UniS Centre for Vision Speech & Signal Processing
University of Surrey, Guildford GU2 7XH.

HMM tutorial 3

by Dr Philip Jackson

- Recap. of α, β and Viterbi
- Re-estimating models
  - Occupation
  - Transition
  - Baum-Welch formulae
- Summary

Problem 1: Forward procedure

Consider $\alpha_t(i) = P(x_t = i, \boldsymbol{o}_1^t | \lambda)$:

1. Initially,
   $\alpha_1(i) = \pi_i \, b_i(o_1)$, for $1 \leq i \leq N$;

2. For $t = 2, 3, \ldots, T$,
   $\alpha_t(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i) \, a_{ij}\right] b_j(o_t)$, for $1 \leq j \leq N$;

3. Finally,
   $P(\mathcal{O}|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$.

(1)

Thus, we can solve Problem 1 efficiently by recursion.

Problem 1: Backward procedure

Define $\beta_t(i) = P(\boldsymbol{o}_{t+1}^T | x_t = i, \lambda)$:

1. Initially,
   $\beta_T(i) = 1,$ for $1 \leq i \leq N$;

2. For $t = T-1, T-2, \ldots, 1$,
   $\beta_t(i) = \sum_{j=1}^N a_{ij}\, b_j(o_{t+1})\, \beta_{t+1}(j),$ for $1 \leq i \leq N$;

3. Finally,
   $P(\mathcal{O}|\lambda) = \sum_{i=1}^N \pi_i\, b_i(o_1)\, \beta_1(i).$

(2)

We now have another efficient way of computing $P(\mathcal{O}|\lambda)$.

Problem 2: Finding best state sequence

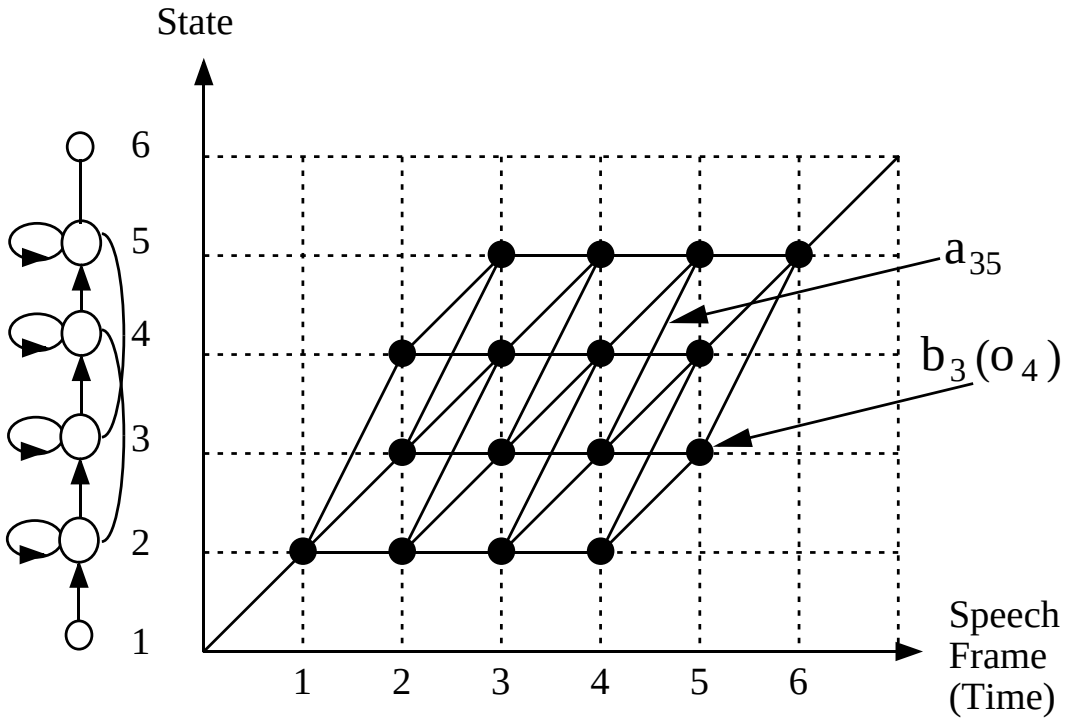


Figure 1.6 Trellis diagram for an Isolated Word Recognition task. From (Young et al. 1997), p. 10.

