Discriminative Linear Transforms

**Goal:**
Develop discriminative versions of existing Maximum Likelihood training procedures.

**Focus on:**
Techniques that incorporate ML estimation of linear transforms during training.

**Prior Work:**
Both MLLT and SAT were developed as ML techniques but have also been used with MMI.

**Estimation Criterion:**
To develop discriminative versions of these techniques, we use Conditional Maximum Likelihood (CML) estimation procedures.

- **CMLLR** developed by Asela Gunawardana
- **Likelihood (CML) estimation procedures**
- **Estimation Criterion:** To develop discriminative versions of these techniques, we use Conditional Maximum Likelihood (CML) estimation procedures.
  - Subsequently using MMI for the estimation of the SI HMM Gaussian parameters.
  - Estimating the SD transforms under ML, and
  - Estimating the SD transforms under MMI by
  
**McDonough et al. (ICASSP'02)** combined SAT with MMI by
- Were then fixed throughout the subsequent iterations of MMI model estimation.
- Feature-based transforms obtained by ML estimation techniques, and
- Rebuilt-based transforms obtained by ML estimation techniques.

**The AT&T LVCSR-2001 System used:**

- **Prior Work:** Both MLLT and SAT were developed as ML techniques but have also been used with MMI.

- **SAT:** Apply speaker-dependent transforms to speaker-independent models.
- **MLLT:** Apply acoustic data to ease diagonal covariance Gaussian model assumption.

- **Goal:** Develop discriminative versions of existing Maximum Likelihood training procedures.

- **Discriminative Linear Transforms**
CMLAuxiliary function

As a result, both transforms and HMM Gaussian parameters are estimated discriminately. The goal is to maximize $P(O | \theta, W)$ by alternately updating the transforms and HMM parameters. CML versions of MLT and SAT are readily obtained.

State dependent distributions are reparametrized to incorporate the linear transforms.

2. Speaker adaptive training

1. Covariance modeling

We apply this criterion to two estimation problems:

$$\{ d = (s) \mathcal{Y} | s \} = \{ \theta \}$$

Parameters values are tied over sets of states, defined by the regression classes $s'$. $\theta$ is the parameter we wish to estimate under the CML criterion.

$$0 = \sum_{s, s} \log P(\theta | s) b \sum_{o} \log \Delta(\theta, s | o) b \int_{\theta} d \sum_{o} + (\theta, s | o) b \sum_{o} \log \Delta \cdot ((\theta)^{\mathcal{Y}, s} - (\theta)^{\mathcal{Y}, s})$$

CML criterion uses a general auxiliary function similar to EM.
Discriminative Likelihood Linear Transforms

**Objective:** We estimate the transforms and HMM parameters under the CML criterion. The transforms obtained under this criterion are termed Discriminative Likelihood Linear Transform (DLLT).

Let $Y$ be the observation vector, and $x$ be the hidden state vector. The discriminative likelihood is given by:

$$p(Y, x) = \frac{p(x, \theta) p(Y|x)}{p(x|\theta)}$$

where $p(x, \theta) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$ and $p(Y|x) = \prod_{i=1}^{T} \frac{1}{\sqrt{2\pi \sigma^2}} \exp\left(-\frac{(y_i - \sum_{j} a_{ij} x_j)^2}{2\sigma^2}\right)$

**Goal:** Transform feature vector to capture the correlation between the vector components.

**Objective:** We estimate the transforms and HMM parameters under the CML criterion. The transforms obtained under this criterion are termed Discriminative Likelihood Linear Transform (DLLT).

$$a = (s) a : s A \gamma = (\gamma - \gamma L) \gamma \gamma \gamma = \gamma L$$

Under the preceding model, the reparameterized emission density of state $s$ is

$$q + \sigma q = \gamma L$$

Apply affine transform matrix $L$ to the (extended) observation vector $Y$.
Effective DLLT Estimation

As in MILT (Gales '97), the row of the transformation matrix is found by

\[ \mathbf{V} \mathbf{X} \mathbf{W} \]

where \( \mathbf{X} \) is the estimate of the full covariance matrix.

**Problem:**

The diagonal terms of \( \mathbf{D} \) dominate when \( \mathbf{D} \) is diagonal.

