Discriminative Cluster Adaptive Training

Kai Yu

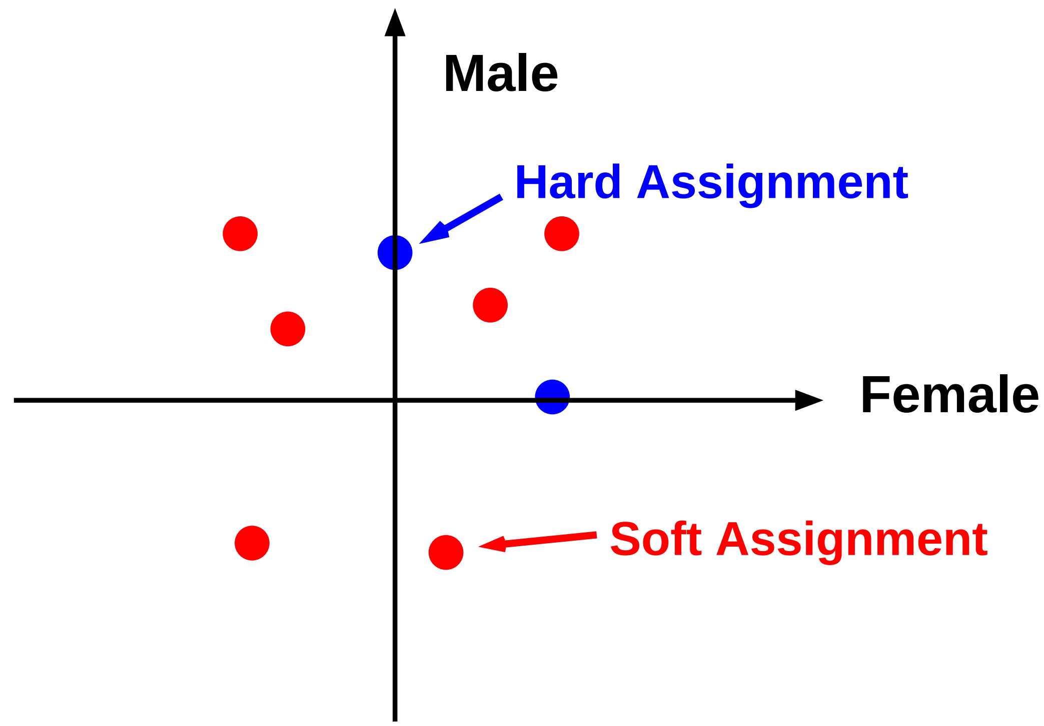
Mar. 2004

Cambridge University Engineering Department

Overview

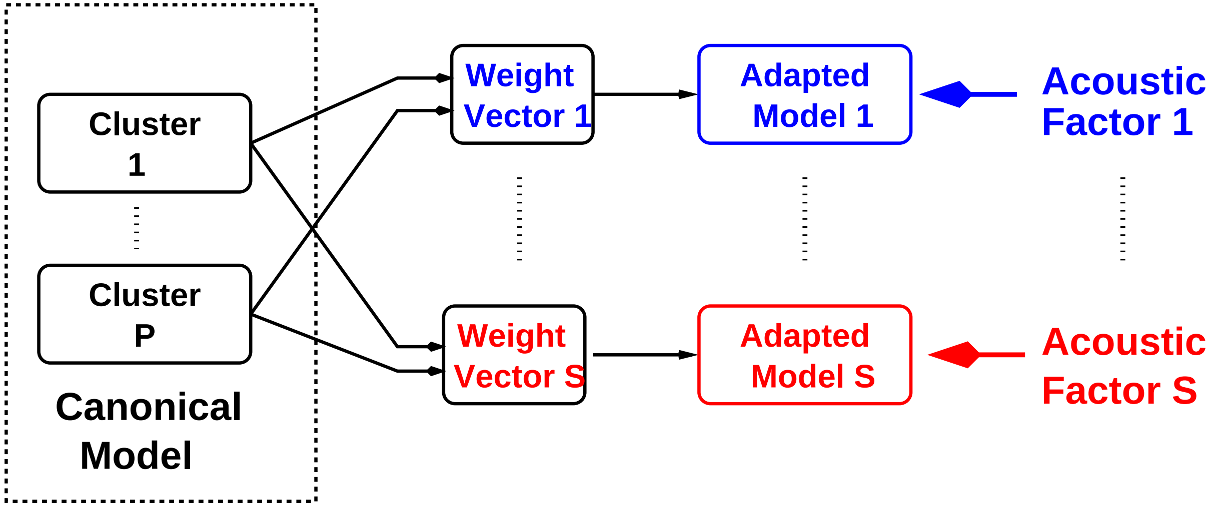
- Multi-cluster system and cluster adaptive training (CAT)
  - ML re-estimation of multi-cluster hmm model
  - ML re-estimation of interpolation weights
  - Initialisation
- MPE training for multi-cluster hmm model
  - form of smoothing function to use
  - nature of prior to use
- MPE training for interpolation weights
- Cluster adaptive training combined with constrained MLLR
- Performance evaluated on CTS English.

