The 5th CHiME Speech Separation and Recognition Challenge

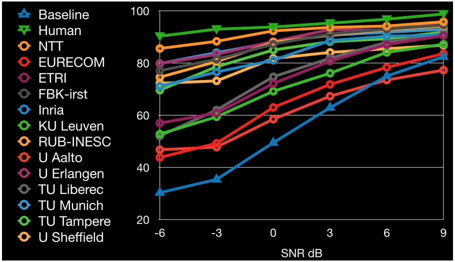
Jon Barker, University of Sheffield
Shinji Watanabe, Johns Hopkins University
Emmanuel Vincent, Inria
Jan Trmal, Johns Hopkins University

Overview

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

CHiME-1, Interspeech 2011

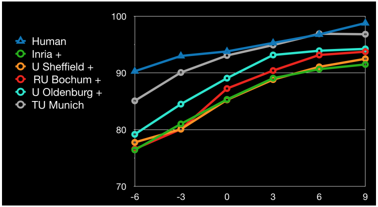
- 50 hours of audio recorded in a family home via a binaural manikin
- Small vocabulary Grid corpus speech artificially added at distance of 2 m
- Range of SNRs -6 to 9 dB
- 13 submissions; best system (NTT) approached human performance



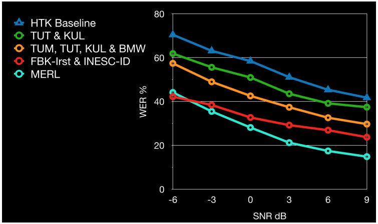
CHiME-2, ICASSP 2013

- Same noise backgrounds and set up 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.

Track 1 - speaker movement

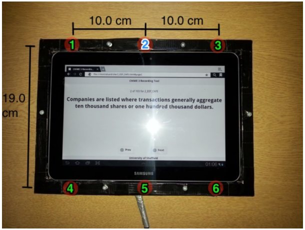


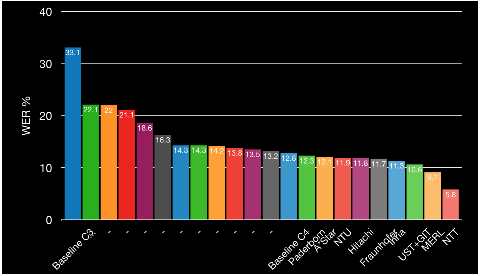
Track 2 - larger vocab



CHiME-3, ASRU 2015

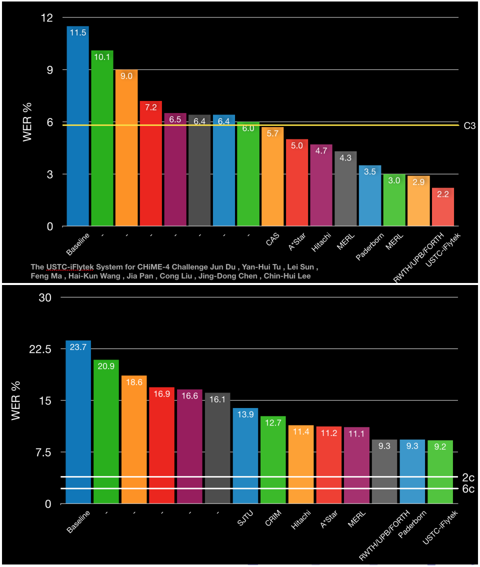
- 6 channel tablet recording device
- WSJ speech recorded live in noisy environments
  - cafe, bus, street, pedestrian
- Baseline system performance 33% WER
- Best system (NTT) reduced WER to 5.8%





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

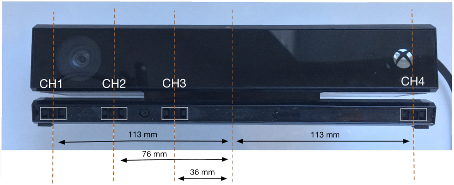
The CHiME-5 is designed around a ‘dinner party’ scenario

