Automatic Speech Recognition (CS753)

Lecture 13: Assignment 1 + Revision

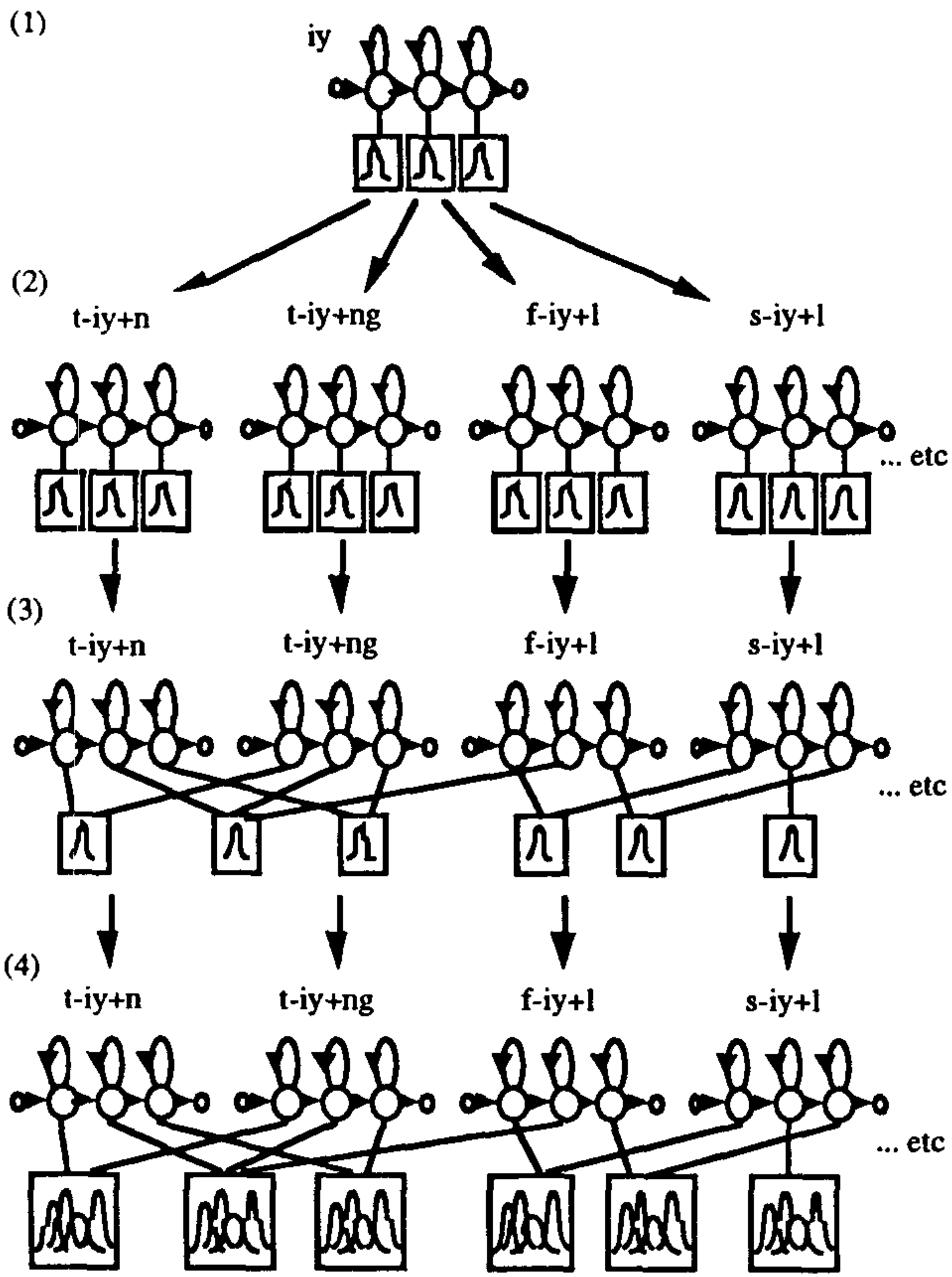
Instructor: Preethi Jyothi
Feb 16, 2017

Assignment 1 Solutions

https://www.cse.iitb.ac.in/~pjyothi/cs753/assgmt1_soln.pdf

Revising Tied State HMMs

Tied state HMMs

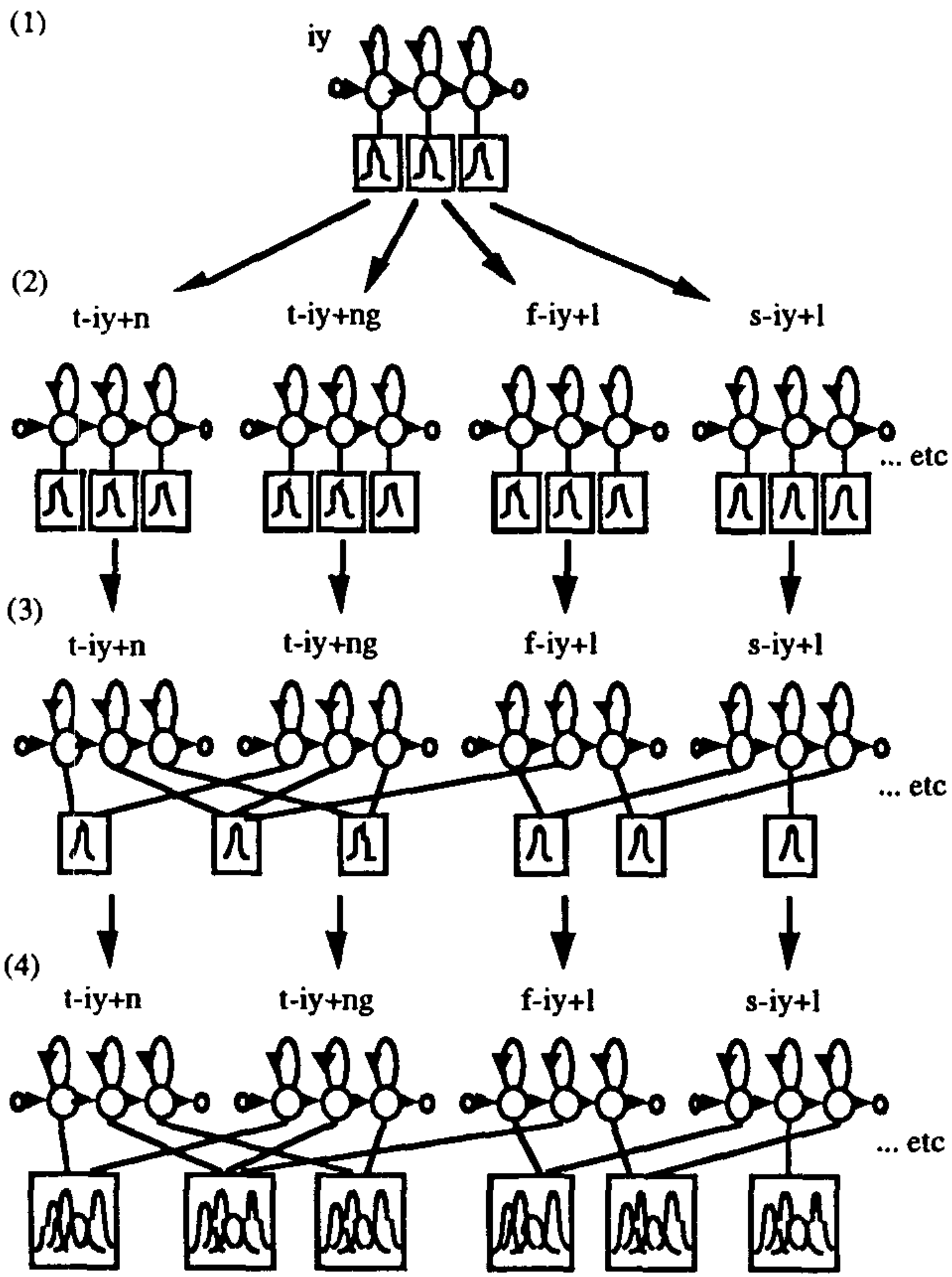


Four main steps in building a tied state HMM system:

1. Create and train 3-state monophone HMMs with single Gaussian observation probability densities
2. Clone these monophone distributions to initialise a set of untied triphone models. Train them using Baum-Welch estimation. Transition matrix remains common across all triphones of each phone.
3. For all triphones derived from the same monophone, cluster states whose parameters should be tied together.
4. Number of mixture components in each tied state is increased and models re-estimated using BW

Image from: Young et al., "Tree-based state tying for high accuracy acoustic modeling", ACL-HLT, 1994

Tied state HMMs



Four main steps in building a tied state HMM system:

1. Create and train 3-state monophone HMMs with single Gaussian observation probability densities
2. Clone these monophone distributions to initialise a set of untied triphone models. Train them using Baum-Welch estimation. Transition matrix remains common across all triphones of each phone.
3. For all triphones derived from the same monophone, cluster states whose parameters should be tied toge
4. Num in each mod

Which states should be tied together? Use decision trees.

