The 6th CHiME Speech Separation and Recognition Challenge

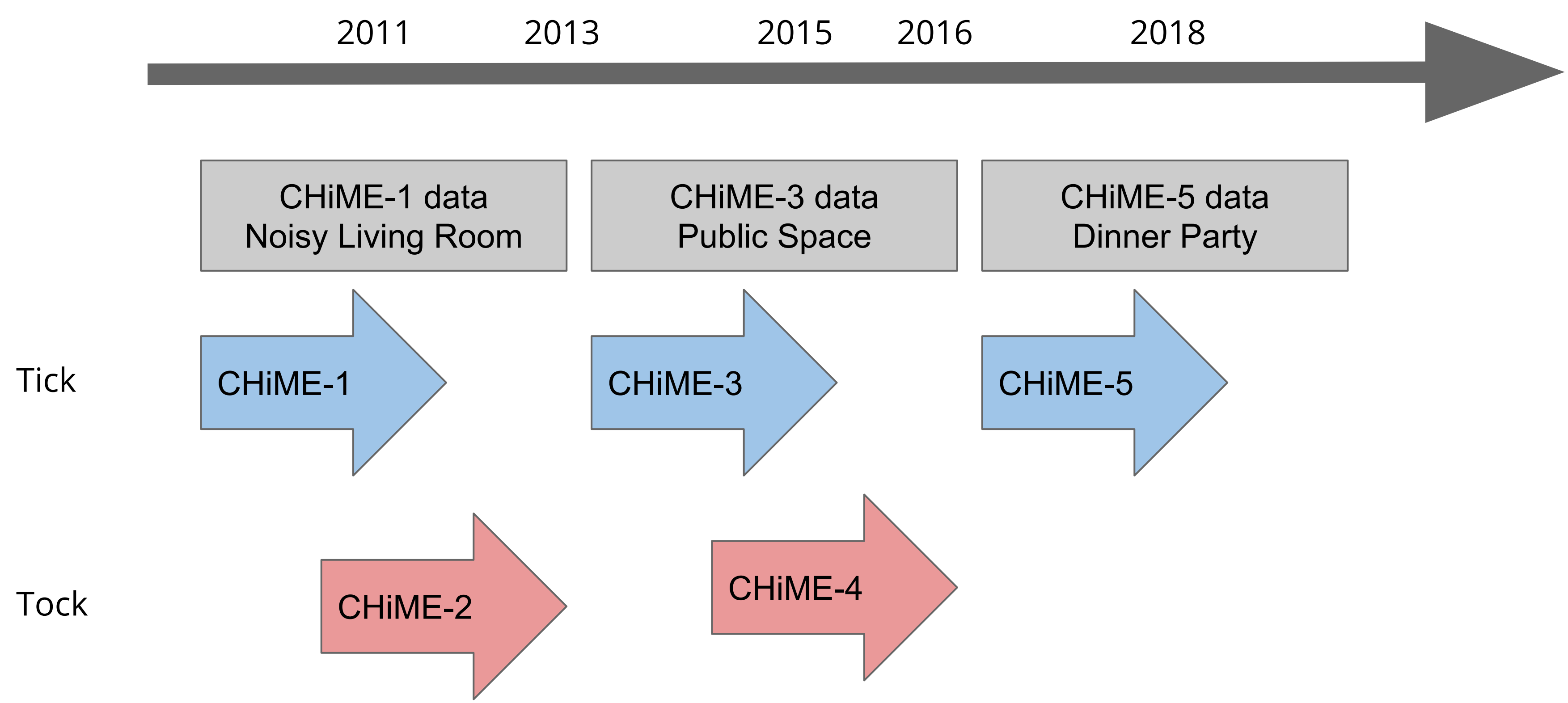
Shinji Watanabe, Johns Hopkins University
Michael Mandel, The City University of New York, USA
Jon Barker, University of Sheffield
Emmanuel Vincent, Inria

