The USTC-iFlytek System For CHiME-4 Challenge

Jun Du
2016.09.13
Team

University of Science and Technology of China (USTC)

Jun Du  Yan-Hui Tu  Lei Sun

iFlytek Research

Feng Ma  Hai-Kun Wang  Jia Pan  Cong Liu

Georgia Institute of Technology

Chin-Hui Lee
Joint Framework For X-channel Tasks

(1)
Joint Framework For X-channel Tasks (II)
Implementation

- The official Kaldi recipe
  - Features: fMLLR and LMFB features
  - DNN-HMM acoustic model: concatenating fMLLR and LMFB
  - Model ensemble and two-pass decoding
- CNTK toolkit: IRM-DNN training
- Self-developed toolkit
  - Beamforming
  - DCNN-HMM acoustic model (only CE training)
  - LSTM language model
Feedback Loop Optimization

Multi-Channel Noisy Speech

Beamforming

Deep Neural Network Based IRM Estimation

Recognizer

T-F IRM

Frame VAD

For the IRM-DNN training:

1. The output T-F IRM is defined between CH5 and clean
2. The input is the LPS feature of CH5
Beamforming

1. CGMM to estimate the noise/noisy covariance matrix
2. Frame VAD and DNN-IRM to improve the masking
3. Frame-level VAD to determine the noise segment
4. DNN-IRM to determine T-F units in speech segments

Spectrogram Comparison

CH5 (F06_446C020B_STR_REAL)

The official 6-channel beamforming

The proposed 6-channel beamforming
Beamforming (Official vs. Ours)

Evaluation on the official baseline DNN system

More effective for more adverse environments and more microphones!
Training Data Augmentation

- Multi-style training
  - A: 1-channel (1,3,4,5,6) noisy speech simulating 1-channel case
  - B: 2-channel beamformed speech simulating 2-channel case
  - C: 6-channel beamformed speech simulating 6-channel case

- Training for 6-channel case
  - A+C for 1 DNN and 4 DCNNs

- Training for 2-channel and 1-channel cases
  - A+C for 1 DNN and 4 DCNNs, A+B for 2 DCNNs
Model Ensemble

- Ensemble via the state posterior average of NN output
- For 6-channel, 5-model ensemble (DNN, DCNN1-4)
- For 2-channel and 1-channel, 7-model ensemble (DNN, DCNN1-6)
Two LSTM-LMs (Forward and Backward) are combined.
Evaluation on real test set for the best configured system

Better front-end and acoustic models, more effective LSTM-LM!
Summary

Table 2: WER (%) per environment for the best system.

<table>
<thead>
<tr>
<th>Track</th>
<th>Envir.</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>real</td>
<td>simu</td>
</tr>
<tr>
<td>1ch</td>
<td>BUS</td>
<td>5.84</td>
<td>4.90</td>
</tr>
<tr>
<td></td>
<td>CAF</td>
<td>5.09</td>
<td>9.84</td>
</tr>
<tr>
<td></td>
<td>PED</td>
<td>2.66</td>
<td>4.84</td>
</tr>
<tr>
<td></td>
<td>STR</td>
<td>4.63</td>
<td>6.86</td>
</tr>
<tr>
<td>2ch</td>
<td>BUS</td>
<td>2.74</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td>CAF</td>
<td>2.18</td>
<td>4.29</td>
</tr>
<tr>
<td></td>
<td>PED</td>
<td>1.73</td>
<td>2.94</td>
</tr>
<tr>
<td></td>
<td>STR</td>
<td>2.65</td>
<td>3.79</td>
</tr>
<tr>
<td>6ch</td>
<td>BUS</td>
<td>2.05</td>
<td>1.64</td>
</tr>
<tr>
<td></td>
<td>CAF</td>
<td>1.50</td>
<td>1.99</td>
</tr>
<tr>
<td></td>
<td>PED</td>
<td>1.50</td>
<td>1.55</td>
</tr>
<tr>
<td></td>
<td>STR</td>
<td>1.71</td>
<td>1.93</td>
</tr>
</tbody>
</table>

The best system for all tasks (1ch: 9.15%, 2ch:3.91%, 6ch:2.24%)
Thanks!

Q&A