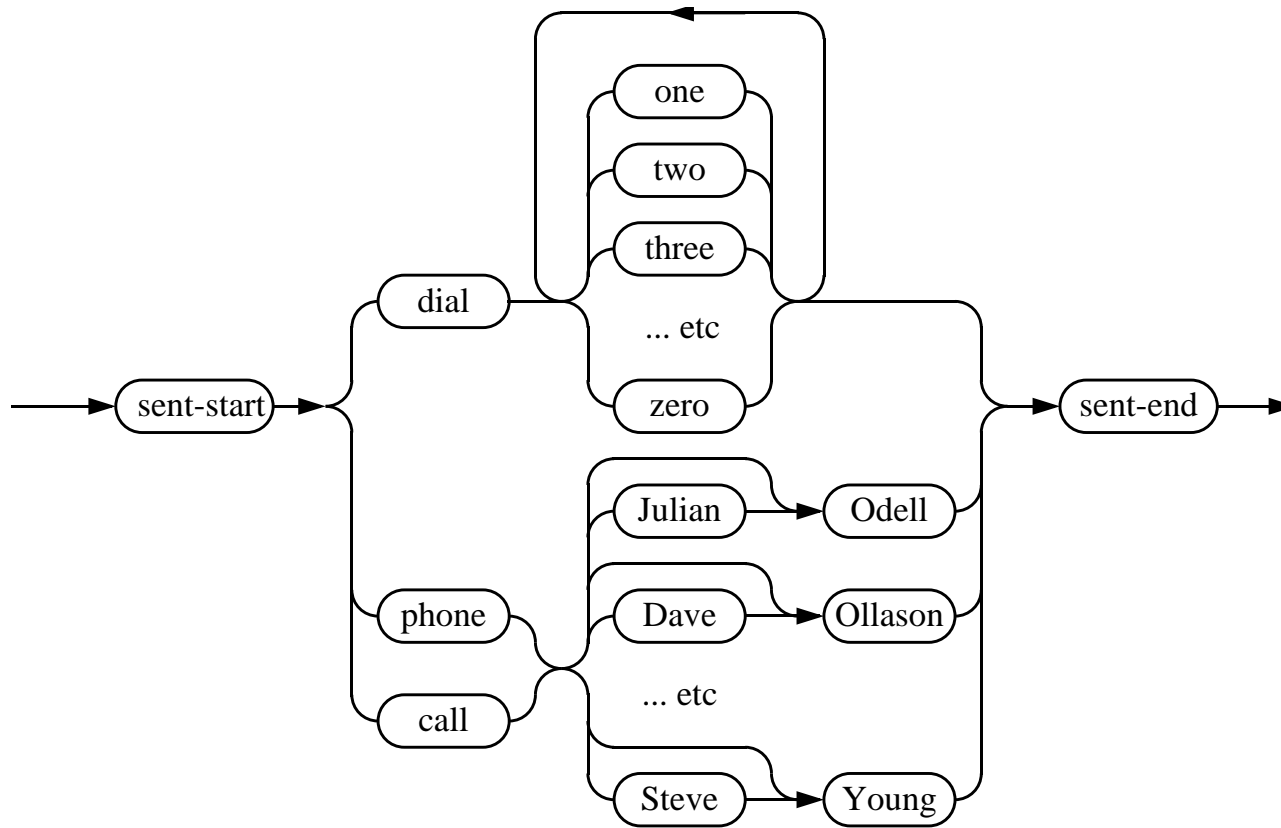
Phone Woodland

Call Steve Young

Task grammar:

```
$digit = ONE | TWO | THREE | FOUR | FIVE |  
        SIX | SEVEN | EIGHT | NINE | OH | ZERO;  
$name  = [ JULIAN ] ODELL |  
        [ DAVE ] OLLASON |  
        [ PHIL ] WOODLAND |  
        [ STEVE ] YOUNG;  
( SENT-START ( DIAL <$digit> | (PHONE|CALL) $name) SENT-END )
```

Key: | alternatives, [...] optional, <.> one or more reps.



