

**Overview**

- **Speaker Verification (SV)**
  - Process of verifying a person’s claimed identity using their enrollment and test utterances
  - 2 Steps: frame-level feature extraction, utterance-level feature aggregation
- **Utterance-level feature aggregation**
  - Aggregate frame-level features into a single utterance-level feature
  - Gated Recurrent Units, Learnable Dictionary Encoding, attention
- **Research background**
  - Sequential information may not be the key in text-independent SV
  - Attention cannot model each frame pair’s intra-relationships

**Proposed Method**

- **Graph attentive feature aggregation**
  - Improved feature aggregation method
  - Utilizing graph attention networks
  - Entire frame-level features are aggregated considering their inter-relationships
- **Our Contributions**
  - Proposed graph attentive feature aggregation
  - First approach using GNN for feature aggregation in SV research
  - Explored various readout and structure
  - Validated the effectiveness of the proposed method using both spectrogram and raw waveform baselines

**Proposed method**

- **Graph formulation**
  - Formulate a graph from a feature map
- **Node projection**
  - Project \( x \) into \( F' \) dimensional space by matrix multiplication with learnable parameter \( W \in R^{F \times F'} \)
- **Edge score calculation**
  - Calculate edge scores \( y \in R^{2 \times F \times F_1} \)
  - The total number of nodes is \( N \), \( N \times N \) scores are calculated
- **Self attention**
  - Performs self-attention on every nodes
- **Graph pooling**
  - Reduce the original graph into a sub-graph by removing less informative nodes
- **Readout**
  - Combines the processed nodes into a single node

**Experiments & Results**

- **Dataset**
  - Train: VoxCeleb development set
  - Test: VoxCeleb 1 test set
- **Baseline**
  - Used two baseline to check the effect according to the input domain
    - SE-ResNet:
      - Input: 40-dimensional mel-filterbank features
      - Modified Clova system
      - RawNet2:
        - Input: raw waveform
        - Modified original RawNet2
- **Results**
  - Table 1: Both systems improved performance with fewer parameters than baselines
    - Proposed graph attentive feature aggregation was effective
  - Table 2: The proposed system showed superior performance in both spectrogram and raw waveform domain
    - Our system achieved state-of-the-art performance

<table>
<thead>
<tr>
<th>Feature extractor</th>
<th>Aggregation</th>
<th># Params</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE-ResNet</td>
<td>SAP</td>
<td>6.0M</td>
<td>1.98</td>
</tr>
<tr>
<td>SE-ResNet</td>
<td>GAT</td>
<td>5.4M</td>
<td>1.86</td>
</tr>
<tr>
<td>RawNet2</td>
<td>GRU</td>
<td>13.2M</td>
<td>2.48</td>
</tr>
<tr>
<td>RawNet2</td>
<td>GAT</td>
<td>9.9M</td>
<td>2.23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input Feature</th>
<th>Front-end</th>
<th>Aggregation</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chung et al.</td>
<td>Spec-257</td>
<td>Thin ResNet-34</td>
<td>SAP</td>
</tr>
<tr>
<td>Yu et al.</td>
<td>Spec-512</td>
<td>ResNet-50</td>
<td>TAP</td>
</tr>
<tr>
<td>Liu et al.</td>
<td>MFR-40</td>
<td>Dense-Residual</td>
<td>ABP</td>
</tr>
<tr>
<td>Jung et al.</td>
<td>MFR-40</td>
<td>Fast ResNet-34</td>
<td>LDE</td>
</tr>
<tr>
<td>Kye et al.</td>
<td>MFR-40</td>
<td>Fast ResNet-34</td>
<td>CAP</td>
</tr>
<tr>
<td>Ours</td>
<td>SE-ResNet</td>
<td>SAP</td>
<td>1.98</td>
</tr>
<tr>
<td>Ours(Proposed)</td>
<td>MFR-40</td>
<td>SE-ResNet</td>
<td>1.75</td>
</tr>
</tbody>
</table>

Table 2: performance comparison with state-of-the-art systems