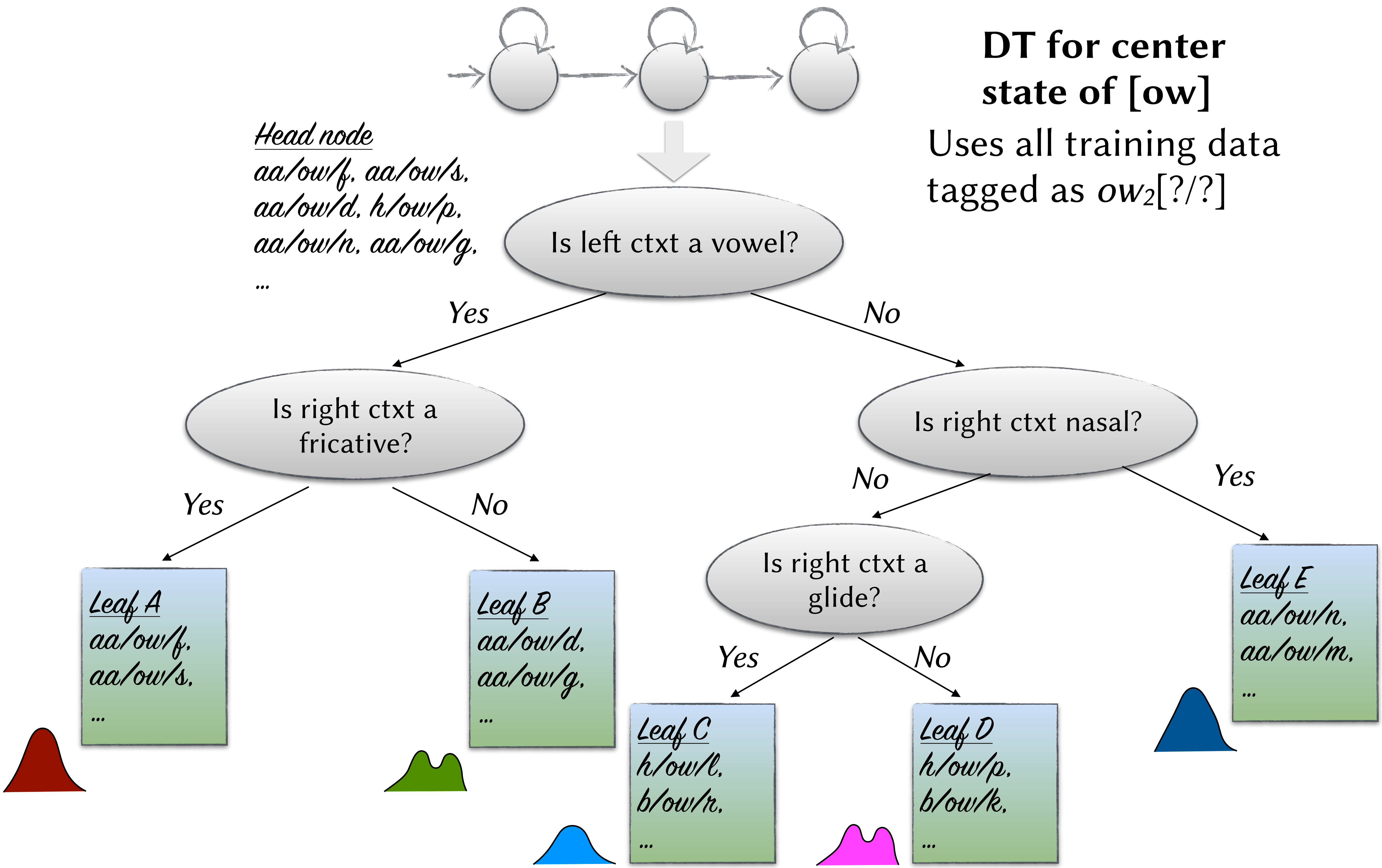
Image from: Young et al., "Tree-based state tying for high accuracy acoustic modeling", ACL-HLT, 1994

Phonetic Decision Trees (DT)

One tree is constructed for each state of each phone to cluster all the corresponding triphone states

DT for center state of [ow]

Uses all training data tagged as ow_2[?/?]



How do we build these phone DTs?

