Variational Bayes Eigenvoice HMM Diarization

Lukáš Burget,Mireia Diez

Speech@FIT, Brno
burget@fit.vutbr.cz, mireia@fit.vutbr.cz

Brno, October 30 2017

BRNO UNIVERSITY OF TECHNOLOGY
speech@fit

BRNO UNIVERSITY OF TECHNOLOGY
speech@fit

The model
Structure of the model

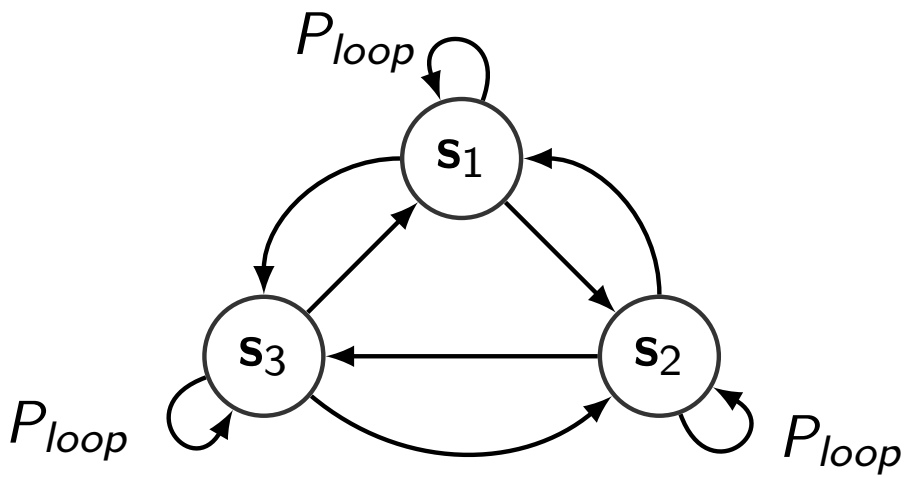
- The input sequence of observed feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ is assumed to be generated from a Hidden Markov Model.
- Each HMM state represents one of S speakers (there can be more than one state per speaker).
- For a particular speaker (i.e. given the HMM state) the distribution of observations is modeled using a GMM[1].

[1] This model was inspired by [1]

The model

Structure of the model

HMM Model for 3 speakers with a single state per speaker.



The Model
Speaker Models

- Robust speaker models:
  - Strong informative prior imposed on the GMM parameters, with the same form as in JFA or the i-vector extraction model

For a speaker s, the super-vector of concatenated component means $\boldsymbol{\mu}_s = [\boldsymbol{\mu}_{s1}^T \boldsymbol{\mu}_{s2}^T \dots \boldsymbol{\mu}_{sC}^T]^T$ is constrained to live in a linear subspace:

$$\boldsymbol{\mu}_s = \mathbf{m}^{ubm} + \mathbf{V}\mathbf{y}_s$$

The low dimensional vectors $\mathbf{y}_s$ are treated as latent random variables with standard normal prior.

