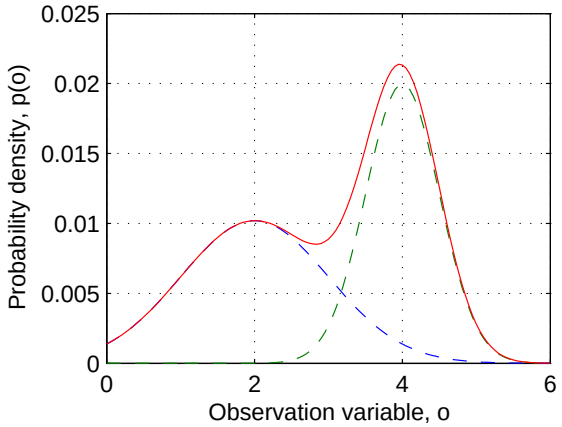
UNIVERSITY OF SURREY
Centre for Vision Speech & Signal Processing
University of Surrey, Guildford GU2 7XH.

HMM part 5

Dr Philip Jackson

- General output pdfs
  - Alternative pdfs
  - Gaussian mixtures
  - Re-estimation with mixtures
- Implementing B-W formulae
  - Review of α, β, γ and ξ
  - Fwd/backward procedures
  - Accumulators & update
- Practical training & recognition



Parametric output pdfs

- For the output or emission probabilities, $B = \{b_i\}$, we have so far considered:

  1. Discrete observations, $b_i(o_t) = P(o_t = k)$

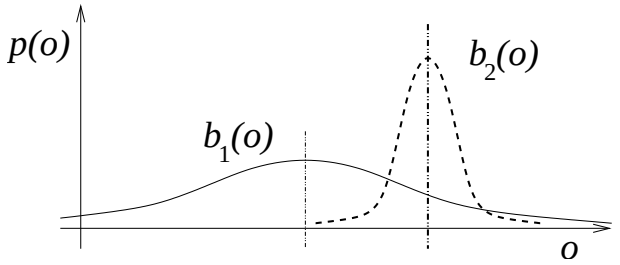
  2. Normally-distributed continuous observations, $b_i(\boldsymbol{o}_t) = p(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

- What if our speech features do not fit either of these assumptions?

Univariate Gaussian (scalar observations)

For a given state i,

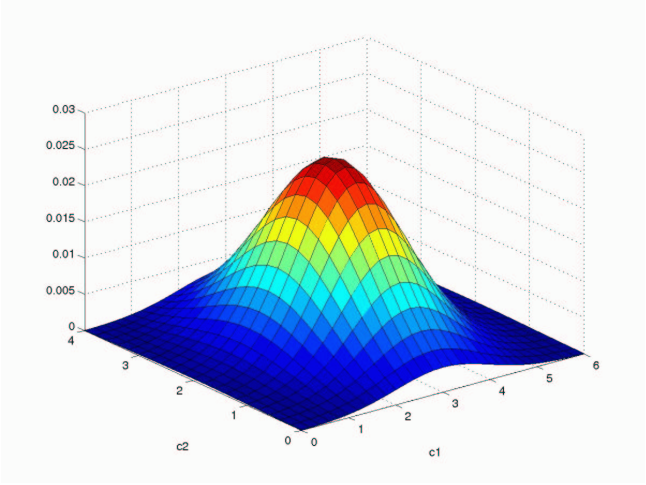
$$b_i(o_t) = \frac{1}{\sqrt{2\pi\Sigma_i}} \exp\left[\frac{-(o_t - \mu_i)^2}{2\Sigma_i}\right]$$



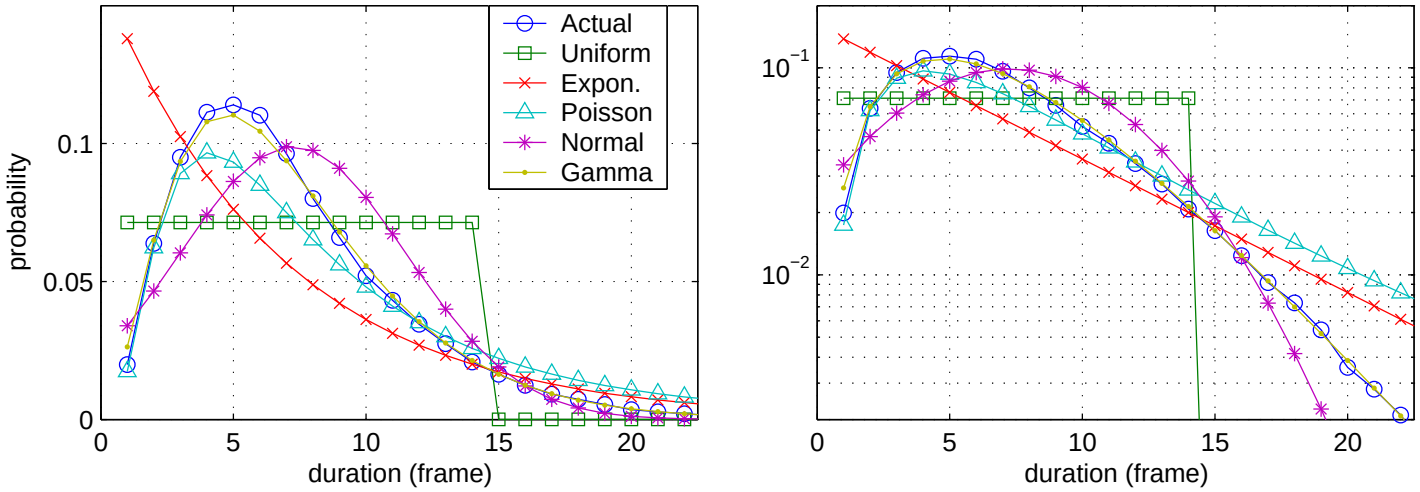
Multivariate Gaussian (vector observations)

$$b_i(\boldsymbol{o}_t) = \frac{1}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma}_i|}} \exp\left[-\frac{1}{2}(\boldsymbol{o}_t - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{o}_t - \boldsymbol{\mu}_i)^{\mathsf{T}}\right]$$

where the dimensionality of the observation space is K.



