The Academia Sinica Systems of Speech Recognition and Speaker Diarization for the CHiME-6 Challenge

Hung-Shin Lee, Yu-Huai Peng, Pin-Tuan Huang, Ying-Chun Tseng, Chia-Hua Wu, Yu Tsao, Hsin-Min Wang

Academia Sinica, Taiwan
Outline

• Track 1: multiple-array ASR
  – Our contributions
  – Results

• Track 2: multiple-array diarization+ASR
  – Our contributions
  – Results
Track 1: Our Contributions

• Compared with the baseline system, we

  – We applied WPE, GSS, and BF to all the **Kinect data** in the training phase

  – Alignment expansion from the **Worn data** to the **Kinect data** was used

  – Four other kinds acoustic models, including our proposed **DcAE** and **FEAM**, were used
## Track 1: Results

Table 1: WERs (%) for Track 1 and Track 2 (Category A only).

<table>
<thead>
<tr>
<th>Model</th>
<th>Track 1</th>
<th></th>
<th>Track 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dev</td>
<td>Eval</td>
<td>Dev</td>
</tr>
<tr>
<td>Baseline</td>
<td>51.32</td>
<td>51.36</td>
<td>84.25</td>
<td>77.94</td>
</tr>
<tr>
<td>TDNN-F</td>
<td>50.12</td>
<td>49.36</td>
<td>75.89</td>
<td>73.68</td>
</tr>
<tr>
<td>RBiLSTM</td>
<td>52.43</td>
<td>50.26</td>
<td>76.90</td>
<td>73.39</td>
</tr>
<tr>
<td>DcAE-B</td>
<td>50.12</td>
<td>49.68</td>
<td>75.90</td>
<td>73.66</td>
</tr>
<tr>
<td>DcAE-U</td>
<td>49.86</td>
<td>49.63</td>
<td>75.78</td>
<td>73.54</td>
</tr>
<tr>
<td>FEAM-U</td>
<td>53.47</td>
<td>52.70</td>
<td>78.70</td>
<td>76.20</td>
</tr>
<tr>
<td>ROVER</td>
<td><strong>47.28</strong></td>
<td><strong>46.82</strong></td>
<td><strong>74.36</strong></td>
<td><strong>71.56</strong></td>
</tr>
</tbody>
</table>
Front-end Data Processing

- Worn Data
  - 1ch L & 1ch R
  - RIRs
- Kinect Data
  - 4 ch / U
  - WPE, GSS, BF
- Worn Set
  - 1st ch / U
  - 400k samp.
- Kinect Set
  - 1 ch
- GMM
- Cleanup
- Alignment Expansion

Cleaned Full Training Set, State-level Alignments
Front-end Data Processing

Worn Data

1ch L & 1ch R

RIRs

Worn Set

1st ch / U

4 ch / U

400k samp.

Kinect Data

GMM

Cleanup

Alignment Expansion

Cleaned Full Training Set,
State-level Alignments

WPE, GSS, BF
Front-end Data Processing

Worn Data

1ch L & 1ch R

RIRs

4 ch / U
400k samp.

Worn Set

Kinect Set

GMM

Cleanup

Alignment Expansion

Kinect Data

1st ch / U

4 ch / U

WPE, GSS, BF

Cleaned Full Training Set,
State-level Alignments
Front-end Data Processing

Worn Data
1ch L & 1ch R
RIRs
Worn Set
GMM
Cleanup
Cleaned Full Training Set, State-level Alignments

Kinect Data
1st ch / U
400k samp.
Kinect Set
Alignment Expansion
4 ch / U
WPE, GSS, BE
1 ch
Back-end Acoustic Modeling

• Data augmentation
  – speed perturbation
  – volume perturbation

• 40-d MFCCs & 100-d i-vectors

• Five kinds of acoustic models
  – TDNN-F
  – RBiLSTM (1ch)
  – DcAE-B
  – DcAE-U
  – FEAM-U
Back-end Acoustic Modeling

Output (MFCCs)

Decoder (FC)

Output (states)

AM (TDNN-F)

Output (MFCCs)

Input (MFCCs)

Input (i-vector)

P-Code

R-Code

Encoder (TDNN-F)

U-Net (add.)

Input (MFCCs)

U-Net (concat.)

FEN (TDNN)

(b) FEAM-U

(a) DcAE-B & -U
Back-end Acoustic Modeling

(a) DcAE-B & -U

(b) FEAM-U
Back-end Acoustic Modeling

(a) DcAE-B & -U

(b) FEAM-U
Track 2: Our Contributions

- Compared with the baseline system, we
  
  - Combined all channels of the Kinect data with BeamformIt (BF)
  
  - Developed a new training scheme for speaker representations using Speaker Change information and CNN-based ResNet-34
  
  - Performed re-segmentation with VB diarization
# Track 2: Results

Table 1: *WERs (%) for Track 1 and Track 2 (Category A only).*

<table>
<thead>
<tr>
<th>Model</th>
<th>Track 1</th>
<th>Track 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
<td>Eval</td>
</tr>
<tr>
<td>Baseline</td>
<td>51.32</td>
<td>51.36</td>
</tr>
<tr>
<td>TDNN-F</td>
<td>50.12</td>
<td>49.36</td>
</tr>
<tr>
<td>RBiLSTM</td>
<td>52.43</td>
<td>50.26</td>
</tr>
<tr>
<td>DcAE-B</td>
<td>50.12</td>
<td>49.68</td>
</tr>
<tr>
<td>DcAE-U</td>
<td>49.86</td>
<td>49.63</td>
</tr>
<tr>
<td>FEAM-U</td>
<td>53.47</td>
<td>52.70</td>
</tr>
<tr>
<td>ROVER</td>
<td>47.28</td>
<td>46.82</td>
</tr>
</tbody>
</table>
Track 2: Results

Table 2: Results for Track 2. The acoustic models are the same.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DER</td>
<td>JER</td>
</tr>
<tr>
<td>Baseline</td>
<td>63.42</td>
<td>70.83</td>
</tr>
<tr>
<td>Proposed</td>
<td>56.77</td>
<td>60.62</td>
</tr>
</tbody>
</table>
Track 2: Front-end Processing

• Our front-end data processing follows the baseline program, except that...

  – We used all channels in the Kinect data

  – (Only one specific Kinect was used in baseline)
Track 2: Speaker Modeling

(a) Traditional

(b) Proposed

samples (segments)

acoustic frames

1

2

3

speaker A

speaker B

speaker C

speaker change
Track 2: Speaker Modeling

(a) Traditional

[1, 0, 0]
[0, 1, 0]
[0, 0, 1]

(b) Proposed

[3/8, 5/8, 0]
[2/8, 3/8, 3/8]
[0, 4/8, 4/8]

samples (segments)

acoustic frames

speaker A

speaker B

speaker C

speaker change
Conclusions & Future Work

• In Track 1, we evaluated newly proposed acoustic models, namely DcAE and FEAM
  – DcAE outperforms TDNN-F
  – FEAM needs some modifications and fine tuning in the future

• In Track 2, our proposed speaker modeling method was proved useful for speaker diarization and downstream ASR