Adaptive Training with Structured Transforms

K. Yu & M.J.F. Gales

5th Dec. 2003

Cambridge University Engineering Department

Overview

- Adaptive training with structured transforms:
  - review of ML cluster adaptive training
  - combination with constrained MLLR;
- MPE training for multiple cluster models:
  - form of smoothing function to use;
  - nature of prior to use.
- Initial performance evaluated on CTS English.

Structured Transforms

- Found data may be highly non-homogeneous:
  - multiple acoustic factors (e.g. gender/channel/style);
  - effects on acoustic signal of each factor varies;
- Multiple transforms:
  - a separate transform for each kind of unwanted variability;
  - nature of transform (should) reflect factor;
  - (possibly) more compact systems.
- Form examined in this work:
  - constrained MLLR (CMLLR) transforms;
  - interpolation weights in cluster adaptive training (CAT);
  - no explicit association of transform with factor.

CMLLR and CAT

- Likelihood of observation given by

$$p(\mathbf{o}(t)|m,s) \quad \propto \quad -\frac{1}{2}\log|\mathbf{\Sigma}^{(m)}| + \frac{1}{2}\log(|\mathbf{A}^{(s)}|^2)$$
$$-\frac{1}{2}(\mathbf{o}^{(s)}(t) - \boldsymbol{\mu}^{(sm)})^T\mathbf{\Sigma}^{(m)-1}(\mathbf{o}^{(s)}(t) - \boldsymbol{\mu}^{(sm)})$$

- Constrained Maximum Likelihood Regression (CMLLR)

$$\mathbf{o}^{(s)}(t) = \mathbf{A}^{(s)}\mathbf{o}(t) + \mathbf{b}^{(\mathbf{s})}$$

- Cluster Adaptive Training (CAT)

$$\boldsymbol{\mu}^{(sm)} = \mathbf{M}^{(m)}\boldsymbol{\lambda}^{(s)} \quad \mathbf{M}^{(m)} = \left[\boldsymbol{\mu}_1^{(m)}, \cdots, \boldsymbol{\mu}_P^{(m)}\right]$$

ML Parameters Estimation

- Multi-cluster canonical model (updates of variances not described)

$$\mathbf{G}^{(m)} = \sum_{s,t} \gamma_m(t) \boldsymbol{\lambda}^{(s)} \boldsymbol{\lambda}^{(s)T}$$

$$\mathbf{K}^{(m)} = \sum_{s,t} \gamma_m(t) \boldsymbol{\lambda}^{(s)} \mathbf{o}^{(s)}(t)^T \qquad \mathbf{M}^{(m)T} = \mathbf{G}^{(m)-1} \mathbf{K}^{(m)}$$

- Structured transforms estimation:
  - CAT interpolation weights for each speaker s

$$\mathbf{G}^{(s)} = \sum_{m,t} \gamma_m(t) \mathbf{M}^{(m)T} \boldsymbol{\Sigma}^{(m)-1} \mathbf{M}^{(m)}$$

$$\mathbf{k}^{(s)} = \sum_{m} \mathbf{M}^{(m)T} \boldsymbol{\Sigma}^{(m)-1} \left(\sum_{t} \gamma_m(t) \mathbf{o}(t) \right); \qquad \boldsymbol{\lambda}^{(s)} = \mathbf{G}^{(s)-1} \mathbf{k}^{(s)}$$

  - CMLLR: standard approach except using the adapted mean $\boldsymbol{\mu}^{(sm)}$

Minimum Phone Error Criterion

- MPE criterion

$$\mathcal{F}(\mathcal{M}) = \frac{\sum_w p(\mathbf{O}|\mathcal{M}_w)^\kappa P(w)\mathrm{RawAccuracy}(w)}{\sum_w p(\mathbf{O}|\mathcal{M}_w)^\kappa P(w)}$$

- Use weak-sense auxiliary function

$$\mathcal{Q}(\mathcal{M}) = \mathcal{Q}^n(\mathcal{M}) - \mathcal{Q}^d(\mathcal{M}) + \mathcal{G}(\mathcal{M}) + \log p(\mathcal{M})$$

  – $\mathcal{Q}^n(\mathcal{M})$ and $\mathcal{Q}^d(\mathcal{M})$ are standard auxiliary function for numerator and denominator

  – $\mathcal{G}(\mathcal{M})$ is smoothing function to improve stability

  – $\log p(\mathcal{M})$ is prior over the model parameters

- Simplified MPE adaptive training using fixed ML estimate of transforms.