Supported by ISCA SIG RoSP

Overview

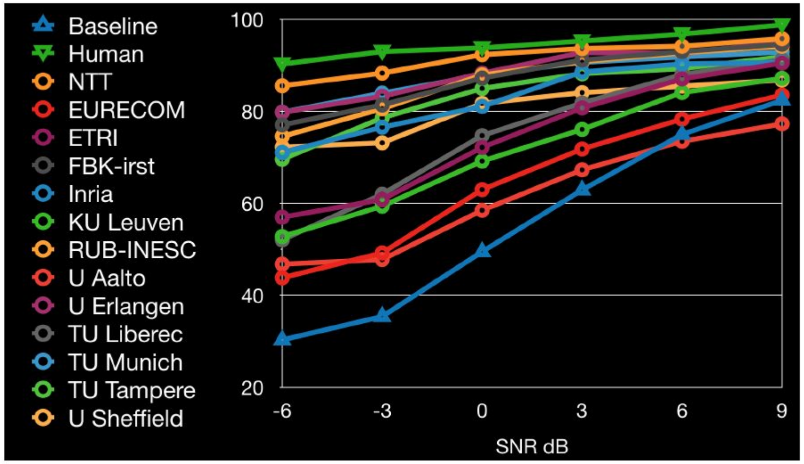
- Background - From CHiME-1 to CHiME-6
- CHiME-6 data and task
- CHiME-6 baseline systems
- CHiME-6 submissions and results

CHiME tick-tock model



CHiME-1, Interspeech 2011

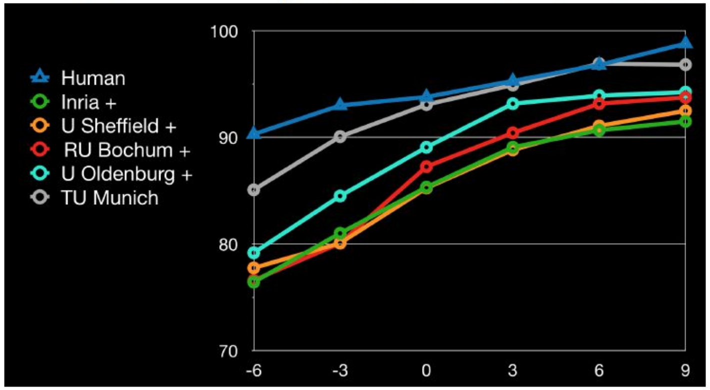
- 50 hours of audio recorded in a family home via a binaural manikin
- Small vocabulary Grid corpus speech artificially added at distance of 2 m
- Range of SNRs -6 to 9 dB
- 13 submissions; best system (NTT) approached human performance



