Diaboliic @ DIHARD 3

Third Dihard Challenge Workshop
2021
Team Structure & Approach

• Team: LIUM, (Herve Bredin) IRIT, Wenda Chen & Sangeeta Ghangam

• Summary –

❖ Team created different systems, each optimized for specific modules - Segmentation, Embedding, Re-segmentation
❖ An ensemble approach was used for both the tracks as part of the final results submission
Segmentation

Based on pyannote.audio [Bredin 2020]

Slightly improved overlapped speech detection thanks to multi-task training

Based on pyannote.audio [Bredin 2020]
Efficient Embeddings

- X-vector model efficiencies for short-duration speech segments [Chen 2020]
- Distance: model_1 0.941; model_2 0.953

<table>
<thead>
<tr>
<th>Models</th>
<th>1s</th>
<th>2s</th>
<th>4s</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model1: AM-Softmax</td>
<td>12.74</td>
<td>6.70</td>
<td>3.99</td>
<td>1.90</td>
</tr>
<tr>
<td>Model2: AM-Softmax-IRL</td>
<td>13.67</td>
<td>7.13</td>
<td>3.69</td>
<td>1.49</td>
</tr>
</tbody>
</table>

- SID to diarization

Voxceleb → VOiCES → DIHARD III Dev
Re-segmentation

- Based on VBx system [Landini 2020]
- Optimized Parameters - Ploop (0.40), Interpolation Alpha (0.75)
- 4% improvement in the DER when the re-segmentation system was combined with the baseline pyannote system

Pyannote Embeddings (up to 4s sliding window) → Create Speech Segments in Kaldi Ark format → Process through VBx re-segmentation system
Final Results Summary

• Track1 - Ensemble of following systems
  • Pyannote (Second DIHARD)
  • Baseline Third DIHARD system

These were combined using Dover-lap [Raj 2021]

• Track2 - Ensemble of following systems
  • Pyannote (Updated with segmentation/embeddings)
  • Resegmentation output
  • VBx Baseline System [Landini 2020v2]

These were combined using Dover [Stolcke]
## Results Summary - 1

<table>
<thead>
<tr>
<th>Track</th>
<th>System</th>
<th>Dataset</th>
<th>DER</th>
<th>JER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Track 1</td>
<td>Pyannote (Second DIHARD)</td>
<td>Third DIHARD Dev</td>
<td>20.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Baseline Third DIHARD system</td>
<td>Third DIHARD Dev</td>
<td>19.4</td>
<td></td>
</tr>
<tr>
<td>Final - Full</td>
<td>Third DIHARD Eval</td>
<td></td>
<td>18.89</td>
<td>42.98</td>
</tr>
<tr>
<td>Final - Core</td>
<td>Third DIHARD Eval</td>
<td></td>
<td>20.61</td>
<td>48.38</td>
</tr>
</tbody>
</table>

These were combined used Dover-lap [Raj 2021]
<table>
<thead>
<tr>
<th>Track</th>
<th>System</th>
<th>Dataset</th>
<th>DER</th>
<th>JER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Track 2</td>
<td>Pyannote Fusion (Updated segmentation/Embedding)</td>
<td>Third DIHARD Dev</td>
<td>28.40</td>
<td>45.63</td>
</tr>
<tr>
<td></td>
<td>Resegmentation</td>
<td>Third DIHARD Dev</td>
<td>24.32</td>
<td>42.72</td>
</tr>
<tr>
<td></td>
<td>VBx Baseline [Landini 2020v2]</td>
<td>Third DIHARD Dev</td>
<td>15.71</td>
<td>34.48</td>
</tr>
<tr>
<td>Final - Full</td>
<td></td>
<td>Third DIHARD Eval</td>
<td>21.18</td>
<td>42.58</td>
</tr>
<tr>
<td>Final - Core</td>
<td></td>
<td>Third DIHARD Eval</td>
<td>22.96</td>
<td>47.17</td>
</tr>
</tbody>
</table>

These were combined using Dover [Stolcke]
References

- [Yin 2017] “Speaker change detection in broadcast TV using bidirectional long short-term memory networks”. InterSpeech 2017
- [Lavechin 2020] “End-to-end domain-adversarial voice activity detection”. InterSpeech 2020
- [Stolcke] “Improving Diarization Robustness using Diversification, Randomization and the DOVER Algorithm”
- [Landini 2020v2] “Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks”