MAP adaptation with SphinxTrain

David Huggins-Daines

dhuggins@cs.cmu.edu

Language Technologies Institute

Carnegie Mellon University

Theory of MAP adaptation

- Standard Baum-Welch training produces a maximum-likelihood estimate of model parameters λ:

$$\lambda_{ML} = \arg\max_{\lambda} P(O|\lambda)$$

- MAP training produces the maximum a-posteriori estimate:

$$\lambda_{MAP} = \arg\max_{\lambda} P(O|\lambda)P(\lambda)$$

- Reduces to ML estimate with a non-informative prior $P(\lambda)$.
- For speaker adaptation, the prior $P(\lambda)$ is derived from a baseline or speaker-independent model.

MAP adaptation in practice

- The simplest method is Bayesian updating of each Gaussian mean, assuming the following (incorrect) prior:

$$\mu \sim N(\mu_{SI}, \sigma^2_{SI})$$

- This reduces to interpolation between SI parameters and ML (forward-backward) estimates from the adaptation data:

$$\hat{\mu}_{MAP} = \frac{\sum_{t=1}^{T} \gamma_t(i,k)\sigma^2_{SI}\mu_{ML} + \sigma^2_{ML}\mu_{SI}}{\sum_{t=1}^{T} \gamma_t(i,k)\sigma^2_{SI} + \sigma^2_{ML}}$$

- The posterior variance can also be computed, but it is not useful.

MAP adaptation in practice

- With a more detailed prior, all HMM/GMM parameters can be updated.
- This is important for semi-continuous models since the mixture weights can be modified.
- Prior is a product of a Dirichlet distribution with hyperparameters $\{\eta, \nu\}$ and a Gamma-Normal distribution with hyperparameters $\{\alpha, \beta, \mu, \tau\}$.
- The τ hyperparameter controls the “speed” of adaptation. Larger τ = less adaptation.
- Estimation of these hyperparameters is tricky. Generally, τ is estimated, then all other hyperparameters derived from it and the SI model.

MAP adaptation in practice

- The τ can be fixed to a global value (e.g. 2.0) or it can be estimated separately for each Gaussian:

$$\tau_{ik} = \frac{p \sum_{t=1}^{T} \gamma_t(i,k)}{\sum_{t=1}^{T} \gamma_t(i,k)(\hat{\mu}_{ik} - \mu_{ik})^T (w_{ik}\Sigma_{ik})(\hat{\mu}_{ik} - \mu_{ik})}$$

- ν_{ik} is then estimated as $w_{ik} \sum_{k=1}^{K} \tau_{ik}$ and the mixture weights are re-estimated as:

$$\hat{w}_{ik} = \frac{\nu_{ik} - 1 + \sum_{t=1}^{T} \gamma_t(i,k)}{\sum_{k=1}^{K} \nu_{ik} - K + \sum_{k=1}^{K} \sum_{t=1}^{T} \gamma_t(i,k)}$$

MAP with SphinxTrain

- MAP interpolation and updating has been implemented in SphinxTrain as the `map_adapt` tool.
- It works similarly to the `norm` tool, except that it produces a MAP re-estimation rather than an ML one.
  1. Collect forward-backward statistics on adaptation data using the baseline models and `bw`.
  2. Run `map_adapt` the same way you would `norm`, specifying the baseline model files and the output MAP model files.
- Works for continuous and semi-continuous models.
- (SCHMM is broken in current version but I'll fix it).

Combining MAP and MLLR

- In theory they are equivalent, if each Gaussian has its own regression class.
- In practice, this never happens, and their effects are additive.
- To combine them:
  1. Compute an MLLR transformation with `bw` and `mllr_solve`
  2. Apply it to the baseline means with `mllr_transform`
  3. Re-run `bw` with the transformed means
  4. Run `map_adapt` to produce a MAP re-estimation

Unsupervised MAP

- This doesn't work. Don't do it.
- The lack of parameter tying in the standard MAP algorithm means that the adaptation is not robust.
- Incorrect transriptions of adaptation data result in the wrong models being updated.
- MLLR alone is a better choice for sparse or noisy adaptation data.

MAP results on RM1 (CDHMM)

1000 CD senones, 8 gaussians, Sphinx 3.x fast decoder

Speaker	Baseline	100	200	400	MLLR+400	Relative
bef0_3	10.40%	8.43%	7.75%	7.75%	7.84%	-24.62%
cmr0_2	8.78%	6.66%	6.40%	6.40%	4.66%	-46.92%
das1_2	9.73%	7.04%	5.92%	4.75%	4.21%	-56.73%
dms0_4	8.31%	5.66%	5.48%	5.01%	4.42%	-46.81%
dtb0_3	9.43%	7.49%	6.63%	5.84%	5.10%	-45.92%
ers0_7	8.19%	7.93%	7.96%	6.34%	5.92%	-27.72%
hxs0_6	16.36%	11.14%	11.08%	7.52%	7.22%	-55.87%
jws0_4	8.81%	6.90%	6.69%	6.40%	5.69%	-35.41%
pgh0_1	7.84%	6.37%	6.54%	5.13%	5.04%	-35.71%
rkm0_5	24.20%	17.83%	17.18%	13.82%	11.97%	-50.54%
tab0_7	6.51%	5.57%	5.33%	4.27%	4.30%	-33.95%

MAP results on RM1 (SCHMM)

4000 CD senones, 256 gaussians, Sphinx 3.x slow decoder

Speaker	Baseline	MAP	MLLR+MAP	Relative
bef0_3	8.87%	8.37%	7.87%	-11.27%
cmr0_2	7.10%	6.63%	6.42%	-9.58%
das1_2	6.07%	5.31%	4.75%	-21.75%
dms0_4	5.87%	5.25%	4.69%	-20.10%
dtb0_3	7.87%	7.13%	6.93%	-11.94%
ers0_7	7.10%	6.93%	6.60%	-7.04%
hxs0_6	10.08%	8.81%	7.60%	-24.60%
jws0_4	6.63%	5.92%	5.63%	-15.08%
pgh0_1	7.93%	7.13%	6.54%	-17.53%
rkm0_5	15.97%	14.09%	11.94%	-25.23%
tab0_7	5.89%	5.04%	4.63%	-21.39%

MAP on SRI CALO scenario meetings

- CALOBIG models, 5000 CD senones, 16 gaussians, Sphinx 3.x fast decoder
- One meeting in a sequence of five was adapted with the other four. Results were averaged for all five meetings.

Speaker	Baseline	MLLR	MLLR+MAP	Best relative WER
bill_deans	50.89%	47.37%	39.39%	-22.60%
lpound	56.54%	51.34%	38.90%	-31.20%
jpark	29.54%	27.71%	25.88%	-12.40%

References

- Chin-Hui Lee and Jean-Luc Gauvain, “Speaker Adaptation Based on MAP Estimation of HMM Parameters”. *Procedings of ICASSP 1993*, pp. 558-561.
- Chin-Hui Lee and Jean-Luc Gauvain, “MAP Estimation of Continuous Density HMM: Theory and Applications”. *Proceedings of DARPA Speech & Nat. Lang. 1992*.
- Qiang Huo and Chorkin Chan, “Bayesian Adaptive Learning of the Parameters of Hidden Markov Model for Speech Recognition”. *IEEE Transactions on Speech and Audio Processing*, 3:5, pp. 334-345.