A Symmetrization of the Subspace Gaussian Mixture Model (SGMM)

Daniel Povey, Martin Karafiat,
Arnab Ghoshal, Petr Schwarz

The Subspace Gaussian Mixture Model (SGMM)

- Will describe the SGMM in stages
- Let the state index be j.
- First, write down a full-covariance GMM:

$$p(\mathbf{x}|j) = \sum_{i=1}^{I_j} w_{ji} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{ji})$$

- Parameters are:

$$w_{ji},\ \boldsymbol{\mu}_{ji},\ \boldsymbol{\Sigma}_{ji}$$

The SGMM in stages

- Enforce that the number of Gaussians is fixed across states:

$$p(\mathbf{x}|j) = \sum_{i=1}^{I} w_{ji}\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{ji})$$

- Parameters are (as before):

$$w_{ji},\ \boldsymbol{\mu}_{ji},\ \boldsymbol{\Sigma}_{ji}$$

The SGMM in stages

- Share the covariances across states (but not across Gaussians):

$$p(\mathbf{x}|j) = \sum_{i=1}^{I} w_{ji}\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_{ji},\boldsymbol{\Sigma}_i)$$

- Parameters are:

$$w_{ji},\ \boldsymbol{\mu}_{ji},\ \boldsymbol{\Sigma}_i$$

The SGMM in stages

- Limit the means to a subspace:

$$p(\mathbf{x}|j) = \sum_{i=1}^{I} w_{ji}\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_{ji},\boldsymbol{\Sigma}_i)$$

$$\boldsymbol{\mu}_{ji} = \mathbf{M}_i\mathbf{v}_j$$

- Parameters are:

$$w_{ji},\ \mathbf{v}_j,\ \mathbf{M}_i,\ \boldsymbol{\Sigma}_i$$

- Note: subspace dimension (dim of $\mathbf{v}_j$) is arbitrary; typically 40-50.

The SGMM in stages

- Make weights dependent on subspace:

$$
\begin{aligned}
p(\mathbf{x}|j) &= \sum_{i=1}^{I} w_{ji}\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_{ji},\boldsymbol{\Sigma}_i) \\
\boldsymbol{\mu}_{ji} &= \mathbf{M}_i\mathbf{v}_j \\
w_{ji} &= \frac{\exp \mathbf{w}_i^T\mathbf{v}_j}{\sum_{i'}\exp \mathbf{w}_{i'}^T\mathbf{v}_j}
\end{aligned}
$$

- Parameters are:

$$\mathbf{w}_i,\ \mathbf{v}_j,\ \mathbf{M}_i,\ \boldsymbol{\Sigma}_i$$

The SGMM in stages

- Add “sub-states” (new level of mixture):

$$p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{jmi}, \boldsymbol{\Sigma}_i)$$

$$\boldsymbol{\mu}_{jmi} = \mathbf{M}_i \mathbf{v}_{jm}$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}}$$

- Parameters are:

$$c_{jm}, \mathbf{w}_i, \mathbf{v}_{jm}, \mathbf{M}_i, \boldsymbol{\Sigma}_i$$

The SGMM in stages

- Add speaker-dependent mean offset:

$$p(\mathbf{x}|j,s) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{jmi}^{(s)}, \boldsymbol{\Sigma}_i)$$

$$\boldsymbol{\mu}_{jmi}^{(s)} = \mathbf{M}_i \mathbf{v}_{jm} + \mathbf{N}_i \mathbf{v}^{(s)}$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}}$$

- Parameters are:

$$c_{jm}, \mathbf{w}_i, \ \mathbf{v}_{jm}, \ \mathbf{M}_i, \ \boldsymbol{\Sigma}_i, \mathbf{N}_i, \mathbf{v}^{(s)}$$

Asymmetry in the SGMM

- The previous page gave the “baseline” SGMM (that we previously described).
- As far as the means are concerned, the model treats speakers and phonetic states symmetrically.
- However, the weights are only dependent on the “phonetic” state (not the speaker).

The symmetrized SGMM

- Symmetrize weight equation:

$$p(\mathbf{x}|j,s) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi}^{(s)} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{jmi}^{(s)}, \boldsymbol{\Sigma}_i)$$

$$\boldsymbol{\mu}_{jmi}^{(s)} = \mathbf{M}_i \mathbf{v}_{jm} + \mathbf{N}_i \mathbf{v}^{(s)}$$

$$w_{jmi}^{(s)} = \frac{\exp(\mathbf{w}_i^T \mathbf{v}_{jm} + \mathbf{u}_i^T \mathbf{v}^{(s)})}{\sum_{i'=1}^{I} \exp(\mathbf{w}_{i'}^T \mathbf{v}_{jm} + \mathbf{u}_{i'}^T \mathbf{v}^{(s)})}$$

- Parameters are:

$$c_{jm}, \mathbf{w}_i, \mathbf{v}_{jm}, \mathbf{M}_i, \boldsymbol{\Sigma}_i, \mathbf{N}_i, \mathbf{v}^{(s)}, \mathbf{u}_i$$

That wasn’t so hard!

