Noise Adaptive Training for Subspace Gaussian Mixture Models

Liang Lu, Arnab Ghoshal, Steve Renals
University of Edinburgh
Introduction
  - Subspace GMM (SGMM) acoustic model

Noise adaptive training
  - Motivation
  - Adaptive training method
  - Experimental results

Conclusion
Subspace Gaussian Mixture Models [Povey, et al. 2010]

- Globally shared
  - $M_i$ is the projection matrix for means
  - $w_i$ is the projection vector for weights
  - $\Sigma_i$ is the covariance matrix
  - $i$ is the subspace component index

- State-dependent
  - $v_{jk}$ is low dimensional sub-state vectors (e.g. 40dim)
  - Gaussian mean: $\mu_{jki} = M_i v_{jk}$
Subspace Gaussian Mixture Models

\[
\begin{bmatrix}
    m_{11} & m_{12} \\
    \vdots & \vdots \\
    m_{i1} & m_{i2}
\end{bmatrix}
\cdot
\begin{bmatrix}
    v_1 \\
    v_2
\end{bmatrix}
\]

Liang Lu, Interspeech, August, 2013.
Subspace Gaussian Mixture Models

\[
\begin{bmatrix}
  m_{11} & m_{12} \\
  \vdots & \vdots \\
  m_{i1} & m_{i2}
\end{bmatrix}
\cdots
\begin{bmatrix}
  v_1 \\
  v_2
\end{bmatrix}
=\begin{bmatrix}
  \mu_1 \\
  \mu_2
\end{bmatrix}
\]
Subspace Gaussian Mixture Models

\[
\begin{bmatrix}
  m_{11} & m_{12} \\
  \vdots & \vdots \\
  m_{i1} & m_{i2}
\end{bmatrix}
\cdot
\begin{bmatrix}
  v_1 \\
  v_2
\end{bmatrix}
= 
\begin{bmatrix}
  \mu_1 \\
  \mu_2
\end{bmatrix}
\]

Liang Lu, Interspeech, August, 2013.
Subspace Gaussian Mixture Models

\[
\begin{bmatrix}
  m_{11} & m_{12} \\
  \vdots & \vdots \\
  m_{i1} & m_{i2}
\end{bmatrix}
\cdot
\begin{bmatrix}
  v_1 \\
  v_2
\end{bmatrix}
= 
\begin{bmatrix}
  \mu_1 \\
  \mu_2
\end{bmatrix}
\]

Liang Lu, Interspeech, August, 2013.
Subspace Gaussian Mixture Models

\[ \mu = Mv \]

Factorisation

\[ \mu = Mv +Ns \]

Speaker subspace
Subspace Gaussian Mixture Models

- Typical features
  - Re-structure the HMM-GMM model parameters
  - Smaller number of free model parameters
  - Large number of Gaussian components
  - Factorize the phonetic and speaker variability

- Outperforms GMM-based systems on several tasks, e.g. [D. Povey 2010, L. Burget 2010, L. Lu 2011]
Noise robustness

- Aurora 4 dataset
- GMM with 50K components
- SGMM with 6.4M components

Clean training speech $x$ → Clean acoustic model → Noisy test speech $y$

WER

GMM-clean, SGMM-clean, GMM-noisy, SGMM-noisy

Liang Lu, Interspeech, August, 2013.
Noise robustness

- SGMM with Joint uncertainty decoding (JUD [H. Liao, 2005])

Liang Lu, Interspeech, August, 2013.
Noise adaptation of clean speech model

- Adaptation with noise dependent transform for a specific noise condition
Noise adaptation of clean speech model

- Aurora 4 dataset
- A, B, C and D denote different noise conditions.