Alternative distributions



Various functions on (left) linear and (right) logarithmic scales.

Many alternatives to Gaussian (normal) distribution exist:

- Exponential
- Poisson
- Gamma
- ...
- Log-normal
- Ricean
- Rayleigh
- Beta
- student-t
- Cauchy

Gaussian mixture pdfs

Univariate Gaussian mixture

$$b_i(o_t) = \sum_{m=1}^{M} c_{im} \mathcal{N}(o_t; \mu_{im}, \Sigma_{im}) \quad (1)$$

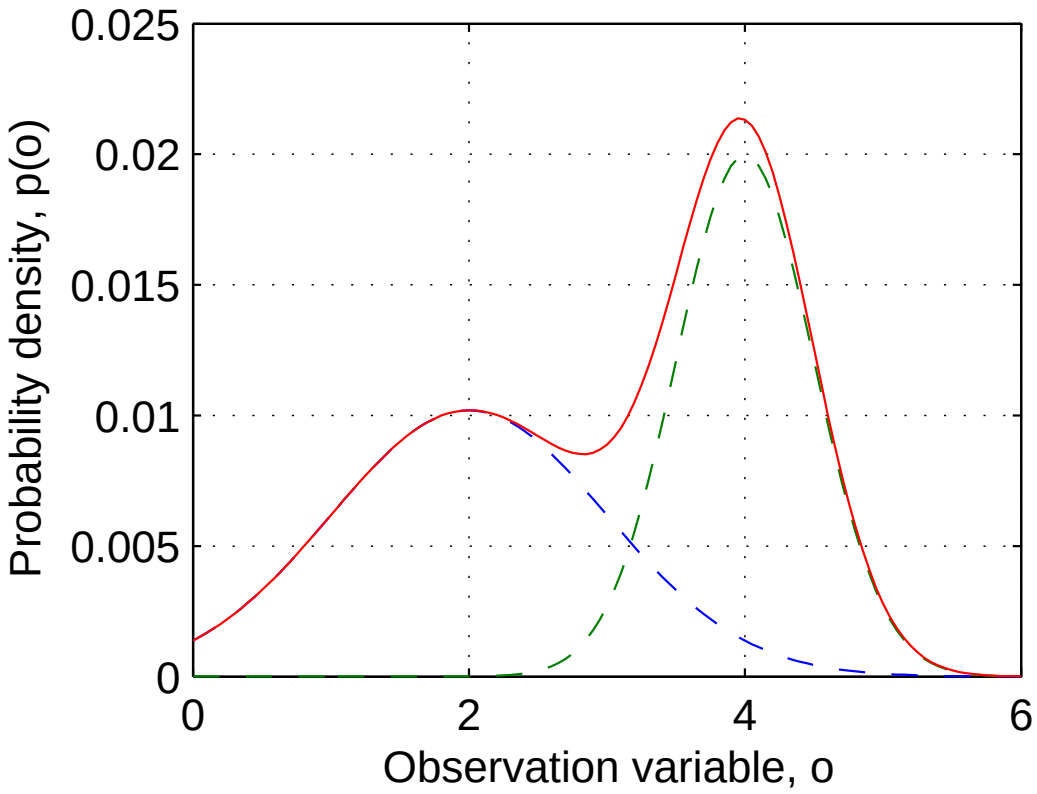
where M is the number of mixture components (M-mix), and the mixture weights sum to one: $\sum_{m=1}^{M} c_{im} = 1$

Multivariate Gaussian mixture

$$b_i(\boldsymbol{o}_t) = \sum_{m=1}^{M} c_{im} \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im}) \quad (2)$$