Hard Assignment and Soft Assignment



- Hard assignment only selects one of the GD models $\lambda_1 + \lambda_2 = 1$, $\lambda_1, \lambda_2 \in \{0, 1\}$
- Soft assignment can construct any linear combination of the two models $\lambda_1 + \lambda_2 = 1$, $\lambda_1, \lambda_2 \in \{-\infty, +\infty\}$, better use of axes

Cluster Adaptive Training



- Canonical model consists of
  - Common covariance, mixture weight and transition matrices
  - Cluster-specific mean vectors $\mathbf{M} = [\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_P]$, P is the number of clusters
- The speaker mean is given by interpolating among means of several clusters $\boldsymbol{\mu}^{(s)} = \mathbf{M}\boldsymbol{\lambda}^{(s)} = \sum_{c=1}^{P} \lambda_c \boldsymbol{\mu}_c$
- Iteratively training multi-cluster model and weights

ML Model Parameters Estimation

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

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

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

- $\mathbf{G}^{(m)}$ is a $P \times P$ matrix, $\mathbf{K}^{(m)}$ is a $P \times D$ matrix, P is cluster number, D is feature vector size

- If $P = 1$ and assume no scaling for speakers, the formula degrades to standard mean update

$$\mathbf{M}^{(m)} = \mu^{(m)} = \frac{\sum_{s,t} \gamma_m(t) \mathbf{o}^{(s)}(t)}{\sum_{s,t} \gamma_m(t)}$$

ML Weights Parameters Estimation

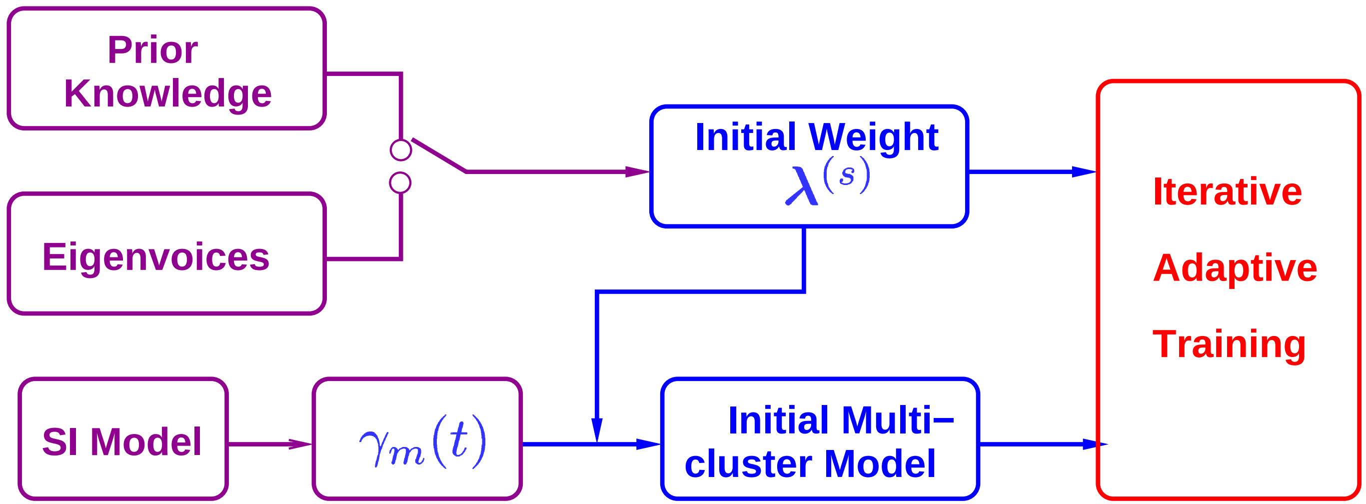
Interpolation weights for each speaker s

