The 5th CHiME Speech Separation and Recognition Challenge

Jon Barker, University of Sheffield
Shinji Watanabe, Johns Hopkins University
Emmanuel Vincent, Inria
Jan Trmal, Johns Hopkins University
Overview

- Background - From CHiME-1 to CHiME-5
- CHiME-5 data and task
- CHiME-5 baseline systems
- CHiME-5 submissions and results
Background - From CHiME-1 to CHiME-5

CHiME-1, Interspeech 2011

- 50 hours of audio recorded in a family home via a binaural manikin
- Small vocabulary Grid corpus speech artificially added at distance of 2 m
- Range of SNRs -6 to 9 dB
- 13 submissions; best system (NTT) approached human performance
CHiME-2, ICASSP 2013

- Same noise backgrounds and setup as CHiME-1
- Difficulty extended in two directions:
  - Track 1 - CHiME-1 + simulated speaker movement
  - Track 2 - CHiME-1 + larger vocab (WSJ)
- Best Track 1 system matches human scores for 0 to 6 dB
- Best Track 2 halved baseline WERs but WERs still much higher than clean WSJ.
Background - From CHiME-1 to CHiME-5

CHiME-3, ASRU 2015

- 6 channel tablet recording device
- WSJ speech recorded live in noisy environments
  - cafe, bus, street, pedestrian
- Baseline system performance
  33% WER
- Best system (NTT) reduced WER to 5.8%
CHiME-4, Interspeech 2016

- Rerun of CHiME-3
- Additional tracks for 2 channel and 1 channel processing
- 6 Channel WER reduced from 5.8% down to 2.2% (USTC-iFlyTek)
- 1 Channel WER 9.2% (USTC-iFlyTek)
Overview

- Background - From CHiME-1 to CHiME-5
- CHiME-5 data and task
- CHiME-5 baseline systems
- CHiME-5 submissions and results
The CHiME-5 scenario

The CHiME-5 is designed around a ‘dinner party’ scenario

- Recordings in people’s actual homes
- Parties of 4 - typically, two hosts and two guests
- All participants are well known to each other
- Collection of 20 parties each lasting 2 to 3 hours
- Each party having three separate stages each of min 30 minutes:
  - Kitchen phase – dinner preparation
  - Dining room phase – eating
  - Sitting room phase – post-dinner socialising
The CHiME-5 recording set up

Data has been captured with 32 audio channels and 6 video channels

- **Participants microphones**
  - Binaural in-ear microphones recorded onto stereo digital recorders
  - Primarily for transcription but also uniquely interesting data
  - Channels: $4 \times 2$

- **Distant microphones**
  - Six separate Microsoft Kinect devices
  - Two Kinects per living area (kitchen, dining, sitting)
  - Arranged so that video captures most of the living space
  - Channel: $6 \times 4$ audio and 6 video
Example recording set ups

S04

S07

S12

S23

Session ID: S04
August 13, 2017 18:03 PST

Session ID: S07
August 13, 2017 12:03 PST

Session ID: S12
August 27, 2017 12:07 PST

Session ID: S23
September 22, 2017 14:53 PST
CHiME-5 kitchen examples
CHiME-5 dinner examples
CHiME-5 data and task

CHiME-5 living room examples 🎵 🎵
CHiME-5 data and task

CHiME-5 data overview

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Parties</th>
<th>Speakers</th>
<th>Hours</th>
<th>Utterances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>16</td>
<td>32</td>
<td>40:33</td>
<td>79,980</td>
</tr>
<tr>
<td>Dev</td>
<td>2</td>
<td>8</td>
<td>4:27</td>
<td>7,440</td>
</tr>
<tr>
<td>Eval</td>
<td>2</td>
<td>8</td>
<td>5:12</td>
<td>11,028</td>
</tr>
</tbody>
</table>

The audio data

- All audio data are distributed as 16 kHz WAV files
- Each session consists of
  - recordings made by the binaural microphones worn by each participant (4 participants per session),
  - 6 microphone arrays with 4 microphones each.
- Total number of microphones per session is 32 (2 × 4 + 4 × 6).
- Total data size: 120 GB
CHiME-5 transcriptions

The transcriptions provided in JSON format. Separate file per session, `<session ID>.json`. The JSON file includes the following pieces of information for each utterance:

- **Session ID** ("session_id")
- **Location** ("kitchen", "dining", or "living")
- **Speaker ID** ("speaker")
- **Transcription** ("words")
- **Start time** ("start_time")
  - For the binaural microphone recording of that speaker ("original")
  - For all array recordings ("U01", etc.)
  - For all binaural microphone recordings ("P01", etc.)
- **End time** ("end_time")
- **Reference microphone array ID** ("ref")
CHiME-5 data and task

CHiME-5 tracks

The challenge has two tracks:

- single array: must use only the reference array (for evaluation),
- multiple array: all arrays can be used.

