DEEP CNN BASED FEATURE EXTRACTION FOR TEXT-PROMPTED SPEAKER RECOGNITION

Sergey Novoselov1,2, Oleg Kudashev2, Vadim Shchemelinin1, Ivan Kremnev2, Galina Lavrentyeva1
1IMU University, St. Petersburg, Russia; 2STC-innovations Ltd., St. Petersubrg, Russia; 3STC Ltd., St. Petersubrg, Russia
{novoselov, kudashev, shchemelinin, kremnev, lavrentyeva}@speechpro.com

1 Introduction

• Text-dependent speaker recognition task [1,2,3,4] is studied
• Deep convolutional neural network based speaker specific features extractor in the text-prompted speaker verification task is presented
• The prompted passphrase is segmented into word states —i.e. digits — to test each digit utterance separately
• A single high-level feature extractor for all states is used and cosine similarity metric is applied for scoring
• Multitask learning scheme is used to train the high-level feature extractor

2 Features

Viterbi segmentation to word states

Input features for the CNN are 64 × 96 log mel power spectra:
• 64 frequency bands
• 96 frames (longest single digit utterance)
• Voice activity detector removes non-speech frames

3 Convolutional Neural Network

Input features are processed with a CNN embedding extractor

Max Feature Mapping

• Max Feature Mapping (MFM) reduces dimensionality and selects features
• Pre-softmax layer produces speaker embeddings
• Last dense layers are included only during training

4 Learning mode

Single-task

• Extractor is trained to discriminate speakers
• N speakers neurons at softmax layer

Multi-task

• Extractor is trained to discriminate speakers and word states
• N speakers × N digits neurons at softmax layer

5 Experiments

We explored 5-digit password verification scenario when the speaker pronounces the correct passphrase. Training/evaluation baselines consist of short digit passphrases

Training Datasets:
• RSR2015 Part 3 train set : 194 speakers (94 Female + 100 Male) - RSR2015tr
• Wells Fargo Bank set: 300 speakers (150 Female + 150 Male) — WF
• STC-Russian-digits train set: 786 speakers (263 Female + 523 Male) — STCRu

Evaluation Datasets:
• RSR2015 Part 3 eval set : 106 speakers (49 Female + 57 Male) - RSR2015ev
• STC-Russian-digits eval set: 92 speakers (42 Female + 50 Male) - STCRu

Results

Table 1. EER (%) and minDCF (C_min = 10, C_max = 1, P_lead = 10^-7) for 5-digit password verification

<table>
<thead>
<tr>
<th>System</th>
<th>Multi-Task mode</th>
<th>Training data</th>
<th>Evaluation data</th>
<th>EER (%)</th>
<th>Min DCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>None</td>
<td>RSR2015tr, WF</td>
<td>RSR2015ev</td>
<td>3.11</td>
<td>0.14</td>
</tr>
<tr>
<td>State-GMM-SVM[2]</td>
<td>None</td>
<td>RSR2015ev</td>
<td></td>
<td>7.83</td>
<td>0.39</td>
</tr>
<tr>
<td>State-CNN</td>
<td>Speaker &amp; Digits</td>
<td>RSR2015tr, WF</td>
<td>STCRu</td>
<td>5.86</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Speaker &amp; Digits</td>
<td></td>
<td>STCRu</td>
<td>2.85</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>Speaker &amp; Digits</td>
<td></td>
<td>STCRu</td>
<td>4.24</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Table 2. Fusion. EER (%) and minDCF for 5-digit password verification

<table>
<thead>
<tr>
<th>System</th>
<th>EER (%)</th>
<th>Min DCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>State-GMM-SVM[2]</td>
<td>2.09</td>
<td>0.1</td>
</tr>
<tr>
<td>State-Plda</td>
<td>1.63</td>
<td>0.07</td>
</tr>
<tr>
<td>State-GMM-Plda[3]</td>
<td>1.57</td>
<td>0.08</td>
</tr>
<tr>
<td>All</td>
<td>1.43</td>
<td>0.07</td>
</tr>
</tbody>
</table>

6 Conclusions

• A deep CNN based speaker feature extractor for speech digits is presented
• Multitask learning mode allows to train effective high-level speaker embeddings extractor for all states (digits)
• Discriminatively trained deep CNN based solution is able to surpass the classic baseline systems in terms of quality
• No complex trainable backend is needed for scoring. Speaker embeddings can be compared simply with cosine similarity metric
• CNN-based method fuses well with our previous methods [2,3]

7 References


8 Acknowledgements

This work was financially supported by the Ministry of Education and Science of the Russian Federation, Contract 14.578.21.0189 (ID RFMEFI57816X0189).