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

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

- $\mathbf{G}^{(s)}$ is a $P \times P$ matrix, $\mathbf{k}^{(s)}$ is a $P \times 1$ matrix
- Only $\gamma_m(t)$ and multi-cluster model are needed to estimate weights
- SI model used for initial alignment when estimating weights in testing adaptation

Initialisation



- Initialise weights
  - Prior knowledge: e.g. 2 cluster initialisation using gender information
  - Eigenvoices:
    Simple speaker-dependent model $\Longrightarrow$ meta-vector for each speaker $\Longrightarrow$ PCA on meta-vectors $\Longrightarrow$ eigen-meta-vectors (eigenvoices) $\Longrightarrow$ initial weights
  - Bias cluster $\boldsymbol{\lambda}^{(s)} = [\lambda_1^{(s)}, \cdots, \lambda_{P-1}^{(s)}, 1]$
- Construct initial multi-cluster model using standard model and initial weights

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 I-smoothing distribution over the model parameters to improve generalisation ability

Multi-Cluster Smoothing Function

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

- Standard smoothing function $\mathcal{G}(\mathcal{M}) = \sum_m \mathcal{G}_m\left(\boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)}; \hat{\boldsymbol{\mu}}^{(m)}, \hat{\boldsymbol{\Sigma}}^{(m)}\right)$

- Multi-cluster version: $\mathcal{G}(\mathcal{M}) = \sum_{s,m} \nu_m^{(s)} \mathcal{G}_m\left(\boldsymbol{\mu}^{(sm)}, \boldsymbol{\Sigma}^{(m)}; \hat{\boldsymbol{\mu}}^{(sm)}, \hat{\boldsymbol{\Sigma}}^{(m)}\right)$

- Difference between multi-cluster and standard smoothing function
  - Defined at speaker level, use $\hat{\boldsymbol{\mu}}^{(sm)} = \hat{\mathbf{M}}^{(m)}\boldsymbol{\lambda}^{(s)}$ and $\boldsymbol{\mu}^{(sm)} = \mathbf{M}^{(m)}\boldsymbol{\lambda}^{(s)}$
  - Add normalised contribution from speaker s - $\nu_m^{(s)}$, though for any $\nu_m^{(s)}$, $\mathcal{G}(\mathcal{M})$ is a valid smoothing function

$$\nu_m^{(s)} = \frac{\sum_t \gamma_m^n(t)^{(s)}}{\sum_s \sum_t \gamma_m^n(t)}$$

- Effective smoothing statistics are $D_m\mathbf{G}_D^{(m)}$ and $D_m\mathbf{K}_D^{(m)}$