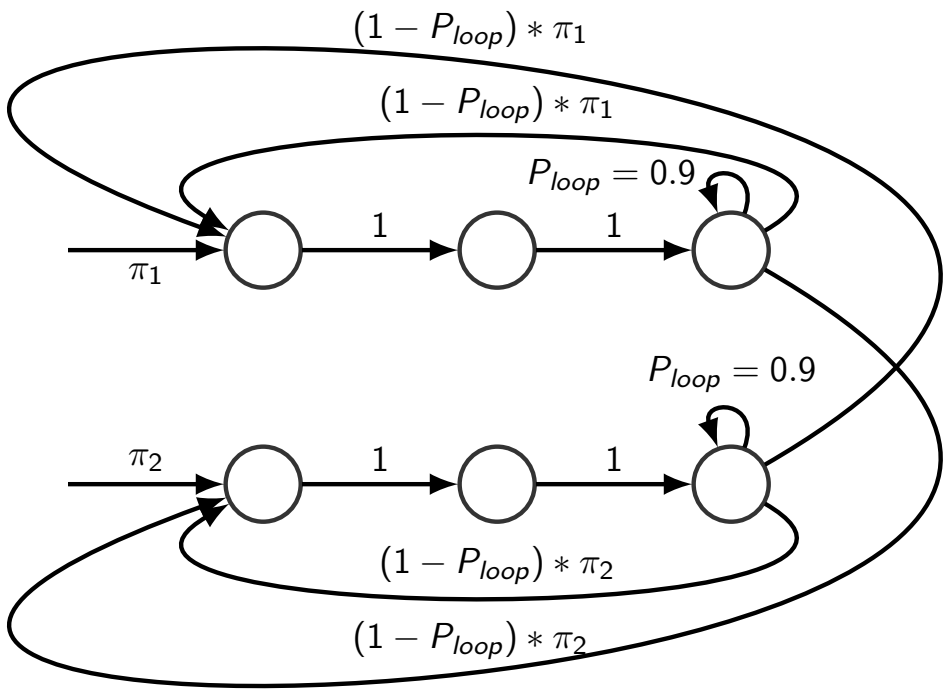
The Model
The HMM

- Each speaker will be modeled by a number of σ states $(u_s = u_{s1}, u_{s2}, ...u_{s\sigma})$
- For all the states u_{si} where $i < \sigma$, the transitions are restricted to the next state of the same speaker
- For the last state of the speakers, we define a loop probability parameter
- A transition probability of $(1 - P_{loop}) * \pi_s$ is assigned to the other arcs

The Model

HMM model example

HMM Model 2 speakers with 3 states per speaker.



The Model
Definition of variables

$\mathbf{z}_t$ set of hidden variables determining the assignment of frames to speakers

	$\mathbf{z}_1$	$\mathbf{z}_2$	$\mathbf{z}_3$	$\mathbf{z}_4$	$\mathbf{z}_5$
s=1	1	1	0	0	0
s=2	0	0	0	1	1
s=3	0	0	1	0	0

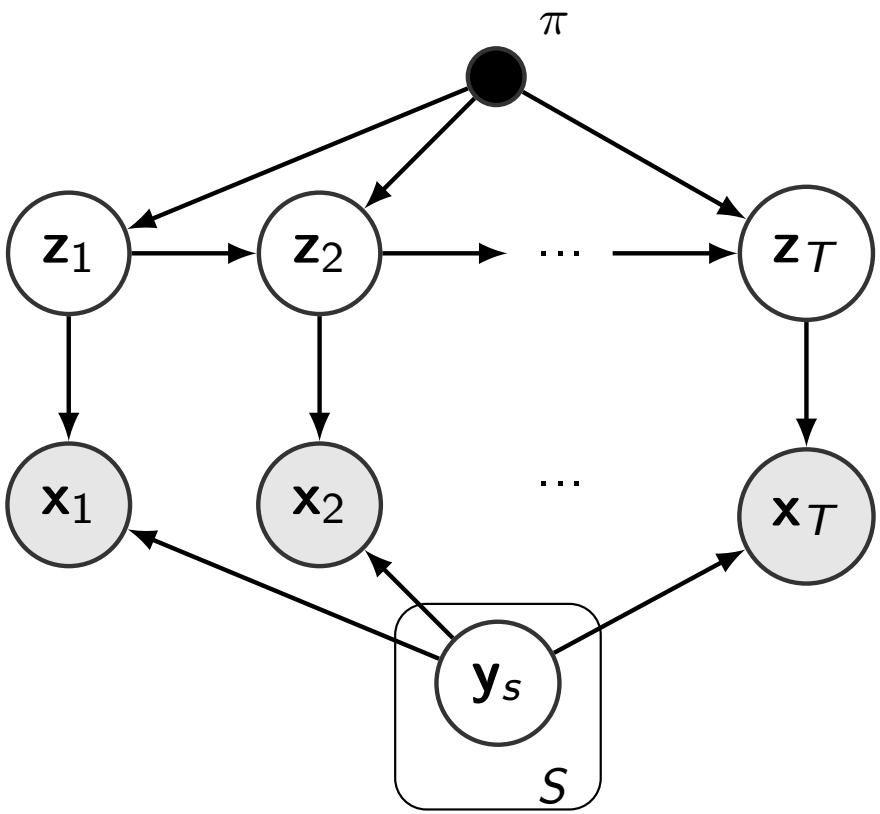
$\mathbf{x}_t$ set of observed variables, generated from a HMM

