Speaker adaptive training in deep neural networks using speaker dependent bottleneck features

Rama Doddipatla [rama.doddipatla@crl.toshiba.co.uk]
Toshiba Research Europe Limited, Cambridge Research Laboratory, Cambridge, UK.

Overview
- Improve the DNN acoustic model using speaker adaptive training (SAT).
  - Transforming input features with VTLN, CMLLR or appending speaker information in the form of speaker codes fall into the frame work of SAT.
- This work focuses on tuning the weights of DNN to implement SAT.
- Proposed approach follows a two-stage architecture to implement SAT.
  - Stage-1 is a bottleneck (BN) feature extractor, where the weights of the BN layer are adjusted using speaker specific data while keeping the weights in rest of the layers fixed.
  - Stage-2 is the SAT-DNN model trained using the speaker dependent bottleneck (SDBN) features from stage-1.
- Unsupervised adaptation using SAT on Aurora4 task provides:
  - 8.6% WERR* on DNN trained with Mel filter-bank (FBANK) features.
  - 10.3% WERR* on DNN trained with CMLLR-FBANK.
- Supervised adaptation using one minute of audio improves the performance when compared with the performance of baseline DNN.

Experimental Setup
- Corpus: Aurora4
  - Train: 7138 Utterances, 83 speakers, multi-condition.
  - Test: 4620 Utterances, 8 speakers per test condition, 14 test conditions.
  - Each test speaker has 40 utterances (approx. 5 min of audio).
- Mel-filter bank features - 40 dimensions (No-LDA)
- Bi-gram language model (Vocabulary - 5K).
- Conventional DNN: 2048 (hid-dim) x 7 (layers)
- Bottleneck DNN: 512 (hid-dim) x 3 (layers), BN-dim : 75
- SI/SAT-DNN: 2048 (hid-dim) x 3 (layers)
- D-vector is obtained by averaging the BN features of a DNN trained using speaker labels as targets over an utterance.
  - DNN : 1024 (hid-dim) x 2 (layers), BN-dim : 40
- CMLLR transforms are estimated from the SAT-GMM model.

Results: Conventional Vs Two-stage DNN
<table>
<thead>
<tr>
<th>%WER</th>
<th>Conventional</th>
<th>Two-stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBANK</td>
<td>14.5</td>
<td>14.5</td>
</tr>
<tr>
<td>+ D-vec</td>
<td>13.9</td>
<td>13.9</td>
</tr>
<tr>
<td>+ CMLLR</td>
<td>12.5</td>
<td>12.5</td>
</tr>
<tr>
<td>+ CMLLR + D-vec</td>
<td>12.3</td>
<td>11.9</td>
</tr>
</tbody>
</table>

- Two-stage DNN seems to perform similar to the conventional DNN.
- Appending speaker information in the form of D-vectors or transforming the features with CMLLR improve the performance.

Results: Appending D-vectors
- D-vectors seem to provide similar gains irrespective of the position where they are introduced into the training, i.e either with FBANK or with BN features.
- Appending D-vectors, both with FBANK and BN features did not provide any gains in performance.

Results: Unsupervised adaptation of the proposed SAT
<table>
<thead>
<tr>
<th>%WER</th>
<th>Baseline</th>
<th>+ SAT-DNN</th>
<th>%WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBANK</td>
<td>14.5</td>
<td>13.2</td>
<td>8.9</td>
</tr>
<tr>
<td>+ D-vec</td>
<td>13.9</td>
<td>13.9</td>
<td>8.9</td>
</tr>
<tr>
<td>+ CMLLR</td>
<td>12.6</td>
<td>12.6</td>
<td>10.3</td>
</tr>
<tr>
<td>+ CMLLR + D-vec</td>
<td>11.9</td>
<td>11.2</td>
<td>5.9</td>
</tr>
</tbody>
</table>

- The proposed SAT consistently improves the performance over the baseline.
- Best gain in performance is obtained when SAT is applied on top of DNN trained with CMLLR-FBANK features.
- Performance of SAT saturates when applied on top of CMLLR-FBANK-D-vector system.

Results: Supervised adaptation of the proposed SAT
- Using as little as 10 utterances (approx. 1 min of audio) for supervised adaptation already improves the performance when compared with baseline.
- Performance improves as the data from a specific speaker increases.
- Less data is required to achieve similar performance of FBANK by applying CMLLR or appending D-vectors.

Conclusion
- Presented an approach to perform SAT in DNNs using a 2-stage architecture.
  - First stage is a BN-DNN, used for deriving SDBN features.
  - Second stage is the SAT-DNN model trained using SDBN features.
- Unsupervised adaptation using SAT on CMLLR-FBANK DNN provided the best performance (10.3% WERR).
- Supervised adaptation using one minute of audio improved the performance when compared with the performance of baseline DNN.

*WERR - Relative word error rate