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

  - D_m is a smoothing constant
  - $\mathbf{G}_D^{(m)}$ is a $P \times P$ matrix, $\mathbf{K}_D^{(m)}$ is a $P \times D$ matrix, P is cluster number, D is feature vector size
  - Sum over all speakers, note $\sum_s \nu_m^{(s)} = 1$
  - If $P = 1$ and assume no scaling for speakers, the formula degrades to standard MPE mean update

$$\mathbf{G}_D^{(m)} = 1; \qquad \mathbf{K}_D^{(m)} = \hat{\boldsymbol{\mu}}^{(m)}$$

Multi-cluster Model I-smoothing Distribution

- Standard I-smoothing distribution

$$\log p(\mathcal{M}) = \sum_m \log p(\boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)}; \tau^I, \tilde{\boldsymbol{\mu}}^{(m)}, \tilde{\boldsymbol{\Sigma}}^{(m)})$$

- Multi-cluster verstion:

$$\log p(\mathcal{M}) = \sum_{s,m} \tilde{\nu}_m^{(s)} \log p(\boldsymbol{\mu}^{(sm)}, \boldsymbol{\Sigma}^{(m)}; \tau^I, \tilde{\boldsymbol{\mu}}^{(sm)}, \tilde{\boldsymbol{\Sigma}}^{(m)})$$

- Main difference
  - Variables of interst are actually $\mathbf{M}^{(m)}$ and $\boldsymbol{\Sigma}^{(m)}$
  - Defined at speaker level, use $\boldsymbol{\mu}^{(sm)} = \mathbf{M}^{(m)}\boldsymbol{\lambda}^{(s)}$
  - Add normalised contribution from speaker s

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

- Key issue is to select appropriate prior $\tilde{\boldsymbol{\mu}}^{(sm)}$

Selection of I-smoothing Prior

- Different forms of prior
  - $\tilde{\boldsymbol{\mu}}^{(sm)} = \tilde{\mathbf{M}}_{ML}{}^{(m)}\boldsymbol{\lambda}^{(s)}$, $\tilde{\mathbf{M}}_{ML}$ is the ML estimates of multi-cluster mean matrix
  - $\tilde{\boldsymbol{\mu}}^{(sm)} = \tilde{\boldsymbol{\mu}}^{(m)}$, $\tilde{\boldsymbol{\mu}}^{(m)}$ is single-cluster prior mean vector, can be
    * Static (existing model parameters): ML-SAT, MPE-SI, etc.
    * Dynamic (from current accumulated statistics): ML-SI, MPE-SAT, etc.
  - $\tilde{\boldsymbol{\mu}}^{(sm)} = \tilde{\mathbf{M}}_{MAP}^{(m)}\boldsymbol{\lambda}^{(s)}$, $\tilde{\mathbf{M}}_{MAP}$ is the MAP estimates of multi-cluster mean matrix, a trade-off of the above two kinds of priors

- This work uses standard static MPE-SI model as the prior

- Effective I-smoothing statistics are $\tau^I\tilde{\mathbf{G}}^{(m)}$ and $\tau^I\tilde{\mathbf{K}}^{(m)}$

$$\tilde{\mathbf{G}}^{(m)} = \sum_s \tilde{\nu}_m^{(s)}\boldsymbol{\lambda}^{(s)}\boldsymbol{\lambda}^{(s)T}; \qquad \tilde{\mathbf{K}}^{(m)} = \left(\sum_s \tilde{\nu}_m^{(s)}\boldsymbol{\lambda}^{(s)}\right)\tilde{\boldsymbol{\mu}}^{(m)T}$$

Multi-cluster Model MPE Estimates

- Complete update based on smoothing all accumulates:

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

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

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

- Multi-cluster model re-estimation based on

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

Interpolation Weights MPE Estimates

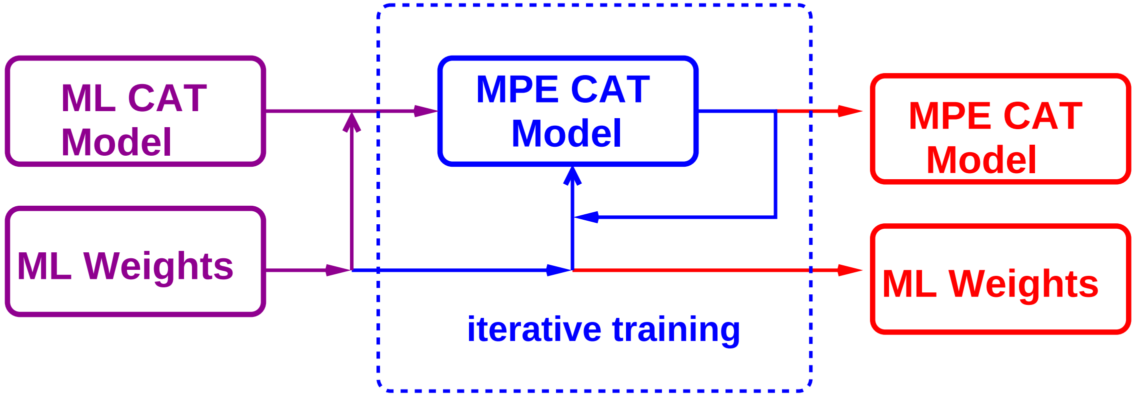
- Selection of smoothing function and I-smoothing distribution
  - Smoothing function has the same form as for multi-cluster model
  - Variable of interest in I-smoothing distribution is $\boldsymbol{\lambda}^{(s)}$, similar prior types can be selected

- Similar form of MPE Estimates: $\boldsymbol{\lambda}^{(s)} = \mathbf{G}^{(s)-1}\mathbf{k}^{(s)}$

$$\begin{aligned}
\mathbf{G}^{(s)} &= \sum_m \left(\left(\sum_t \gamma_m^{mpe}(t) \right) + D_m \mathbf{g}_D + \tau^I \tilde{\mathbf{g}} \right) \mathbf{M}^{(m)T} \boldsymbol{\Sigma}^{(m)-1} \mathbf{M}^{(m)} \\
\mathbf{k}^{(s)} &= \sum_m \mathbf{M}^{(m)T} \boldsymbol{\Sigma}^{(m)-1} \left(\left(\sum_t \gamma_m^{mpe}(t) \mathbf{o}(t) \right) + D_m \mathbf{k}_D + \tau^I \tilde{\mathbf{k}} \right)
\end{aligned}$$

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

Simplified MPE-CAT Training Procedure



- Multi-cluster model and weights are ML estimated
- Fix weights for further MPE training
- Only multi-cluster model is MPE updated

Structured Transforms

- Found data may be highly non-homogeneous
  - multiple acoustic factors (e.g. gender/channel/style);
  - effects on acoustic signal of each factor vary;
- 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