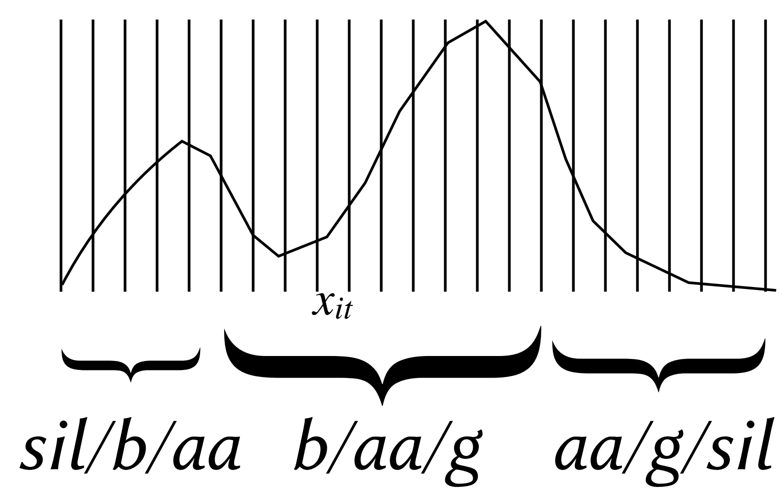
1. What questions are used?

   Linguistically-inspired binary questions: “Does the left or right phone come from a broad class of phones such as vowels, stops, etc.?” “Is the left or right phone [k] or [m]?”

2. What is the training data for each phone state, p_j? (root node of DT)

Training data for DT nodes

- Align training data, $x_i = (x_{i1}, \ldots, x_{iT_i})$ $i=1\ldots N$ where $x_{it} \in \mathbb{R}^d$, against a set of triphone HMMs
- Use Viterbi algorithm to find the best HMM state sequence corresponding to each x_i
- Tag each x_{it} with ID of current phone along with left-context and right-context



x_{it} is tagged with ID aa_2[b/g] i.e. x_{it} is aligned with the second state of the 3-state HMM corresponding to the triphone b/aa/g

- For a state j in phone p, collect all x_{it}'s that are tagged with ID p_j[?/?]

How do we build these phone DTs?

1. What questions are used?

   Linguistically-inspired binary questions: "Does the left or right phone come from a broad class of phones such as vowels, stops, etc.?" "Is the left or right phone [k] or [m]?"

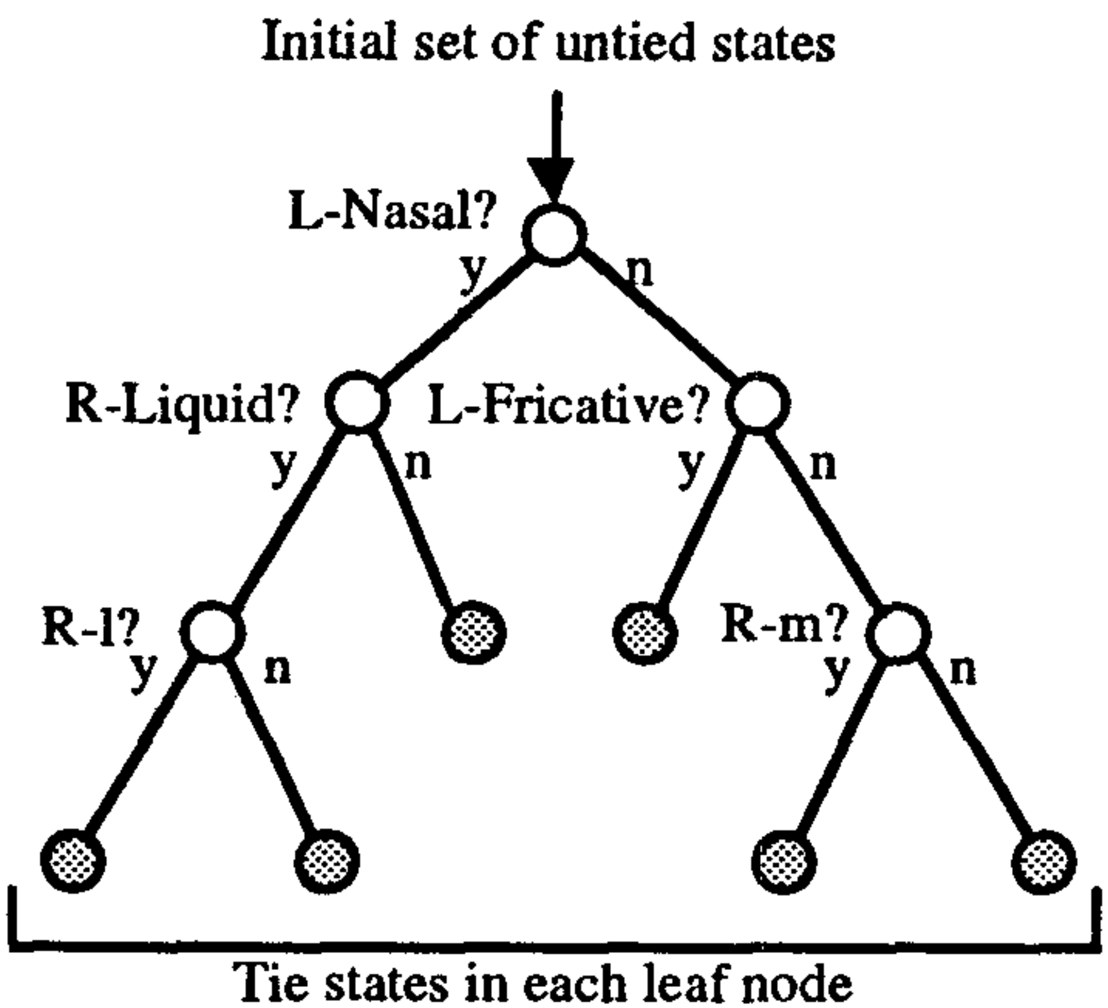
2. What is the training data for each phone state, p_j? (root node of DT)

