The USTC-NELSLIP Systems for CHiME-6 Challenge

Jun Du
National Engineering Lab for Speech and Language Information Processing (NELSLIP)
University of Science and Technology of China (USTC)
05/04/2020
# CHiME-5 vs. CHiME-6

<table>
<thead>
<tr>
<th></th>
<th>CHiME-5</th>
<th>CHiME-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data/Label Quality</td>
<td>New Array Synchronization</td>
<td></td>
</tr>
<tr>
<td>Front-End</td>
<td>Two-Stage SD-SS[1]</td>
<td>MGSS, BGSS</td>
</tr>
<tr>
<td>System Fusion</td>
<td>State Posterior Average Lattice Fusion [2]</td>
<td>MBR Fusion</td>
</tr>
<tr>
<td>Speaker Diarization</td>
<td>N/A</td>
<td>NBSS ResNet Based x-vector Spectral Clustering</td>
</tr>
</tbody>
</table>


Track 1:
Multiple-array Speech Recognition
System Overview (I)

Training Stage

Front-end and Data Augmentation

Official Training Data

Worn Data

Far-field Data

GSS

MGSS

Time Annotations

Forced-alignment

Acoustic Model Training Data

Worn Data

Data Cleanup

GSS Data

Speed Perturbation

MGSS Data

Volume Perturbation

Back-end Acoustic Modeling

Acoustic Model

Single-feature AM
- ResNet-TDNN-RBiLSTM
- ResNet-SelfAttention-TDNNF

Multi-feature AM
- ResNet-TDNN-RBiLSTM
- ResNet-SelfAttention-TDNNF
- ResNet-TDNNF
- ResNet-TDNNF-Dilation
System Overview (II)

Recognition Stage

Official Multiple-array Test Data

Far-field Data

Time Annotations

Acoustic Models

Rank A: Recognized Transcriptions

ASR Decoding and Alignment

ASR Decoding and Model Fusion

RNN LM Rescoring

Rank B: Recognized Transcriptions

Two-Stage Mask Estimation

MGSS

BGSS

SS Data
Implementation Platform

• The official Kaldi toolkit
  • Guided source separation (GSS)
  • Acoustic models
  • Language models
  • Model ensemble

• The Pytorch toolkit
  • Neural network based speech separation models

• Self-developed toolkit
  • cACGMM
  • Beamforming
Modified GSS (MGSS)

- cACGMM with 5 Gaussian mixtures
  - Corresponding to four speaker sources and one noise source
- The main difference between MGSS and GSS [4]
  - GEVD beamforming: offline + online [5]
  - Processing for selected-array data based on SINR

\[
\text{SINR} = 10 \log_{10} \frac{\sum_i \sum_f |\hat{S}(t,f)|^2}{\sum_i \sum_f (|\hat{N}(t,f)|^2 + |\hat{I}(t,f)|^2)}
\]

\[\hat{S}\]

Beamforming GSS (BGSS)

- Motivation: improving the mask estimation of MGSS

\[ \mathbf{\hat{S}}_{\text{BF}}^{T1} \] denotes the beamformed STFT features of Target 1.
\[ \mathbf{LPS}_{\text{BF}}^{T1} \] denotes the beamformed LPS features of Target 1.
\[ M_{\text{IRM}}^{T1} \] denotes the learning mask of Target 1.
\[ \phi_{\text{BF}}^{D1} \] denotes the IPD (inter-phase difference) between \[ \mathbf{\hat{S}}_{\text{BF}}^{T1} \] and \[ \mathbf{\hat{S}}_{\text{BF}}^{T2} \].
The above procedure generates the inputs/outputs for one target speaker.

The four speakers in one session are in turn considered as target speakers.
Model Optimization for BGSS

• Architecture
  • BLSTM
  • Input layer: 5130=513*10
  • Hidden layers: 1024*2
  • Output layer: 2052=513*4

• Objective function

\[ Err = (M_{BGSS}^{T1} - M_{IRM}^{T1})^2 + (M_{BGSS}^{T2} - M_{IRM}^{T2})^2 + (M_{BGSS}^{T3} - M_{IRM}^{T3})^2 + (M_{BGSS}^{T4} - M_{IRM}^{T4})^2 \]
Speech Demo

Original, channel-1

Worn

BeamformIt (interfering speaker is still existing)

GSS (Good suppression of interference, residual noises are still existing)

Our MGSS+BGSS (Good suppression of interference, better denoising)
Front-end (GSS vs. Ours)

Results on development sets using the official baseline AM

Acoustic Data Augmentation

• Worn data:
  • Left-channel and right-channel with data cleanup
  • Speed perturbation
  • Data size: (32+32)*3=192 hours

• GSS data:
  • Speed perturbation
  • Data size: 32*3=96 hours

• MGSS data:
  • Multi-array and selected-array
  • Data size: 32+32=64 hours

• Volume perturbation and SpecAugment for all data
• Total training data: 352 hours
Acoustic Models (AMs)