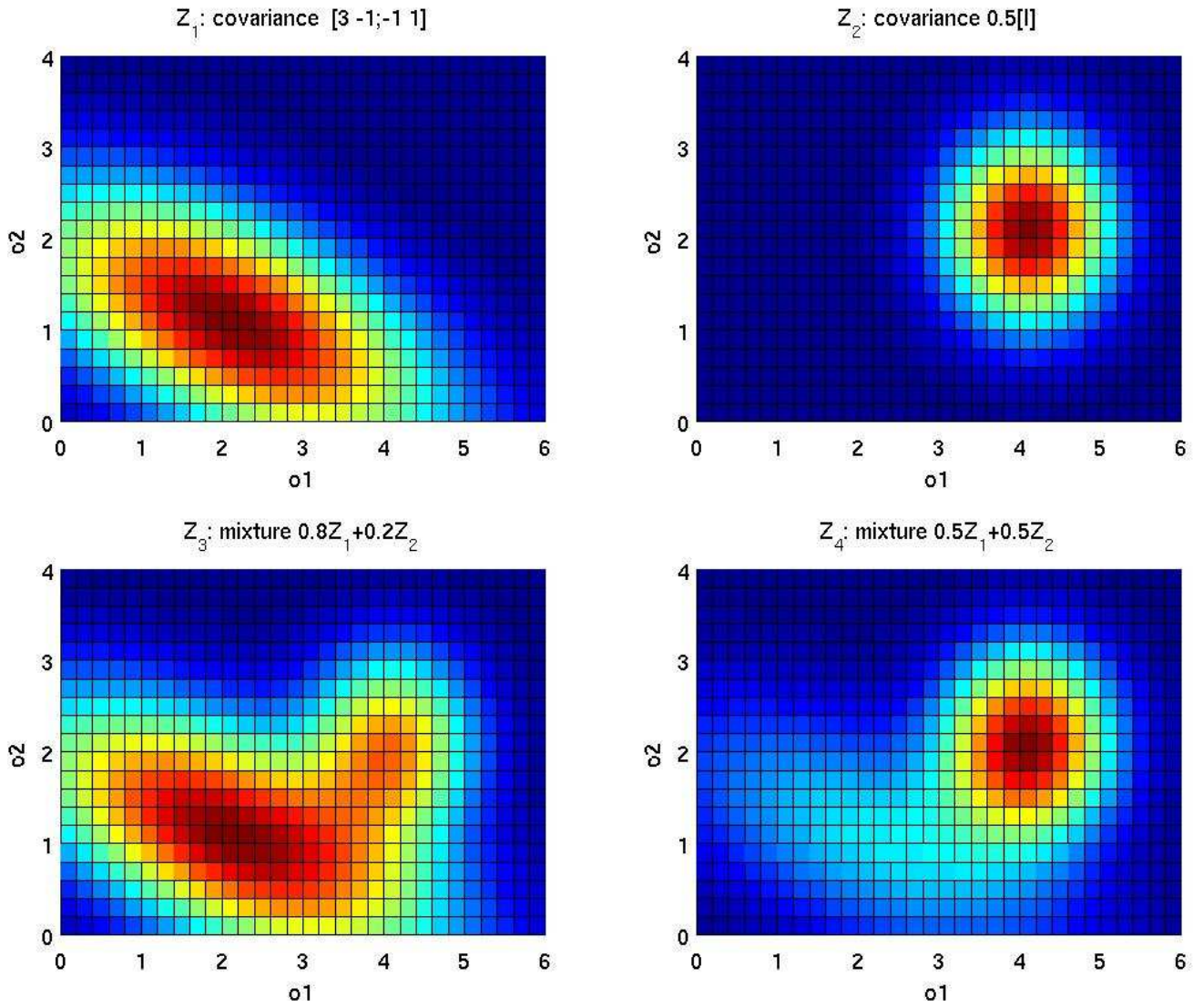
where $\mathcal{N}(\cdot)$ is the multivariate normal distribution with vector mean $\boldsymbol{\mu}_{im}$ and covariance matrix $\boldsymbol{\Sigma}_{im}$, evaluated at $\boldsymbol{o}_t$

Univariate mixture of Gaussians



Mixture of two univariate Gaussian components

Multivariate mixtures of Gaussians



Examples of two bivariate (2D) Gaussian pdfs (upper), two Gaussian mixtures with different weights (lower)

Viterbi training with mixtures

Mean: $$\hat{\boldsymbol{\mu}}_{jm} = \frac{\sum_{t=1}^{T} q_t(j,m)\boldsymbol{o}_t}{\sum_{t=1}^{T} q_t(j,m)}$$

Covariance: $$\hat{\boldsymbol{\Sigma}}_{jm} = \frac{\sum_{t=1}^{T} q_t(j,m)(\boldsymbol{o}_t-\boldsymbol{\mu}_{jm})(\boldsymbol{o}_t-\boldsymbol{\mu}_{jm})^{\mathsf{T}}}{\sum_{t=1}^{T} q_t(j,m)}$$

Weights: $$\hat{c}_{jm} = \frac{\sum_{t=1}^{T} q_t(j,m)}{\sum_{t=1}^{T} q_t(j)}$$

where $q_t(j,m)$ and $q_t(j)$ are occupation indicators:

$$q_t(j) = \begin{cases} 1 & \text{for } j = x_t \\ 0 & \text{otherwise} \end{cases}$$

$$q_t(j,m) = \begin{cases} 1 & \text{for } j = x_t,\ m = \hat{m} \\ 0 & \text{otherwise} \end{cases}$$

and the occupied mixture is

$$\hat{m} = \arg\max_m \left[c_{jm}\, \mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}\right)\right]$$

Baum-Welch training with mixtures

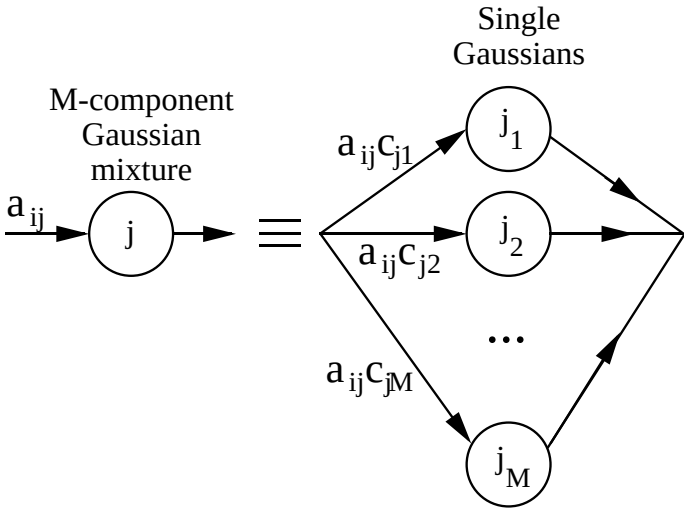
We define the **mixture-occupation likelihood**:

$$\gamma_t(j,m) = \frac{\alpha_t(j,m)\ \beta_t(j)}{p(\mathcal{O}|\lambda)} \tag{3}$$

where $\gamma_t(j) = \sum_{m=1}^{M} \gamma_t(j,m)$

$$\alpha_t(j,m) = \begin{cases} \pi_j\ , b_{jm}(\boldsymbol{o}_t) & \text{for } t{=}1 \\ \sum_{i=1}^{N} \alpha_{t-1}(i)\, a_{ij}\ b_{jm}(\boldsymbol{o}_t) & \text{otherwise} \end{cases}$$

$$b_{j,m}(\boldsymbol{o}_t) = c_{jm}\,\mathcal{N}\left(\boldsymbol{o}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}\right)$$



Occupations within a mixture considering the output from each component (Young et al., 2002)

Baum-Welch re-estimation of mixture parameters

Using soft assignment of observations to mixture components given by mixture-occupation likelihood $\gamma_t(j,m)$, we train our parameters with revised update equations.

Mean vector:

$$\hat{\boldsymbol{\mu}}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j,m)\boldsymbol{o}_t}{\sum_{t=1}^{T} \gamma_t(j,m)} \tag{4}$$

Covariance matrix:

$$\hat{\boldsymbol{\Sigma}}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j,m)(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})(\boldsymbol{o}_t - \boldsymbol{\mu}_{jm})^{\mathsf{T}}}{\sum_{t=1}^{T} \gamma_t(j,m)} \tag{5}$$

Mixture weights:

$$\hat{c}_{jm} = \frac{\sum_{t=1}^{T} \gamma_t(j,m)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{6}$$

Implementing Baum-Welch re-estimation

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

(a) State-transition probabilities,

$$\widehat{a}_{ij} = \frac{\sum_{r=1}^{R} \sum_{t=2}^{T} \xi_t^r(i,j)}{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_t^r(i)} = \frac{a_{ij}}{\bar{a}_i} \qquad \text{for } 1 \leq i, j \leq N;$$

(b) Gaussian output probabilities,

$$\widehat{b}_i(\boldsymbol{o}_t) \begin{cases} \widehat{\boldsymbol{\mu}}_i = \dfrac{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_t^r(i) \mathbf{o}_t^r}{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_t^r(i)} = \dfrac{\underline{\boldsymbol{\mu}}_i}{\bar{b}_i} \\ \widehat{\boldsymbol{\Sigma}}_i = \dfrac{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_t^r(i) (\mathbf{o}_t^r - \boldsymbol{\mu}_i)(\mathbf{o}_t^r - \boldsymbol{\mu}_i)^{\mathsf{T}}}{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_t^r(i)} = \dfrac{\underline{\boldsymbol{\Sigma}}_i}{\bar{b}_i} \end{cases}$$

These updated parameters define our re-estimated model $\widehat{\lambda} = \{\widehat{A}, \widehat{B}\}$.

Re-estimation procedure (one training iteration)

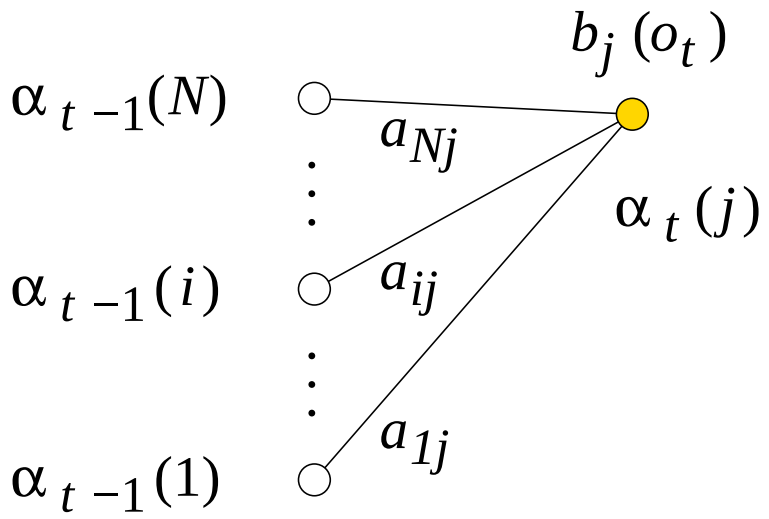
To update the models using multiple training files:

- **Forward**: compute likelihoods $\alpha_t^r(i)$
- **Backward**: compute likelihoods $\beta_t^r(i)$
- **Parallel**:
  - compute occupation $\gamma_t^r(i)$
  - compute transition $\xi_t^r(i,j)$
  - accumulate $\underline{a}_{ij}$ and $\bar{a}_i$
  - accumulate $\underline{\boldsymbol{\mu}}_i$, $\underline{\boldsymbol{\Sigma}}_i$ and $\bar{b}_i$
- **Repeat** for all files in the training set, $r \in \{1, 2, \ldots, R\}$.

