Ensemble combination between different time segmentations
Jeremy H. M. Wong, Dimitrios Dimitriadis, Kenichi Kumatani, Yashesh Gaur, George Polovets, Partha Parthasarathy, Eric Sun, Jinyu Li, and Yifan Gong
Microsoft speech and language group
jeremy.wong@microsoft.com

MOTIVATION

**Task:** Speech recognition

**Problem**
Hypothesis-level combination requires all models to use the same input time segmentations.

**Proposal**
Allow different time segmentations between models by splitting and re-joining the hypothesis N-best lists.

**Applications**
Allow combinations between:
• Different voice activity detection front-ends.
• Different unsynchronised recording devices.
• Overlapping inference.
• 1st pass used to refine time segmentation of 2nd pass.

MEETING TRANSCRIPTION SETUP

1st pass streaming ASR → diarisation → 2nd pass offline ASR

1st pass streaming ASR uses VAD segments.
2nd pass ASR uses per-speaker segments from diarisation.
Want to combine 1st pass and 2nd pass hypotheses.

**Data:**
- dev - 51 meetings, 23 hours
- eval - 60 meetings, 35 hours
- Average of 7 participants per meeting

MULTI-PASS COMBINATION

CONFUSION NETWORK SPLITTING

**Steps:**
1. Convert N-best list into confusion network.
2. Estimate start and end times of confusion sets.
3. Estimate confusion set speaker from 1-best hypothesis.
4. Split up confusion network into separate confusion sets.
5. Re-join consecutive confusion sets of the same speaker.
6. Do Confusion Network Combination (CNC).

**Advantages:**
• 1-best is preserved after re-joining.

**Disadvantages:**
• Confusion set times are approximate.
• Context of language model scores is not preserved.

N-BEST LIST SPLITTING

**Steps:**
1. Distribute hypothesis scores to words.
2. Estimate speakers for N-best words from 1-best hypothesis.
3. Split up N-best lists according to segment time and speaker.
4. Re-join N-best lists according to segment time and speaker.
5. Do Minimum Bayes' Risk (MBR) combination.

**Advantages:**
• Exact word start and end times are preserved.
• Context of language model scores is preserved.

**Disadvantages:**
• Hypothesis rank order may not be preserved after re-joining.

EXPERIMENTS

Distribution of hypothesis scores to words, on 1st pass eval

<table>
<thead>
<tr>
<th>Split</th>
<th>Per-word scores</th>
<th>Speaker-attributed WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>original</td>
<td>20.43</td>
</tr>
<tr>
<td>yes</td>
<td>original</td>
<td>22.09</td>
</tr>
<tr>
<td></td>
<td>language model re-score</td>
<td>22.09</td>
</tr>
<tr>
<td></td>
<td>prefix tree</td>
<td>20.62</td>
</tr>
<tr>
<td></td>
<td>suffix tree</td>
<td>20.60</td>
</tr>
<tr>
<td></td>
<td>average</td>
<td>20.55</td>
</tr>
</tbody>
</table>

• Best performance with average of prefix and suffix trees.

Multi-pass combination (Speaker-attributed WER (%))

<table>
<thead>
<tr>
<th>Eval</th>
<th>dev</th>
<th>eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st pass streaming hybrid</td>
<td>21.43</td>
<td>20.43</td>
</tr>
<tr>
<td>2nd pass offline hybrid</td>
<td>19.93</td>
<td>19.13</td>
</tr>
<tr>
<td>2nd pass offline LAS</td>
<td>19.91</td>
<td>19.04</td>
</tr>
<tr>
<td>CNC streaming hybrid + offline hybrid</td>
<td>20.01</td>
<td>19.10</td>
</tr>
<tr>
<td>CNC streaming hybrid + offline LAS</td>
<td>19.71</td>
<td>18.71</td>
</tr>
<tr>
<td>MBR streaming hybrid + offline hybrid</td>
<td>19.83</td>
<td>19.00</td>
</tr>
<tr>
<td>MBR streaming hybrid + offline LAS</td>
<td>19.30</td>
<td>18.43</td>
</tr>
<tr>
<td>MBR offline hybrid + offline LAS</td>
<td>19.11</td>
<td>18.24</td>
</tr>
</tbody>
</table>

• N-best list splitting outperforms confusion network splitting.
• Combination with no increase in 2nd pass computational cost.
• Hybrid + LAS outperforms hybrid + hybrid.

CONCLUSION

• Distribute hypothesis scores to words using trees.
• Combine different time segments by splitting N-best lists.