Summary

- We show the fusion result of a GMM i-vector system and five DNN x-vector systems.
- PolyU has one i-vector system (S1) and two x-vector systems (S2 and S3) which are all trained on 8kHz speech data.
- The HKUST system uses 8kHz x-vectors for CMN2 (S4-CMN2) and 16kHz x-vectors for VAST (S4-VAST), applying the self-attention mechanism.
- VoiceAI implemented a larger 16kHz DNN (S5) to extract 1024-dimensional x-vectors for VAST data. The VAD is based on a bidirectional long short-term memory network.

X/I-Vector/PLDA Training

- S1 (i-vector system) is based on gender-independent UBM with 2048 mixtures and 600 dimensional total variability matrix.
- The DNN in S2 was retrained using the Kaldi SRE16 x-vector recipe.
- S3 uses the pre-trained DNN from the Kaldi repository.
- In S4, the pooling-layer in the standard x-vector extractor was replaced by a self-attentive layer.
- S5 has a larger DNN architecture than the standard x-vector extractor. The dimension of the x-vector is 1024 rather than 512.
- We trained gender-independent out-of-domain PLDA models first, and then adapted them using suitable in-domain dataset for S1, S2, S3 and S4-CMN2. For S4-VAST and S5, we directly trained PLDA models using in-domain data.

X/I-vector pre-processing: centering + LDA + length-norm

- S1: 600  300
- S2: 512  300
- S3, S4: 512  150
- S5: 1024  150

Dataset used for PLDA training, PLDA adaptation, and PLDA scoring

<table>
<thead>
<tr>
<th>Sys</th>
<th>Embedder</th>
<th>Sub-task</th>
<th>PLDA Training</th>
<th>PLDA adaptation</th>
<th>Score</th>
<th>Adapt.source</th>
<th>Scoring source</th>
<th>Centering</th>
<th>Adapt.source</th>
<th>Scoring source</th>
<th>Centering</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>i-vector</td>
<td>CMN2</td>
<td>sre18-10</td>
<td>aug, aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
</tr>
<tr>
<td>S2</td>
<td>x-vector</td>
<td>CMN2</td>
<td>sre18-10</td>
<td>aug, aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
</tr>
<tr>
<td>S3</td>
<td>x-vector</td>
<td>CMN2</td>
<td>sre18-10</td>
<td>aug, aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
</tr>
<tr>
<td>S4</td>
<td>x-vector</td>
<td>CMN2</td>
<td>sre18-10</td>
<td>aug, aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
</tr>
<tr>
<td>S5</td>
<td>x-vector</td>
<td>CMN2</td>
<td>sre18-10</td>
<td>aug, aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
<td>aug</td>
</tr>
</tbody>
</table>

The column "Adapt source" indicates the data source for computing the S-norm parameters.

Sre18-unlabeled, unlabeled data in SRE18, sre18/unlabeled. 16kHz upsampled version of sre18-unlabeled. x-vector, and x-vector: speaker embedding based on 8kHz speech data. x-vector16k: speaker embedding based on 16kHz speech data.

System Fusion

- We have three fused systems denoted as Fuse1, Fuse2 and Fuse3, respectively.
- Fuse1 is a linear weighted fusion of S1+S2 (S1 and S2 were first fused by Bosaris) and S4.
- Fuse2 is a linear weighted fusion of Fuse1 and S3.
- Fuse3 is formed by fusing Fuse1 and S5 using Bosaris.

References