Two separate rankings have been produced:

- Ranking A: conventional acoustic model + official language model (‘acoustic robustness’),
- Ranking B: all other systems (including end-to-end).
Overview

- Background - From CHiME-1 to CHiME-5
- CHiME-5 data and task
- CHiME-5 baseline systems
- CHiME-5 submissions and results
The baseline system software

Fully open-source baseline systems are provided,

- Array synchronisation (Python/OpenCV code)
- Enhancement (BeamformIt)
- Conventional ASR (Kaldi)
  - GMM
  - LF-MMI TDNN
- Enhancement and end-to-end ASR (ESPnet)

Alignment - Headworn mics

Delay w.r.t to reference mic (P17) plotted against time.

‘clock drift’ - drift about ± 25 ms per hour
Alignment - Kinect mics

Delay w.r.t to reference mic (P17) plotted against time.

Kinect’s exhibit clock drift but also ‘frame dropping’ events.
Enhancement + Conventional ASR (GMM)

GMM system - Kaldi recipe based mostly on the TED-LIUM and Switchboard recipe:

1. Data preparation, language model generation
2. Weighted delay-and-sum beamforming using BeamformIt
3. MFCC feature extraction
4. Triphone acoustic modelling
5. LDA transform
6. Maximum likelihood linear transform (MLLT)
7. Feature space MLLR with speaker adaptive training
Enhancement + Conventional ASR (GMM)

Notes:

□ Language Model
  ▶ No external text used
  ▶ Uses a 3gram language model - built with SRILM
  ▶ Vocabulary size: 127,712
  ▶ Many OOV words w.r.t CMU dict
    ● Arjan, Netflix, pesto, thrones, prolly, konichiwa, betterer
    ● phonetisaurus G2P to generate pronunciations - k aa n ih ch iy w ah

□ Beamforming
  ▶ Perform beamforming with 4 microphones for an entire audio in each array
  ▶ Pick up beamformed signals from the reference array

□ Training data
  ▶ Trained using unenhanced signals
  ▶ 100k randomly selected Kinect utterances
  ▶ 75k left channel of worn mic utterances
Enhancement + Conventional ASR (DNN)

- Data cleaning
  - removes irregular utterances from the obtained GMM model
  - totally 15% of utterances in the training data are excluded

- LF-MMI TDNN - advanced DNN baseline that runs as last step of the Kaldi recipe.
  Requires,
  - multiple CPUs for i-vector and alignment lattice generation
  - multiple GPUs for TDNN training
  - applies data augmentation (‘speed perturbation’)

End-to-End ASR

ESPnet: open source end-to-end ASR toolkit using Chainer and PyTorch

1. Date preparation: similar to Kaldi except (but no lexicons)
2. Feature extraction: similar to Kaldi
3. Character-based LSTM language modeling
4. Hybrid CTC/attention training, character based
5. Combining LSTM language model and end-to-end ASR during decoding
Baseline dev set performances

<table>
<thead>
<tr>
<th>Session Location</th>
<th>S02 Kit</th>
<th>S02 Din</th>
<th>S02 Liv</th>
<th>S09 Kit</th>
<th>S09 Din</th>
<th>S09 Liv</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM</td>
<td>93.9</td>
<td>91.4</td>
<td>90.8</td>
<td>90.8</td>
<td>90.9</td>
<td>88.1</td>
<td>91.7</td>
</tr>
<tr>
<td>DNN</td>
<td>86.6</td>
<td>79.1</td>
<td>78.4</td>
<td>82.6</td>
<td>81.67</td>
<td>77.9</td>
<td>81.1</td>
</tr>
<tr>
<td>Worn</td>
<td>51.8</td>
<td>53.2</td>
<td>46.7</td>
<td>46.3</td>
<td>49.7</td>
<td>42.5</td>
<td>47.9</td>
</tr>
</tbody>
</table>
## Baseline system performance