Problem 2: Viterbi algorithm

1. Initially,
   $\delta_1(i) = \pi_i b_i(o_1)$
   $\psi_1(i) = 0$ for $1 \leq i \leq N$;

2. For $t = 2, \ldots, T$,
   $\delta_t(j) = \max_i \left[\delta_{t-1}(i) a_{ij}\right] b_j(o_t)$
   $\psi_t(j) = \arg\max_i \left[\delta_{t-1}(i) a_{ij}\right]$ for $1 \leq j \leq N$;

3. Finally,
   $\Delta^* = \max_i [\delta_T(i)]$
   $x_T^* = \arg\max_i [\delta_T(i)]$;

4. Trace back, for $t = T-1, T-2, \ldots, 1$,

$$x_t^* = \psi_{t+1}\left(x_{t+1}^*\right), \quad \text{and} \quad X^* = \{x_1^*, x_2^*, \ldots, x_T^*\}. \tag{3}$$

Problem 3: Re-estimation of λ

Using the Viterbi algorithm (eq. 3), we compute the likelihood of the best path through the trellis:

$$\begin{aligned} \Delta^* = \Delta(X^*) &= P(\mathcal{O}, X^*|\lambda) \\ &\approx P(\mathcal{O}|\lambda), \end{aligned} \tag{4}$$

which approximates the total likelihood of the observations considering all paths $\sum_{\{\text{all } X\}} P(\mathcal{O}, X|\lambda)$.

We could use the occupation of states, and transitions between them, to train the parameters of our model $\lambda = \{\pi, A, B\}$.

Viterbi re-estimation, aka. Segmental K-means

Hence:

(a) Initial-state probabilities,

$$\widehat{\pi}_i = q_1(i) \qquad \text{for } 1 \le i \le N;$$

where state indicator $q_t(i) = \begin{cases} 1 & \text{for } i = x_t; \\ 0 & \text{otherwise.} \end{cases}$

(b) State-transition probabilities,

$$\widehat{a}_{ij} = \frac{\sum_{t=2}^{T} q_{t-1}(i) q_t(j)}{\sum_{t=2}^{T} q_{t-1}(i)} \qquad \text{for } 1 \le i, j \le N;$$

(c) Discrete output probabilities,

$$\widehat{b}_j(k) = \frac{\sum_{t=1}^{T} \omega_t(k) q_t(j)}{\sum_{t=1}^{T} q_t(j)} \qquad \text{for } 1 \le j \le N;$$

and $1 \le k \le K$.

where event indicator $\omega_t(k) = \begin{cases} 1 & \text{for } k = o_t; \\ 0 & \text{otherwise.} \end{cases}$

Viterbi re-estimation (multiple files)

Using a set of training examples, $r \in \{1, \ldots, R\}$:

(a) Initial-state probabilities,

$$\widehat{\pi}_i = \frac{1}{R} \sum_{r=1}^{R} q_1^r(i) \qquad \text{for } 1 \leq i \leq N;$$

(b) State-transition probabilities,

$$\widehat{a}_{ij} = \frac{\sum_{r=1}^{R} \sum_{t=2}^{T} q_{t-1}^r(i) q_t^r(j)}{\sum_{r=1}^{R} \sum_{t=2}^{T} q_{t-1}^r(i)} \qquad \text{for } 1 \leq i, j \leq N;$$

(c) Discrete output probabilities,

$$\widehat{b}_j(k) = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T} \omega_t^r(k) q_t(j)}{\sum_{r=1}^{R} \sum_{t=1}^{T} q_t^r(j)} \qquad \text{for } 1 \leq j \leq N;$$

and $1 \leq k \leq K$.

