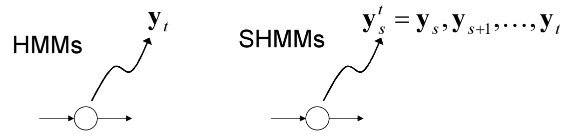
Comparison of pruning strategies for segmental HMMs

Y. Shiga and P. J. B. Jackson
Centre for Vision, Speech and Signal Processing
University of Surrey

Overview

Segment Models

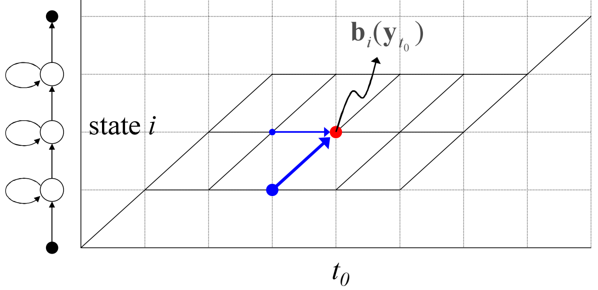
- Segmental HMMs (SHMMs) generate a **feature-vector trajectory** per state, for speech recognition or synthesis.



- However, expanding the state space for the trajectory makes SHMMs **computationally costly**.

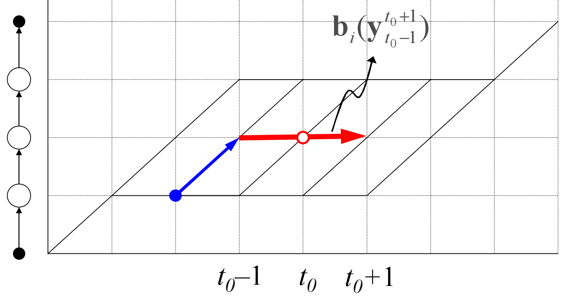
Computational load problem

- Standard HMMs $\quad \alpha_t(i) = \max_j \alpha_{t-1}(j)\, \mathbf{a}_{ji}\, \mathbf{b}_i(\mathbf{y}_t)$



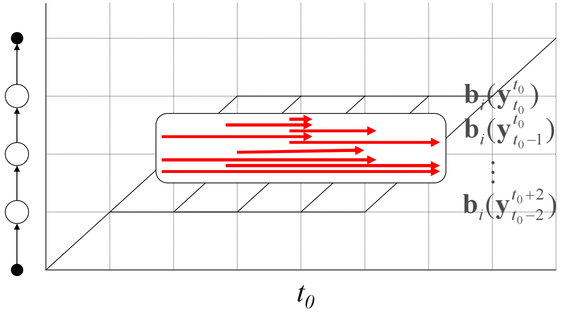
Computational load problem

- SHMMs $\quad \alpha_t(i) = \max_j \max_{d=1,\ldots,D_{\max}} \alpha_{t-d}(j)\, \mathbf{a}_{ji}\, \mathbf{b}_i(\mathbf{y}_{t-d+1}^{t})$



Computational load problem

- SHMMs $\alpha_t(i) = \max_j \max_{d=1,\ldots,D_{\max}} \alpha_{t-d}(j)\, \mathbf{a}_{ji}\, \mathbf{b}_i(\mathbf{y}_{t-d+1}^t)$



Computational load problem

- Therefore…
  - Efficient search is essential for a recognizer to perform decoding within a reasonable time, for training and recognition.

Pruning

- But, before that,…
  - Decoding algorithm for SHMMs, derived from the Viterbi algorithm

The Decoding Algorithm

- More elegant way of decoding
  - Introduction of *Start Node* (SN) and *End Node* (EN), and their probabilities SNP and ENP

