The OPPO System for CHiME-6 Challenge

Xiaoming Ren, Huifeng Zhu, Liwei Wei, Linju Yang, Ming Yu, Chenxing Li, Dong Wei, Jie Hao
2020-04-28
System Framework for Track1 (ASR only)

Training Stage:
- Worn data
- Frontend processing
- Data Augmentation
- Acoustic Modeling
- Far field data

Acoustic Models:
- AM1
- AM2
- AM3
- ... (continued)
- AMN

Testing Stage:
- Test Data
- Frontend Processing
- Decoding
- N-gram LM

Lattice Files:
- Lattice1
- Lattice2
- Lattice3
- ... (continued)
- LatticeN

Combination:
- Lattice Rescore
- Lattice Combination
- MBR Decoding
- Final Result
System Description

- Data Preparation
- Frontend processing
- Acoustic modeling
- Language modeling
- Decoding
System Description

- Data Preparation
  - For the worn (L+R) microphone training data, realign original utterance segmentation using ASR model.
  - Clean up the training data by filtering out segments which are less than 1 second.
  - Remove noises which can be recognized as words from noises used in Room Impulse Responses (RIR) [1] convolution.
  - Apply only speed perturbation for the training data without the volume perturbation.

Finally, we obtain about 1400 hours of training data, which contains the following dataset:

- The realigned worn (L+R) training data.
- The far field enhanced by GSS module.
- The worn data and enhanced far field data both convolved with RIRs.
- The augmented previous dataset by speed perturbation.

System Description

- Frontend processing
  - Compared to the official baseline setup, apply the GSS[2] not only in testing stage but also in training stage.

- Acoustic modeling
  - Use two different kinds of acoustic model structures (CNN-TDNN-F and TDNN-F[3]) based on LF-MMI training
  - Train 8 acoustic models with different parameters using Kaldi[4] toolkit

References


System Description

- Acoustic modeling

  - All the acoustic models is described as follows

    - CNN-TDNN-F\{1, 2, 3, 4\}: GSS module with \{10, 15\} context, 40-dim MFCC, 6-layer CNN + 19-layer TDNN-F with bottleneck-dim = 512, NUM-PDFS = \{2500, 3500\}
    - CNN-TDNN-F\{5, 6\}: GSS module with \{10, 15\} context, 40-dim MFCC, 6-layer CNN + 19-layer TDNN-F with bottleneck-dim = 768, NUM-PDFS = 3500
    - CNN-TDNN-F7: GSS module with 15 context, 80-dim MFCC, 6-layer CNN + 19-layer TDNN-F with bottleneck-dim = 512, NUM-PDFS = 3500
    - TDNN-F8: GSS module with 15 context, 40-dim MFCC, 25-layer TDNN-F with bottleneck-dim = 512, NUM-PDFS = 2500
System Description

- Language modeling
  - Build a 2-layer LSTM-based language model.
  - Rescore the lattice using the score of LSTM-base LM and official n-gram LM with a weighting of 0.55 and 0.45.
System Description

- Decoding
  - STEP1: generate lattices using eight acoustic models described in acoustic modeling section
  - STEP2: for Category B, rescore the lattice with LSTM-LM and official n-gram LM, for Category A, skip this step
  - STEP3: Combine all the lattices and apply MBR[5] decoding to get the final result

Experimental evaluation

- Acoustic models
- Frontend
- Data augmentation
- Feature
- System combination
Experimental evaluation

- Acoustic models
  - Compared to official TDNNF model with 15 layers, deeper TDNNF model with 25 layers and CNN-TDNNF model which has 6 convolution layers and followed by 19 TDNNF layers can reduce WER obviously

<table>
<thead>
<tr>
<th>AM</th>
<th>DEV(%)</th>
<th>EVAL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Official TDNNF</td>
<td>51.76</td>
<td>51.29</td>
</tr>
<tr>
<td>Deeper TDNNF</td>
<td>50.77</td>
<td>50.30</td>
</tr>
<tr>
<td>CNN-TDNNF</td>
<td>48.53</td>
<td>48.15</td>
</tr>
</tbody>
</table>

Table 1: WER of different acoustic models on the dev and eval sets
Experimental evaluation

- Frontend
  - Based on CNN-TDNNF model, in order to match the data in testing stage, apply GSS for all multi-array data in training stage
  - Compared to the official baseline which select 400K utterance from multi-array data, apply GSS in training stage can reduce WER by 2% absolutely on the dev set

<table>
<thead>
<tr>
<th>AM</th>
<th>Frontend</th>
<th>DEV(%)</th>
<th>EVAI(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-TDNNF</td>
<td>baseline</td>
<td>48.53</td>
<td>48.15</td>
</tr>
<tr>
<td>CNN-TDNNF</td>
<td>apply GSS in training stage</td>
<td>46.54</td>
<td>48.02</td>
</tr>
</tbody>
</table>