Maximum likelihood re-estimation

Parameters of the model $\lambda = \{\pi, A, B\}$ are adjusted to maximise $P(\mathcal{O}|\lambda)$, which is as in eq. 4 from Tutorial 2:

$$P(\mathcal{O}|\lambda) = \sum_{X} P(\mathcal{O}, X|\lambda). \tag{5}$$

We can perform this maximisation according to the ML criterion, which considers all possible state sequences, by using the *Baum-Welch* formulae.

Baum-Welch re-estimation (occupation)

Consider

$$\gamma_t(i) \;=\; P\left(x_t = i|\mathcal{O},\lambda\right). \tag{6}$$

By Bayes law,

$$\begin{aligned}\gamma_t(i) \;&=\; \frac{P\left(x_t = i,\mathcal{O}|\lambda\right)}{P(\mathcal{O}|\lambda)}\\ &=\; \frac{\alpha_t(i)\,\beta_t(i)}{P(\mathcal{O}|\lambda)},\end{aligned} \tag{7}$$

where α_t and β_t are computed using the forward and backward procedures, as in eqs. 1 & 2, and the solution to Problem 1 gives $P(\mathcal{O}|\lambda)$.

Baum-Welch re-estimation (transition)

Define

$$\xi_t(i,j) \;=\; P\left(x_{t-1}=i, x_t=j|\mathcal{O},\lambda\right). \tag{8}$$

Similarly, by Bayes law,

$$\begin{aligned}
\xi_t(i,j) \;&=\; \frac{P\left(x_{t-1}=i, x_t=j, \mathcal{O}|\lambda\right)}{P(\mathcal{O}|\lambda)} \\
&=\; [\, P\left(x_{t-1}=i, o_1, \ldots, o_{t-1}|\lambda\right) \times \\
&\qquad P\left(x_t=j, o_t, \ldots, o_T|x_{t-1}=i, \lambda\right)] \,/P(\mathcal{O}|\lambda) \\
&=\; [\, \alpha_{t-1}(i)\; P\left(x_t=j, o_t|x_{t-1}=i, \lambda\right) \times \\
&\qquad P\Big(o_{t+1}, \ldots, o_T|x_t=j, \lambda\Big)] \,/P(\mathcal{O}|\lambda) \\
&=\; \frac{\alpha_{t-1}(i)\; a_{ij}\, b_j(o_t)\; \beta_t(j)}{P(\mathcal{O}|\lambda)}.
\end{aligned} \tag{9}$$

Baum-Welch re-estimation formulae

(a) Initial-state probabilities,

$$\hat{\pi}_i = \gamma_1(i) \qquad \text{for } 1 \leq i \leq N;$$

(b) State-transition probabilities,

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} \xi_t(i,j)}{\sum_{t=2}^{T} \gamma_{t-1}(i)} \qquad \text{for } 1 \leq i, j \leq N;$$

(c) Discrete output probabilities,

$$\hat{b}_j(k) = \frac{\sum_{t=1}^{T}\Big|_{o_t=k} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \qquad \text{for } 1 \leq j \leq N;$$

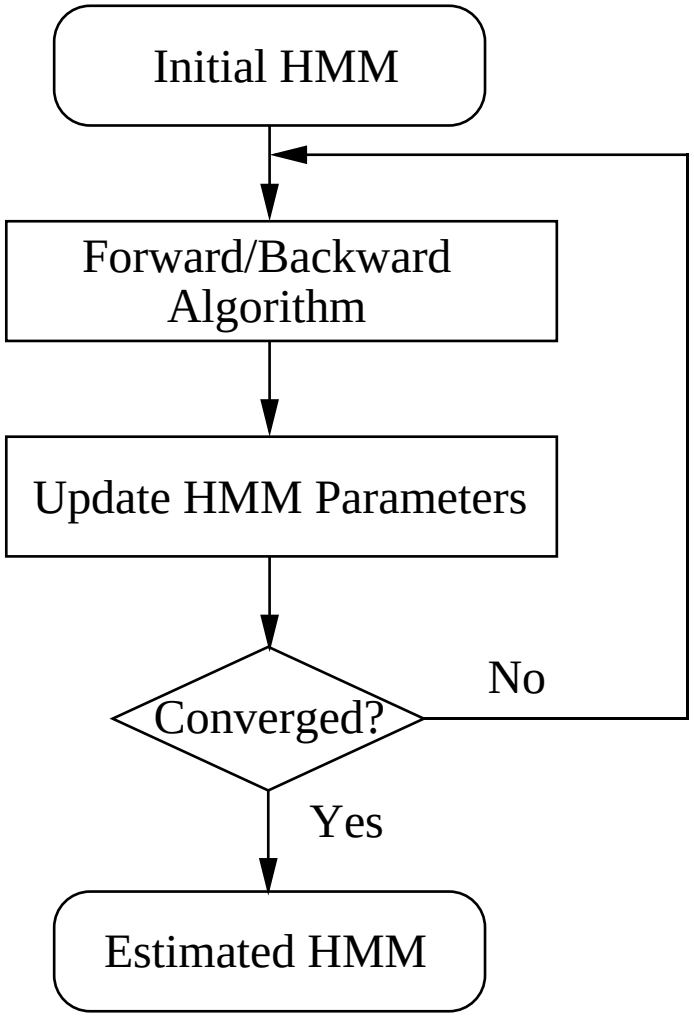
and $1 \leq k \leq K$.