Forward likelihood

The forward likelihood is the joint probability,

$$\alpha_t(i) = P\left(x_t = i, \mathbf{o}_1^t | \lambda\right) \tag{7}$$

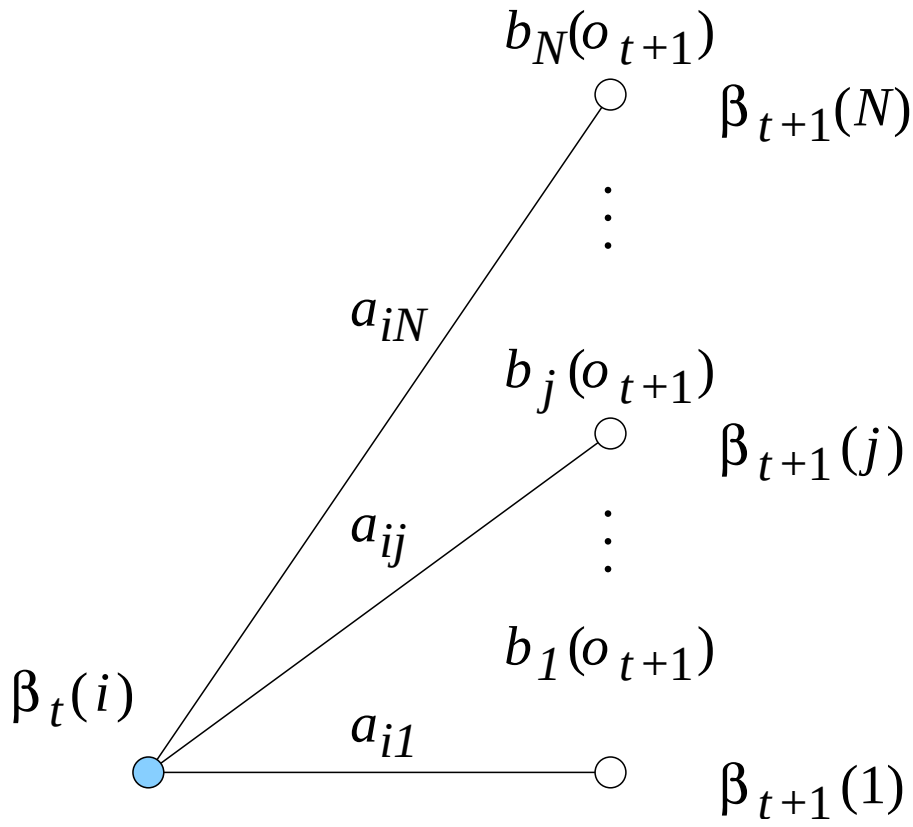


Trellis fragment depiction of forward likelihood.

Backward likelihood

The backward likelihood is the conditional probability,

$$\beta_t(j) = P\left(\mathbf{o}_{t+1}^T | x_t = j, \lambda\right) \tag{8}$$

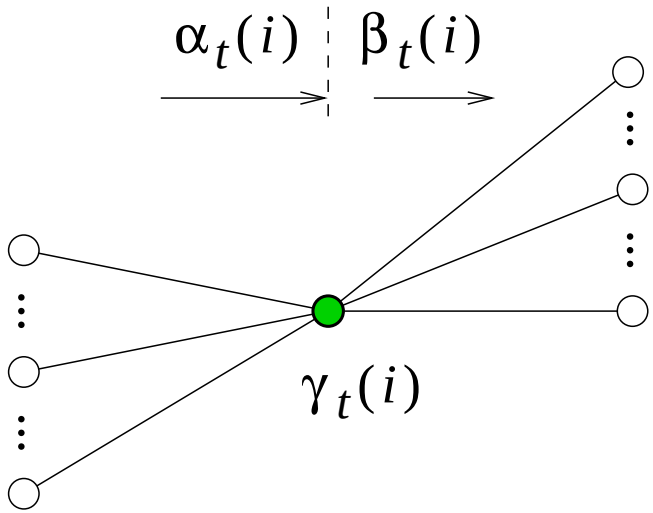


Trellis fragment depiction of backward likelihood.

Occupation likelihood

The occupation likelihood is the conditional probability,

$$\begin{aligned} \gamma_t(i) &= P\left(x_t = i|\mathcal{O}, \lambda\right) \\ &= \frac{\alpha_t(i)\,\beta_t(i)}{P\left(\mathcal{O}|\lambda\right)} \end{aligned} \tag{9}$$

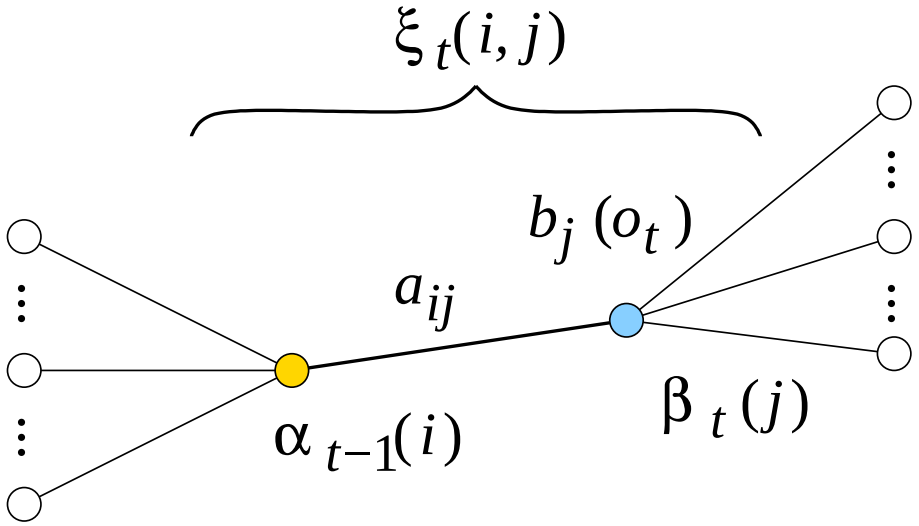


Trellis fragment depiction of occupation likelihood.