$$P(\mathbf{z}_t|\mathbf{z}_{t-1}) = \prod_b \prod_e a_{be}^{z_{t,b}\ z_{t-1,e}}$$
$$p(\mathbf{x}_t|\mathbf{z}_t) = \prod_s p(\mathbf{x}_t|\mathbf{y}_s)^{z_{ts}}$$
$$p(\mathbf{x}_t|\mathbf{y}_s) = GMM(\mathbf{x}_t|\{\boldsymbol{\mu}_c\}_s, \{\boldsymbol{\Sigma}_c^{ubm}\}, \{w_c^{ubm}\})$$
$$[\boldsymbol{\mu}_{s1}^T \boldsymbol{\mu}_{s2}^T \dots \boldsymbol{\mu}_{sC}^T]^T = \boldsymbol{\mu}_s = \mathbf{m}^{ubm} + \mathbf{V}\mathbf{y}_s$$
$$p(\mathbf{y}_s) = \mathcal{N}(\mathbf{y}_s|0, \mathbf{I})$$

BRNO UNIVERSITY OF TECHNOLOGY
speech@fit

The Model

Bayesian Network representation of the Model



The Model

Assumed generative process:

For s in S:
$\mathbf{y}_s \sim \mathcal{N}(0, \mathbf{I})$
$\boldsymbol{\mu}_s = \mathbf{m}^{ubm} + \mathbf{V}\mathbf{y}_s$
For t in T:
$\mathbf{z}_t \sim p(\mathbf{z}_t|\mathbf{z}_{t-1})$
$\mathbf{x}_t \sim p(\mathbf{x}_t|\mathbf{z}_t) = \prod_s p(\mathbf{x}_t|\mathbf{y}_s)^{z_{ts}}$

The joint probability distribution of all the random variables:

$$\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) = \ln \left[p(\mathbf{X}|\mathbf{Z}, \mathbf{Y})p(\mathbf{Z})p(\mathbf{Y})\right]$$
$$= \sum_t \sum_s z_{ts} \ln p(\mathbf{x}_t|\mathbf{y}_s) + \sum_t \ln p(z_t|z_{t-1}) + \sum_s \ln p(\mathbf{y}_s)$$

Variational Bayes

The problem that we need to solve is:

- Infer the posterior distribution $p(\mathbf{Z}|\mathbf{X}) = \int p(\mathbf{Z}, \mathbf{Y}|\mathbf{X}) d\mathbf{Y}$, where $\mathbf{Z}$ defines the assignments of frames to speakers
  - Find the speaker models, (i.e. infer $p(\mathbf{Y}|\mathbf{Z})$)

We will use a Variational Bayes approach to perform the inference

Variational Bayes
Basics

- We need to infer $p(\mathbf{Z}, \mathbf{Y}|\mathbf{X})$, which is intractable.
- We choose to approximate it by $q(\mathbf{Z}, \mathbf{Y})$
- We use the mean-field approximation: $q(\mathbf{Z}, \mathbf{Y}) = q(\mathbf{Z})q(\mathbf{Y})$
- With VB we attempt to minimize the Kullback-Liebler divergence between both distributions:

$$KL(q||p) = -\int q(\mathbf{Z}, \mathbf{Y}) \; ln \left\{ \frac{p(\mathbf{Z}, \mathbf{Y}|\mathbf{X})}{q(\mathbf{Z}, \mathbf{Y})} \right\} d\mathbf{Z} d\mathbf{Y}$$

Variational Bayes
Basics

With Variational Bayes we have[2]:

$$ln\ p(\mathbf{X}) = \mathcal{L}(q) + KL(q||p)$$

We will maximize:

$$\mathcal{L}(q) = \int q(\mathbf{Z}, \mathbf{Y})\ ln \left\{ \frac{p(\mathbf{X}, \mathbf{Y}, \mathbf{Z})}{q(\mathbf{Z}, \mathbf{Y})} \right\} d\mathbf{Y}$$

