Optimal Feature Spaces for Noise-Robust Speech Recognition

Rogier van Dalen

Presentation for Toshiba, 18 Sep 2007
Outline

Noise

Joint uncertainty decoding

Covariance modelling
  CMMLR
  Semi-tied covariance matrices

Results

Conclusion
The effects of noise on the feature space

- Different means;
- different variances;
- different covariances.

Thanks to Hank Liao for the figure and Toshiba Research Ltd. for the data.
Joint uncertainty decoding compensates models for noise.

\[ p(o_t | m) \propto \mathcal{N}(A^{(r)} o_t + b^{(r)}; \mu^{(m)}, \Sigma^{(m)} + \Sigma^{(r)}_{\text{bias}}) \]

Parameters estimated from channel noise \( \mu_h \) and additive noise \( (\mu_n, \Sigma_n) \).

\( \Sigma^{(r)}_{\text{bias}} \) can be

- diagonal: fast decoding, correlation not modelled;
- full: slow decoding, correlation modelled.
How to model the changing covariance . . .

. . . without the cost?

\[ p(o_t|m) \propto \mathcal{N}(\tilde{o}_t; \mu^{(m)}, \Sigma^{(m)} + \Sigma^{(r)}_{\text{bias}}) \]

where \( \tilde{o}_t = A^{(r)}o_t + b^{(r)} \).

- Diagonalise the covariance; or
- use a feature space transform, e.g. constrained MLLR; or
- use a structured precision matrix transform, e.g. semi-tied covariance matrices.
Predictive CMLLR

\[
p(o_t|m) \propto N(\tilde{o}_t; \mu^{(m)}, \Sigma^{(m)} + \Sigma^{(r)}_{\text{bias}}) \\
\approx N(A^{(r)}\tilde{o}_t + b^{(r)}; \mu^{(m)}, \Sigma^{(m)})
\]

Parameters are estimated

- iteratively;
- using statistics predicted by the Joint transform.
- Gathering statistics is fast;
- no model parameter adaptation;
- correlations not modelled well.
Predictive semi-tied covariance matrices

\[ p(o_t|m) \propto \mathcal{N}(\bar{o}_t; \mu^{(m)}, \Sigma^{(m)} + \Sigma^{(r)}_{\text{bias}}) \]
\[ \approx \mathcal{N}(A^{(r)}\bar{o}_t; A^{(r)}\mu^{(m)}, \tilde{\Sigma}^{(m)}_{\text{diag}}) \]

Parameters are estimated

- iteratively;
- using statistics predicted by the Joint transform.
- Gathering statistics is slower;
- model parameters have to be updated;
- correlations are modelled.
Predictive semi-tied covariance matrices: estimation

\[ p(o_t|m) \propto \mathcal{N}(A^{(r)}\tilde{o}_t; A^{(r)}\mu^{(m)}, \tilde{\Sigma}^{(m)}_{\text{diag}}) \]

Parameter estimation:
Start with \( A^{(r)} = I \) and \( \tilde{\Sigma}^{(m)}_{\text{diag}} = \text{diag}(\Sigma^{(m)} + \Sigma^{(r)}_{\text{bias}}) \)
Repeat \( J \) times:
1. Update \( A^{(r)} \).
2. \( \tilde{\Sigma}^{(m)}_{\text{diag}} = \text{diag} \left( A^{(r)}(\Sigma^{(m)} + \Sigma^{(r)}_{\text{bias}})A^{(r)^T} \right) \)

- Half an iteration, performing step 1 only, works well;
- 10 iterations works very well.
Experiments

Task:
- AURORA 2;
- small vocabulary digit recognition;
- low signal-to-noise ratios (20 to 5 dB).

System:
- Whole word digit models;
- 39 features: energy, 12 MFCCs, deltas and delta-deltas;
- 16 transforms;
- Joint transforms are estimated on stereo data (clean — noise-corrupted).
Results

On AURORA, SNR of 5dB.

<table>
<thead>
<tr>
<th></th>
<th>Word error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>59.2</td>
</tr>
<tr>
<td>Joint</td>
<td></td>
</tr>
<tr>
<td>Full</td>
<td>9.9</td>
</tr>
<tr>
<td>Diagonal</td>
<td>16.6</td>
</tr>
<tr>
<td>CMLLR</td>
<td>11.3</td>
</tr>
<tr>
<td>Predictive</td>
<td></td>
</tr>
<tr>
<td>Semi-tied half iteration</td>
<td>10.1</td>
</tr>
<tr>
<td>Semi-tied 10 iterations</td>
<td>9.9</td>
</tr>
</tbody>
</table>

- Predictive CMLLR performs much better than diagonal Joint;
- half-iteration semi-tied works almost as well as 10 iterations;
- 10-iteration semi-tied performs as well as full Joint uncertainty decoding;
- correlation modelling is important.
Conclusion

- Joint uncertainty decoding transforms can be approximated with
  - CMLLR; or
  - semi-tied covariance matrices.
- This gives a gain in accuracy without the loss in speed.

Future work:

- Using predictive Joint transforms on real-world data.
- Improving Uncertainty Decoding, e.g. using non-Gaussian distributions.
References I