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

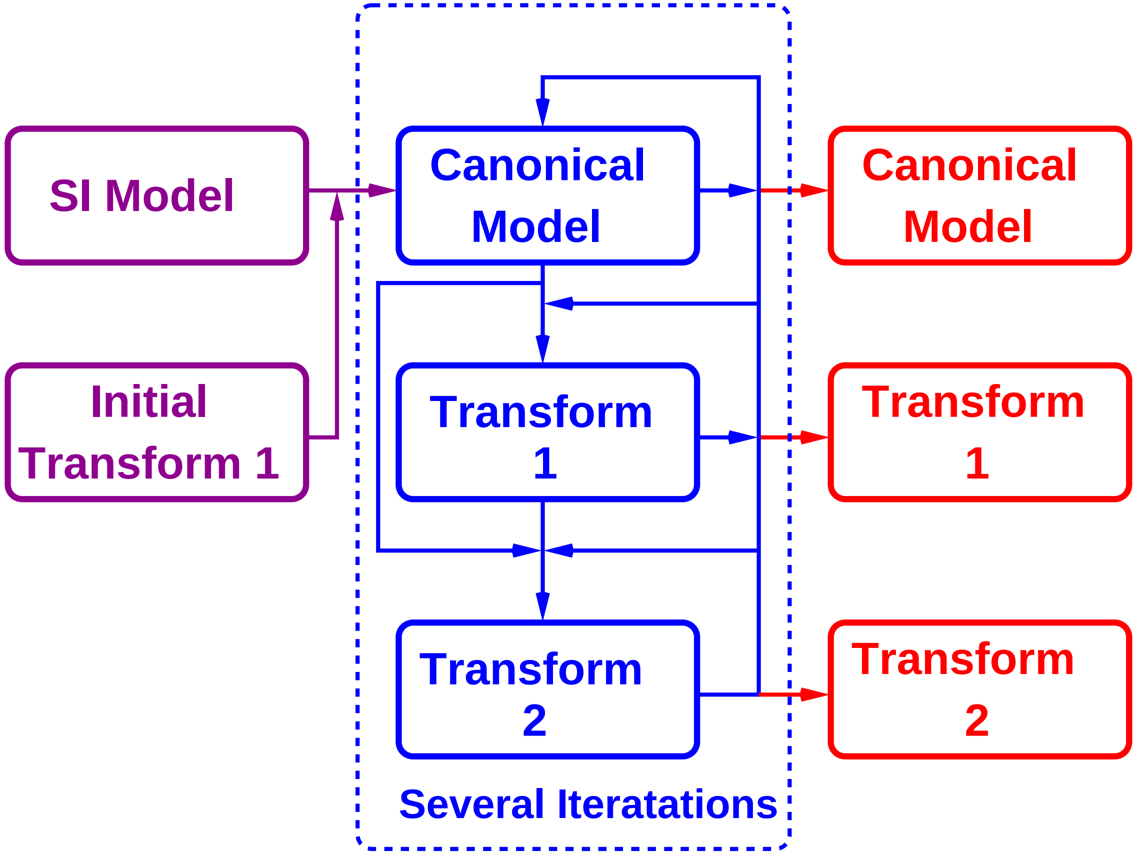
- Constrained Maximum Likelihood Regression (CMLLR)

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

- Cluster Adaptive Training (CAT)

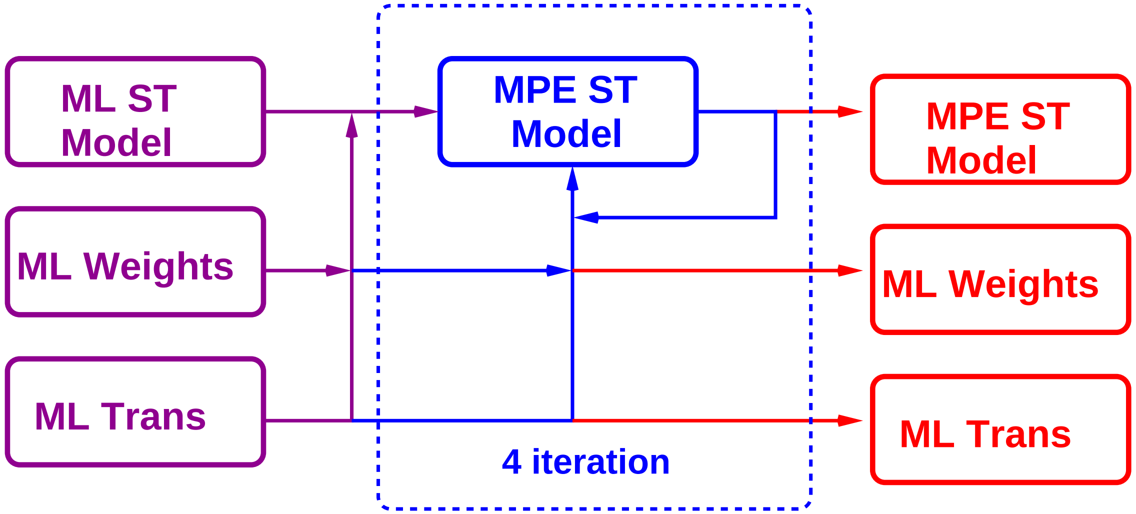
$$\boldsymbol{\mu}^{(sm)} = \mathbf{M}^{(m)}\boldsymbol{\lambda}^{(s)} \quad \mathbf{M}^{(m)} = \left[\boldsymbol{\mu}_1^{(m)}, \cdots, \boldsymbol{\mu}_P^{(m)}\right]$$