- 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

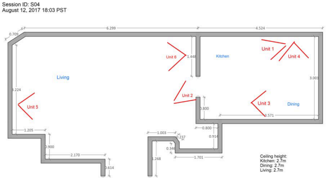
Data has been captured with 32 audio channels and 6 video channels

- Participants microphones
  - Binaural in-ear microphones recorded onto stereo digital recorders
  - Primarily for transcription but also uniquely interesting data
  - Channels: 4×2
- 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: 6×4 audio and 6 video

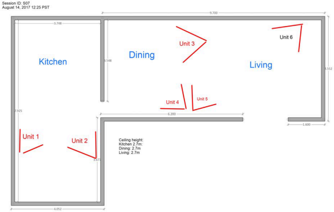


Example recording set ups

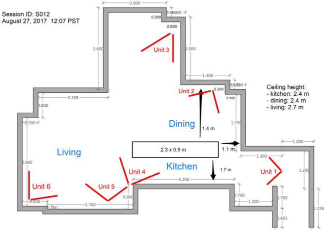
S04



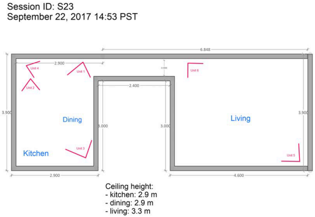
S07



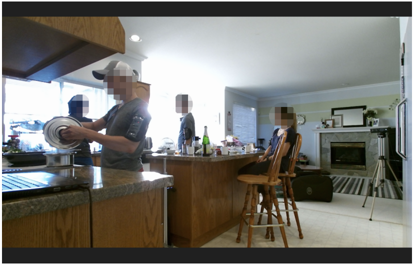
S12



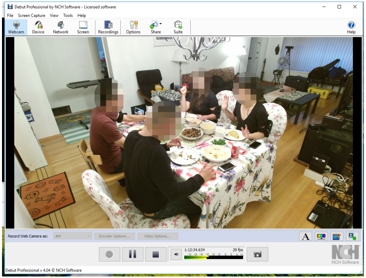
S23



CHiME-5 kitchen examples



CHiME-5 dinner examples



CHiME-5 living room examples

CHiME-5 data overview

Dataset	Parties	Speakers	Hours	Utterances
Train	16	32	40:33	79,980
Dev	2	8	4:27	7,440
Eval	2	8	5:12	11,028

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 x 4 + 4 x 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 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

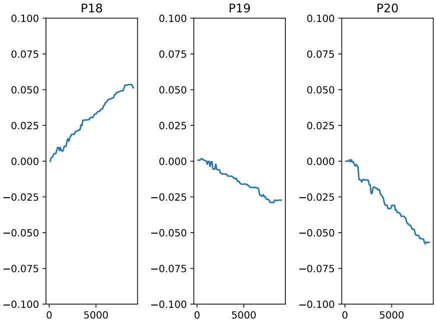
Fully open-source baseline systems are provided,

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

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal, "The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," accepted for Interspeech, 2018.

Alignment - Headworn mics

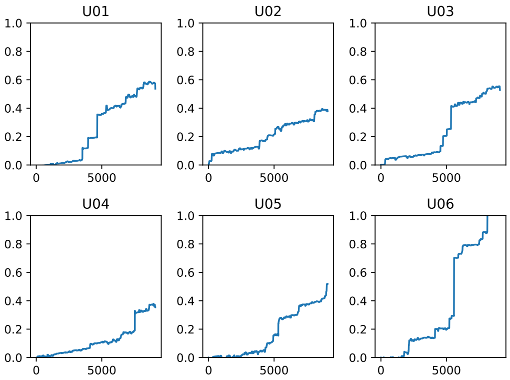
Delay w.r.t to reference mic (P17) plotted against time.