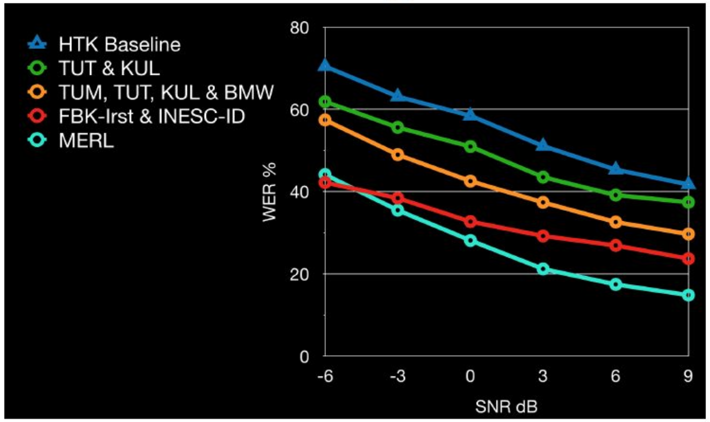
CHiME-2, ICASSP 2013

- 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

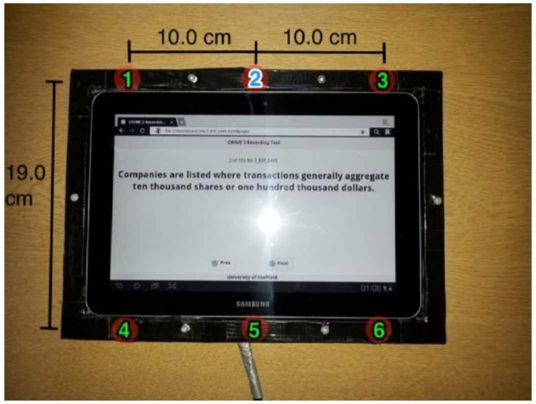


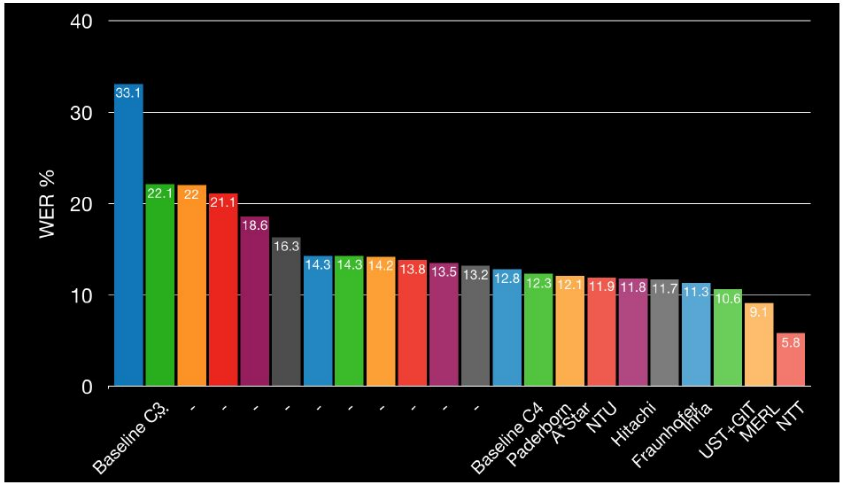
Track 2 - larger vocab



CHiME-3, ASRU 2015

- Jumped to the real data
- 6 channel tablet recording device (~50cm between source and mic)
- WSJ speech recorded live in noisy public environments
  - cafe, bus, street, pedestrian
- Baseline system performance 33% WER
- Best system (NTT) reduced WER to 5.8%

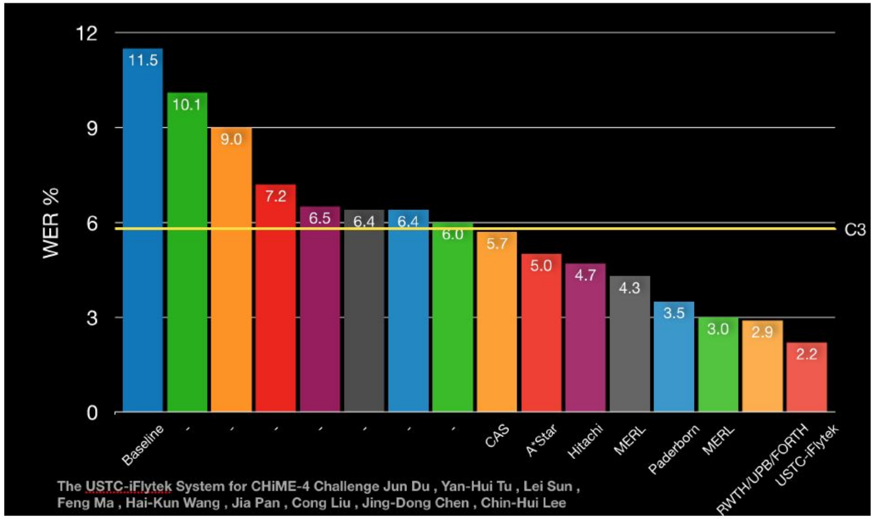


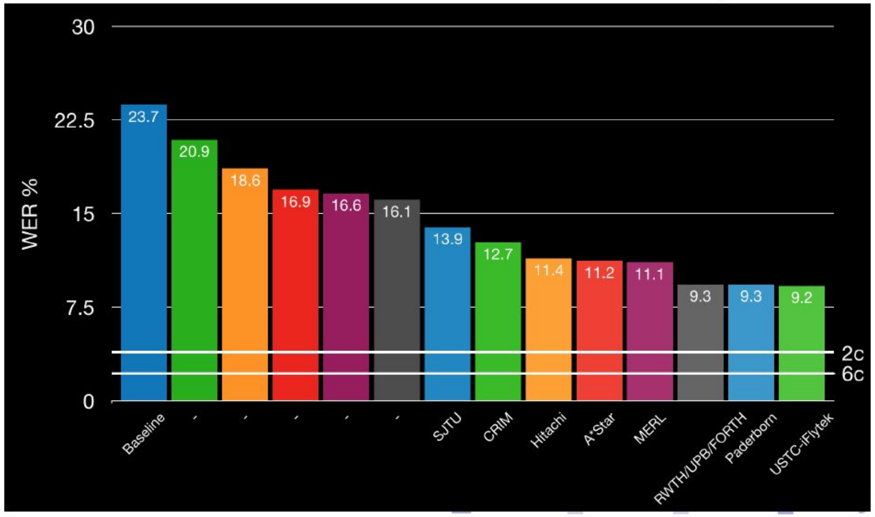


CHiME-4, Interspeech 2016

- 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)