Multi-Cluster Smoothing Function

- Smoothing function satisfies $\frac{\partial}{\partial \mathcal{M}}\mathcal{G}(\mathcal{M})\big|_{\hat{\mathcal{M}}} = 0$

- ML estimates for model given by $(\hat{\boldsymbol{\mu}}^{(sm)} = \hat{\mathbf{M}}^{(m)}\hat{\boldsymbol{\lambda}}^{(m)})$

$$\begin{aligned}\mathcal{G}(\mathcal{M}) &= -\sum_{m,s}\frac{D_m\nu_m^{(s)}}{2}(\log|\boldsymbol{\Sigma}^{(m)}| + \mathrm{tr}((\hat{\boldsymbol{\Sigma}}^{(m)} + \hat{\boldsymbol{\mu}}^{(sm)}\hat{\boldsymbol{\mu}}^{(sm)T})\boldsymbol{\Sigma}^{(m)-1}) \\ &\quad +\hat{\boldsymbol{\lambda}}^{(s)T}\mathbf{M}^{(m)T}\boldsymbol{\Sigma}^{(m)-1}(\mathbf{M}^{(m)}\hat{\boldsymbol{\lambda}}^{(s)} - 2\hat{\boldsymbol{\mu}}^{(sm)}))\end{aligned}$$

- Effective smoothing statistics

$$\mathbf{G}_D^{(m)} = \sum_s \nu_m^{(s)}\hat{\boldsymbol{\lambda}}^{(s)}\hat{\boldsymbol{\lambda}}^{(s)T}; \qquad \mathbf{K}_D^{(m)} = \mathbf{G}_D^{(m)}\hat{\mathbf{M}}^{(m)T}$$

- $\nu_m^{(s)}$ is normalised contribution from speaker s - $\nu_m^{(s)} = \frac{\sum_t \gamma_m^n(t)^{(s)}}{\sum_s\sum_t \gamma_m^n(t)}$

Multi-cluster Model Prior Function

- A possible form of prior is

$$
\begin{aligned}
\log p(\mathcal{M}) \quad = \quad & K - \frac{\tau^I}{2} \sum_{s,m} \tilde{\nu}_m^{(s)} \Big(\log |\boldsymbol{\Sigma}^{(m)}| + \mathrm{tr}(\tilde{\boldsymbol{\Sigma}}^{(m)} \boldsymbol{\Sigma}^{(m)-1}) \\
& + (\mathbf{M}^{(m)} \hat{\boldsymbol{\lambda}}^{(s)} - \tilde{\mathbf{M}}^{(m)} \hat{\boldsymbol{\lambda}}^{(s)})^T \boldsymbol{\Sigma}^{(m)-1} (\mathbf{M}^{(m)} \hat{\boldsymbol{\lambda}}^{(s)} - \tilde{\mathbf{M}}^{(m)} \hat{\boldsymbol{\lambda}}^{(s)}) \Big)
\end{aligned}
$$

- τ^I determines contribution of the prior;
- $\tilde{\nu}_m^{(s)}$ is the normalised contribution from speaker s -

$$
\tilde{\nu}_m^{(s)} = \frac{\sum_t \gamma_m^{ml}(t)^{(s)}}{\sum_s \sum_t \gamma_m^{ml}(t)}
$$

- Need to determine prior parameters $\tilde{\mathbf{M}}^{(m)}$ and $\tilde{\boldsymbol{\Sigma}}^{(m)}$

Multi-Cluster Prior Model Parameters

- Possible sources of prior information:
  - ML CAT model parameters (derived from $\mathbf{G}^{(m)}, \mathbf{K}^{(m)}$);
  - ML/MPE SI model parameters;
  - ML/MPE SAT model parameters;

