RawNeXt: Speaker verification system for variable-duration utterances with deep layer aggregation and extended dynamic scaling policies

Ju-ho Kim, Hye-jin Shim, Jungwoo Heo, and Ha-Jin Yu
School of Computer Science, University of Seoul, South Korea

Overview

- Speaker verification (SV): The task of determining whether the identity of an anonymous voice matches the target speaker
- Problems: Variable-duration input utterance degrades the reliability of SV system
- Insufficient speaker-specific information of short utterance
- SV systems operating in a fixed way with manually designed layers

Proposed model: RawNeXt

- Apply deep layer aggregation: Enhance speaker information by iteratively and hierarchically aggregating features
- Propose extended dynamic scaling policy: Process features according to the length of the utterance

1. Deep layer aggregation (DLA

- Apply to derive speaker embeddings by fusing features in a more iterative and hierarchical manner for utterances of various lengths
- Iterative deep aggregation module: Enhance temporal context information by merging the different time resolution features
- Hierarchical deep aggregation module: Enhance spectral context information by combining the feature channels of different levels
- Aggregation block: Learn to select important information from the multiple inputs and project it into a single output

2. Extend dynamic scaling policy (EDSP)

- Propose for utterance of arbitrary lengths based on Elastic
- Utilize three resolution branches and a gate module
- Low, original, and high resolution branches: Feature extraction with receptive fields of different sizes

Baseline architecture with raw waveform

- Input feature of models: Raw waveform
- Proposed model: RawNeXt

<table>
<thead>
<tr>
<th>Model</th>
<th>Input Feature</th>
<th>Loss Function</th>
<th>Test utterance length (EER, %)</th>
<th>Test utterance length (EER, %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RawNeXt</td>
<td>Waveform</td>
<td>Softmax</td>
<td>6.12</td>
<td>3.68</td>
</tr>
<tr>
<td>RawNeXt</td>
<td>Waveform</td>
<td>Softmax</td>
<td>6.47</td>
<td>2.58</td>
</tr>
<tr>
<td>ResNet4x</td>
<td>Waveform</td>
<td>Softmax</td>
<td>5.92</td>
<td>3.38</td>
</tr>
<tr>
<td>ResNet4x</td>
<td>Waveform</td>
<td>Softmax</td>
<td>6.12</td>
<td>3.38</td>
</tr>
<tr>
<td>ResNet4x</td>
<td>Waveform</td>
<td>Softmax</td>
<td>5.92</td>
<td>3.38</td>
</tr>
<tr>
<td>ResNet4x</td>
<td>Waveform</td>
<td>Softmax</td>
<td>6.12</td>
<td>3.38</td>
</tr>
</tbody>
</table>

Experiments & Results

- Experiment configurations
  - Batch size: 320
  - Training dataset: VoxCeleb2
  - Training epoch: 80
  - Weight decay: 10^-4
  - Learning rate (LR): Equal error rate (EER) 10^-3 to 10^-4

- Exp1: Comparison with recently proposed SV system for variable-duration utterances
  - Proposed RawNeXt outperforms other models for all test conditions
  - Compared to baseline, 28.7% improvement for full-length test / 28.4% improvement for mean result of 1.2, and 5s-lengths
  - RawNeXt demonstrates superior generalization and robustness to variable-length utterances

- Exp2: Ablation experiments of RawNeXt components
  - #1: ResNetXt (Baseline), #7: RawNeXt (Proposed)
  - Performance improves as each module is applied
  - The motivations of each method are well aligned with the goal of variable-duration utterance SV

- Exp3: Variation score for mean activation of each resolution path according to the input utterance length
  - Score at each resolution branch by differences of mean activations between l and a 1-second utterance

- RawNeXt extracts speaker information with appropriate resolutions by dynamically applying scaling policies according to the length of the utterance

![RawNeXt block architecture](image)

![Overview](image)

Performance comparison:

<table>
<thead>
<tr>
<th>Test utterance length (EER, %)</th>
<th>1s</th>
<th>2s</th>
<th>5s</th>
<th>full</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1 x x x</td>
<td>6.12</td>
<td>3.68</td>
<td>2.45</td>
<td>2.16</td>
</tr>
<tr>
<td>#2 o x x</td>
<td>4.82</td>
<td>2.98</td>
<td>2.08</td>
<td>1.93</td>
</tr>
<tr>
<td>#3 o x x</td>
<td>5.39</td>
<td>3.19</td>
<td>2.16</td>
<td>1.95</td>
</tr>
<tr>
<td>#4 o x x</td>
<td>4.66</td>
<td>2.94</td>
<td>2.13</td>
<td>1.94</td>
</tr>
<tr>
<td>#5 o o o</td>
<td>4.67</td>
<td>3.01</td>
<td>2.08</td>
<td>1.88</td>
</tr>
<tr>
<td>#6 o o x</td>
<td>4.65</td>
<td>2.81</td>
<td>1.94</td>
<td>1.82</td>
</tr>
<tr>
<td>#7 o o o</td>
<td>4.47</td>
<td>2.58</td>
<td>1.72</td>
<td>1.54</td>
</tr>
</tbody>
</table>

2. F. Yu et al., Deep layer aggregation, CVPR 2018.
5. Y. Jung et al., Improving multi-scale aggregation using feature pyramid module for robust speaker verification of variable-duration utterances, Interspeech 2020.