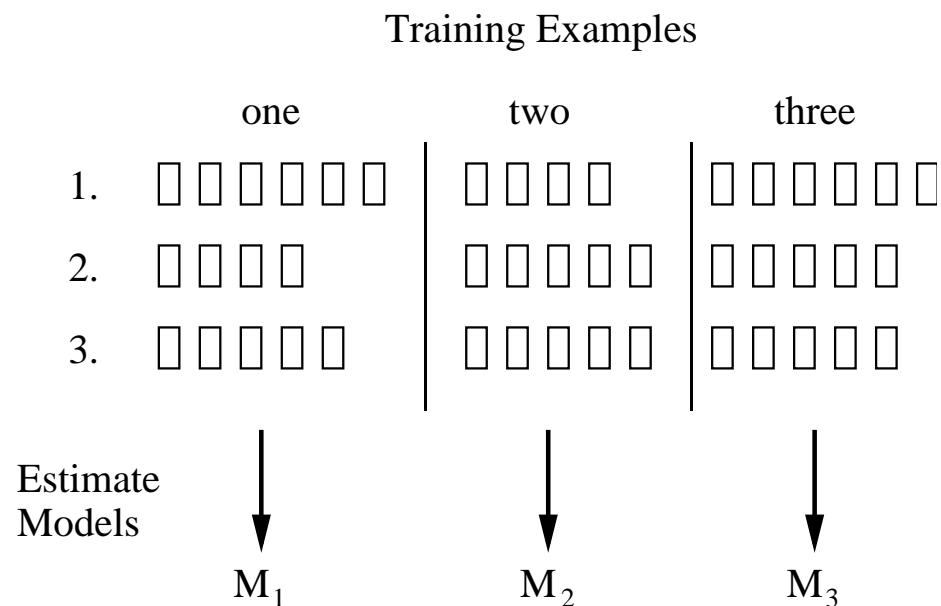
Grammar for voice dialling.*

Dictionary

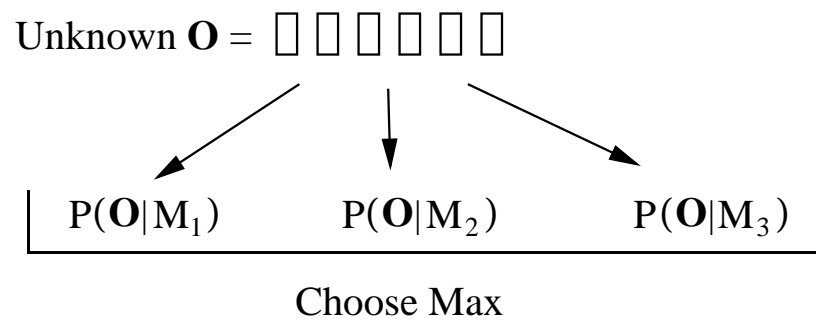
ONE	w ah n sp
TWO	t uw sp
THREE	th r iy sp
FOUR	f ao sp
FOUR	f ao r sp
...	