   All speech frames that align with the j^{th} state of every triphone HMM that has p as the middle phone

3. What criterion is used at each node to find the best question to split the data on?

   Find the question which partitions the states in the parent node so as to give the maximum increase in log likelihood

Likelihood criterion



Given a phonetic question, let the initial set of untied states S be split into two partitions S_{yes} and S_{no}

Each partition is clustered to form a single Gaussian output distribution with mean $\mu_{S_{yes}}$ and covariance $\Sigma_{S_{yes}}$

Use the likelihood of the parent state and the subsequent split states to determine which question a node should be split on

Image from: Young et al., "Tree-based state tying for high accuracy acoustic modeling", ACL-HLT, 1994

Likelihood of a cluster of states

- If a cluster of HMM states, $S = \{s_1, s_2, \ldots, s_M\}$ consists of M states and a total of K acoustic observation vectors are associated with S, $\{x_1, x_2 \ldots, x_K\}$, then the log likelihood associated with S is:

$$\mathcal{L}(S) = \sum_{i=1}^{K} \sum_{s \in S} \log \Pr(x_i; \mu_S, \Sigma_S) \gamma_s(x_i)$$

- For a question that splits S into S_{yes} and S_{no}, compute the following quantity:

$$\Delta = \mathcal{L}(S_{\text{yes}}) + \mathcal{L}(S_{\text{no}}) - \mathcal{L}(\mathcal{S})$$

- Go through all questions, find Δ for each and choose the question for which Δ is the biggest

- Terminate when: Final Δ is below a threshold or data associated with a split falls below a threshold

Revising EM and Baum Welch training

Recall EM: Fitting Parameters to Data

Parameter θ determines $\Pr(x, z; \theta)$ where x is observed and z is hidden

Observed data: i.i.d samples x_i, i=1, …, N

Goal: Find $\arg\max_\theta \mathcal{L}(\theta)$ where $\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \Pr(x_i; \theta)$

Initial parameters: θ^0

Iteratively compute θ^ℓ as follows:

$$Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \Pr(z|x_i; \theta^{\ell-1}) \log \Pr(x_i, z; \theta)$$

$$\theta^\ell = \arg\max_\theta Q(\theta, \theta^{\ell-1})$$

Estimate θ^ℓ cannot get worse over iterations because for all θ:

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^{\ell-1}) \geq Q(\theta, \theta^{\ell-1}) - Q(\theta^{\ell-1}, \theta^{\ell-1})$$

EM is guaranteed to converge to a local optimum [Wu83]

Coin example to illustrate EM

$\rho_1 = \Pr(\mathrm{H}) = 0.3$ $\rho_2 = \Pr(\mathrm{H}) = 0.4$ $\rho_3 = \Pr(\mathrm{H}) = 0.6$

Repeat:

Toss Coin 1 privately
if it shows H:
    Toss Coin 2 twice
else
    Toss Coin 3 twice

The following sequence is observed: "HH, TT, HH, TT, HH"

How do you estimate ρ_1, ρ_2 and ρ_3?

