Multiple-Cluster Adaptive Training Schemes

Mark Gales

mjfg@eng.cam.ac.uk

CAMBRIDGE UNIVERSITY
ENGINEERING DEPARTMENT

Adaptive Training

Adaptive training used in **all** state-of-the-art LVCSR system

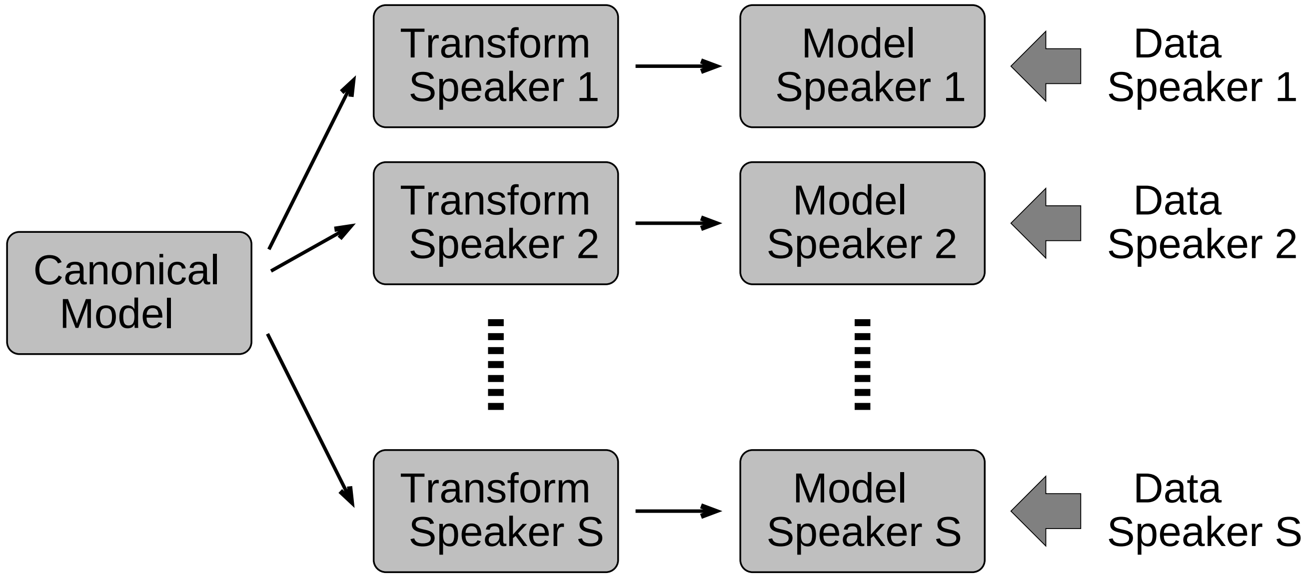
- removes "inter-speaker" variability - models "intra-speaker" variability.

Forms are:

- **Model independent**: no explicit use of any model information - CMN, Variance Normalisation;
- **ML Feature transformation**: VTLN, constrained MLLR;
- **ML Model transformation**: the model parameters, means and possibly variances, are transformed - speaker adaptive training (SAT), and cluster adaptive training (CAT).

This paper examines the use of *multiple clusters* with model-transformation adaptive training schemes.