Given this result, we moved to the next more realistic challenge

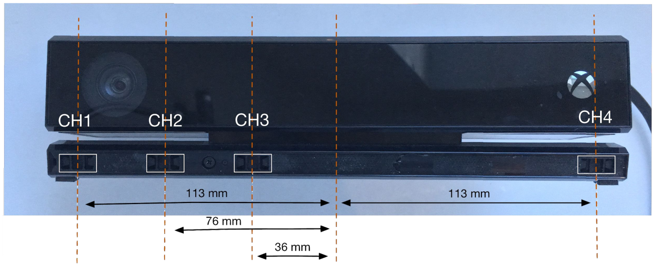


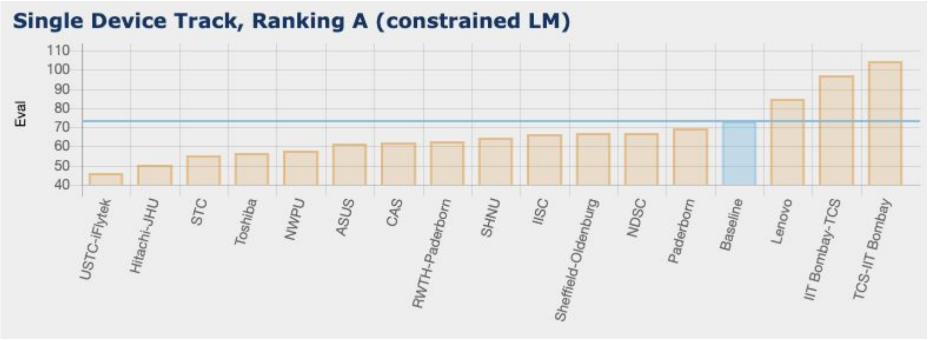


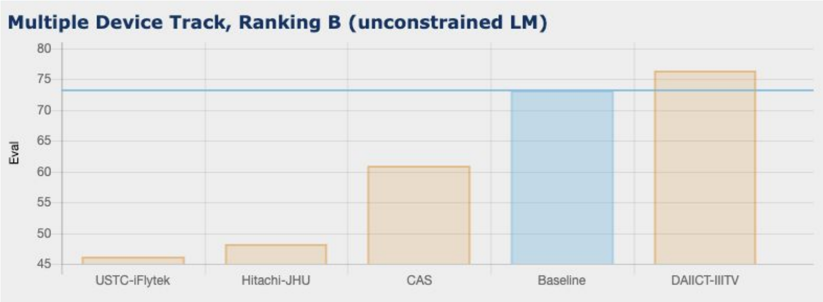
CHiME-5, Interspeech 2018

- Dinner party scenario
- Multiple microphone arrays
  - Binaural mics for participants
  - 6 Kinect devices located on multiple rooms
- Two tracks (single array vs. multiple array)