Coin example to illustrate EM

Our observed data is: {HH, TT, HH, TT, HH}

Let's use EM to estimate $\theta = (\rho_1, \rho_2, \rho_3)$

[EM Iteration, E-step]

Compute quantities involved in

$$Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \gamma(z, x_i) \log \Pr(x_i, z; \theta)$$

where $\gamma(z, x) = \Pr(z \mid x ; \theta^{\ell-1})$

i.e., compute $\gamma(z, x_i)$ for all z and all i

Compute $\gamma(\mathrm{H}, \mathrm{HH})$, $\gamma(\mathrm{H}, \mathrm{TT})$, $\gamma(\mathrm{T}, \mathrm{TT})$ and $\gamma(\mathrm{T}, \mathrm{HH})$

E-step

ρ_1 = Pr(H)

ρ_2 =Pr(H)

ρ_3 = Pr(H)

What is γ(H, HH)?

$$\gamma(\mathrm{H}, \mathrm{HH}) = \Pr(z{=}\mathrm{H} | x{=}\mathrm{HH}; \theta^{\ell-1})$$

$$= \Pr(x{=}\mathrm{HH}|z{=}\mathrm{H})\Pr(z{=}\mathrm{H}) \Big/ \Big(\Pr(x{=}\mathrm{HH}|z{=}\mathrm{H})\Pr(z{=}\mathrm{H}) + \Pr(x{=}\mathrm{HH}|z{=}\mathrm{T})\Pr(z{=}\mathrm{T})\Big)$$

$$= \rho_1\rho_2^2/(\rho_1\rho_2^2 + (1 - \rho_1)\rho_3^2)$$

Similarly compute γ(H, TT), γ(T, TT) and γ(T, HH)

where $\Pr(z;\theta) = \begin{cases} \rho_1 & \text{if } z = \mathrm{H} \\ 1 - \rho_1 & \text{if } z = \mathrm{T} \end{cases}$

$$\Pr(x|z;\theta) = \begin{cases} \rho_2^h(1-\rho_2)^t & \text{if } z = \mathrm{H} \\ \rho_3^h(1-\rho_3)^t & \text{if } z = \mathrm{T} \end{cases}$$

h : number of heads, t : number of tails

M-step

Our observed data is: {HH, TT, HH, TT, HH}

Let's use EM to estimate $\theta = (\rho_1, \rho_2, \rho_3)$

[EM Iteration, M-step]

Find θ which maximises

$$Q(\theta, \theta^{\ell-1}) = \sum_{i=1}^{N} \sum_{z} \gamma(z, x_i) \log \Pr(x_i, z; \theta)$$

$$\rho_1 = \frac{\sum_{i=1}^{N} \gamma(\mathrm{H}, x_i)}{N}$$

$$\rho_2 = \frac{\sum_{i=1}^{N} \gamma(\mathrm{H}, x_i) h_i}{\sum_{i=1}^{N} \gamma(\mathrm{H}, x_i)(h_i + t_i)}$$

$$\rho_3 = \frac{\sum_{i=1}^{N} \gamma(\mathrm{T}, x_i) h_i}{\sum_{i=1}^{N} \gamma(\mathrm{T}, x_i)(h_i + t_i)}$$

M-step

Let us derive an estimate for ρ_1

$$Q(\theta, \theta^{l-1}) = \Sigma_i \gamma(H,x_i)\log[\rho_2^{hi}(1-\rho_2)^{ti}\rho_1] + \Sigma_i \gamma(T,x_i)\log[\rho_3^{hi}(1-\rho_3)^{ti}(1-\rho_1)]$$

$$\partial Q/\partial \rho_1 = 0 \Longrightarrow \Sigma_i \gamma(H,x_i)/\rho_1 - \Sigma_i \gamma(T,x_i)/(1 - \rho_1) = 0$$

$$\Longrightarrow (1 - \rho_1)/\rho_1 = \Sigma_i \gamma(T,x_i)/\Sigma_i \gamma(H,x_i)$$

$$\Longrightarrow \rho_1 = \Sigma_i \gamma(H,x_i)/\big(\Sigma_i \gamma(H,x_i) + \Sigma_i \gamma(T,x_i)\big)$$

$$\Longrightarrow \rho_1 = \Sigma_i \gamma(H,x_i)/N$$

Similarly, estimate ρ_2 and ρ_3

Baum-Welch Algorithm as EM

Observed data: N *sequences*, $x_i = (x_{i1}, \ldots, x_{iT_i})$, $i=1 \ldots N$ where $x_{it} \in \mathbb{R}^d$

Parameters θ : transition matrix A, observation probabilities B

[EM Iteration, E-step]

Compute quantities involved in $Q(\theta, \theta^{\ell-1})$

$\gamma_{i,t}(j) = \Pr(z_t = j \mid x_i ; \theta^{\ell-1})$

$\xi_{i,t}(j,k) = \Pr(z_{t-1} = j, z_t = k \mid x_i ; \theta^{\ell-1})$

Baum-Welch Algorithm as EM

Observed data: N *sequences*, $x_i = (x_{i1}, \ldots, x_{iT_i})$, $i=1 \ldots N$ where $x_{it} \in \mathbb{R}^d$