Training and test data

(a) Training

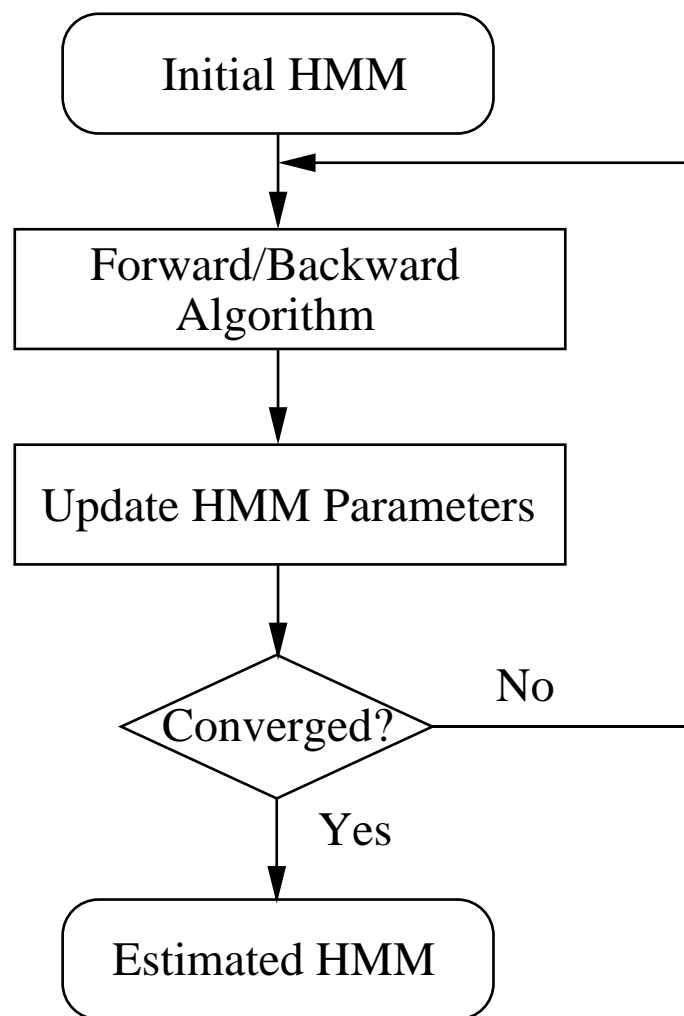


(b) Recognition



Using HMMs for isolated word recognition.*

Re-estimation process

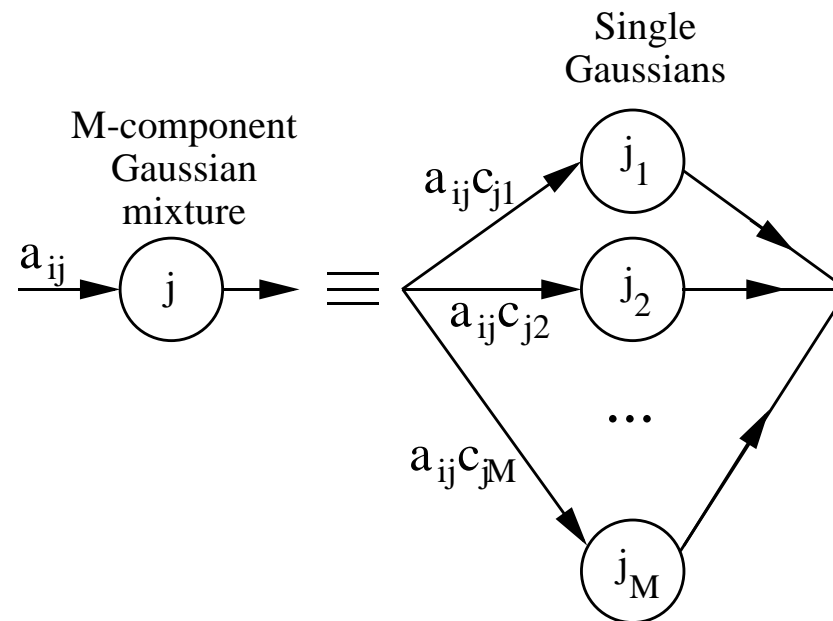


Isolated unit re-estimation.*

Output probabilities

Recall, for a Gaussian mixture, the output probability

$$b_j(\mathbf{o}_t) = \sum_{m=1}^M c_{jm} \mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}) \quad (3)$$



Representing a mixture of Gaussians.*

Viterbi re-estimation:

$$\text{means} \quad \hat{\boldsymbol{\mu}}_{jm} = \frac{\sum_{t=1}^T \psi_t(j,m) \mathbf{o}_t}{\sum_{t=1}^T \psi_t(j,m)},$$

$$\text{variances} \quad \hat{\boldsymbol{\Sigma}}_{jm} = \frac{\sum_{t=1}^T \psi_t(j,m) (\mathbf{o}_t - \boldsymbol{\mu}_{jm})(\mathbf{o}_t - \boldsymbol{\mu}_{jm})'}{\sum_{t=1}^T \psi_t(j,m)},$$

$$\text{weights} \quad \hat{c}_{jm} = \frac{\sum_{t=1}^T \psi_t(j,m)}{\sum_{t=1}^T \psi_t(j)},$$

where $\psi_t(j, m)$ and $\psi_t(j)$ are binary indicator functions.

Baum-Welch re-estimation:

$$\text{means} \quad \hat{\boldsymbol{\mu}}_{jm} = \frac{\sum_{t=1}^T \gamma_t(j,m) \mathbf{o}_t}{\sum_{t=1}^T \gamma_t(j,m)},$$