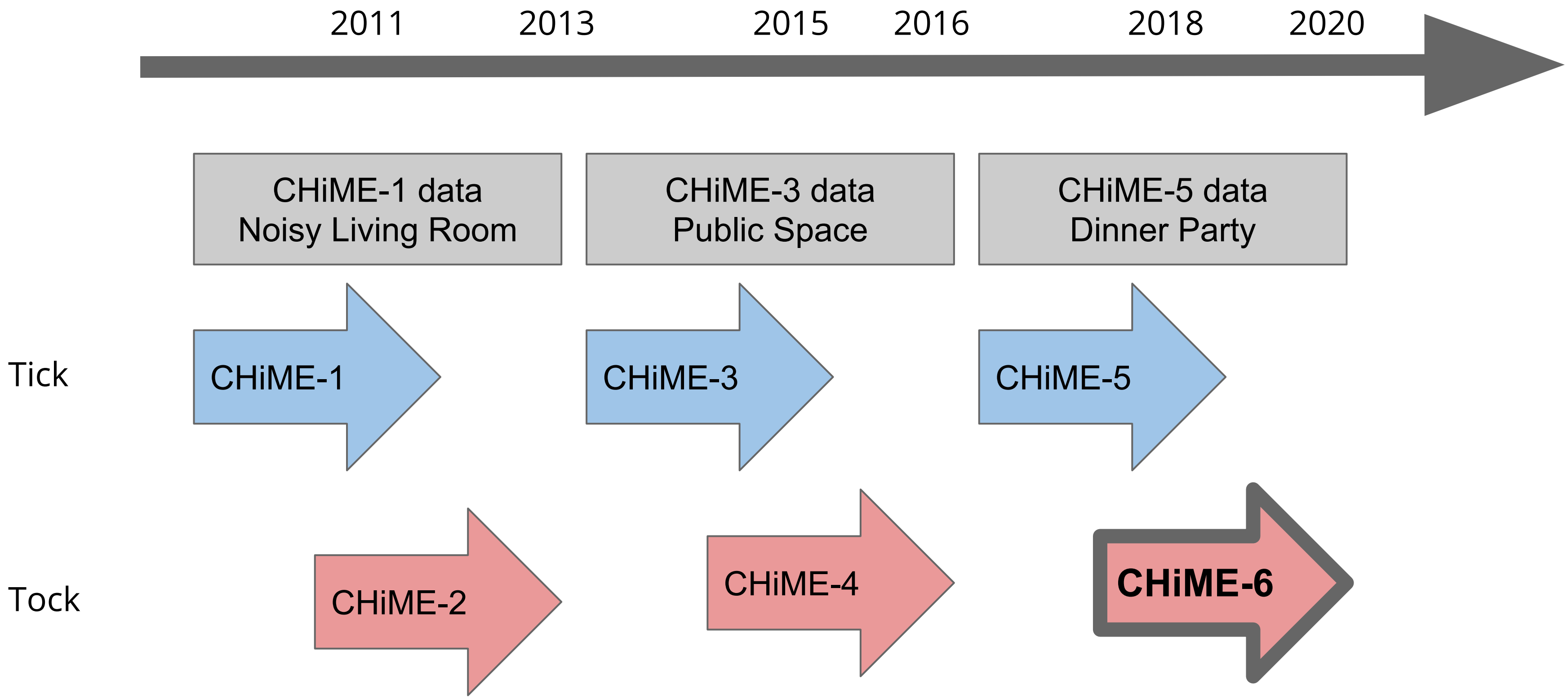
- Kaldi Baseline: 73.3%
- Best system (USTC-iFlyTek): 46.1%







CHiME tick-tock model



Overview

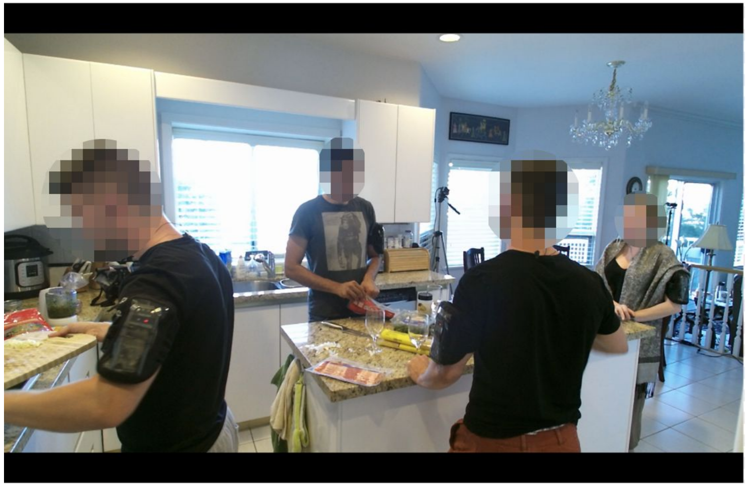
- Background - From CHiME-1 to CHiME-6
- CHiME-6 data and task
- CHiME-6 baseline systems
- CHiME-6 submissions and results

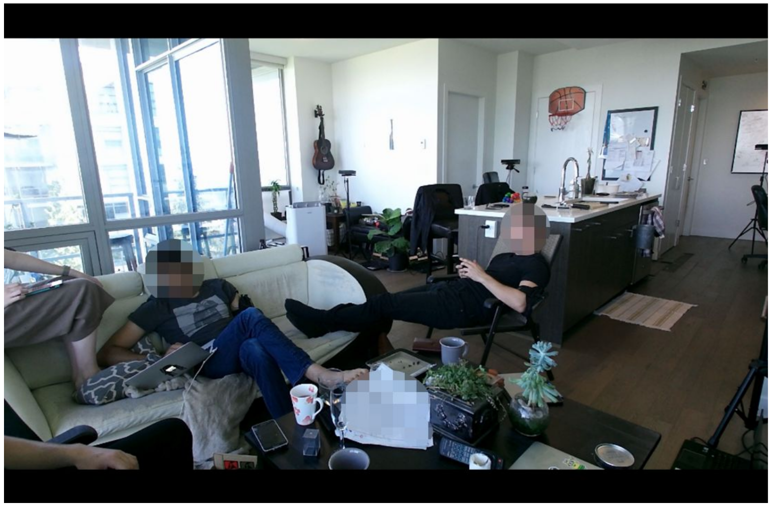
The CHiME-6 scenario

Revisiting the CHiME-5 `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 has three separate stages each of at least 30 minutes:
  - Kitchen phase - dinner preparation
  - Dining room phase - eating
  - Sitting room phase - post-dinner socialising