Parameters θ : transition matrix A, observation probabilities B

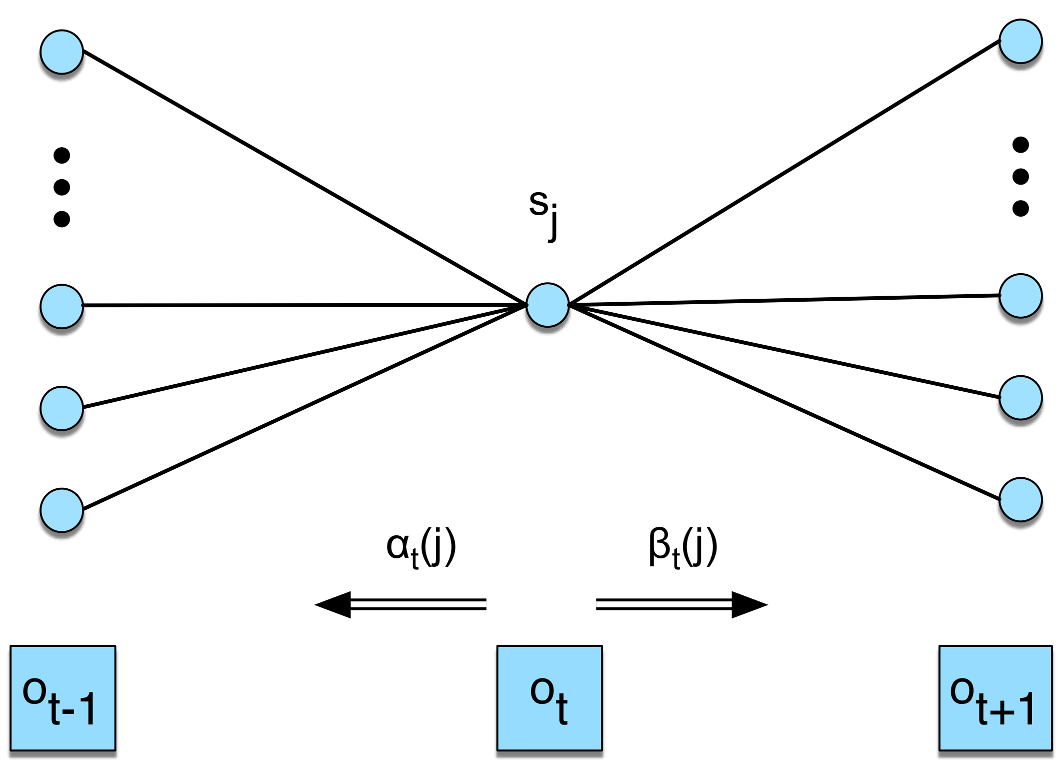
[EM Iteration, E-step]

Compute quantities involved in $Q(\theta, \theta^{\ell-1})$

$$\gamma_{i,t}(j) = \Pr(z_t = j \mid x_i ; \theta^{\ell-1})$$

$$\xi_{i,t}(j,k) = \Pr(z_{t-1} = j, z_t = k \mid x_i ; \theta^{\ell-1})$$

$$\begin{aligned}\gamma_{i,t}(j) &= \Pr(z_t = j \mid x_i ; \theta^{\ell-1}) \\ &= \alpha_t(j)\beta_t(j)/\Pr(x_i ; \theta^{\ell-1})\end{aligned}$$



Baum-Welch Algorithm as EM

Observed data: N *sequences*, $x_i = (x_{i1}, \ldots, x_{iT_i})$, $i=1..N$ where $x_{it} \in \mathbb{R}^d$

Parameters θ : transition matrix A, observation probabilities B

[EM Iteration, E-step]