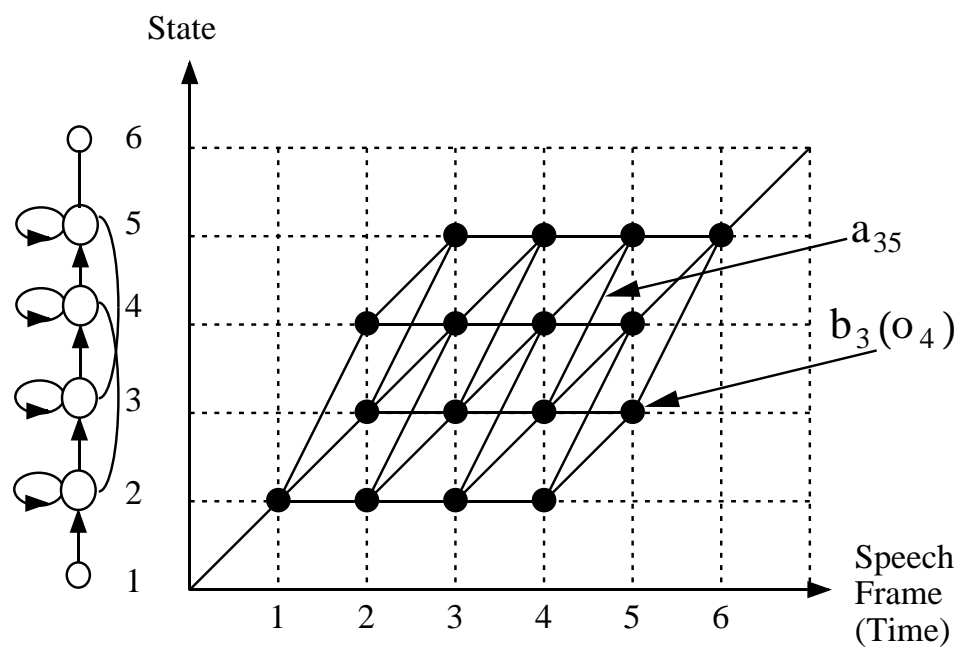
$$\text{variances} \quad \hat{\boldsymbol{\Sigma}}_{jm} = \frac{\sum_{t=1}^T \gamma_t(j,m) (\mathbf{o}_t - \boldsymbol{\mu}_{jm})(\mathbf{o}_t - \boldsymbol{\mu}_{jm})'}{\sum_{t=1}^T \gamma_t(j,m)},$$

$$\text{weights} \quad \hat{c}_{jm} = \frac{\sum_{t=1}^T \gamma_t(j,m)}{\sum_{t=1}^T \gamma_t(j)},$$

where $\gamma_t(j, m)$ is as before and $\gamma_t(j) = \frac{\alpha_t(j) \beta_t(j)}{P(\mathcal{O}|\lambda)_i}$.

Training procedure

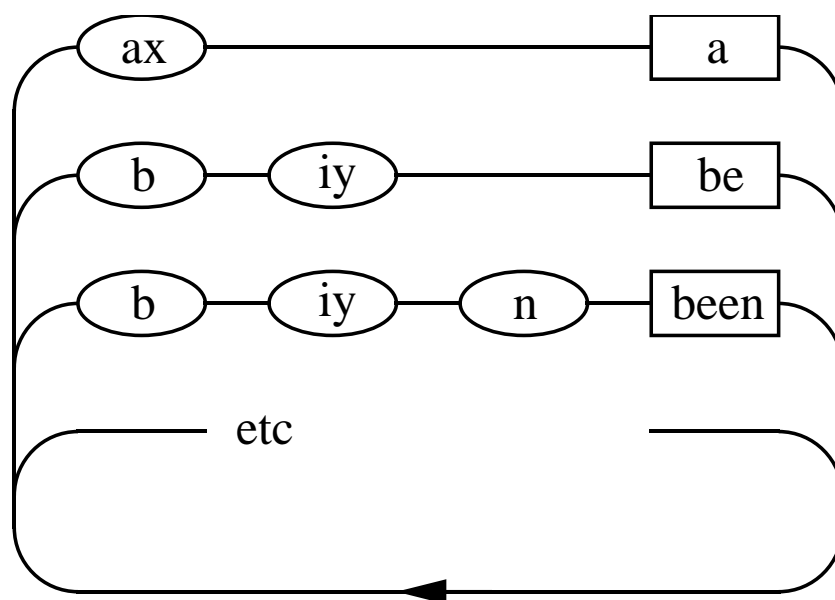
1. initialise accumulators for all HMMs' parameters
2. read the next training utterance
3. join HMMs in sequence to make composite HMM
4. calculate forward & backward probabilities