Speaker Adaptive Training

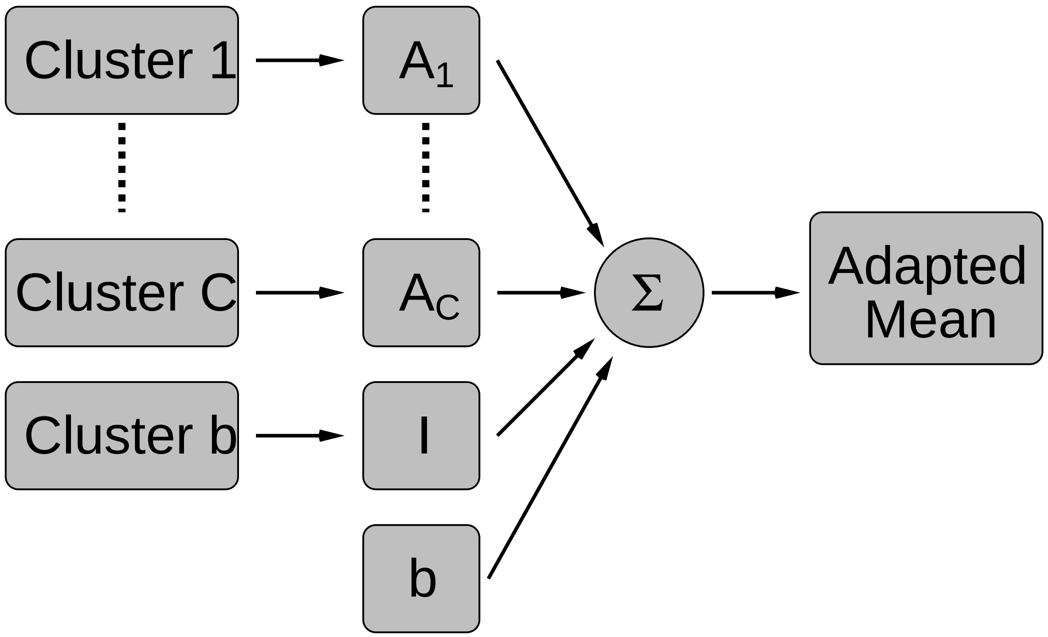


SAT a Maximum Likelihood based Model Transformation scheme:-

- **Canonical model** is a standard HMM model set;
- **Transformation** is MLLR.

This paper examines systems where the canonical model consists of multiple clusters.

Multiple-Cluster Transforms



Different systems (only considering adaptation of the means) described by:

- restricting number of clusters
- removing bias cluster
- restricting form of transform

Transform Descriptions

System	Transform Restrictions	Cluster Restrictions	Bias Cluster
SI	—	$C = 1$	No
SAT	—	$C = 1$	No
GD	$\mathbf{A}_c = \lambda_c \mathbf{I}$ $\lambda_c \in \{0, 1\}$, $\sum_c \lambda_c = 1$	$C = 2$	No
CAT	$\mathbf{A}_c = \lambda_c \mathbf{I}$	—	—
BSAT	—	$C = 1$	Yes
ESAT	—	—	—

General form for mean transform

$$\hat{\boldsymbol{\mu}}^{(sm)} = \mathbf{W}^{(s)} \boldsymbol{\xi}^{(m)}$$

where $\mathbf{W}^{(s)}$ is an $n \times p$ matrix and $\boldsymbol{\xi}^{(m)}$ is a $p \times 1$ vector, p is determined by the system.

CAT vs Eigenvoices

Same transform used for CAT and Eigenvoices:

- Difference in how clusters trained (CAT-ML, Eigenvoices - PCA);
- Difference in initialisation (CAT-clustering, Eigenvoices - PCA), CAT can use PCA;
- Original Eigenvoices only applicable for small tasks;
- Extension to Eigenvoices for LVCSR - MLES - approximation to model-based CAT;
- Eigenvoices ML projection same as ML interpolation of clusters and transforms;
- (Transform-based CAT for compact systems (Eigen-transforms!!);
- Eigenvoices is a far better name than CAT.

Bias SAT

Interesting extension to SAT

$$\begin{aligned}\hat{\boldsymbol{\mu}}^{(sm)} &= \mathbf{A}^{(s)}\boldsymbol{\mu}^{(m)} + \boldsymbol{\mu}_{\mathbf{b}}^{(m)} + \mathbf{b}^{(s)} \\ &= \begin{bmatrix} \mathbf{A}^{(s)} & \mathbf{I} & \mathbf{b}^{(s)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}^{(m)} \\ \boldsymbol{\mu}_{\mathbf{b}}^{(m)} \\ 1 \end{bmatrix}\end{aligned}$$

BSAT Features:

- Regression based around component specific point in acoustic space;
- Bias cluster capture aspects that are speaker independent;
- No increase in number of transform parameters over standard SAT system.

Extended SAT

General extension to SAT (and cluster transformations)

$$\hat{\boldsymbol{\mu}}^{(sm)} = \sum_{c} \mathbf{A}_c^{(s)} \boldsymbol{\mu}_{\mathbf{c}}^{(mc)} + \boldsymbol{\mu}_{\mathbf{b}}^{(m)} + \mathbf{b}^{(s)}$$

