An Investigation into Speaker Informed DNN Front-end for LVCSR

Yulan Liu¹, Penny Karanasou², Thomas Hain¹

¹The University of Sheffield, UK; ²University of Cambridge, UK
¹{acp12y1, t.hain}@sheffield.ac.uk ²p.karanasou@eng.cam.ac.uk

Abstract

- Considerable interest in “informed training” of DNNs: DNN input is augmented with auxiliary codes carrying speaker information.
- This work - shows mathematical equivalence between speaker informed DNN training and “bias adaptation”;
- analyses influential factors such as dimension, discrimination and stability of auxiliary codes;
- compares different speaker informed DNN training methods in LVCSR task;
- introduces a system based on speaker classification followed by speaker informed DNN for short utterances.

This work was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

Background

Signal and feature level

<table>
<thead>
<tr>
<th>VTLN (150ms/500ms)</th>
<th>MCLAR (µs)</th>
</tr>
</thead>
</table>
| Integration into DNN
| LNN (Nat)1986 | LON (LJU10) |
| Speaker informed DNN training
| Eigenvectors (20020), speaker code (Sil2011), speaker code (Sil2011), UBIC (2014) |
| Encoding speaker information in DNN topology |

Speaker informed DNN training

Auxiliary codes:

- eigenvectors in speaker space
- speaker i-vectors
- speaker codes
- speaker separation bottleneck features (SSBN)

Equivalent overall bias: \( \beta(t) = \sum_{k=1}^{L} c(t)_k \beta, \beta = h(t) \)

with speaker dependent auxiliary codes →

Speaker dependent equivalent bias: \( \beta'_s = \sum_{k=1}^{L} c(t)_k \beta'_s \)

- Assume optimal bias \( \beta'_s \), ideally \( \beta'_s = \beta' \)

- force it to be true for all speakers →

  \[
  \begin{pmatrix}
  c_1 & c_2 & \cdots & c_L \\
  \vdots & \ddots & \vdots & \vdots \\
  \end{pmatrix}
  \begin{pmatrix}
  \beta_1 \\
  \beta_2 \\
  \vdots \\
  \beta_L \\
  \end{pmatrix} = \begin{pmatrix}
  \beta' - \beta' \\
  \vdots \\
  \end{pmatrix}
  \]

- Special case: auxiliary code matrix is a unit matrix: Unique Binary Index Codes (UBIC)

Frame-wise auxiliary codes

- AMI corpus, IVM, DNN-HMM-GMM (6 layered DNN);
- Training: 77.5h from 170 speakers;
- Test: 6.9h from 27 speakers (2.5h: 10 seen speakers; 4.4h: 17 unseen speakers);
- Average utterance length: 4.2 in training, 5s in test.

Results

<table>
<thead>
<tr>
<th>Dim</th>
<th>Seen Spkr</th>
<th>Unseen Spkr</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>23.8</td>
<td>23.6</td>
<td>23.7</td>
</tr>
<tr>
<td>8 dim codes</td>
<td>21.6</td>
<td>22.2</td>
<td>21.9</td>
</tr>
<tr>
<td>UBIC</td>
<td>25.8</td>
<td>26.4</td>
<td>26.1</td>
</tr>
<tr>
<td>SSBN</td>
<td>26.1</td>
<td>26.7</td>
<td>26.4</td>
</tr>
</tbody>
</table>

Experiments

- Speaker informed DNN training:
  - common mathematical framework for auxiliary codes in DNN input.
  - Equivalent to using speaker dependent biases;
  - The dimension, discriminability and stability of auxiliary codes all influence the performance in practice.

  - Performance
    - i-vectors and UBIC based methods achieved equivalent and the best performance;
    - SSBN-UBIC structure enables fast adaptation on short utterances without performance degradation.
    - Unseen speakers: no improvement potentially because: i-vectors: insufficient speaker diversity in training data; rest methods: lacking information about unseen speakers, system overfits to training.
    - Overall: i-vectors achieved best performance, followed by 8 dim hand-crafted binary index codes.