The Viterbi algorithm for isolated word recognition.*

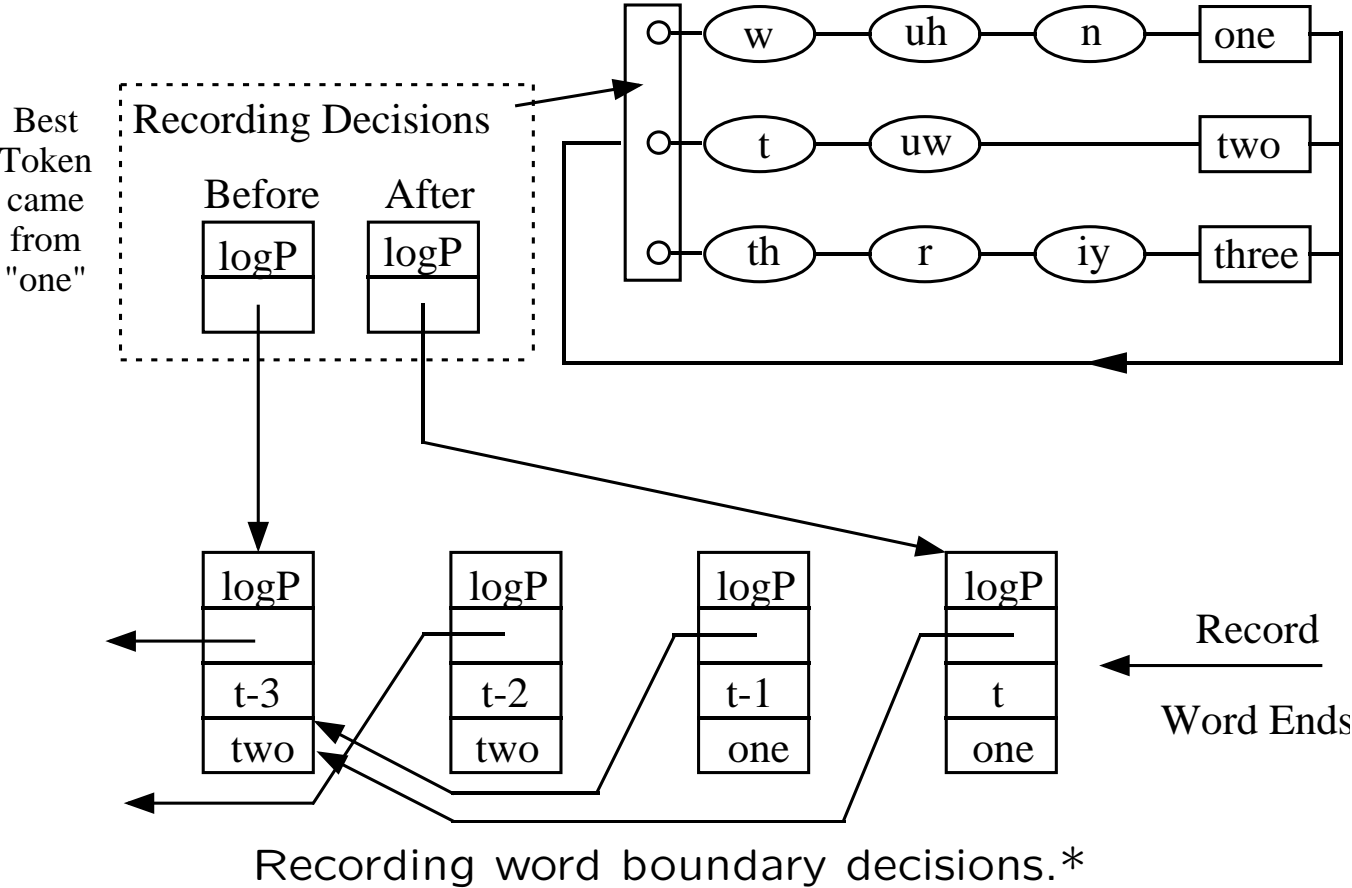
for Continuous speech recognition

5. use forwd/backwd probs to increment accumulators
6. repeat from step 2 until all utterances processed
7. use accumulators to update parameters for all HMMs