Transition likelihood

The transition likelihood is the joint probability of states i and j, conditioned on the observations,

$$\begin{aligned} \xi_t(i,j) &= P\left(x_{t-1}=i, x_t=j|\mathcal{O},\lambda\right) \\ &= \frac{\alpha_{t-1}(i)\ a_{ij}\, b_j(\boldsymbol{o}_t)\ \beta_t(j)}{P\left(\mathcal{O}|\lambda\right)} \end{aligned} \tag{10}$$



Trellis fragment depiction of transition likelihood.

Forward procedure (Task 1)

1. Initially, for the rth file,
   $\alpha_1^r(i) = \pi_i \, b_i(\boldsymbol{o}_1^r)$, for $1 \leq i \leq N$;

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

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

Backward procedure (Task 1)

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

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

3. Finally, we get the same
   $P^r \;=\; \sum_{i=1}^N \pi_i \, b_i(\boldsymbol{o}_1^r) \, \beta_1^r(i)$.

Parallel procedure

Occupation likelihoods:

$$\gamma_t^r(i) = \frac{\alpha_t^r(i)\,\beta_t^r(i)}{P^r}$$

Transition likelihoods:

$$\xi_t^r(i,j) = \frac{\alpha_{t-1}^r(i)\;\; a_{ij}\, b_j(\boldsymbol{o}_t^r)\;\; \beta_t^r(j)}{P^r}$$

Transition accumulators:

$$\underline{a}_{ij} \;\; \mapsto \;\; \underline{a}_{ij} + \sum_{t=2}^{T} \xi_t^r(i,j)$$

$$\bar{a}_i \;\; \mapsto \;\; \bar{a}_i + \sum_{t=1}^{T} \gamma_t^r(i)$$

Parallel (continued)

Output accumulators:

$$\underline{\boldsymbol{\mu}}_i \;\mapsto\; \underline{\boldsymbol{\mu}}_i + \sum_{t=1}^{T} \gamma_t^r(i)\boldsymbol{o}_t^r$$

$$\underline{\boldsymbol{\Sigma}}_i \;\mapsto\; \underline{\boldsymbol{\Sigma}}_i + \sum_{t=1}^{T} \gamma_t^r(i)(\boldsymbol{o}_t^r - \boldsymbol{\mu}_i)(\boldsymbol{o}_t^r - \boldsymbol{\mu}_i)^{\mathsf{T}}$$

$$\bar{b}_i \;\mapsto\; \bar{b}_i + \sum_{t=1}^{T} \gamma_t^r(i)$$

Repeat

For all files in the training set:

1. recompute the forward and backward likelihoods
2. recompute the occupation and transition likelihoods
3. increment the transition and output accumulators

Update (Task 3)

Finally, we update the models:

$$\widehat{a}_{ij} = \frac{a_{ij}}{\overline{a}_i}$$

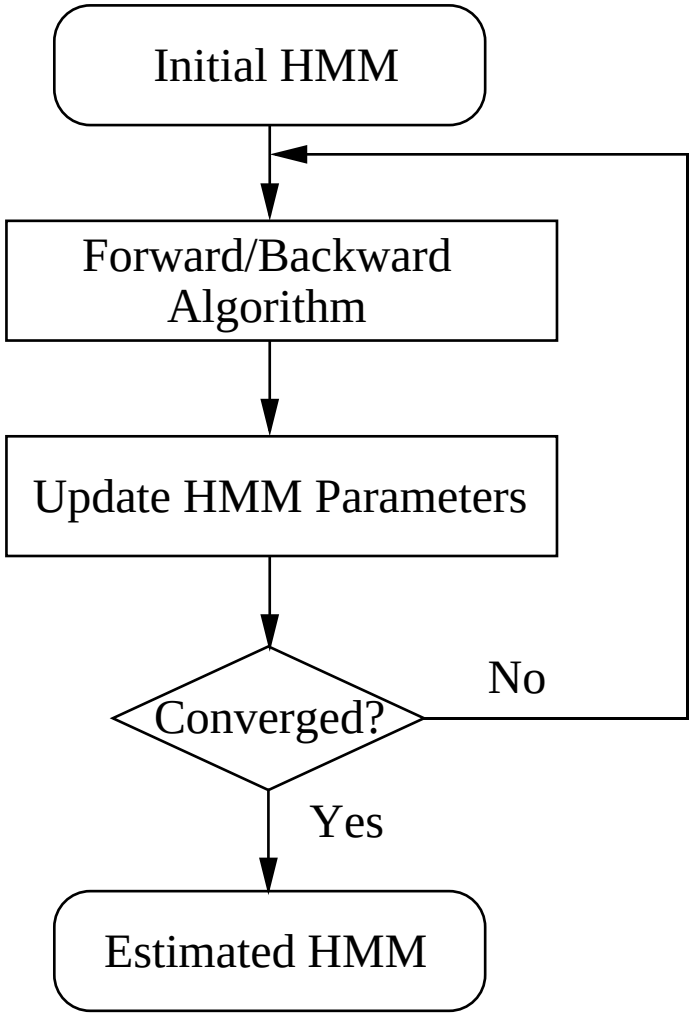
$$\widehat{b}_i \begin{cases} \widehat{\boldsymbol{\mu}}_i = \dfrac{\overline{\boldsymbol{\mu}}_i}{\overline{b}_i} \\ \\ \widehat{\boldsymbol{\Sigma}}_i = \dfrac{\overline{\boldsymbol{\Sigma}}_i}{\overline{b}_i} \end{cases}$$

where $\widehat{}$ denotes the re-estimated values of the model, $\widehat{\lambda}$

Procedure for one iteration of training

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
5. use occupation & transition likelihoods to increment accumulators
6. repeat from step 2 until all utterances processed
7. use accumulators to update parameters for all HMMs

Overview of training process



Procedure for re-estimating the parameters of an HMM (Young et al., 2002).