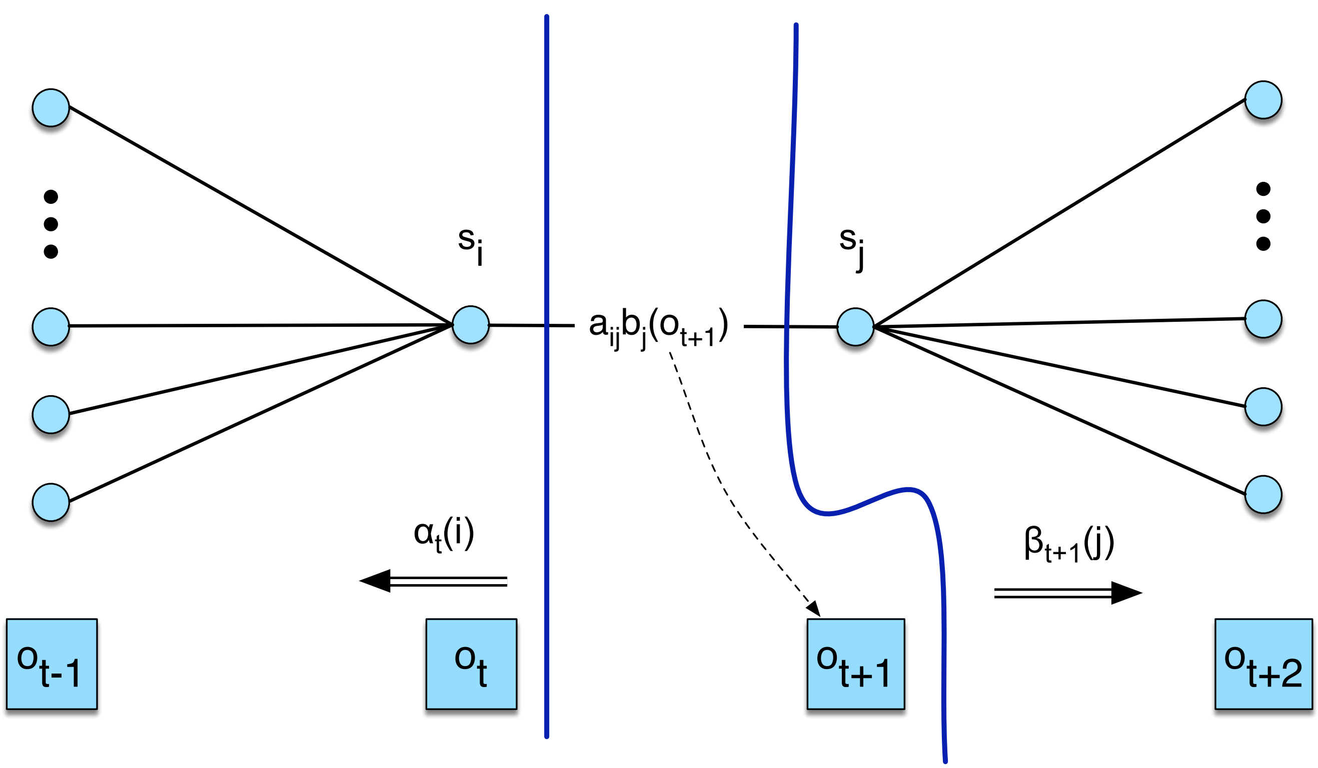
Compute quantities involved in $Q(\theta,\theta^{\ell-1})$

$\gamma_{i,t}(j) = \Pr(z_t = j \mid x_i ;\theta^{\ell-1})$

$\xi_{i,t}(j,k) = \Pr(z_{t-1} = j, z_t = k \mid x_i ;\theta^{\ell-1})$

$$\gamma_{i,t}(j) = \Pr(z_t = j \mid x_i ;\theta^{\ell-1})$$
$$= \alpha_t(j)\beta_t(j)/\Pr(x_i ;\theta^{\ell-1})$$

$$\xi_{i,t}(j, k) = \Pr(z_{t-1} = j, z_t = k \mid x_i ;\theta^{\ell-1})$$
$$= \alpha_t(j)a_{jk}b_k(x_{it+1})\beta_{t+1}(k)/\Pr(x_i ;\theta^{\ell-1})$$



BW for Gaussian Mixture Model

Observed data: N *sequences*, $x_i = (x_{i1}, \ldots, x_{iT_i})$, $i=1\ldots N$ where $x_{it} \in \mathbb{R}^d$

Parameters θ : transition matrix A, observation prob. $B = \{(\mu_{jm}, \Sigma_{jm}, c_{jm})\}$ for all j,m

[EM Iteration, M-step]

Find θ which maximises $Q(\theta, \theta^{\ell-1})$

$$A_{j,k} = \frac{\sum_{i=1}^{N} \sum_{t=2}^{T_i} \xi_{i,t}(j,k)}{\sum_{i=1}^{N} \sum_{t=2}^{T_i} \sum_{k'} \xi_{i,t}(j,k')}$$

$$\mu_{jm} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m) x_{it}}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m)}$$

$$\Sigma_{jm} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m)(x_{it} - \mu_{jm})(x_{it} - \mu_{jm})^T}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m)}$$

$\gamma_{i,t}(j,m)$
$= \gamma_{i,t}(j) c_{il} b_{il}(x_{it}) / b_i(x_{it})$

$$c_{jm} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j,m)}{\sum_{i=1}^{N} \sum_{t=1}^{T_i} \gamma_{i,t}(j)}$$

Mixing probabilities