Recognition network for continuously spoken word recognition.*

Recognition and traceback



Annotation results

ice	cream
-----	-------



(a) 1-alternative, 1-level

ice		cream			
ay	s	k	r	iy	m



(b) 1-alternative, 2-level

I	scream
ice	cream
eyes	cream



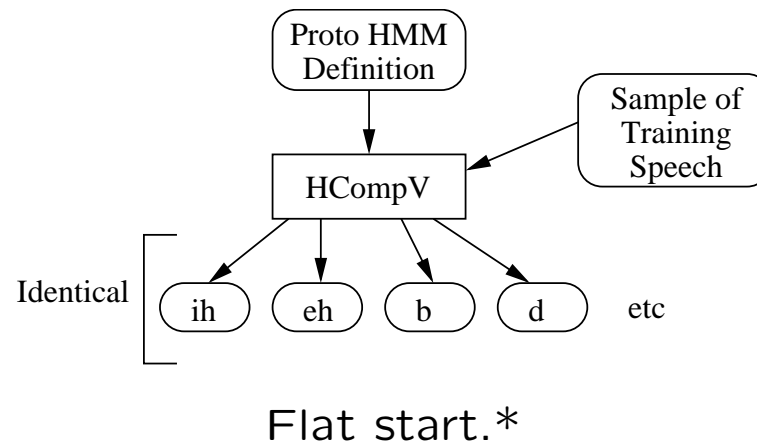
(c) 3-alternative, 1-level

Example transcriptions.*

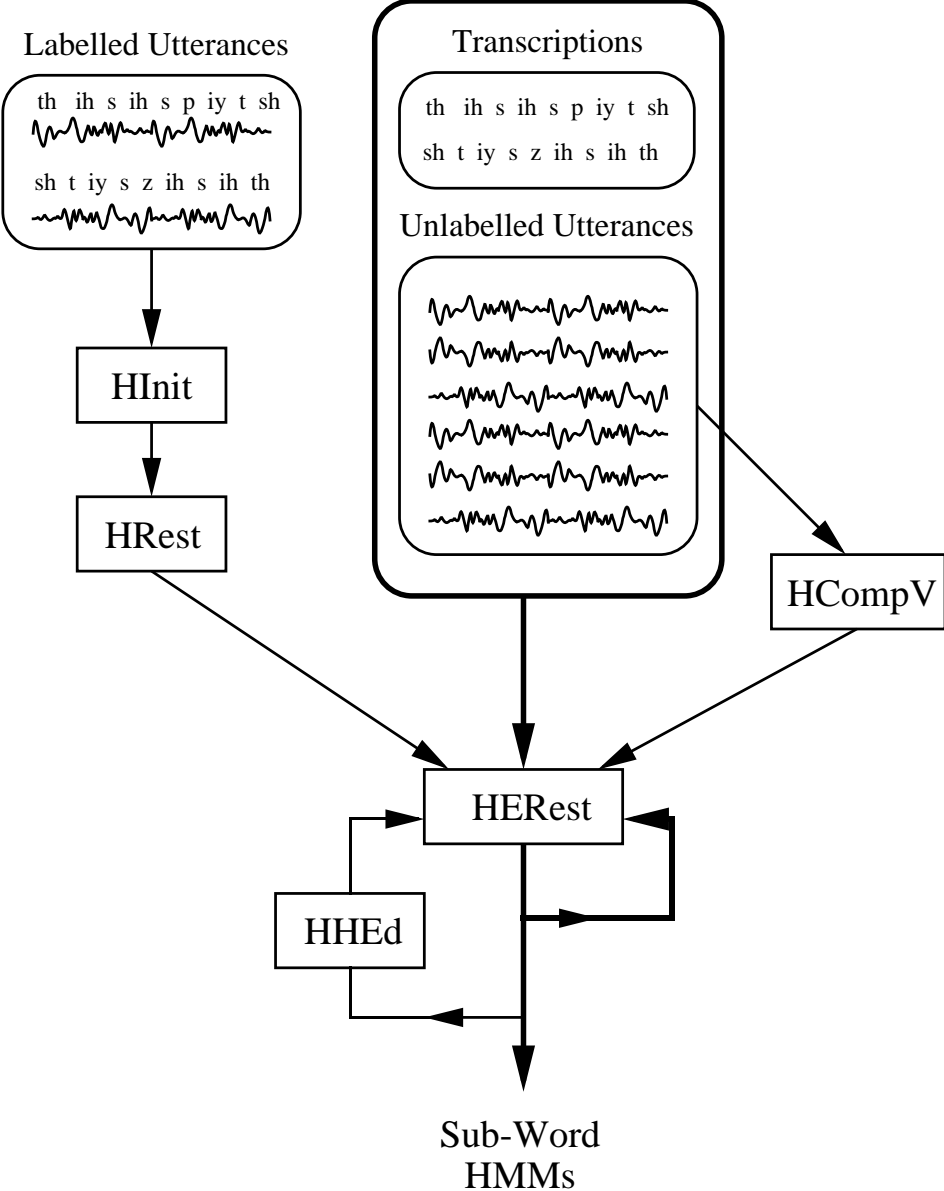
Practical issues

Initialisation

1. Random
2. Flat start
3. Least squares
4. Viterbi
5. Baum-Welch (supervised)
6. Baum-Welch (unsupervised)



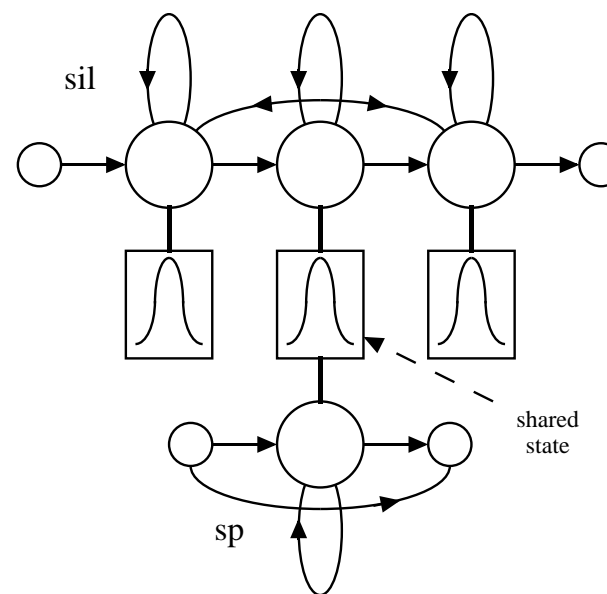
Re-estimation and embedded re-estimation