- The large values of \( \mathbf{D} \) as used in MILT further exaggerate this effect.

- The resulting DLLT transformation is effectively identity.

**Solution:**

Replace \( \mathbf{D} \) by the estimate of its full covariance matrix.

\[
\begin{pmatrix}
\mathbf{V} \\
\mathbf{X} \\
\mathbf{W}
\end{pmatrix}
\begin{pmatrix}
\mathbf{X} \\
\mathbf{Y} \\
\mathbf{Z}
\end{pmatrix}
= \mathbf{X}
\]

\[
\begin{pmatrix}
\mathbf{X} \\
\mathbf{Y} \\
\mathbf{Z}
\end{pmatrix}
\begin{pmatrix}
\mathbf{A} \\
\mathbf{B} \\
\mathbf{C}
\end{pmatrix}
= \mathbf{A}
\]

Where

\[ \mathbf{I} \mathbf{A} (\mathbf{I} \mathbf{A} + \mathbf{I} \mathbf{B}) = \mathbf{I} \mathbf{A} \]
Discriminative Speaker Adaptive Training

Goal:
Reduce the inter-speaker variability within the training set.

Objective:
Compute the speaker dependent transformations and speaker independent parameters of the state dependent distribution under the CML criterion. We call this procedure Discriminative Speaker Adaptive Training (DSAT).

This estimation is performed as a two-stage iterative procedure:

1) We first maximize the CML criterion with respect to the speaker dependent affine transformations while keeping the speaker independent means fixed to their current values.

2) Subsequently, we compute the speaker independent means and variances using the updated values of the speaker dependent affine transformations.

Under the preceding model, the reparametrized emission density for state $s$ and speaker $s'$ is

$$f(x|\gamma, \theta, s') = \frac{|\Sigma_i|^{-\frac{1}{2}} |W_i^{-\frac{1}{2}}|}{\Gamma(\frac{\nu}{2})} \left( \frac{1}{\nu} \right)^{\frac{\nu}{2}} \left( \frac{|x - W_i \mu|}{\nu/2} \right)^{-\frac{1}{2}}$$

Apply speaker dependent transformations to speaker independent means.
System Description

Acoustic Models
– Standard HTK half-start training procedure
– Tied state, cross-word, context-dependent triphones
– 4000 unique triphone states
– 6 mixtures per speech state
– Tagged acoustic clustering to incorporate inflection and word-boundary info
– 866 utterances from the 2000 Hub-5 Switchboard-1 evaluation set (Swbd1) and 913 utterances from the 1998 Hub-5 Switchboard-2 evaluation set (Swbd2)
– Training: 16.4 hours from Switchboard-1 and 0.5 hour from Callhome English data
– Test: 866 utterances from the 2000 Hub-5 Switchboard-1 evaluation set (Swbd1) and 913 utterances from the 1998 Hub-5 Switchboard-2 evaluation set (Swbd2)

Training/Test Set
– The collection defined the minimum & maximum set for the 2001 JHU LVCSR system
Discriminative training requires alternate word sequences that are representative of the recognition errors.

- Obtain triphone lattices generated on the training data, using the AT&T FSM decoder.
- Obtain lattices generated on the training data, using the AT&T FSM decoder.
- Use a variation of the HTK regression class implementation.
- Use the Viterbi procedure over lattices generated with the same monophone are assigned to the same states of all context-dependent phones associated with the same monophone are assigned to the same initial class.
- Apply the HTK splitting algorithm to each of the initial classes.
- Constraint: all the mixture components associated with the same state belong to the same regression class.
- Assignments of Gaussians into classes:
  - These triphone segments are fixed throughout MMI training.
  - Backward procedure at the word level.
  - Use the Viterbi procedure over lattices generated with the same monophone are assigned to the same states of all context-dependent phones associated with the same monophone are assigned to the same initial class.
  - Apply the HTK splitting algorithm to each of the initial classes.
  - Constraint: all the mixture components associated with the same state belong to the same regression class.