- Possible form interpolate two sources - yields count smoothing:

$$\tilde{\mathbf{G}}^{(m)} = \frac{\mathbf{G}^{(m)} + \tau^M \sum_s \tilde{\nu}_m^{(s)} \hat{\boldsymbol{\lambda}}^{(s)} \hat{\boldsymbol{\lambda}}^{(s)T}}{\sum_m \gamma_m^{ml} + \tau^M}$$

$$\tilde{\mathbf{K}}^{(m)} = \frac{\mathbf{K}^{(m)} + \tau^M (\sum_s \tilde{\nu}_m^{(s)} \boldsymbol{\lambda}^{(s)}) \tilde{\boldsymbol{\mu}}_{SI}^{(m)T}}{\sum_m \gamma_m^{ml} + \tau^M}$$

- τ^M - balance between CAT-ML estimate and (approximate) SI MPE model.
  - $\tau^M \rightarrow \infty$ used for experiments (v.roughly tuned).

Multi-cluster MPE Estimates

- Complete update based on smoothing all accumulates:

$$\mathbf{G}^{(m)} = \sum_{s,t} \gamma_m^{mpe}(t)\hat{\boldsymbol{\lambda}}^{(s)}\hat{\boldsymbol{\lambda}}^{(s)T} + D_m\mathbf{G}_D^{(m)} + \tau^I\tilde{\mathbf{G}}^{(m)}$$

$$\mathbf{K}^{(m)} = \sum_{s,t} \gamma_m^{mpe}(t)\hat{\boldsymbol{\lambda}}^{(s)}\mathbf{o}^{(s)}(t)^T + D_m\mathbf{K}_D^{(m)} + \tau^I\tilde{\mathbf{K}}^{(m)}$$

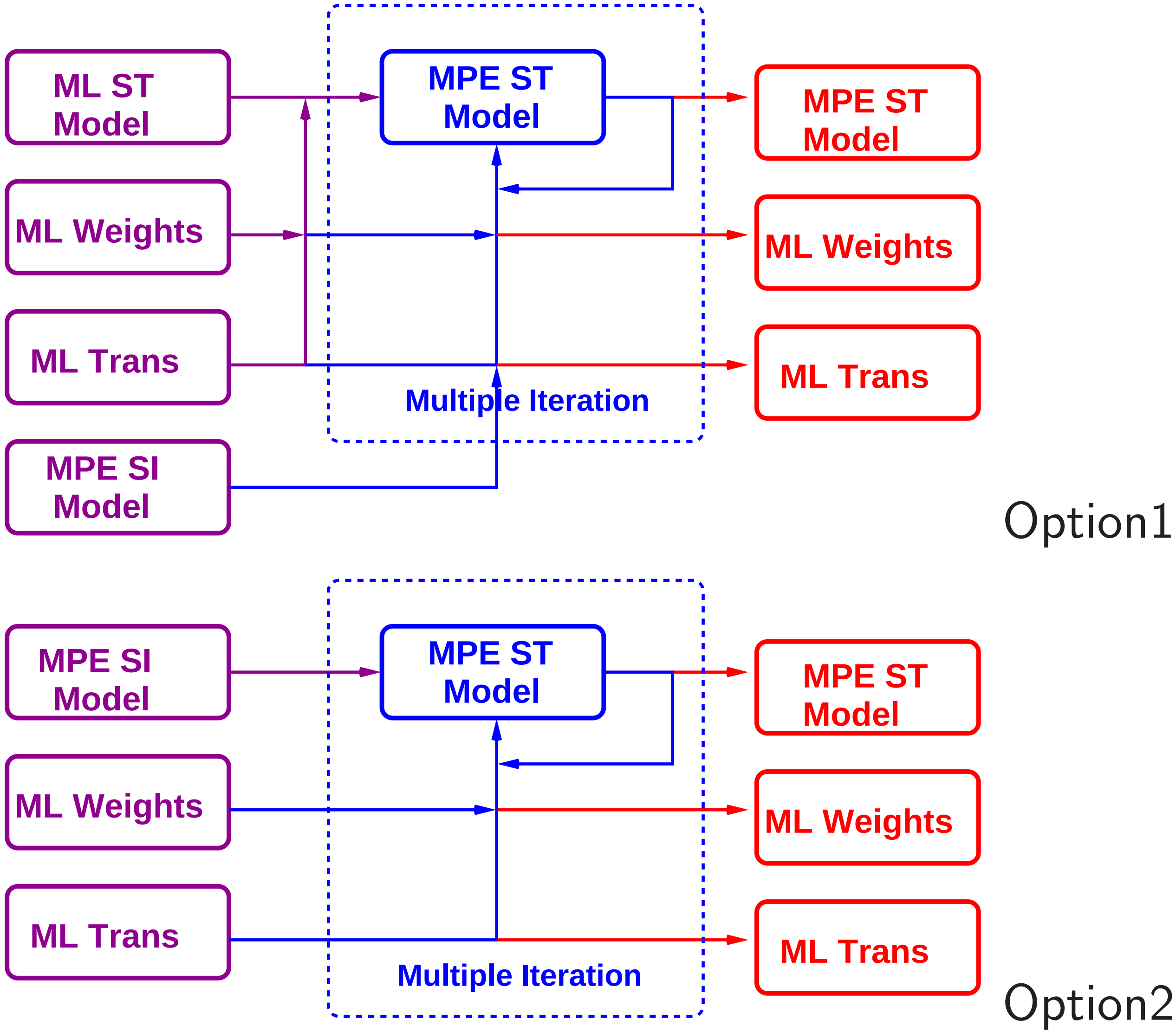
where $\gamma_m^{mpe}(t) = \gamma_m^n(t) - \gamma_m^d(t)$