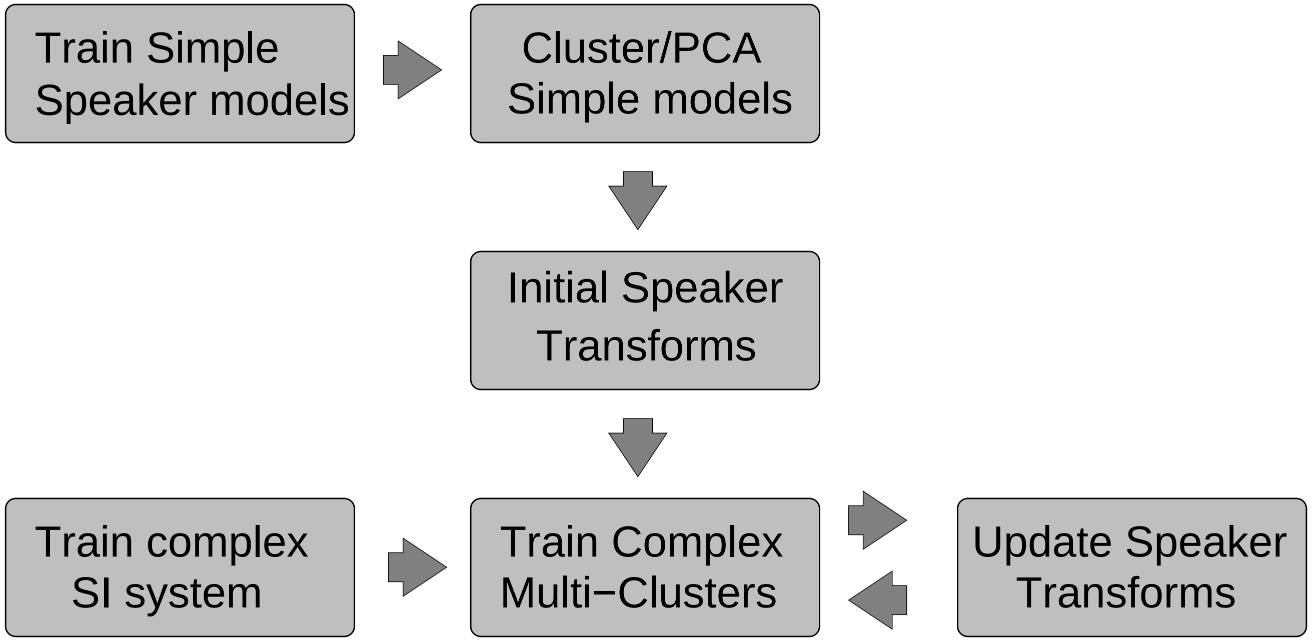
$$= \begin{bmatrix} \mathbf{A}_1^{(s)} & \ldots & \mathbf{A}_C^{(s)} & \mathbf{I} & \mathbf{b}^{(s)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{c}}^{(m1)} \\ \vdots \\ \boldsymbol{\mu}_{\mathbf{c}}^{(mC)} \\ \boldsymbol{\mu}_{\mathbf{b}}^{(m)} \\ 1 \end{bmatrix}$$

ESAT Features:

- Linear transform acts over all clusters;
- Optional bias clusters;
- Single bias vector.

ML Training

Training is a modified version of SAT training:



All the multi-cluster model parameters (cluster means, variances and weights) are updated (see paper for equations).

Initial Speaker Transforms

Two basic choices after generating simple speaker models:

- **Speaker clustering**:
  1. Using some distance measure cluster the simple speaker models.
  2. Yields a series of initial cluster assignments $\lambda_c^{(s)} \in \{0, 1\}$, $\sum_c \lambda_c^{(s)} = 1$;
  3. No bias cluster generated.
- **Eigenvoices**:
  1. Form a meta-vector from the means of each of the speaker models;
  2. Using PCA reduce the dimensionality to the required dimension;
  3. Project to obtain real weights, $\lambda_c^{(s)}$, in the reduced space;
  4. Bias cluster always generated.

Initial transforms generated using:

$$\mathbf{A}_c^{(s)} = \lambda_c^{(s)}\mathbf{I}, \qquad \mathbf{b}^{(s)} = \mathbf{0}$$

Preliminary Experiments

Baseline system:

- Standard LVCSR task - Wall Street Journal (1994);
- Cross-word, state-clustered, triphone system;
- 12 Gaussian components per state ($\approx 75K$);
- MFCC + log-energy + delta + delta-delta, CMN (39-dimensional);
- Trigram language model, 64K vocabulary;
- 2 Block Diagonal MLLR transforms (speech & silence);
- Supervised adaptation (40 sentences).

Multiple cluster systems:

- Two clusters for all systems (only means differ);
- Bias cluster used with CAT and BSAT;
- No bias cluster with GD and ESAT.

Results

System	Adapt Transform	Error Rate (%) H1 Dev	H1 Eval	Average
SI	—	9.38	10.01	9.71
	MLLR	8.24	9.24	8.77
SAT	MLLR	7.62	8.36	8.01
GD	GD	8.99	9.39	9.20
	ESAT	8.12	8.61	8.28
CAT	CAT	8.84	9.13	8.99
	ESAT	7.77	8.60	8.20
BSAT	BSAT	7.63	8.23	7.95
	ESAT	7.51	8.33	7.94
ESAT	ESAT	7.16	8.25	7.74

Conclusions

- Introduced general training of linear transform multiple cluster systems;
- Considered specific forms
  - GD and CAT - rapid adaptation;
  - SAT and BSAT - same complexity as MLLR;
  - ESAT - most complex - (slightly) better results.
- Described initialisation of such schemes
  - Speaker clusters - “hard” assignments;
  - Eigenvoices - “soft” assignment.
- From preliminary results - form of adaptive training of clusters important.