CHiME-6 examples

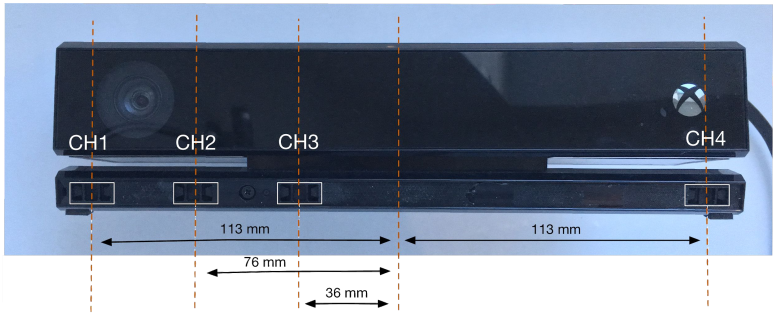




The CHiME-6 recording setup

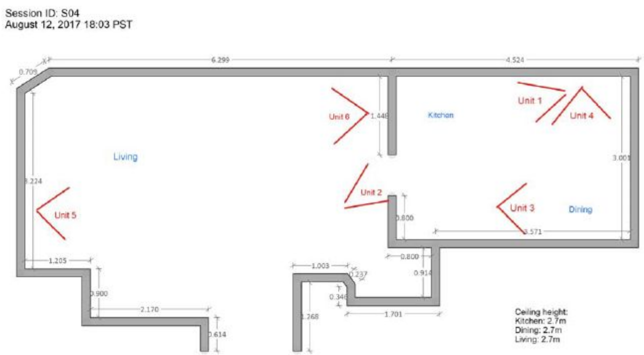
Data has been captured with 32 audio channels and 6 video channels