'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 BeamformtIt
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')

  D. Povey et al., "Purely sequence-trained neural networks for ASR based on lattice-free MMI", in Proc. Interspeech 2016.

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

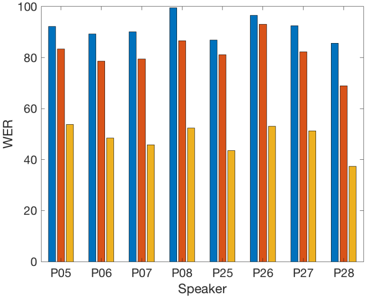
Session	S02			S09			Overall
Location	Kit	Din	Liv	Kit	Din	Liv	
GMM	93.9	91.4	90.8	90.8	90.9	88.1	91.7
DNN	86.6	79.1	78.4	82.6	81.67	77.9	81.1
Worn	51.8	53.2	46.7	46.3	49.7	42.5	47.9

Baseline system performance

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	87.4	79.1	78.8	81.1
		S09	81.7	80.6	77.6	
	Eval	S01	82.6	67.2	81.6	73.3
		S21	77.6	65.8	70.4	

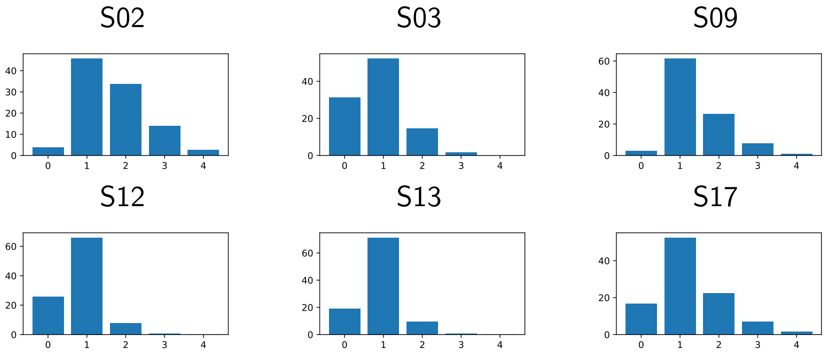
Baseline result analysis: WER by speaker

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



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.