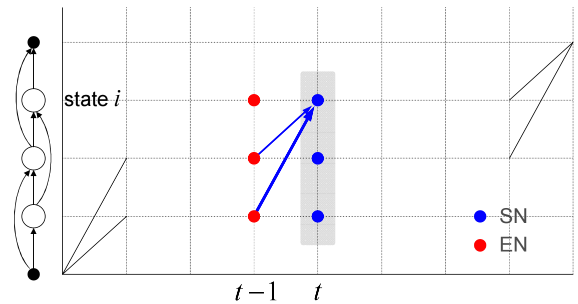
$$\alpha_t(i) = \max_j \max_{d=1,\ldots,D_{\max}} \alpha_{t-d}(j)\, \mathbf{a}_{ji} \mathbf{b}_i(\mathbf{y}_{t-d+1}^t)$$

$$= \max_{d=1,\ldots,D_{\max}} \left[\max_j \alpha_{t-d}(j)\, \mathbf{a}_{ji} \right] \mathbf{b}_i(\mathbf{y}_{t-d+1}^t)$$

$$\begin{cases} \text{SNP:} & \beta_t(i) = \max_j \alpha_{t-1}(j)\, \mathbf{a}_{ji} \\ \text{ENP:} & \alpha_t(i) = \max_{d=1,\ldots,D_{\max}} \beta_{t-d+1}(i)\, \mathbf{b}_i(\mathbf{y}_{t-d+1}^t) \end{cases}$$

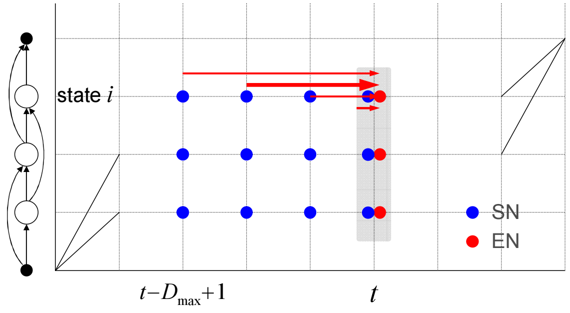
The Decoding Algorithm

- SNP calculation $\beta_t(i) = \max_j \alpha_{t-1}(j)\, \mathbf{a}_{ji}$
  - Finding best state-transition



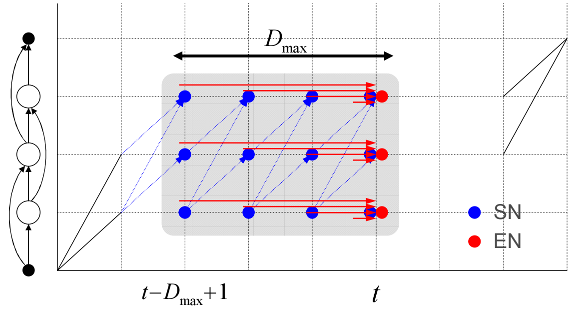
The Decoding Algorithm

- ENP calculation $\alpha_t(i) = \max_{d=1,\ldots,D_{\max}} \beta_{t-d+1}(i)\, \mathbf{b}_i(\mathbf{y}_{t-d+1}^{t})$
  - Finding best segment-duration



The Decoding Algorithm

- SNP calculation --- seeding
- ENP calculation --- harvesting



Pruning Strategies

- Russell (2005) proposes:

(3) – beam pruning (for EN)

(4) – state-duration pruning

- We add:

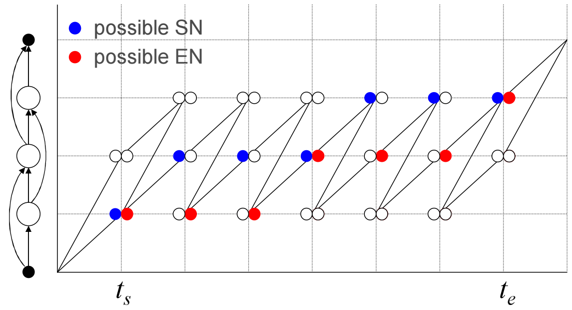
(1) – pre-cost partition

(2) – beam pruning for SN

Pruning Strategies

1. Pre-cost partition

in the case of $D_{\max} = 3$



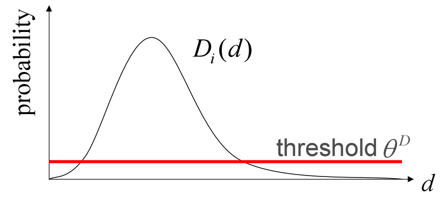
Pruning Strategies

2. SN beam pruning
   - Pruning before output probability calculation
     - Let $\beta_t(i_{\max})$ denote the maximal SNP at time t. If $|\log \beta_t(i_{\max}) - \log \beta_t(i)| > \theta^{\mathrm{S}}$, the start node of state i at time t is pruned.
3. EN beam pruning (Russell,2005)
   - Pruning after output probability calculation
     - Let $\alpha_t(j_{\max})$ denote the maximal ENP at time t. If $|\log \alpha_t(j_{\max}) - \log \alpha_t(j)| > \theta^{\mathrm{E}}$, the end node of state j at time t is pruned.

Pruning Strategies

4. State-duration pruning (Russell,2005)

$$\mathbf{b}_i(\mathbf{y}_{t-d+1}^{t}) = \underline{D_i(d)}^F \prod_{r=t-d+1}^{t} \mathcal{N}(f_i(r), \sigma_i; \mathbf{y}_r)$$



Experiments

- Condition
  - *Monophone*, 3-state SHMMs
  - Linear segment-trajectory
  - Parametric duration model using Gamma distribution
  - Phone-level bigram language model
- Data
  - TIMIT male speaker training set
  - 3180 and 80 sentences for training and evaluation, respectively
  - 13 MFCCs including C_0, 25ms width, 10ms spacing window

Experiments

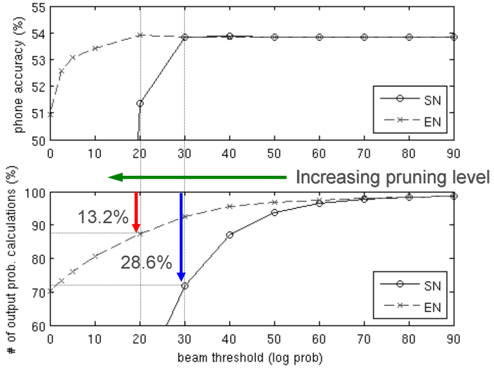
1. Pre-cost partition

Reduction of number of output-prob calculations (%)

	training	recognition
supervised (phone-level)	18.9	42.8
embedded (sentence-level)	0.1	0.4

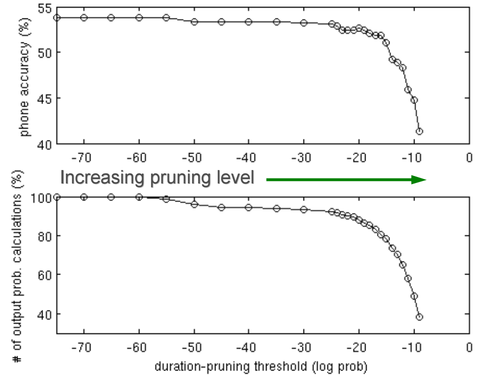
Experiments

2&3. SN and EN beam pruning



Experiments

4. State-duration pruning



Summary

- Pre-cost partition reduced output-prob computation for **supervised** training and recognition by 18.9% and 42.8%.
- The result of beam pruning showed that **SN beam pruning is more efficient** than EN beam pruning.
- Recognition accuracy was **sensitive to duration probability threshold** θ^D, unlike the experiment by Russell (2005).

Summary

Recognition (embedded)

	accuracy(%)	computational reduction (%)
no pruning	53.8	0.0
pre-cost part.	53.8	0.4
pre-cost part. + SN (θ^S=30)	53.8	28.6
pre-cost part. + EN (θ^E=20)	53.9	13.2
pre-cost part. + SN (θ^S=30) +EN (θ^E=20)	**54.0**	**30.9**

Conclusions

- SHMM decoder based on SNP and ENP
- Experiments on TIMIT with four pruning strategies

What's next?

- Introducing context-sensitive models
- SN beam pruning for standard HMMs?

Thank you very much for your attention

DANSA project is funded by EPSRC (GR/S85511/01)
http://www.ee.surrey.ac.uk/Personal/P.Jackson/Dansa