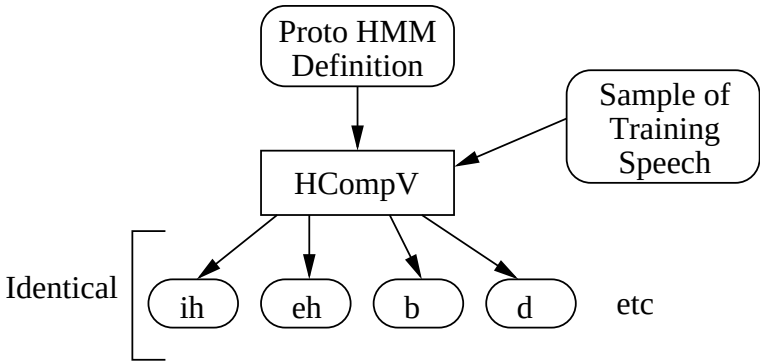
Practical issues

Decoding

- Probabilities stored as log probs to avoid underflow
- Paths propagated through trellis by token passing
- Search space kept tractable by beam pruning

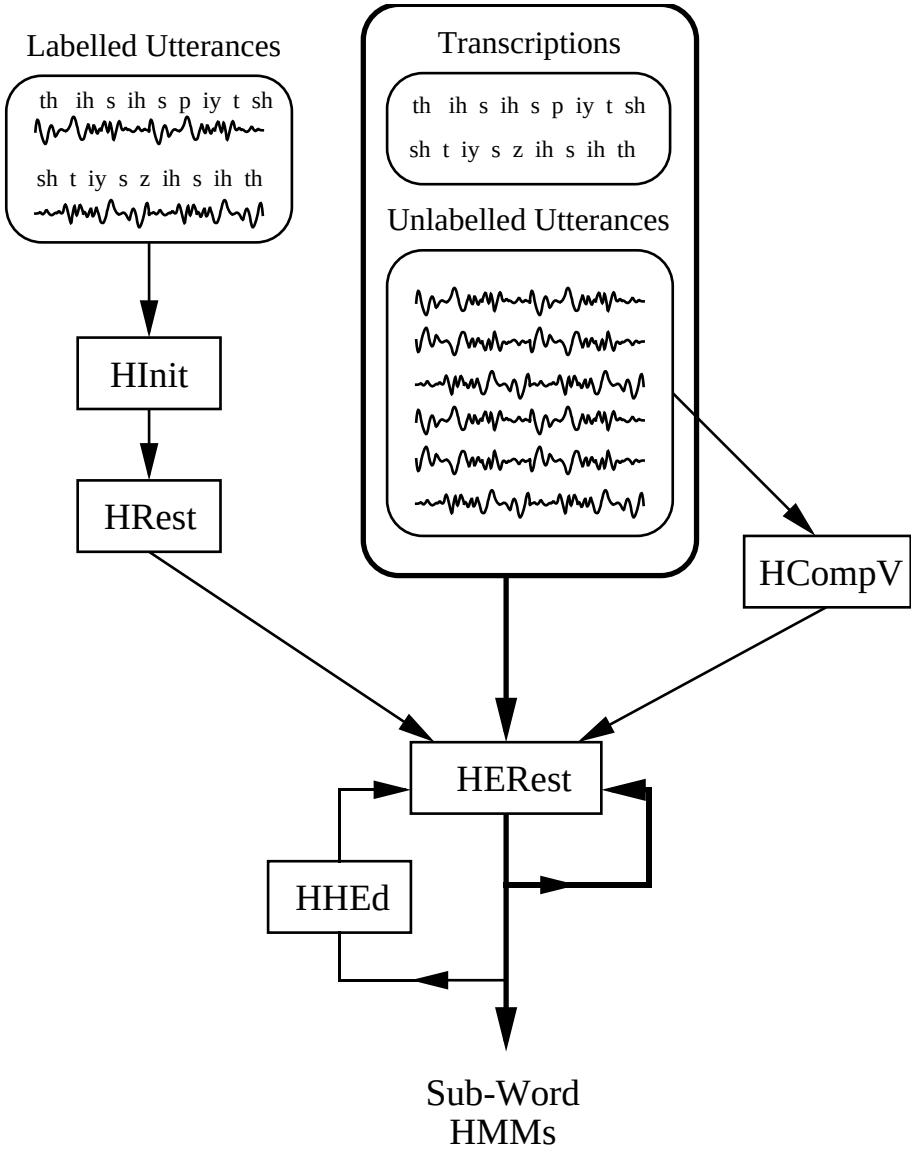
Model initialisation

1. Random
2. Flat start
3. Viterbi alignment (supervised/ unsupervised)



Flat start (Young et al., 2002)

Re-estimation and embedded re-estimation



HMM training with variously labelled data (Young et al., 2002)