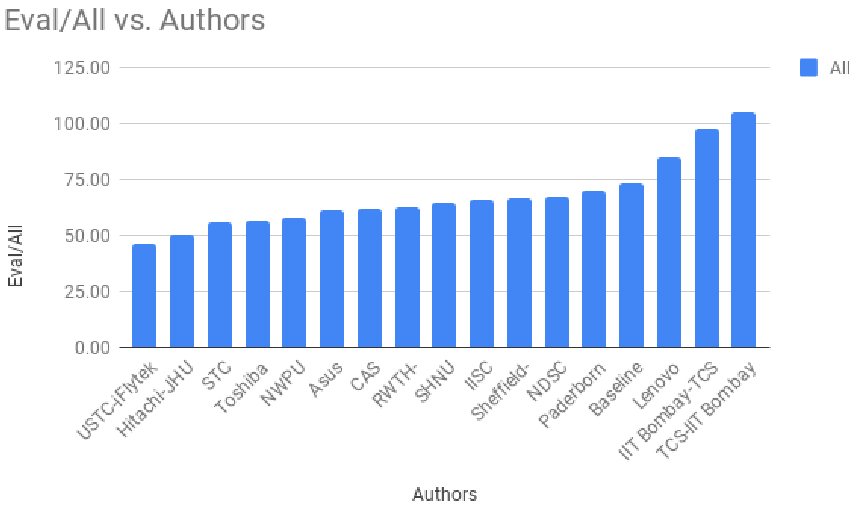
Overview

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

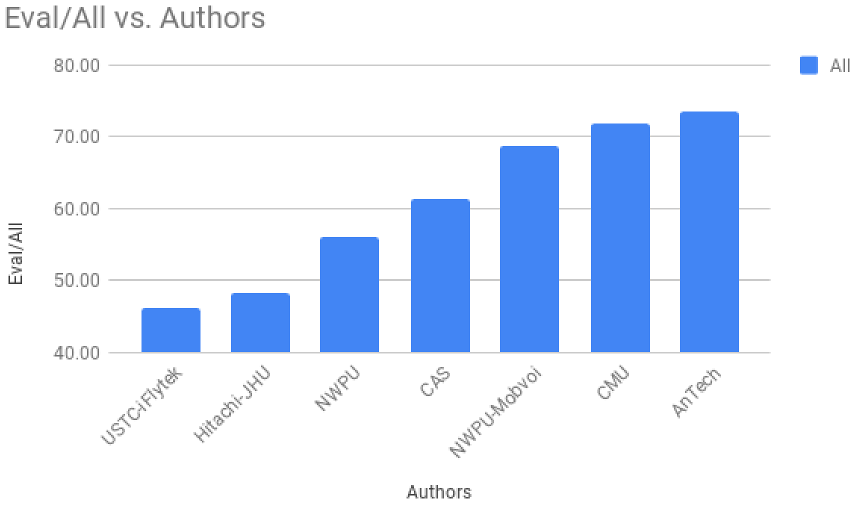
Submissions

- distributed to **381** research groups! (cf. CHiME-3/4: $\sim$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

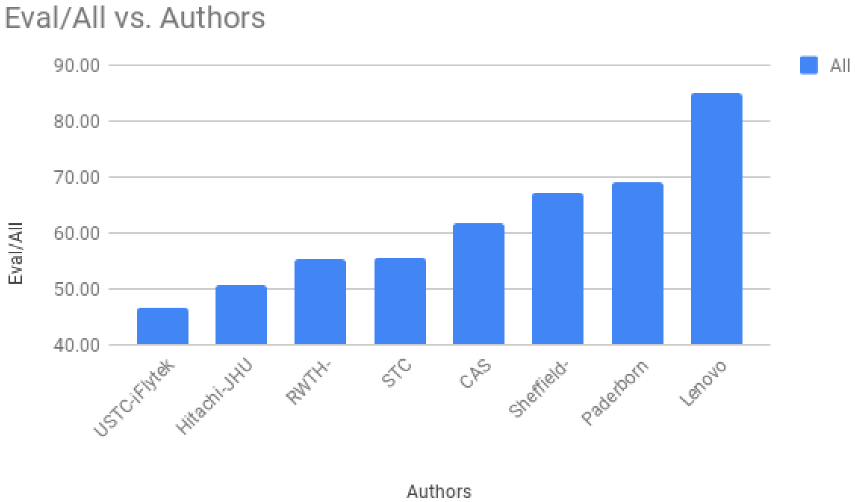
Results: Single-Device A



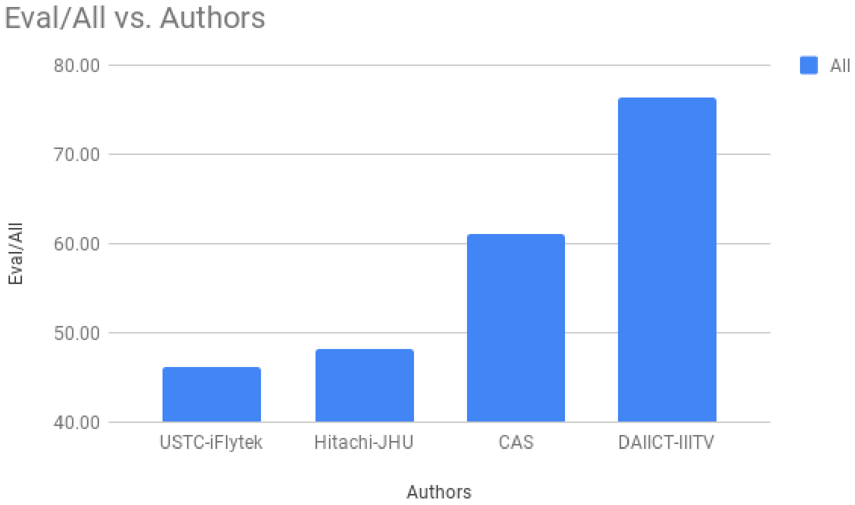
Results: Single-Device B



Results: Multiple-Device A



Results: Multiple-Device B



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**

Naoyuki Kanda, Rintaro Ikeshita, Shota Horiguchi, Yusuke Fujita, Kenji Nagamatsu, Xiaofei Wang, Vimal Manohar, Nelson Enrique Yalta Soplin, Matthew Maciejewski, Szu-Jui Chen, Aswin Shanmugam Subramanian, Ruizhi Li, Zhiqi Wang, Jason Naradowsky, L. Paola Garcia-Perera and Gregory Sell

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