Procedure of ML Training with ST



- Canonical model consists of P sets of cluster means;
- Transform1 is interpolation weight, transform2 is CMLLR transform

Simplified MPE Training Procedure



- MPE training with ST for multi-cluster model parameters: The same as MPE-CAT training except for using transformed features:

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

- CMLLR transforms and interpolation weights are not discriminatively updated

Experiments on SwitchBoard System

- Switchboard (English): conversational telephone speech task
  - Training dataset: h5etrain03, 290hr, 5446spkr
  - Test dataset: dev01sub (3hr) and eval03 (6hr)
  - Front-end: PLP_0_D_A_T, HLDA and VTLN are used
  - Full decoding with trigram language model
- System description
  - 16 components and 28 components
  - All systems employed 4 ML iterations
- Initialisation
  - CMLLR transforms initialised to identity transforms.
  - Interpolation weights initialised using gender information/corpus information/eigenvoices;

Comparison of Different Initialisation

System	Initialization	Bias	#Cluster	dev01sub	eval03
MPE-SI	—	—	—	30.4	29.9
MPE-CAT	gender info.	no	2	29.3	29.1
	corpus info.	no	3	29.2	28.9
	eigenvoices	no	3	29.0	28.9
	eigenvoices	yes	2	29.3	29.2
	eigenvoices	yes	3	29.0	28.9
	eigenvoices	yes	4	29.0	28.9

- 16 component systems with 4 MPE iterations
- Most gain over MPE-SI was obtained by 2 cluster systems due to adaptive training
- 3 cluster systems outperforms 2 cluster systems, but more clusters did not help
- No wer difference between bias and non-bias 3 cluster eignvoices systems
- Eigenvoices initialised systems slightly outperformed corpus initialised systems

Results on 16-Component Systems (8 Iter.)

System	dev01sub		eval03	
	ML	MPE	ML	MPE
GI	33.4	29.5	33.3	29.5
GD	32.7	29.8	32.9	29.9
GD(MPE-MAP)	—	29.6	—	29.6
CAT(GD-Init)	32.6	29.1	32.6	29.0

- MPE systems significantly outperformed ML systems 3-4 percent absolute
- ML-GD system significantly outperformed ML-SI system,
- GD MPE-MAP needs tuning parameter, though outperforms MPE-GD
- MPE-CAT system still significantly outperformed MPE-SI system and gained 3.5 percent absolute over ML-CAT system

Results on 28-Component MPE Systems (8 Iter.)

System	Training Adaptation	Test Adaptation	dev01sub	eval03
SI	—	CMLLR	27.1	26.9
GD-MAP	gender info	CMLLR	27.5	26.7
SAT	CMLLR	CMLLR	26.9	26.6
CAT(GDInit)	CAT	ST	27.1	26.4
ST	ST	ST	26.7	26.2

- ST is CAT+CMLLR
- GD-MAP still needs tuning parameters
- Performance of SAT and CAT with ST in adaptation were similar;
- Adaptive training with ST obtained statistically significant gain on eval03

Summary

- MPE training for multi-cluster model and interpolation weights
  - Redefine smoothing function and I-smoothing distribution
  - Select appropriate priors of I-smoothing distribution
- Adaptive training with structured transforms: CAT+CMLLR
- Simplified MPE-training for CAT and ST-based systems
- Gains over other systems after adaptation
- Possibly more useful as amount of data increasing