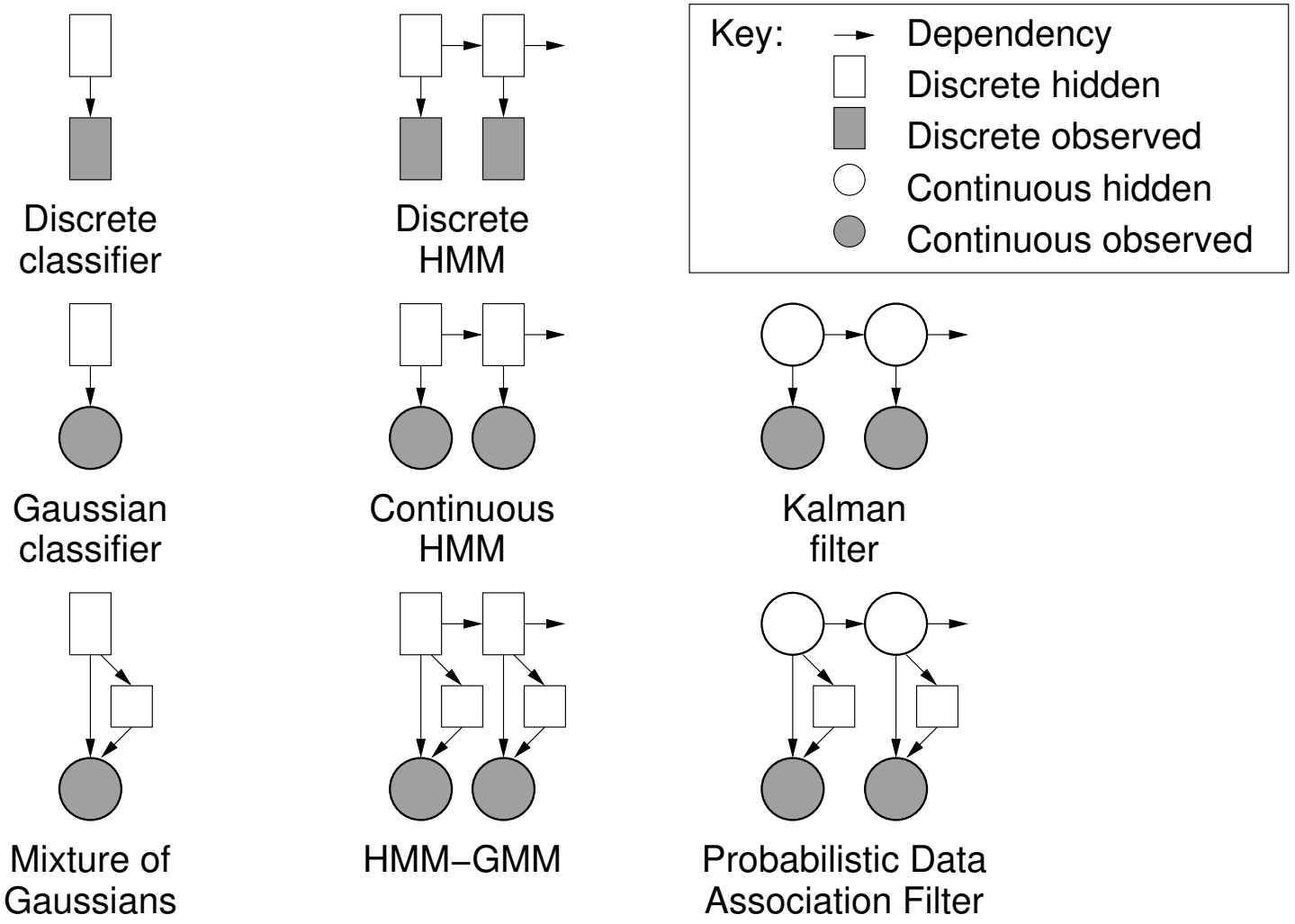
Number of parameters & Regularisation

- Context sensitivity
- Size of database
- Parsimonious models
  - Occam's razor
- variance floor
- parameter tying
  - agglomerative clustering
  - decision trees

Related topics for further reading

- Language modeling
- Noise robustness
  - factorial HMMs
  - loosely-coupled HMMs
- Model adaptation
  - MLLR/MAP
- Discriminative training
- Probabilistic modeling
  - Graphical model
  - Bayesian network
  - Markov random field
  - Markov decision process
  - Markov chain Monte Carlo

Relationship of HMM to other estimation methods



Conditional dependencies for various probabilistic methods, drawn using graphical model notation

Part 5 summary

- General output pdfs
  - Alternative distribution functions
  - Multivariate Gaussian mixtures
- Implementation of B-W formulae:
  - review of likelihoods α_t, β_t, γ_t and ξ_t
  - forward, backward, accumulate & update
- Practical issues
  - Annotations, initialisation & tying
  - Exploiting the training data
  - Computationally efficient recognition

References

B. Gold, N. Morgan & D. Ellis, *Speech and Audio Signal Processing*, New York: Wiley-Blackwell, 2nd ed., 2011 [ISBN-13: 978-0470195369].

J. N. Holmes & W. J. Holmes, *Speech Synthesis and Recognition*, CRC Press, 2nd ed., 2001 [ISBN-13: 978-0748408573].

F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, 1998 [ISBN-13: 978-0262100663].

D. Jurafsky & J. H. Martin, *Speech and Language Processing*, Prentice Hall, 2nd ed., 2003 [ISBN-13: 978-0131873216]

L. R. Rabiner. *A tutorial on HMM and selected applications in speech recognition*. In *Proc. IEEE*, Vol. 77, No. 2, pp. 257–286, 1989.

S. J. Young, et al., *The HTK Book*, Cambridge Univ. Eng. Dept. (v3.4), 2009 [http://htk.eng.cam.ac.uk/].