Discriminative Adaptation & Adaptive Training

Lan Wang & Phil Woodland

December 5th 2003

Cambridge University Engineering Department
Introduction

- Investigate linear transform parameter estimation for
  - Adaptive training
  - Unsupervised test set adaptation
  - Supervised adaptation (enrollment)

- Use MPE and MMI optimisation

- For speaker adaptive training estimate with consistent training criterion both
  - Linear transforms
  - Canonical models
Adaptive Training

Data for our current tasks contains much variability

- thousands of speakers,
- noisy background (environment),
- diversity of channels.

Adaptive training tries to remove some variability from the data during training

- Common model “independent” approaches:
  - Cepstral mean/variance normalization (CMN/CVN)
  - vocal tract length normalization (VTLN).

- Common model dependent approach: estimate linear transforms for each speaker/condition in both training and adaptation.
Speaker Adaptive Training

- SAT: speaker-specific train-set transforms are applied to the HMM parameter so as to construct a **canonical** HMM set.

- The canonical models with testing adaptation perform better than non-SAT models.
Speaker Adaptive Training (II)

- Maximum likelihood (ML) framework for transform estimation.
  - maximum likelihood linear regression (MLLR):
    \[
    \hat{\mu}_m = A\mu_m + b = W\xi_m
    \]
  - constrained MLLR:
    the same transforms are used to adapt means and covariances
    \[
    \hat{o}(t) = Ao(t) + b = W\zeta(t)
    \]
- Canonical model parameter re-estimation under ML criterion.
Discriminative Training

- Maximum mutual information (MMI) criterion.

\[ F_{MMI}(\lambda) = \sum_{r=1}^{R} \log \frac{P_\lambda(O_r | M^{w_r})^\kappa P(w_r)^\kappa}{\sum_{\hat{w}} P_\lambda(O_r | M^{\hat{w}})^\kappa P(\hat{w})^\kappa} \]

- Minimum phone error (MPE) criterion.

\[ F_{MPE}(\lambda) = \sum_{r=1}^{R} \frac{\sum_{s} p_\lambda(O_r | M^{w_s})^\kappa P(w_s)^\kappa \text{Raw Accuracy}(w_s, w_r)}{\sum_{u} p_\lambda(O_r | M^{w_u})^\kappa P(w_u)^\kappa}, \]

where \( \text{Raw Accuracy}(w_s, w_r) \) measures the accuracy of hypothesis \( w_s \).

- Lattice-based framework.
Optimization of Discriminative Criteria

- **Strong-sense auxiliary function** for ML optimization.

\[ Q(\lambda, \hat{\lambda}) - Q(\hat{\lambda}, \hat{\lambda}) \leq \mathcal{F}(\lambda) - \mathcal{F}(\hat{\lambda}) \]

- **Weak-sense auxiliary function** for the optimization of discriminative criteria.

\[ \frac{\partial}{\partial \hat{\lambda}} Q(\lambda, \hat{\lambda}) \bigg|_{\hat{\lambda}=\lambda} = \frac{\partial}{\partial \hat{\lambda}} \mathcal{F}(\hat{\lambda}) \bigg|_{\hat{\lambda}=\lambda} \]
Discriminative SAT

- **Discriminative SAT (DSAT):**
  
  discriminative criteria are used consistently to construct the canonical models in two sequential steps:
  
  - discriminative linear transform (DLT) generation,
  - discriminative model (canonical model) parameter re-estimation.

- **DSAT implementations:**
  
  - DLT & canonical model re-estimation.
  - constrained DLT & canonical model re-estimation.

- **A simplified implementation:**
  
  ML-based transform is used for discriminative model parameter re-estimation.
DLT Estimation

- Applying linear transforms to tune Gaussian components.
  - mean transform $\hat{W}$.
  - diagonal (full) variance transform: $\hat{\Sigma}_m = H^T \Sigma_m H$

- Weak-sense auxiliary function for the optimization MMI/MPE-based DLT.
  \[
  Q_{MMI}(W, \hat{W}) = Q^{num}(W, \hat{W}) - Q^{den}(W, \hat{W}) + Q_{sm}(W, \hat{W})
  \]
  \[
  Q^{num}(W, \hat{W}) = \sum_m \sum_t \gamma_{m}^{num}(t) \log N(o(t), \hat{W} \xi_m, \Sigma_m)
  \]

- The smoothing term should satisfy: \[
  \left. \frac{\partial Q_{sm}(W, \hat{W})}{\partial \hat{W}} \right|_{\hat{W}=W} = 0
  \]
  \[
  Q_{sm}(W, \hat{W}) = \sum_m D_m \left[ -\frac{1}{2} \left( \log |\hat{\Sigma}_m| + (W \xi_m - \hat{W} \xi_m)^T \hat{\Sigma}_m^{-1} (W \xi_m - \hat{W} \xi_m) + \Sigma_m \hat{\Sigma}_m^{-1} \right) \right]
  \]
**DLT Estimation (II)**

- Transform estimation for each row: $w^{(i)} = G^{(i)^{-1}} k^{(i)}$
  - ML accumulator:
    $$G^{(i)} = \sum_m \frac{1}{\sigma^2_{m(i)}} \gamma_m \xi_m \xi_m^T$$
  - MMI/MPE accumulator:
    $$G^{(i)} = \sum_m \frac{1}{\sigma^2_{m(i)}} \left( (\gamma_m^{\text{num}} - \gamma_m^{\text{den}}) + D_m \right) \xi_m \xi_m^T$$
    where $D_m = E \gamma_m^{\text{den}}$ with constant $E$.