<table>
<thead>
<tr>
<th>Track</th>
<th>Session</th>
<th>Kitchen</th>
<th>Dining</th>
<th>Living</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>S02</td>
<td>87.4</td>
<td>79.1</td>
<td>78.8</td>
<td>81.1</td>
</tr>
<tr>
<td></td>
<td>S09</td>
<td>81.7</td>
<td>80.6</td>
<td>77.6</td>
<td></td>
</tr>
<tr>
<td>Eval</td>
<td>S01</td>
<td>82.6</td>
<td>67.2</td>
<td>81.6</td>
<td>73.3</td>
</tr>
<tr>
<td></td>
<td>S21</td>
<td>77.6</td>
<td>65.8</td>
<td>70.4</td>
<td></td>
</tr>
</tbody>
</table>
## Baseline result analysis: WER by speaker

<table>
<thead>
<tr>
<th>Spkr</th>
<th>P05</th>
<th>P06</th>
<th>P07</th>
<th>P08</th>
<th>P25</th>
<th>P26</th>
<th>P27</th>
<th>P28</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM</td>
<td>92.18</td>
<td>89.25</td>
<td>90.18</td>
<td>99.44</td>
<td>86.82</td>
<td>96.58</td>
<td>92.56</td>
<td>85.65</td>
</tr>
<tr>
<td>DNN</td>
<td>83.34</td>
<td>78.59</td>
<td>79.45</td>
<td>86.61</td>
<td>81.07</td>
<td>93.05</td>
<td>82.22</td>
<td>68.99</td>
</tr>
<tr>
<td>Worn</td>
<td>53.73</td>
<td>48.50</td>
<td>45.75</td>
<td>52.42</td>
<td>43.53</td>
<td>53.03</td>
<td>51.21</td>
<td>37.38</td>
</tr>
</tbody>
</table>

![Bar chart showing WER by speaker](chart.png)
Simultaneous speech

Proportion of parties in which 0, 1, 2, 3 or 4 speakers are active.

Some parties are more ‘overlappy’ than others. Big variations in amount of non-speech audio.
Overview

- Background - From CHiME-1 to CHiME-5
- CHiME-5 data and task
- CHiME-5 baseline systems
- CHiME-5 submissions and results
CHiME-5 Submissions and Results

**Submissions**

- distributed to **381** research groups! (cf. CHiME-3/4: ~100)
- challenge submissions
  - totally 35 submissions by 20 papers (cf. CHiME-4: 43 submissions by 19 papers)
    - Single-A: 17, Single-B: 7, Multiple-A: 8, Multiple-B: 4
  - totally 132 authors, 6.6 authors per paper
  - totally 37 groups, 1.8 groups per paper
  - academia 24 vs. Industry 13
  - Asia 21, Europe 11, North America 5
  - CHiME 1-4 participants versus new participants: 21 vs 111
Results: Single-Device A

Eval/All vs. Authors

Authors

USTC-Flytek
Hitachi-JHU
STC
Toshiba
NWPU
Asus
CAS
RWTH
SHNU
IISC
Sheffield
NDSC
Paderborn
Baseline
Lenovo
IIT Bombay-TCS
TCS-IIT Bombay
Results: Single-Device B

Eval/All vs. Authors

Authors

USTC-iFlytek
Hitachi-JHU
NWPV
CAS
NWPV-Mobvoi
CMU
AnTech

Eval/All

40.00
50.00
60.00
70.00
80.00

All
Results: Multiple-Device A

Eval/All vs. Authors

Authors

USTC-iFlytek  Hitachi-JHU  RWTH-  STC  CAS  Sheffield-  Paderborn  Lenovo
Results: Multiple-Device B

Eval/All vs. Authors

- USTC-iFlytek
- Hitachi-JHU
- CAS
- DAIICT-IIITV

Authors
Oral presentations from top 3 teams

The next three talks will be presentations from the top 3 teams:

3rd **The STC System for the CHiME 2018 Challenge**
Ivan Medennikov, Ivan Sorokin, Aleksei Romanenko, Dmitry Popov, Yuri Khokhlov, Tatiana Prisyach, Nikolay Malkovskiy, Vladimir Bataev, Sergei Astapov, Maxim Korenevsky and Alexander Zatvornitskiy

2nd **The Hitachi/JHU CHiME-5 system: Advances in speech recognition for everyday home environments using multiple microphone arrays**

1st **The USTC-iFlytek systems for CHiME-5 Challenge**
Jun Du, Tian Gao, Lei Sun, Feng Ma, Yi Fang, Di-Yuan Liu, Qiang Zhang, Xiang Zhang, Hai-Kun Wang, Jia Pan, Jian-Qing Gao, Chin-Hui Lee and Jing-Dong Chen