- It’s one thing to write down the equation for the likelihood…
- We also need efficient equations for likelihood evaluation and update.
- That’s what the paper is about.
- In the paper we write down these equations.
- Derivation in separate technical report.
- Won’t discuss the details here.

Exact vs. inexact update

- We describe two types of update formula for the $\mathbf{u}_i$ quantities (the "new" quantities we introduced).
- Exact update:
  - Mirror-image of the update for $\mathbf{w}_i$ (which is the "phonetic-subspace" version).
  - Requires storing a list of speaker-specific quantities with the accumulators (not very scalable).
  - Guaranteed to converge
- Inexact update
  - Avoid storing lists of speaker-specific quantities.
  - Not guaranteed to converge.

Results (CallHome, no CMLLR)

GMM:	52.5					
	#Substates					
	2700	4k	6k	9k	12k	16k
SGMM:	48.8	48.2	48.0	47.7	**47.4**	47.5
+spk-vecs:	47.6	47.0	46.4	46.4	46.1	**45.9**
+symmetric,exact:	46.3	45.6	45.2	44.8	44.5	**44.4**
+symmetric,inexact	46.5	45.6	45.0	44.6	**44.4**	

Table 1. CallHome English: WERs without CMLLR adaptation

- Symmetrization helps considerably here
- No difference exact vs. inexact

Results (CallHome +CMLLR)

GMM:	49.7					
+SAT:	46.0					
	#Substates					
	2700	4k	6k	9k	12k	16k
SGMM+spk-vecs:	46.5	45.5	45.2	45.4	44.8	**44.7**
+symmetric,exact	44.9	44.4	44.1	43.2	**42.8**	42.9
+symmetric,inexact	45.2	44.1	43.5	43.4	**43.3**	

Table 2. CallHome English: WERs with CMLLR adaptation

- Symmetrization still helps a lot
- Exact vs. inexact: probably insignificant

Results (Switchboard, with VTLN)

GMM	#Gauss per state						
	20	26	32	34	36	38	40
-		36.8	36.6	36.4	36.4	36.4	36.4
CMLLR		34.8	34.5	34.4	34.3	34.5	34.3
STC		35.4	35.3		35.2		
+CMLLR		33.1	32.9		32.9		
SGMM	#Substates						
	30k	40k	50k	75k	100k	150k	200k
unadapted	35.7	35.7	35.1	34.7	34.3	33.9	33.7
CMLLR			32.2				
+spk-vecs	32.0	31.7	31.4	31.2	30.8		
+symmetric	31.9	31.7	31.3	31.0	30.6		

Table 3. Switchboard: WERs, with VTLN

Results (Switchboard, no VTLN)

GMM	#Gauss per state				
					36
-					39.2
CMLLR					37.0
STC					38.0
+CMLLR					35.2
SGMM	#Substates				
	30k	40k	50k	75k	100k
unadapted	37.9	37.5	37.1	36.6	36.3
CMLLR+spk-vecs	33.9	33.5	33.4		
+symmetric	33.8	33.0	33.2		

Table 4. Switchboard: WERs, no VTLN

Results (Switchboard: summary)

- On this setup, improvements from symmetrization are very small (e.g. ~0.2%)
- This is true with or without VTLN.
- We looked at the likelihood changes from symmetrization…
- There were similar likelihood improvements in Switchboard and CallHome: this doesn't help to explain the difference.
- No obvious explanation why different from CallHome.

Summary

- Symmetrization helped considerably on CallHome, and hardly at all on Switchboard
- Probably the “normal” case is somewhere in between
- In the near future we will test this on other setups.
- Note: symmetrization hardly affects speed, but nearly doubles memory requirements (for the model).
- Whether it’s worth it probably depends on the WER improvement and other factors.
- Further work: training UBM with speaker adaptation?

Kaldi

- We have now released software that implements SGMMs.
- Kaldi is also a general-purpose speech recognition toolkit that uses OpenFst for FST-based training and decoding.
- Apache 2.0 license; on Sourceforge.
- Example scripts (RM, WSJ) that run from LDC databases.
- see http://kaldi.sf.net
- Presentation on Friday after lunch (1:45), room 3.3 upstairs