The DKU-TVM-SYSU System for the VoxCeleb Speaker Recognition Challenge

Xiaoyi Qin¹, Weicheng Cai², Yunbo Gong³, Ming Li¹

¹Data Science Research Center, Duke Kunshan University, Kunshan, China
²School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China
³TVMining Media Technology Ltd.,
1. Open training data task
   • Introduction
   • Data Usage
   • System Description

2. Fix training data task
   • Introduction
   • Data Usage
   • System Description

3. System performance

Data augmentation
Acoustic features
Front-end modeling
Back-end scoring
System fusion
1. Open training data task

• Introduction

For the task, we develop two systems which used the same network structure but different strategies.

The final submission result fused two systems and the EER is 1.69%.
1. Open training data

- Data Usage
  - Training Set

  VoxCeleb 1 & VoxCeleb 2

- Development Set

  VoxCeleb 1 test data (40 Speaker)

  We do not use Voice Activity Detection (VAD) for the training data because the utterances in VoxCeleb contain a little silence.
1. Open training data

- System Description
- Data augmentation

- We adopt Kaldi’s data augmentation strategy. It employs additive noises and reverberation.
- For additive noise, the music and noise part of the MUSAN dataset is used.
- Fixed Filter Number.

- Dynamic filter number
1. Open training data

**Acoustic Features**

- **Fixed Filter Number**
  - 16k Wav
  - 64-dim Mel-fbank energies

- **Dynamic Filter Number**
  - 16k Wav
  - K-dim Mel-fbank energies $K \in [58, 70]$
1. Open training data

- Front-end modeling

The first part is a deep CNN structure based on the well known ResNet-34.

It learns high-level abstract local patterns from raw filter bank energies.
A global statistics pooling (GSP) layer is then designated on top of the CNN, and it transforms the local feature maps into an utterance-level representation.

The output of GSP is normalized by means of mean and standard deviation.
• Front-end modeling

A fully-connected layer then processes the utterance-level representation.

Dropout with a rate of 0.5 is added before the output layer to prevent over-fitting.
1. Open training data

- Front-end modeling

Finally connected with a classification output layer.

Each unit in the output layer is represented as a target speaker identity.
1. Open training data

- Front-end modeling

System 1 and System 2 take the same structure

**Softmax**

**FC-Drop (128 dim)**

**Global Statistics Pooling (512 dim)**

**log Mel-Fbank 16k**
1. Open training data

- Back-end Scoring cosine similarity

- System Fusion

  BOSARIS toolkit
2. Fixed training data task

- **Introduction**: Single system (No fusion) EER is 2.05%
- **Data Usage**: VoxCeleb 2 (5994 speakers)
- **System Description**

  The system adopts 64-dim Mel-fbank as input feature and ResNet34 + GSP + FC-Drop + softmax network structure.
3. System Performance

Table 1: System performance on different fusion system

<table>
<thead>
<tr>
<th>System</th>
<th>Development (EER[%])</th>
<th>Evaluation (EER[%])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open training data system</td>
<td>System 1: 1.72</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>System 2: 1.64</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Fusion System: 1.55</td>
<td>1.69</td>
</tr>
<tr>
<td>Fix training data system</td>
<td>2.00</td>
<td>2.05</td>
</tr>
</tbody>
</table>

- Data augmentation accumulates large storage space and takes a lot of time to generate data, which is not cost-effective for the system. Therefore, the dynamic filter is proposed as an attempt to expend data. According to the results, it is not difficult to find that dynamic filter is works.

- We have investigated to jointly adopt these two techniques together in an online augmentation framework which achieves good performance (submitted, under review). We will report the performance in the near feature.

- Although the fusion system obtains a noticeable performance in the EER of development set, the test set is 1.69% and much higher than 1.55%. This may be over-fitting the development set.
Thanks

ming.li369@dukekunshan.edu.cn
xiaoyi.qin@dukekunshan.edu.cn