- Multi-cluster model re-estimation based on

$$\mathbf{M}^{(m)T} = \mathbf{G}^{(m)-1}\mathbf{K}^{(m)}$$

- Initialisation of MPE training required.

MPE Training Configurations



Experiments on SwitchBoard System

- Switchboard (English): conversational telephone speech task
  - Training dataset: h5etrain03, 290hr, 5446spkr
  - Test dataset: dev01sub, 3hr, 59spkr
  - Front-end: PLP_0_D_A_T, HLDA and VTLN are used
  - Full decoding with trigram language model
- System based on option 1
  - 16 components and 28 components
  - 2 Sets of cluster means
  - adaptive training with structured transforms (CMLLR+CAT)
  - 4 MPE iterations used (possibly sub-optimal)
- Initialisation
  - Interpolation weights initialised using gender information;
  - CMLLR transforms initialised to identity transforms.

Initial Results on 16-Component System

System	Training Adaptation	Test Adaptation	Estimation MLE	Estimation MPE
GI	—	—	33.4	30.4
		CMLLR	31.5	28.3
GD	gender info	—	32.7	30.3
		CMLLR	30.9	28.4
GD (MPE-MAP)	gender info	—	—	29.7
		CMLLR	—	27.8
SAT	CMLLR	CMLLR	31.0	27.8
CAT	CAT	CAT	32.6	29.3
		ST	30.8	27.7
ST	ST	ST	30.6	27.3

- MPE-GD similar to MPE-GI, MPE-MAP significantly outperforms MPE-GD
- ML-CAT performs similar to ML-GD, MPE-CAT performs similar to MPE-MAP
- Adaptive training with ST significantly outperforms SAT

Initial Results on 28-Component System

System	Training Adaptation	Test Adaptation	Estimation	
			MLE	MPE
GI	—	—	32.2	29.4
		CMLLR	30.3	27.5
SAT	CMLLR	CMLLR	29.7	27.2
ST	ST	ST	29.4	26.9

- Adaptive training with ST slightly outperforms SAT;
- Gain on 28 components system is smaller than 16 components system;
- Further investigation of smoothing/model priors required.

Summary

- Structured transforms used in adaptive training;
- Current system combines;
  - Interpolation of multiple cluster means (CAT);
  - Contrained MLLR (SAT);
- MPE extended for multi-cluster models (CAT and eigenvoices);
- Small gains over CMLLR adaptive training system;
- Further work required on:
  - number of CAT cluster means and form of initialisation;
  - form of prior for multi-cluster models;
  - appropriate adaptive training scheme.