- The num/den occupancies are computed as in MMI/MPE training.

**DLT for DSAT Model Re-estimation**

- Optimize MMI/MPE objective functions by applying transforms to the models.

- DLT for MMI/MPE-SAT model parameter re-estimation (e.g. for means).

\[
\hat{\mu}_m = M_m^{-1}V_m + \mu_m
\]

- ML statistics:

\[
M_m = \sum_{s} \gamma_m A(s)^T \Sigma_m^{-1} A(s)
\]

- MMI/MPE statistics:

\[
M_m = \sum_{s=1}^{S} \left( (\gamma_m^{num} - \gamma_m^{den}) + D_m \right) A(s)^T \Sigma_m^{-1} A(s)
\]

- Computational expensive.
Constrained DLT Estimation

- Weak-sense auxiliary function for the optimization of MMI/MPE-based constrained DLT.

- An iterative optimization to estimate transforms, like constrained MLLR.
  - ML accumulator:
    \[ G^{(i)} = \sum_m \frac{1}{\sigma^2_{m(i)}} \sum_t \gamma_m(t) \zeta(t) \zeta(t)^T \]
  - MMI/MPE accumulator:
    \[ G^{(i)} = \sum_m \frac{1}{\sigma^2_{m(i)}} \left( \sum_t \left( \gamma_m^{num}(t) - \gamma_m^{den}(t) \right) \zeta(t) \zeta(t)^T + D_m \begin{bmatrix} 1 & \tilde{\mu}_m \\ \hat{\Sigma}_m + \tilde{\mu}_m \tilde{\mu}_m^T \end{bmatrix} \right) \]

- The num/den occupancies are computed as in MMI/MPE training.

- Baseclass I-smoothing technique for MPE-based constrained DLT.
Constrained DLT for DSAT Model Re-estimation

- Constrained DLT for MMI/MPE-SAT model parameter re-estimation.
- Applying to the features makes canonical model parameter re-estimation more straightforward.
- The same updating formulas for MMI/MPE training can be used with adapted observations.

\[
\hat{\mu}_m = \frac{\theta_m^{num}(\hat{O}) - \theta_m^{den}(\hat{O}) + D_m\mu_m}{\{\gamma_m^{num} - \gamma_m^{den}\} + D_m} \\
\hat{\sigma}_m^2 = \frac{\theta_m^{num}(\hat{O}^2) - \theta_m^{den}(\hat{O}^2) + D_m(\sigma_m^2 + \mu_m^2)}{\{\gamma_m^{num} - \gamma_m^{den}\} + D_m} - \hat{\mu}_m^2
\]

- The num/den statistics are calculated in the same way as MMI/MPE training.
**CTS Experimental setup**

- Experiments on conversational telephone speech (CTS) transcription.
  - Training set: 76 hours CTS data/1118 conversation sides.
  - Test set (*dev01*): 6 hours CTS data/118 conversation sides.

- The front-end:
  - MF-PLP cepstral parameter (+Δ, +ΔΔ + ΔΔΔ),
  - HLDA projection (52 dim to 39 dim),
  - VTLN analysis.

- Basic GI HMM sets: 5920 tied-states/12 Gaussian components.
Training setup

- Start with the HLDA-ML model.

- Five iterations of interleaved (global) transform estimation and model parameter updating for ML-SAT models.

- Re-estimate MMI/MPE-SAT models with MLLR/CMLLR.

- Generate unconstrained (or constrained) MMI/MPE-based DLT (3 transforms per side).

- Seven iterations model parameter update with fixed transforms for MMI/MPE-SAT models.
Testing setup

- Unsupervised style testing adaptation in a sequential process.
  - 1-best CMLLR: for DSAT models with constrained linear transforms.
  - 1-best MLLR + lattice MLLR: for all DSAT systems.

- Lattice rescoring to evaluate the discriminative SAT models.

- Lattice generation in the same way as CTS RT03 system.
**DSAT with Constrained Linear Transform**

<table>
<thead>
<tr>
<th>System</th>
<th>Transform generation/parameter re-estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMI-SAT(+CMLLR)</td>
<td>constrained MLLR/MMI</td>
</tr>
<tr>
<td>MMI-SAT(+MMI_CDLT)</td>
<td>MMI-based constrained DLT/MMI</td>
</tr>
<tr>
<td>MPE-SAT(+CMLLR)</td>
<td>constrained MLLR/MPE</td>
</tr>
<tr>
<td>MPE-SAT(+MMI_CDLT)</td>
<td>MMI-based constrained DLT/MPE</td>
</tr>
<tr>
<td>MPE-SAT(+MPE_CDLT)</td>
<td>MPE-based constrained DLT/MPE</td>
</tr>
</tbody>
</table>

DSAT systems with constrained MLLR/DLT.

