I-Vector DNN Scoring and Calibration for Noise Robust Speaker Verification

Zhili TAN and Man-Wai MAK
Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University

Introduction

• Observing that adverse acoustic conditions and duration variability in utterances could have detrimental effect on PLDA scores, a number of score calibration methods have been proposed to compensate for the effect by modeling it as a shift in the PLDA scores.

• We propose to estimate the score shifts or the ideal clean scores by multitask DNNs using noisy i-vector pairs and their corresponding PLDA scores as input.

• Results based on noise contaminated speech in NIST 2012 SRE suggest that the multi-task DNNs can effectively calibrate the scores produced by a PLDA model, leading to superior performance as compared to the conventional linear calibration method.

Motivation

• Quality measure function (QMF)-based calibrated score:
  \[ S'_1 = w_0 + w_1 S + Q_{\text{SNR}_u, \text{SNR}_{ref}} = w_0 + w_1 S + w_2 \text{SNR}_u + w_3 \text{SNR}_{ref} \]
  where \( S \) is the uncalibrated score and \( Q \) is the approximated score shift.

• Ideal Score Shift \( \delta_{\text{score}} \equiv \text{PLDA}(x^{\text{in}}, x^{\text{ref}}) - \text{PLDA}(x^{\text{in}}, x^{\text{ref}}) \)

• However, the relationship between score shifts and utterances' SNR is fairly complex and definitely non-linear (see Fig. 2).

• At low SNR, the score shifts will become more difficult to estimate, which is a drawback of the methods that entirely rely on SNR of utterances.

DNN-Based Score Calibration

• DNN Score Compensation: Estimating Score Shifts by DNNs:
  \[ \text{DNN}_c(x_{\text{test}}, x_{\text{ref}}, S) = \delta_{\text{score}}, \quad S'_2 = S + \text{DNN}_c(x_{\text{test}}, x_{\text{ref}}, S) \]

• DNN Score Transformation: Recovering Clean PLDA Scores by DNNs:
  \[ S'_3 = \text{DNN}_c(x_{\text{test}}, x_{\text{ref}}, S) = S_{\text{cal}} \]

Multi-task DNNs for Score Calibration

• Having a single source of errors makes the backpropagation (BP) of error gradients very inefficient.

• One possible solution is to introduce some auxiliary tasks for the network to learn:
  \[ \text{DNN}_c(x_{\text{test}}, x_{\text{ref}}, S) \approx \begin{bmatrix} \delta_{\text{score}} \\ \text{SNR} \text{ of Test Speech} \\ \text{SNR} \text{ of Target Speech} \\ \text{Duration of Test Speech} \\ \text{Duration of Target Speech} \end{bmatrix} \]

Fig. 1: Multitask DNN with classification and regression tasks.

• Recover Clean Scores by Multi-task DNN: \( S'_4 = \text{DNN}_c(x_{\text{test}}, x_{\text{ref}}, S)[3] \equiv S_{\text{cal}} \)

• Estimate Score Shifts by Multi-task DNN: \( S'_5 = S + \text{DNN}_c(x_{\text{test}}, x_{\text{ref}}, S)[4] \equiv S_{\text{cal}} \)

• Posterior odds by Multi-task DNN: \( S'_6 = \log \left( \frac{\text{DNN}_c(x_{\text{test}}, x_{\text{ref}}, S)[3]}{\text{DNN}_c(x_{\text{test}}, x_{\text{ref}}, S)[2]} \right) = \log \left( \frac{S_{\text{cal}}}{S_{\text{in}}} \right) \)

DNN Scoring Machine

• A multi-task DNN without using the noisy PLDA scores as input.

• Advantage: the PLDA model is not necessary during the scoring stage, i.e., given an i-vector pair, we can obtain the approximated clean score or score shift from the DNN’s outputs:
  \[ \text{DNN}_c(x_{\text{test}}, x_{\text{ref}}) \approx \begin{bmatrix} \text{Classification} \\ \text{Regression} \end{bmatrix} \]

• Recover Clean Scores by DNN scoring machine: \( S'_7 = \text{DNN}_c(x_{\text{test}}, x_{\text{ref}})[3] \equiv S_{\text{cal}} \)

• Estimate Score Shifts by DNN scoring machine: \( S'_8 = S + \text{DNN}_c(x_{\text{test}}, x_{\text{ref}})[4] \equiv S_{\text{cal}} \)

• Posterior odds DNN scoring machine: \( S'_9 = \log \left( \frac{\text{DNN}_c(x_{\text{test}}, x_{\text{ref}})[3]}{\text{DNN}_c(x_{\text{test}}, x_{\text{ref}})[2]} \right) = \log \left( \frac{S_{\text{cal}}}{S_{\text{in}}} \right) \)

Results on CC4 of NIST 2012 SRE (male)

• Test utterances are contaminated with different levels of babble noise.

Fig. 2: The distributions of ideal score shifts with respect to the SNR of test utterances when the target-speaker utterances is clean.

• The uncalibrated PLDA scores play an important role in the calibration DNN.

• The classification task plays an important role in the training of the multi-task DNN.

Table 1: Performance of various DNN-based score calibration methods.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Score Calibration Method</th>
<th>EER(%)</th>
<th>minDCF</th>
<th>EER(%)</th>
<th>minDCF</th>
<th>EER(%)</th>
<th>minDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S'_7 ) Estimate SNR-dep Score Shift</td>
<td>1.68</td>
<td>0.209</td>
<td>2.28</td>
<td>0.269</td>
<td>5.35</td>
<td>0.754</td>
<td></td>
</tr>
<tr>
<td>( S'_8 ) Recover Clean Scores by DNN</td>
<td>1.56</td>
<td>0.193</td>
<td>2.21</td>
<td>0.239</td>
<td>3.58</td>
<td>0.430</td>
<td></td>
</tr>
<tr>
<td>( S'_9 ) Estimate Score Shifts by DNN</td>
<td>1.54</td>
<td>0.192</td>
<td>2.21</td>
<td>0.238</td>
<td>3.57</td>
<td>0.428</td>
<td></td>
</tr>
<tr>
<td>( S'_{10} ) Use Posterior Odds as Scores</td>
<td>1.70</td>
<td>0.193</td>
<td>2.23</td>
<td>0.245</td>
<td>3.56</td>
<td>0.426</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Performance of multi-task DNNs without using the noisy PLDA scores as input.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Scoring Method</th>
<th>Score Calibration Method</th>
<th>EER(%)</th>
<th>minDCF</th>
<th>EER(%)</th>
<th>minDCF</th>
<th>EER(%)</th>
<th>minDCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S'_7 ) Recover Clean Score</td>
<td>N/A</td>
<td>2.51</td>
<td>0.308</td>
<td>3.02</td>
<td>0.349</td>
<td>3.61</td>
<td>0.456</td>
<td></td>
</tr>
<tr>
<td>( S'_8 ) PLDA Score Shift</td>
<td>1.39</td>
<td>0.166</td>
<td>1.96</td>
<td>0.230</td>
<td>3.80</td>
<td>0.571</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( S'_{10} ) Posterior Odds</td>
<td>N/A</td>
<td>3.37</td>
<td>0.415</td>
<td>4.52</td>
<td>0.445</td>
<td>5.54</td>
<td>0.660</td>
<td></td>
</tr>
</tbody>
</table>