### Deep Speaker Embedding Learning with Multi-Level Pooling for Text-Independent Speaker Verification

**Yun Tang, Guohong Ding, Jing Huang, Xiaodong He, Bowen Zhou**

**JD AI Research**

---

**WHAT’S NEW**

- A hybrid neural network structure using both TDNN and LSTM
- A multi-level pooling strategy to collect speaker information from both TDNN and LSTM layers
- A regularization scheme on the speaker embedding extraction layer to make the extracted embeddings suitable for the following fusion step

**DATA**

**Test Sets**

- NIST SRE16 eval set
- NIST SRE18 dev set (CMN2)

**Training data sets:**

- SRE data (2004-2006, 2008, and 2010), Switchboard, all Fisher data (1 & 2), all VoxCeleb data
- 13,564 hours data from 20,803 speakers
- Data augmentation to deal with different noise conditions

**LDA/PLDA adaptation:**

- SRE16 unlabelled data is used for SRE16 LDA/PLDA adaptation;
- SRE18 unlabelled data is employed for SRE18 LDA/PLDA adaptation.

**EXPERIMENTAL SETUP**

**From X-vector model to multiple-level pooling model (MP)**

- **Model Configurations**
  - **x-vector**: TDNN1-TDNN2-TDNN3-P
  - **A**: TDNN1-TDNN2-TDNN3-LSTM-P
  - **B**: TDNN1-TDNN2-TDNN3-LSTM-P
  - **MP**: TDNN1-TDNN2-TDNN3-P-LSTM-P

- **Loss Function**
  
  \[
  L = - \sum_{i=1}^{M} \log \frac{\exp{w^T x_i + b_i}}{\sum_j \exp{w^T x_i + b_j}} + \lambda ||z_i||_2
  \]

- **X-vector Baseline**
  - Frame level: 3 TDNN layers
  - Speaker Level: Statistic Pooling + 3 fully connected layers

- **Multi-level Pooling**
  - TDNN focuses on the local feature representation
  - LSTM focuses on sequential and global feature representation.
  - Multi-level pooling collects different level representations to model the target speaker
  - Regularization on the embedding extraction layer helps to extract robust representation for the backend process.

**SRE16 RESULTS**

<table>
<thead>
<tr>
<th>Model</th>
<th>λ = 0</th>
<th>λ = 0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>x-vector</td>
<td>7.61</td>
<td>10.98</td>
</tr>
<tr>
<td>A</td>
<td>8.17</td>
<td>11.78</td>
</tr>
<tr>
<td>B</td>
<td>6.64</td>
<td>9.80</td>
</tr>
<tr>
<td>MP</td>
<td>6.68</td>
<td>9.74</td>
</tr>
</tbody>
</table>

**SRE18 RESULTS**

<table>
<thead>
<tr>
<th>λ</th>
<th>model</th>
<th>EER</th>
<th>DCF(0.01)</th>
<th>DCF(0.005)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>x-vector</td>
<td>7.29</td>
<td>0.593</td>
<td>0.651</td>
</tr>
<tr>
<td>A</td>
<td>7.90</td>
<td>0.581</td>
<td>0.656</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>7.16</td>
<td>0.567</td>
<td>0.632</td>
<td></td>
</tr>
<tr>
<td>MP</td>
<td>7.46</td>
<td>0.594</td>
<td>0.673</td>
<td></td>
</tr>
<tr>
<td>0.001</td>
<td>x-vector</td>
<td>7.77</td>
<td>0.525</td>
<td>0.586</td>
</tr>
<tr>
<td>B</td>
<td>7.61</td>
<td>0.506</td>
<td>0.571</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2:** Results on SRE16 eval test.

**Table 3:** DCF scores for SRE16 (pooled) test set.

**Table 4:** Evaluation results on the SRE18 (CMN2) dev set.

- TDNN + LSTM helps to reduce EER by 12% in SRE16
- Regularization improves the verification performance on the backend
- Multiple-pooling from different sources gives the best results on both SRE16 and SRE18 test.