Table 2: WER of different frontends on the dev and eval sets
Experimental evaluation

- Data augmentation
  - Apply GSS module in training stage, reduce the amount of training data
  - Replace the L channel worn data with L+R channel worn data
  - Realign L+R channel worn data and make RIR data augmentation, and it can greatly improves the performance

<table>
<thead>
<tr>
<th>DATA</th>
<th>DEV(%)</th>
<th>EVAL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>multi-array GSS + worn(L)</td>
<td>46.54</td>
<td>48.02</td>
</tr>
<tr>
<td>multi-array GSS + worn(L+R)</td>
<td>45.55</td>
<td>47.30</td>
</tr>
<tr>
<td>Multi-array GSS + aligned worn(L+R) +RIR</td>
<td>45.31</td>
<td>45.81</td>
</tr>
</tbody>
</table>

Table 3: WER of different datasets on the dev and eval sets
Experimental evaluation

• Feature
  ✓ In addition, we find that using 80-dimension MFCC can slightly improve the performance of the system

<table>
<thead>
<tr>
<th>Feature</th>
<th>DEV(%)</th>
<th>EVAL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>40-dim MFCC</td>
<td>45.31</td>
<td>45.81</td>
</tr>
<tr>
<td>80-dim MFCC</td>
<td>44.99</td>
<td>45.28</td>
</tr>
</tbody>
</table>

Table 4: WER of different features on the dev and eval sets
Experimental evaluation

- System combination
  - Combine lattice produced by 8 acoustic models, WER of each model are presented in Table 5
  - For category A and B in Track1, apply MBR decoding on the combined lattice and get the best result of the system

<table>
<thead>
<tr>
<th>Category</th>
<th>AMs</th>
<th>LM</th>
<th>DEV(%)</th>
<th>EVAL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8 AMs</td>
<td>official n-gram LM</td>
<td>41.99%</td>
<td>42.41%</td>
</tr>
<tr>
<td></td>
<td>8 AMs</td>
<td>+ LSTM-LM</td>
<td>41.18%</td>
<td>42.02%</td>
</tr>
</tbody>
</table>

Table 6: WER of system combination on the dev and eval sets

<table>
<thead>
<tr>
<th>AM</th>
<th>DEV(%)</th>
<th>EVAL(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-TDNN-F1(10,2500)</td>
<td>45.20%</td>
<td>45.76%</td>
</tr>
<tr>
<td>CNN-TDNN-F2(10,3500)</td>
<td>45.00%</td>
<td>45.58%</td>
</tr>
<tr>
<td>CNN-TDNN-F3(15,2500)</td>
<td>45.61%</td>
<td>45.74%</td>
</tr>
<tr>
<td>CNN-TDNN-F4(15,3500)</td>
<td>45.31%</td>
<td>45.81%</td>
</tr>
<tr>
<td>CNN-TDNN-F5(10)</td>
<td>45.46%</td>
<td>45.80%</td>
</tr>
<tr>
<td>CNN-TDNN-F6(15)</td>
<td>45.50%</td>
<td>46.17%</td>
</tr>
<tr>
<td>CNN-TDNN-F7</td>
<td>44.99%</td>
<td>45.28%</td>
</tr>
<tr>
<td>TDNN-F8</td>
<td>46.66%</td>
<td>47.14%</td>
</tr>
</tbody>
</table>

Table 5: WER of different acoustic models on the dev and eval sets
Experimental evaluation

- Results summary

<table>
<thead>
<tr>
<th>Category</th>
<th>Session</th>
<th>Kitchen</th>
<th>Dining</th>
<th>Living</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Dev</td>
<td>47.42</td>
<td>45.57</td>
<td>38.31</td>
<td>41.99</td>
</tr>
<tr>
<td></td>
<td>S09</td>
<td>39.28</td>
<td>43.22</td>
<td>38.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eval</td>
<td>S01</td>
<td>58.00</td>
<td>35.83</td>
<td>47.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S21</td>
<td>51.49</td>
<td>35.09</td>
<td>34.43</td>
</tr>
<tr>
<td>B</td>
<td>Dev</td>
<td>46.66</td>
<td>45.00</td>
<td>37.47</td>
<td>41.18</td>
</tr>
<tr>
<td></td>
<td>S09</td>
<td>38.48</td>
<td>41.99</td>
<td>38.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Eval</td>
<td>S01</td>
<td>57.56</td>
<td>35.47</td>
<td>47.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S21</td>
<td>50.95</td>
<td>34.95</td>
<td>34.75</td>
</tr>
</tbody>
</table>

Table 7: WER of the system in Track1 (ASR only) for category A and category B
Conclusion

- Apply GSS module in training stage can improve the performance
- Data augmentation is helpful to boost the system performance
- Compared to TDNNF model, CNN-TDNNF can get better result
- Combine various acoustic model lattices, do MBR decoding and it can greatly reduce WER
Thanks for listening
Q&A