For the new model $\hat{\lambda}$, it can be shown that,

$$P(\mathcal{O}|\hat{\lambda}) \geq P(\mathcal{O}|\lambda), \tag{10}$$

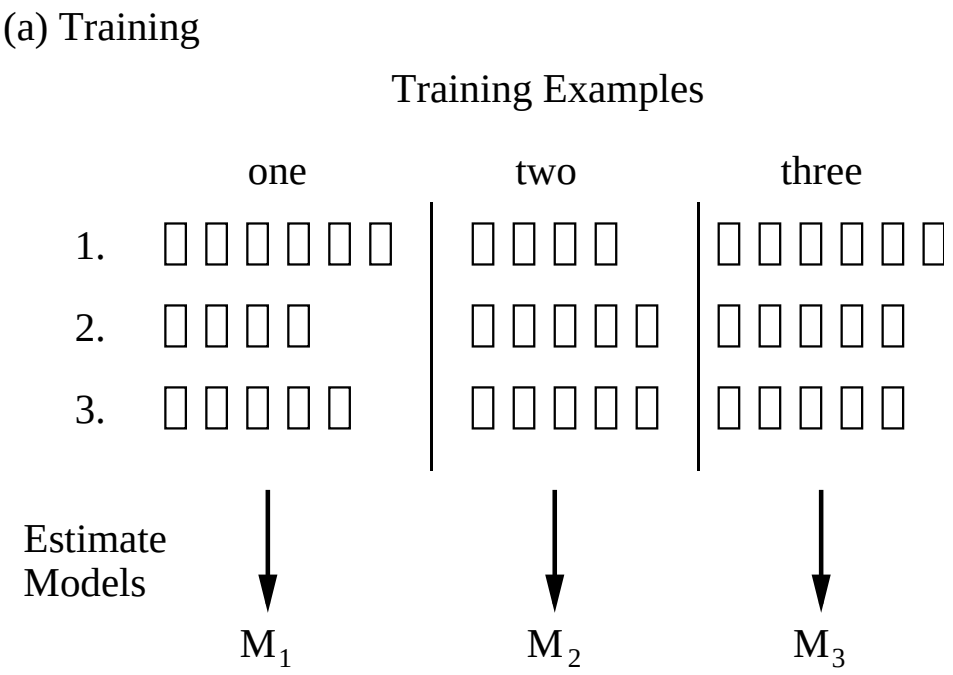
although it does not guarantee a *global* maximum.

Re-estimation process



Parameter re-estimation based on a single training example.*

Training and test procedures



Training and recognition using HMMs for an IWR task.*

Tutorial 3 summary

- Recap. of likelihoods α_t and β_t
- Recap. of Viterbi algorithm
- Re-estimating models, $\Lambda = \{\lambda\}$
  - Occupation and transition
  - Baum-Welch formulae

To follow

- Worked example illustrating Baum-Welch algorithm
- Output pdfs: univariate Gaussian, multivariate Gaussian, Gaussian mixtures, etc.