Open-set Speaker Verification: (determines whether a pair of utterances belongs to the same person)
- Speaker identities in testing set are usually disjoint from the ones in training set, which makes the speaker verification more challenging yet closer to practice.
- Since it is impossible to classify testing utterances to known identities in training set, we need to map speakers to a discriminative feature space.
- In this scenario, open-set speaker verification is essentially a metric learning problem, where the key is to learn discriminative large-margin features.

**Deep Length Normalization**

**Motivation:**
Is it possible to learn the deep speaker embeddings being length-normalized in an end-to-end manner within common classification network?

We add a length normalization layer followed by a scale layer before the output layer of the common classification network.

\[ y_i = a \times \frac{f(x_i)}{||f(x_i)||} \]

- In total, only a single scalar parameter \( a \) is introduced, and it can be inherently trained with other components of the network together.
- This scalar parameter \( a \) has a crucial impact on the performance since it determines the radius of the length-normalized hyperspace.
- The network could have stronger I2-constraint on the small radius hyperspace with smaller \( a \), but faces the risk of not convergent.

**Network Setup:**
- The model is trained with a mini-batch size of 128 using typical stochastic gradient descent with momentum 0.9 and weight decay 1e-4.
- The learning rate is set to \( 0.1, 0.01, 0.001 \) and is switched when the training loss plateaus.
- For each training step, an integer \( l \) within [300,800] interval is randomly generated, and each data in the mini-batch is cropped or extended to \( l \) frames.
- After model training finished, the 128-dimensional speaker embeddings are extracted after the penultimate layer of neural network.

**Traditional Length Normalization**

Length normalization on i-vector has been the de facto standard before back-end modeling.

For open-set SV task, cosine similarity or length normalization followed by probabilistic linear discriminant analysis (PLDA) scoring modeling is widely used to get the final pairwise scores.

The cosine similarity is a similarity measure which is independent of magnitude, it can be seen as the length-normalized version of inner-product of two vectors.

**Length normalization in classical i-vector approach.**

Once speaker embeddings (such as x-vectors) are extracted, just the same as in i-vector approach, cosine similarity or length normalization followed by PLDA is commonly adopted to get the final pairwise scores.

**Length normalization in typical speaker embedding approach.**

**Experimental Results and Discussion**

**Analysis of Length Normalization in End-to-End Speaker Verification System**

Weicheng Cai1,2, Jinkun Chen2 and Ming Li1

1. Data Science Research Center, Duke Kunshan University, Kunshan, China
2. School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China

ming.li36@dukekunshan.edu.cn

**Table 1:** Baseline end-to-end system architecture

<table>
<thead>
<tr>
<th>System Description</th>
<th>minDCF</th>
<th>Averaged</th>
<th>Equal Error Rate (EER)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i-vector + inner-product</td>
<td>N/A</td>
<td>N/A</td>
<td>37%</td>
</tr>
<tr>
<td>i-vector + PLDA</td>
<td>0.524</td>
<td>0.739</td>
<td>5.90</td>
</tr>
<tr>
<td>Deep embedding + inner-product</td>
<td>0.681</td>
<td>0.771</td>
<td>13.80</td>
</tr>
<tr>
<td>Deep embedding + PLDA</td>
<td>0.484</td>
<td>0.627</td>
<td>5.41</td>
</tr>
<tr>
<td>i-vector + PLDA, DNN</td>
<td>0.488</td>
<td>0.639</td>
<td>5.48</td>
</tr>
<tr>
<td>Deep embedding + PLDA, DNN</td>
<td>0.503</td>
<td>0.637</td>
<td>5.46</td>
</tr>
<tr>
<td>Deep embedding + PLDA, DNN, large margin</td>
<td>0.499</td>
<td>0.596</td>
<td>5.32</td>
</tr>
</tbody>
</table>

Experiments on different \( a \):
- We first investigate the setting of scale parameter \( a \). For those systems in Table 1 and Fig. 4, the cosine similarity or equivalently 12-normalized inner-product is adopted to measure the similarities between speaker embeddings.
- From Fig. 4, we can observe the proposed L2-normalized deep embedding system achieves the best minDCF of 0.475, 0.586 and EER of 5.101%, which outperforms the baseline system significantly.
- The best \( a \) is too small and stable with \( a \) higher. The best \( a \) in our experiment is 12.

We further compare the effect of deep length normalization strategy and traditional extra length normalization in the whole SV pipeline. The results are shown in Table 2.

1. No matter in i-vector or baseline deep speaker embedding systems, extra length normalization step followed by PLDA scoring achieves the best performance.
2. When it turns into L2-normalized deep embedding systems, the extra length-normalization introduced from neural network have already been normalized to unit length, we need no more extra length normalization step.
3. In testing stage, a simple inner-product achieves the best performance, which is slightly better than the PLDA scoring result. It might be the reason that our L2-normalized speaker embedding is highly optimized, which could not compete with the objective function introduced by PLDA.