Which in turn will minimize the KL:

$$KL(q(\mathbf{Z}, \mathbf{Y})||p(\mathbf{Z}, \mathbf{Y}|\mathbf{X}))$$

For the mean-field approximation, we iteratively update:

$$\ln q(\mathbf{Z}) = E_Y[\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{Y})] + C$$
$$\ln q(\mathbf{Y}) = E_Z[\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{Y})] + C$$

Variational Bayes
Parameter updates for $q(\mathbf{Y})$

$$\ln q(\mathbf{y}_s) = E_Z[\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{Y})] + C = E_Z[\ln p(\mathbf{X}|\mathbf{Z}, \mathbf{Y})] + \ln p(\mathbf{Y}) + C$$

i-vector like update formulas:

$$q(\mathbf{y}_s) = \mathcal{N}\left(\mathbf{y}_s | \boldsymbol{\alpha}_\mathbf{s}, \mathbf{L}_s^{-1}\right)$$

where $\boldsymbol{\alpha}_\mathbf{s} = \mathbf{L}_s^{-1} \sum_t \gamma_{ts} \boldsymbol{\rho}_t$ $\mathbf{L}_s = \mathbf{I} + \sum_t \gamma_{ts} \boldsymbol{\phi}_t$

γ_{ts} are the responsibilities of speakers (HMM states) for frames as defined by $q(\mathbf{Z})$

$$\boldsymbol{\rho}_\mathbf{t} = \sum_c \gamma_{tc}^{ubm} \left(\mathbf{x}_t - \mathbf{m}_c^{ubm}\right)^T \boldsymbol{\Sigma}_c^{ubm^{-1}} \mathbf{V}_c$$

$$\boldsymbol{\phi}_\mathbf{t} = \sum_c \gamma_{tc}^{ubm} \mathbf{V}_c^T \boldsymbol{\Sigma}_c^{ubm^{-1}} \mathbf{V}_c$$

Variational Bayes
Parameter updates for $q(\mathbf{Z})$

$$\begin{aligned}\ln q(\mathbf{Z}) =& E_Y[\ln p(\mathbf{X}, \mathbf{Z}, \mathbf{Y})] + C = E_Y\left[\ln p(\mathbf{X}|\mathbf{Z}, \mathbf{Y})\right] + \ln p(\mathbf{Z}) + C \\ =& \sum_t \sum_s z_{ts} \ln \overline{p}(\mathbf{x}_t|s) + \sum_t \ln p(z_t|z_{t-1}) + C\end{aligned}$$

$$\text{where } \ln \overline{p}(\mathbf{x}_t|s) = \boldsymbol{\rho}_t^T \boldsymbol{\alpha}_s - \frac{1}{2} tr\left(\boldsymbol{\phi}_t \left[\mathbf{L}_s^{-1} + \boldsymbol{\alpha}_s \boldsymbol{\alpha}_s^T\right]\right) + Const$$

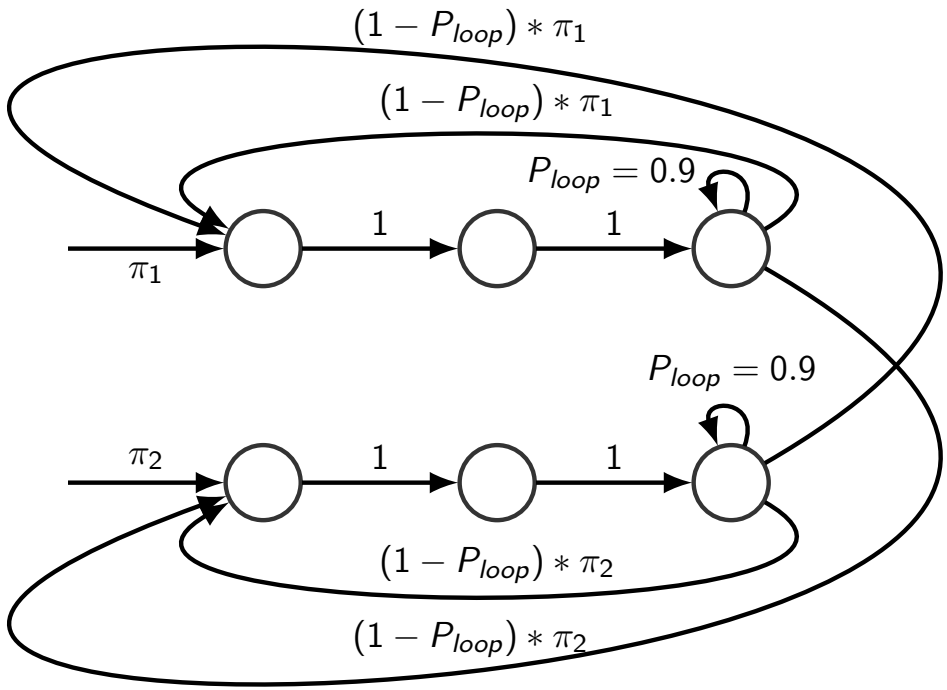
Which results in the same inference as in the standard HMM:

$$\ln P(\mathbf{Z}|\mathbf{X}) + C = \ln P(\mathbf{X}, \mathbf{Z}) = \sum_t \sum_s z_{ts} \ln p(\mathbf{x}_t|s) + \sum_t \ln p(z_t|z_{t-1})$$

What we are interested in are the γ_{ts}, which we obtain by the standard forward-backward algorithm

Variational Bayes
Re-estimating speaker priors

We re-estimate speaker priors π_s. With our Bayesian model, the solution tends to be sparse $\Rightarrow$ the model is able to determine the number of speaker.



Variational Bayes
Re-estimating speaker priors

For the re-estimation of the speaker priors we explicitly maximize the Lower bound:

$$\frac{\partial\ \mathcal{L}(q)}{\partial \pi_s} + \lambda \left(\sum_s \pi_s - 1 \right) = 0$$

$$\frac{\partial\ E_{\mathbf{Z}} \left[\ln p(\mathbf{Z}) \right]}{\partial \pi_s} + \lambda \left(\sum_s \pi_s - 1 \right) = 0$$

$$\pi_s \propto \sum_i \sum_t \xi \left(t, i, s_i \right)$$

Where ξ denotes the posterior probability of making the transition from state i to the initial state of speaker s at time t (calculated again with the forward-backward algorithm)

Variational Bayes
Lower Bound

The lower bound is:

$$\begin{aligned}
\mathcal{L}\left(q(\mathbf{X})\right) =& E_{\mathbf{Y},\mathbf{Z}}\left\{\ln\left(\frac{p(\mathbf{X},\mathbf{Y},\mathbf{Z})}{q(\mathbf{Y},\mathbf{Z})}\right)\right\} \\
=& E_{\mathbf{Y},\mathbf{Z}}\left[\ln p(\mathbf{X}|\mathbf{Y},\mathbf{Z})\right] + E_{\mathbf{Z}}\left[\ln\frac{p(\mathbf{Z})}{q(\mathbf{Z})}\right] + E_{\mathbf{Y}}\left[\ln\frac{p(\mathbf{Y})}{q(\mathbf{Y})}\right] \\
=& \left[\sum_t\sum_s \gamma_{ts}\ \ln\overline{p}(\mathbf{x}_t|s)\right] \\
&+ \left[\sum_i\sum_e\left(\sum_t \xi(t,i,e)\right)\ln\ a_{ie} - E_{\mathbf{Z}}\left[q(\mathbf{Z})\right]\right] \\
&+ \left[\sum_s\left(\frac{R}{2} + \frac{1}{2}\ln|\mathbf{L}_s^{-1}| - \frac{1}{2}tr(\mathbf{L}_s^{-1}) - \frac{1}{2}\boldsymbol{\alpha}_s^T\boldsymbol{\alpha}_s\right)\right]
\end{aligned}$$

Results

Results of different approaches for the CALLHOME dataset.

	DER	
System	Alone	As init. for VB
Random init x1	-	12.7
Random init x5	-	8.6
MFCC ivector-PLDA AHC [3]	13.7	9.7
Oracle labels	0	4.0

Random initx5 refers to initializing the system 5 times and choosing the output that achieves the lowest lower bound.


P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," tech. rep., Montreal: CRIM, 2008.


C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*.
Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.


G. Sell and D. Garcia-Romero, "Diarization resegmentation in the factor analysis subspace," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4794–4798, April 2015.

BRNO UNIVERSITY OF TECHNOLOGY
speech@fit