<table>
<thead>
<tr>
<th>Methods</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean model</td>
<td>5.2</td>
<td>58.2</td>
<td>50.7</td>
<td>72.1</td>
<td>59.9</td>
</tr>
<tr>
<td>+JUD</td>
<td>5.1</td>
<td>13.1</td>
<td>12.0</td>
<td>23.2</td>
<td>16.8</td>
</tr>
</tbody>
</table>

Liang Lu, Interspeech, August, 2013.
Noise adaptation of multi-condition model

- If the training data is from the same types of noise condition
We obtain better baseline system

However, we obtain worse results with adaptation

<table>
<thead>
<tr>
<th>Methods</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean model</td>
<td>5.2</td>
<td>58.2</td>
<td>50.7</td>
<td>72.1</td>
<td>59.9</td>
</tr>
<tr>
<td>+JUD</td>
<td>5.1</td>
<td>13.1</td>
<td>12.0</td>
<td>23.2</td>
<td>16.8</td>
</tr>
<tr>
<td>MST model</td>
<td>6.8</td>
<td>15.2</td>
<td>18.6</td>
<td>32.3</td>
<td>22.2</td>
</tr>
<tr>
<td>+JUD</td>
<td>7.4</td>
<td>13.3</td>
<td>14.7</td>
<td>24.1</td>
<td>17.6</td>
</tr>
</tbody>
</table>
Noise adaptive training scheme

- Iterative update of acoustic models $\mathcal{M}$ and noise transforms $\mathcal{T}$
- Optimization of $Q(\mathcal{T}; \hat{T})$ in [Lu, et al, 2013]
- Optimization of $Q(\mathcal{M}; \hat{\mathcal{M}})$ in this paper

Noise adaptive training - optimization

Optimization of $Q(\mathcal{M} ; \hat{\mathcal{M}})$


$$\theta = \tilde{\theta} - \zeta \left[ \left( \frac{\partial^2 Q(\cdot)}{\partial^2 \theta} \right)^{-1} \left( \frac{\partial Q(\cdot)}{\partial \theta} \right) \right]_{\theta = \tilde{\theta}} \quad (1)$$

- EM-based approach, e.g. noisy-CMLLR [Kim, et al 2011]

$$y_t = H^{(r)} x_t + g^{(r)} + e^{(r)}_{t} \quad \rightarrow \quad P(x_t | y_t, r) \quad (2)$$

- We used the EM-based approach for simplicity
With adaptive training, we obtained better results compared to the clean acoustic model.

<table>
<thead>
<tr>
<th>Methods</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean model</td>
<td>5.2</td>
<td>58.2</td>
<td>50.7</td>
<td>72.1</td>
<td>59.9</td>
</tr>
<tr>
<td>+JUD</td>
<td>5.1</td>
<td>13.1</td>
<td>12.0</td>
<td>23.2</td>
<td>16.8</td>
</tr>
<tr>
<td>MST model</td>
<td>6.8</td>
<td>15.2</td>
<td>18.6</td>
<td>32.3</td>
<td>22.2</td>
</tr>
<tr>
<td>+JUD</td>
<td>7.4</td>
<td>13.3</td>
<td>14.7</td>
<td>24.1</td>
<td>17.6</td>
</tr>
<tr>
<td>NAT model</td>
<td>6.5</td>
<td>20.3</td>
<td>19.8</td>
<td>39.7</td>
<td>27.6</td>
</tr>
<tr>
<td>+JUD</td>
<td>6.1</td>
<td>11.3</td>
<td>11.9</td>
<td>22.4</td>
<td>15.7</td>
</tr>
</tbody>
</table>
Experiments - noise adaptive training

- Effect of phase factor in the extended mismatch function
  \[ y = f\{x, n, h, \alpha\} \] [Deng, et al, 2004]

![Graph showing the relationship between word error rate and the value of phase factor. The graph includes a line labeled 'JUD-SGMM with Clean model'.]
Experiments - noise adaptive training

The value of phase factor

Word Error Rate (%)

JUD–SGMM with Clean model

JUD–SGMM with MST model

JUD–SGMM with NAT model

Liang Lu, Interspeech, August, 2013.
Summary

- Overview of subspace Gaussian mixture models
- Joint uncertainty decoding for noise robustness
- Adaptive training for multi-condition training data
- Experimental results demonstrate the effectiveness of this approach
- To integrate the noise robustness technique with more advanced system

Liang Lu, Interspeech, August, 2013.
Thanks for your attention!