<table>
<thead>
<tr>
<th>Systems</th>
<th>SW-I</th>
<th>SW-II</th>
<th>Cell</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMI</td>
<td>21.1</td>
<td>33.4</td>
<td>33.1</td>
<td>29.2</td>
</tr>
<tr>
<td>MMI-SAT(+MMI_CDLT)</td>
<td>20.3</td>
<td>32.9</td>
<td>32.6</td>
<td>28.6</td>
</tr>
<tr>
<td>MPE</td>
<td>20.2</td>
<td>33.0</td>
<td>32.7</td>
<td>28.6</td>
</tr>
<tr>
<td>MPE-SAT(+MPE_CDLT)</td>
<td>20.1</td>
<td>31.8</td>
<td>31.8</td>
<td>27.8</td>
</tr>
</tbody>
</table>

%WER on test set *dev01* after (1-best) constrained MLLR adaptation

- After 1-best CMLLR, MMI/MPE-SAT give 0.6%-0.8% abs lower WER than non-SAT MMI/MPE systems.
MPE-SAT with Constrained Linear Transform

Average phone accuracy during MPE-SAT training.

- During training, MPE-SAT with MPE_CDLT outperforms the simplified implementation.

- After lattice MLLR, MPE-SAT with MPE_CDLT just improves the WER by abs 0.1% over MPE-SAT with CMLLR.

%WER for MPE-SAT systems on dev01 after lattice-based MLLR adaptation.
**MPE-SAT with Unconstrained Linear Transform**

- During training, MPE-SAT with MPE_DLT significantly outperforms the simplified implementation.

- After lattice MLLR, MPE-SAT with MPE_DLT gets almost same performance as MPE-SAT with MLLR.
DLT for Supervised Adaptation

- Supervised adaptation on WSJ task.

- The front-end: 39 dimensional MF-PLP features.

- The cross-word triphone HMMs.
  - ML training.
  - 6399 states/12 Gaussians.

- Testing set: NAB Spoke 3 (s3-dev/s3-eval) with enrollment set.

- H-criterion DLT (a version of MMI criterion) and MPE-based DLT:
  - mean + diagonal variance transforms.
  - regression tree with 16 baseclasses for sp/1 baseclass for sil.
# DLT for Supervised Adaptation (II)

<table>
<thead>
<tr>
<th>Test sets</th>
<th>iterations</th>
<th>MLLR</th>
<th>H-cri</th>
<th>MPE-DLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>s3-dev</td>
<td>1 ite</td>
<td>13.2</td>
<td>12.4</td>
<td>12.2</td>
</tr>
<tr>
<td>s3-eval</td>
<td>1 ite</td>
<td>11.1</td>
<td>10.3</td>
<td>10.1</td>
</tr>
<tr>
<td>s3-dev</td>
<td>3 ite</td>
<td>12.4</td>
<td>11.9</td>
<td>11.8</td>
</tr>
<tr>
<td>s3-eval</td>
<td>3 ite</td>
<td>10.4</td>
<td>10.1</td>
<td>10.0</td>
</tr>
</tbody>
</table>

%WER on NAB Spoke 3 after MLLR, H-criterion and MPE-based DLT adaptation.

- MPE-based DLT achieves 1% abs WER reduction over MLLR and 0.2% over H-criterion DLT (after 1 iteration).

- MPE-based DLT converges fast.
MPE-based DLT for Unsupervised Adaptation

- The cross-word triphone HMMs built with MPE training on CTS transcription.

- MLLR and MPE-DLT: 2 mean+diagonal variance transforms

- Using the hypothesis as supervision:
  - 1-best Viterbi outputs after lattice MLLR adaptation and confusion network (CN) decoding (WER: 27.0%).

<table>
<thead>
<tr>
<th>Adaptation</th>
<th>hypothesis</th>
<th>true trans</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLLR</td>
<td>27.7</td>
<td>(+CN) 27.0</td>
</tr>
<tr>
<td>MPE-DLT</td>
<td>27.3</td>
<td>(+CN) 26.9</td>
</tr>
</tbody>
</table>

%WER on dev01sub for MPE system.

- For unsupervised style, MPE-based DLT gets 0.1% gain after CN decoding over MLLR.
Discussions and Conclusions

- MMI/MPE-SAT can improve the performance by 0.7%-1.0% compared with non-SAT MMI/MPE training.

- Using MLLR/CMLLR to build MMI/MPE-SAT models is a simplified implementation.

- Using consistent discriminative criteria for MMI/MPE-SAT can give slight improvements under current testing adaptation scheme.

- DLT to adapt discriminative SAT models with unsupervised estimation

- Can also use for supervised adaptation which shows good performance