• Single-feature AMs
  • 40-dim MFCC with 100-dim i-vector

• Multi-feature (from 4 speakers) AMs
  • 100-dim i-vector

Target speaker

Part I

Part II

IPD: Inter-phase difference
Architecture and Optimization

• Four types:
  • ResNet-TDNNF (Multi-feature AM)
  • ResNet-TDNNF-Dialation (Multi-feature AM)
  • ResNet-TDNN-RBiLSTM (Single/Multi-feature AMs)
  • ResNet-SelfAttention-TDNNF (Single/Multi-feature AMs)

• Lattice-Free MMI [1]

ResNet-TDNN-RBiLSTM (Single-feature)
Resnet-TDNN-RBiLSTM (Multi-feature)
ResNet-SelfAttention-TDNNF
ResNet-TDNNF-Dialation & ResNet-TDNNF

ResNet-TDNNF-Dialation

ResNet-TDNNF

Combination of feature maps

Batch normalization

SpecAugment

Linear transformation

idct

MFCC

i-vector

Batch normalization

ResNet

ResNet

12TDNNF

3 ResBlock

3 × 3, 256 Conv

2 ResBlock

3 × 3, 128 Conv

2 ResBlock

3 × 3, 64 Conv

Batch normalization

2048

512

2048

-3, 0, 3

-3, 0, 3

-3, 0, 3

2048

512

2048

× 0.66

× 0.66
AMs with Our Best Front-end

- Four multi-feature AMs with different architectures
- MBR fusion of 4 multi-feature AMs and 2 single-feature AMs

Results on development sets

**Multi-Array Rank A**

NO data-augmentation
## Submitted Results

<table>
<thead>
<tr>
<th>Category</th>
<th>Session</th>
<th>Dining</th>
<th>Kitchen</th>
<th>Living</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>Dev</td>
<td>S02</td>
<td>34.95</td>
<td>35.13</td>
<td>29.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S09</td>
<td>29.60</td>
<td>28.10</td>
<td>27.65</td>
</tr>
<tr>
<td></td>
<td>Eval</td>
<td>S01</td>
<td>25.55</td>
<td>42.75</td>
<td>38.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S21</td>
<td>25.49</td>
<td>34.97</td>
<td>26.05</td>
</tr>
<tr>
<td>B</td>
<td>Dev</td>
<td>S02</td>
<td>34.66</td>
<td>34.86</td>
<td>29.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S09</td>
<td>29.10</td>
<td>27.74</td>
<td>27.22</td>
</tr>
<tr>
<td></td>
<td>Eval</td>
<td>S01</td>
<td>25.01</td>
<td>42.66</td>
<td>37.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S21</td>
<td>25.14</td>
<td>34.84</td>
<td>25.34</td>
</tr>
</tbody>
</table>
Track 2: Multiple-array Diarization and Recognition
Neural Beamforming for SS (NBSS)

- Mask estimation for cACGMM with 5 Gaussian mixtures
- The “target” selection based on speech duration of beamformed data
Training Data Generation for NBSS

- Official Multiple-array Training Data
  - Far-field Data
  - Time Annotations

- Non-overlap Selection
  - Target Speech
  - Interference Speech
  - Estimated Noise (LSA)

- Data Simulation
- WPE
- cACGMM

- Learning Targets

Note: LSA likely stands for Linearly-Scaled Approximation.
Speaker Diarization

- ResNet based x-vector extractor [1]

<table>
<thead>
<tr>
<th>Layer name</th>
<th>Structure</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>–</td>
<td>40 x 200 x 1</td>
</tr>
<tr>
<td>Conv2D-1</td>
<td>3 x 3, Stride 1</td>
<td>40 x 200 x 32</td>
</tr>
<tr>
<td>ResNetBlock-1</td>
<td>[3 x 3, 32] x 3, Stride 1</td>
<td>40 x 200 x 32</td>
</tr>
<tr>
<td>ResNetBlock-2</td>
<td>[3 x 3, 64] x 4, Stride 2</td>
<td>20 x 100 x 64</td>
</tr>
<tr>
<td>ResNetBlock-3</td>
<td>[3 x 3, 128] x 6, Stride 2</td>
<td>10 x 50 x 128</td>
</tr>
<tr>
<td>ResNetBlock-4</td>
<td>[3 x 3, 256] x 3, Stride 2</td>
<td>5 x 25 x 256</td>
</tr>
<tr>
<td>StatsPooling</td>
<td>–</td>
<td>5 x 256</td>
</tr>
<tr>
<td>Flatten</td>
<td>–</td>
<td>2560</td>
</tr>
<tr>
<td>Dense1</td>
<td>–</td>
<td>256</td>
</tr>
<tr>
<td>Dense2 (Softmax)</td>
<td>–</td>
<td>N</td>
</tr>
</tbody>
</table>

- Spectral clustering [2]

Submitted Results

<table>
<thead>
<tr>
<th>Category</th>
<th>Development set</th>
<th></th>
<th>Evaluation set</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DERs</td>
<td>JERs</td>
<td>WER</td>
<td>DERs</td>
</tr>
<tr>
<td>A</td>
<td>56.69</td>
<td>58.49</td>
<td>68.22</td>
<td>65.37</td>
</tr>
<tr>
<td>B</td>
<td>56.69</td>
<td>58.49</td>
<td>68.15</td>
<td>65.37</td>
</tr>
</tbody>
</table>
Thanks

Q&A