---

MMI training & Regression Class Selection

Assignment of Gaussians into classes:

- These triphone segments are fixed throughout MMI training.
- Backward procedure at the word level.
- Use the Viterbi procedure over lattices generated with the same monophone are assigned to the same states of all context-dependent phones associated with the same monophone are assigned to the same initial class.
- Apply the HTK splitting algorithm to each of the initial classes.
- Constraint: all the mixture components associated with the same state belong to the same regression class.

---
Goals of the Experiments

- Proper seeding of DLLT and DSAT turns out to be crucial.

- Identify a proper initialization point for our discriminative techniques.

- Investigate fully discriminative training compared to ML training.

- Compare ML training techniques (MLLT, SAT) to their fully discriminative counterparts.

- Validate CML as a modeling procedure.

- Test whether CML transforms improve over ML transforms.

- Gaussian parameters are fixed throughout transform updates.

- Compare ML trained transforms to CML trained transforms.
Throughout the experiments we use a fixed set of regression classes.

**Table A**: Estimation of transforms under ML (MLLT) and CML (DLLT). No mean and variance update.

**Table B**: CML update of transforms and Gaussian parameters when seeded from the ML baseline.

**Table C**: CML update of transforms and Gaussian parameters when seeded from a well trained MLLT system.

**Observations**: 
- DLLT in isolation is better than MLLT (A).
- DLLT works best when initialized by MLLT (B vs. C).

<table>
<thead>
<tr>
<th>Transformed Regression Only</th>
<th>ML and MLLT</th>
<th>DLLT and MLLT</th>
<th>ML and DLLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML+DLLT-1it</td>
<td>39.9</td>
<td>44.9</td>
<td>37.5</td>
</tr>
<tr>
<td>ML+DLLT-2it</td>
<td>38.5</td>
<td>49.9</td>
<td>36.8</td>
</tr>
<tr>
<td>ML+DLLT-3it</td>
<td>37.4</td>
<td>48.9</td>
<td>35.9</td>
</tr>
<tr>
<td>ML+DLLT-4it</td>
<td>36.5</td>
<td>47.4</td>
<td>34.9</td>
</tr>
<tr>
<td>ML+DLLT-5it</td>
<td>35.6</td>
<td>46.0</td>
<td>34.0</td>
</tr>
<tr>
<td>ML+DLLT-6it</td>
<td>34.7</td>
<td>44.5</td>
<td>33.0</td>
</tr>
</tbody>
</table>

Throughout the experiments we use a fixed set of regression classes.
Decoding results include unsupervised MLR adaptation.
Throughout the experiments we used a fixed set of 2 regression classes (speech & silence).

<table>
<thead>
<tr>
<th>DSAT</th>
<th>DSAT Experiments - WER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML-SAT</td>
<td>34.1</td>
</tr>
<tr>
<td>ML-SAT</td>
<td>34.0</td>
</tr>
<tr>
<td>ML-SAT</td>
<td>34.3</td>
</tr>
<tr>
<td>ML-SAT</td>
<td>34.5</td>
</tr>
<tr>
<td>ML-SAT</td>
<td>43.4</td>
</tr>
<tr>
<td>ML-SAT</td>
<td>43.2</td>
</tr>
<tr>
<td>ML-SAT</td>
<td>43.4</td>
</tr>
<tr>
<td>ML-SAT</td>
<td>43.7</td>
</tr>
<tr>
<td>ML-SAT</td>
<td>43.9</td>
</tr>
</tbody>
</table>

**Conclusion:** Discriminative estimation improves over ML estimation of speaker dependent transforms and speaker independent mean parameters.
Conclusions

Integrated discriminative linear transforms into MMI estimation for LVCSR

Developed estimation procedures that find discriminative transforms in conjunction with:

- speaker-adaptive training
- feature normalization
- speaker-adaptive training

We have found that discriminative versions of speaker-adaptive training and feature normalization outperform MLLR estimation. Each technique gives approximately 0.8% absolute WER improvement on the Switchboard corpus over the MLLR estimation procedures.

Future work:
- DSAT and DLLT may yield complementary improvements in performance when used together if in fact they are capturing different acoustic phenomena.
- DSAT and DLLT were used in the JHU LVCSR-2002 Evaluation System.

For more information see:

http://www.clsp.jhu.edu/research/