Training sub-word HMMs.*

Number of parameters

- Parsimony
 - Occam's razor
- Amount of training data



Linking the silence models.*

Regularisation

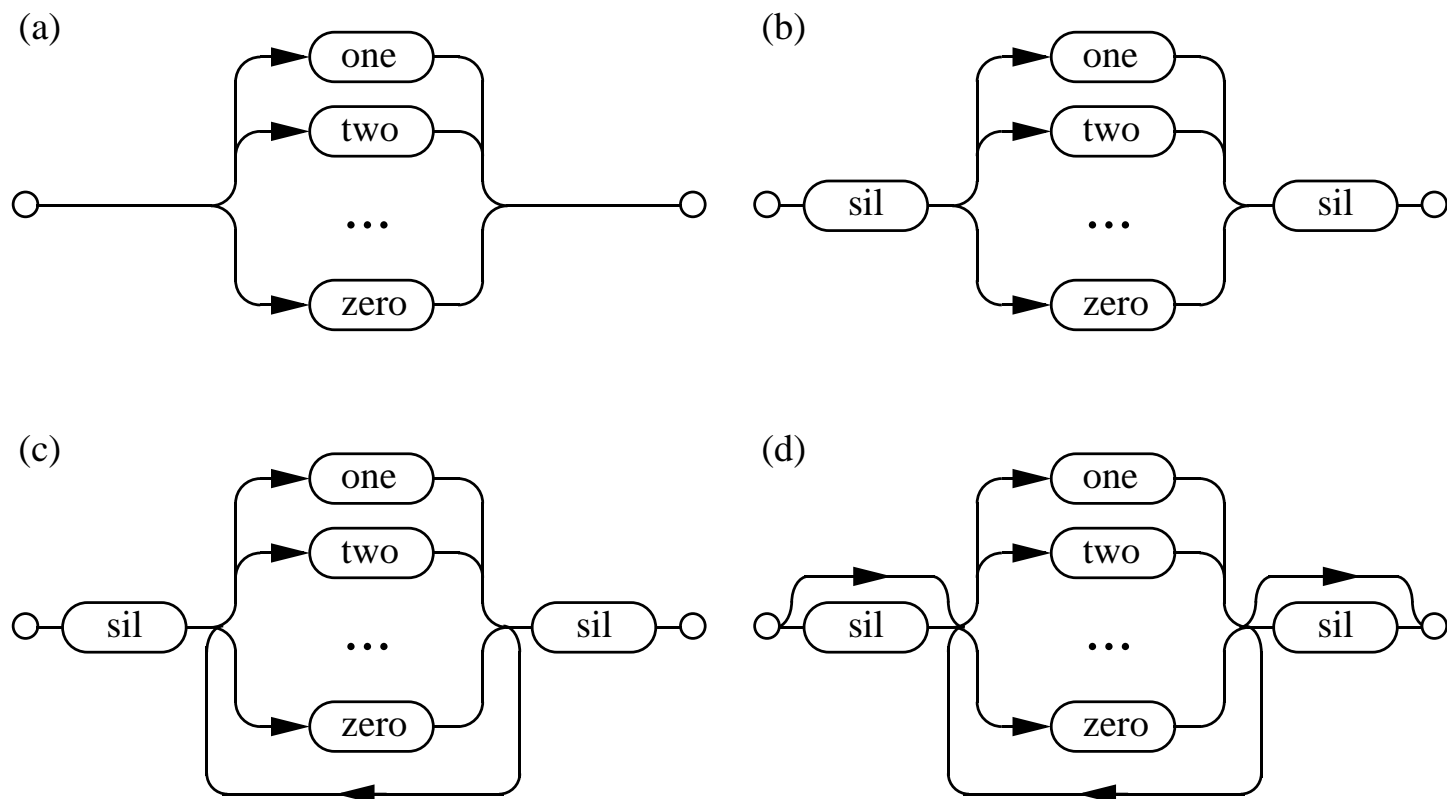
- variance floor
- parameter tying

Search strategies

- Depth first: stack decoder/A*
- Breadth first: beam pruning

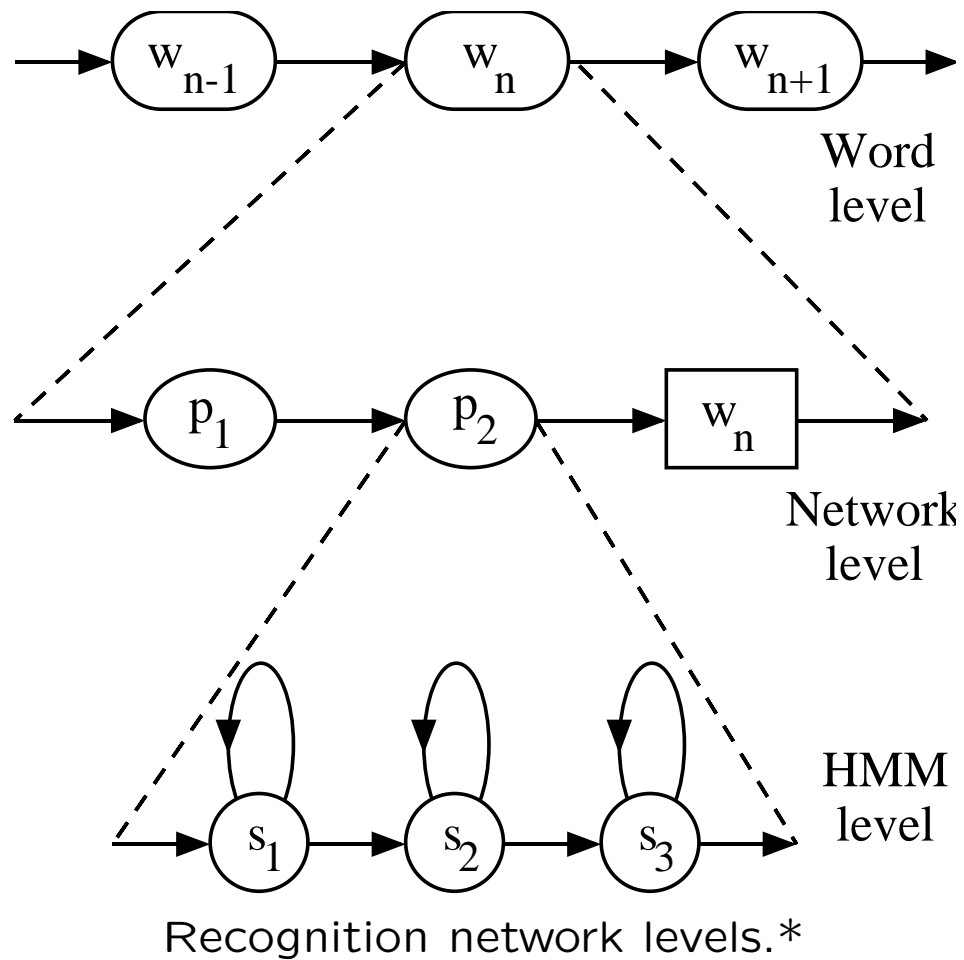
Applications in ASR

Isolated word recognition



Example digit recognition networks.*

Hierarchy of recognition networks



Today's summary

- Example of a simple labelling task
- Intro to HMM toolkit & practical issues
- Applications in automatic speech recognition

Further reading

- F. Jelinek. *Statistical methods for speech recognition*. MIT, Cambridge MA, 1998. [ISBN 0-2621-0066-5].
- S. J. Young, J. Odell, et al. *The Fundamentals of HTK*, chapter 1 in *The HTK Book*, pp. 2–13, Entropic, Cambridge, UK, 1997. [<http://htk.eng.cam.ac.uk/>].
- S. J. Young. Large vocabulary continuous speech recognition. *IEEE Sig. Proc. Mag.* 13(5): 45–57, 1996.