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

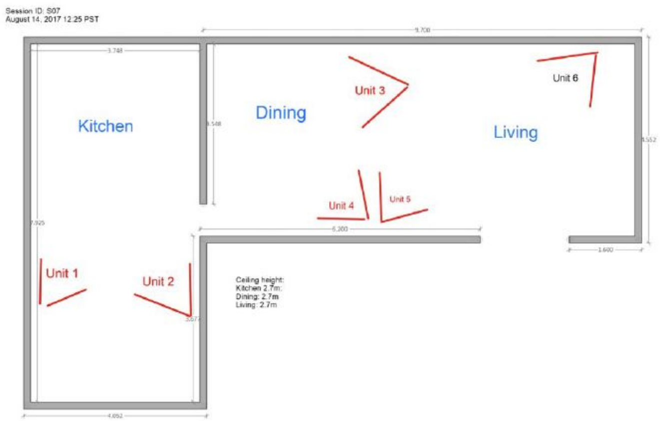


Example recording setups

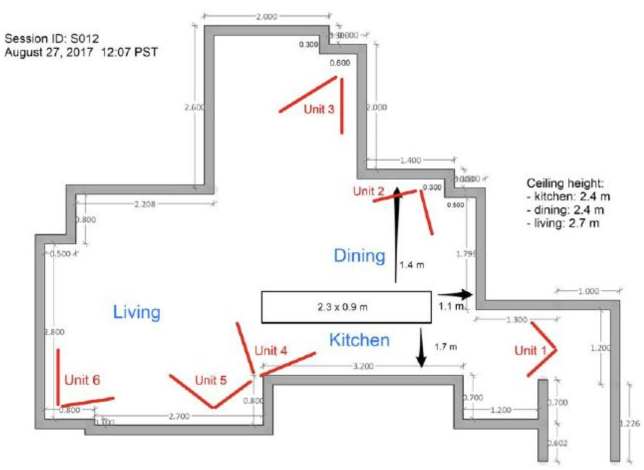
S04



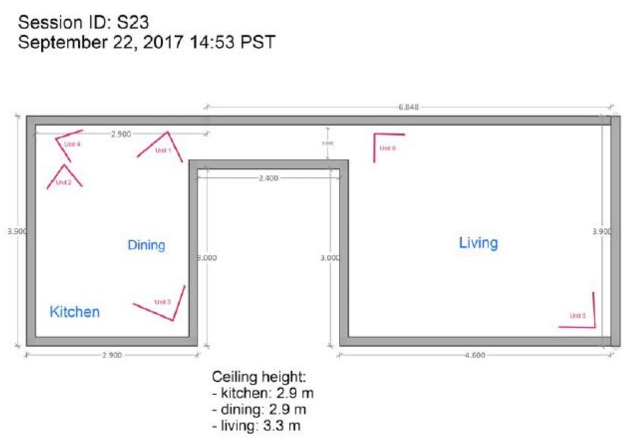
S07



S12



S23



CHiME-6 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	11,208

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-6 transcriptions

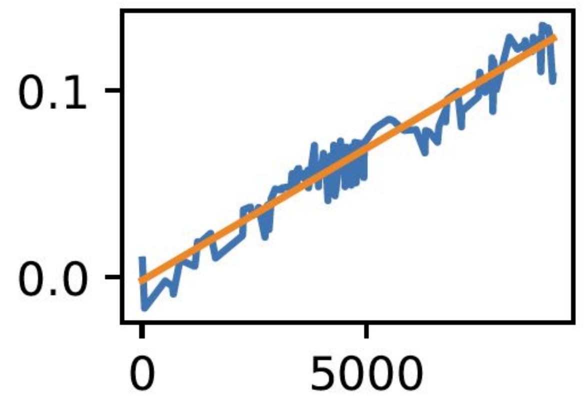
Transcriptions are provided in JSON format. Separate file per session, <session ID>.json. The JSON file includes the following pieces of information for each utterance:

- Session ID (“session id")
- Location (“kitchen", “dining", or “living")
- Speaker ID (“speaker")
- Transcription (“words")
- Start time (“start time")
- End time (“end time”)

Array synchronization

Desynchronisation in CHiME-5 data due to audio **frame dropping** and **clockdrift**.

- Frame dropping (Kinect signals only)
  - Detected by matching to an uncorrupted 1-channel audio signal captured by the video software.
  - Corrected by inserting 0's into signal
  - Typically 1-2 seconds per session.
- Clockdrift:
  - Fix a reference channel. Compute lag at 10 second intervals throughout the session and perform linear fit.
  - Signal speed can then be corrected by sox resampling.
  - Typically ~100 ms per session.



An example clockdrift plot.

Estimated Kinect delay (seconds) versus session time (seconds) and linear fit.

CHiME-6 tracks

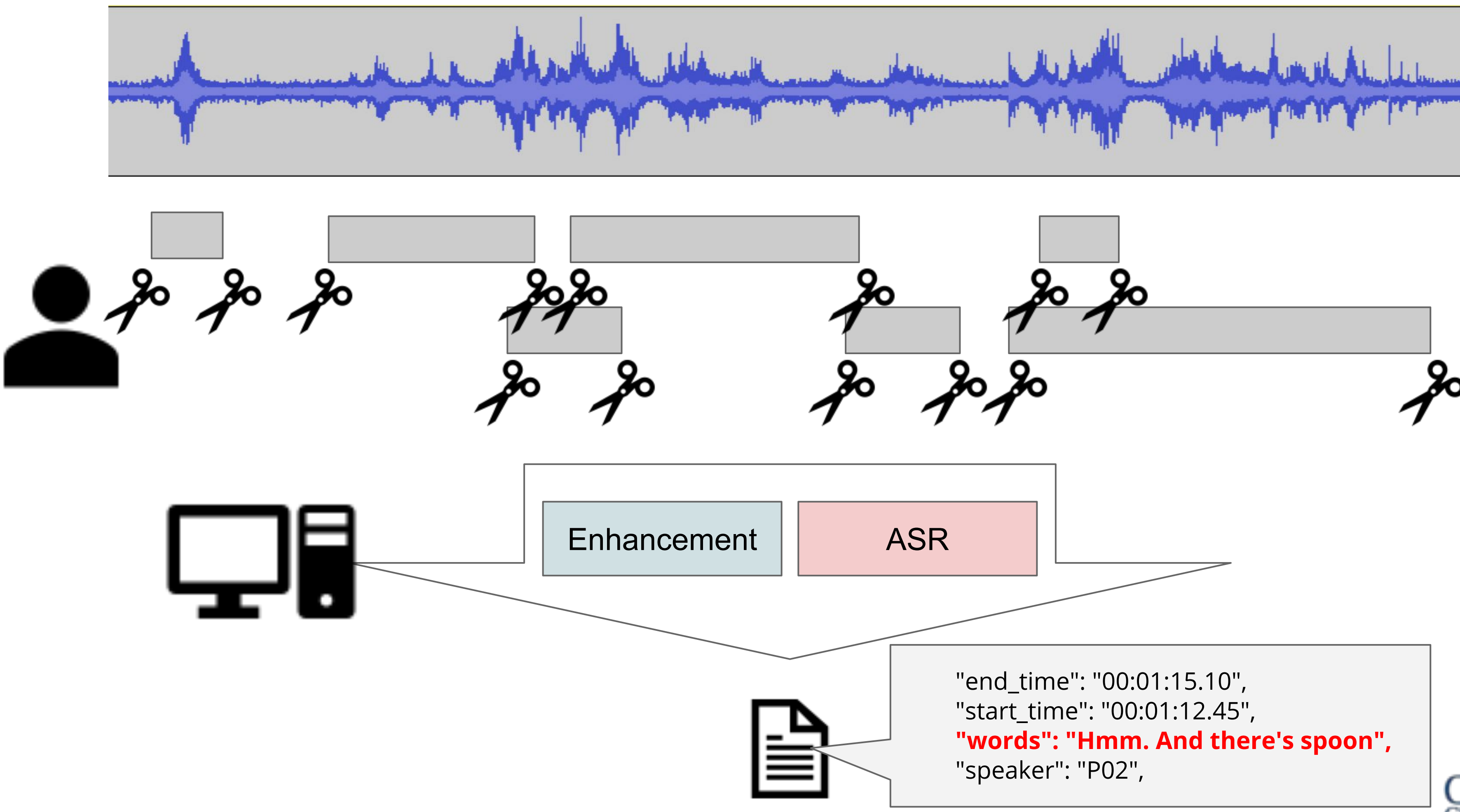
The challenge has two tracks:

- **Track 1**: oracle segmentation (equivalent to CHiME-5 multiple array track)
- **Track 2:** no segmentation

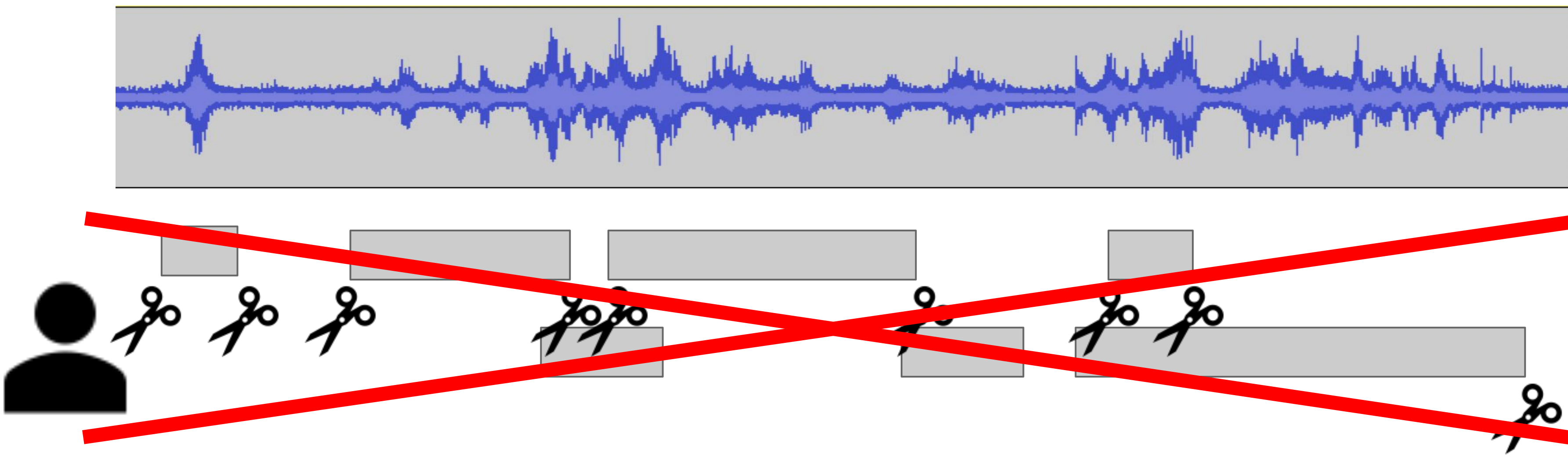
Two separate rankings have been produced:

- Ranking A: conventional acoustic model + official language model (`acoustic robustness')
- Ranking B: all other systems

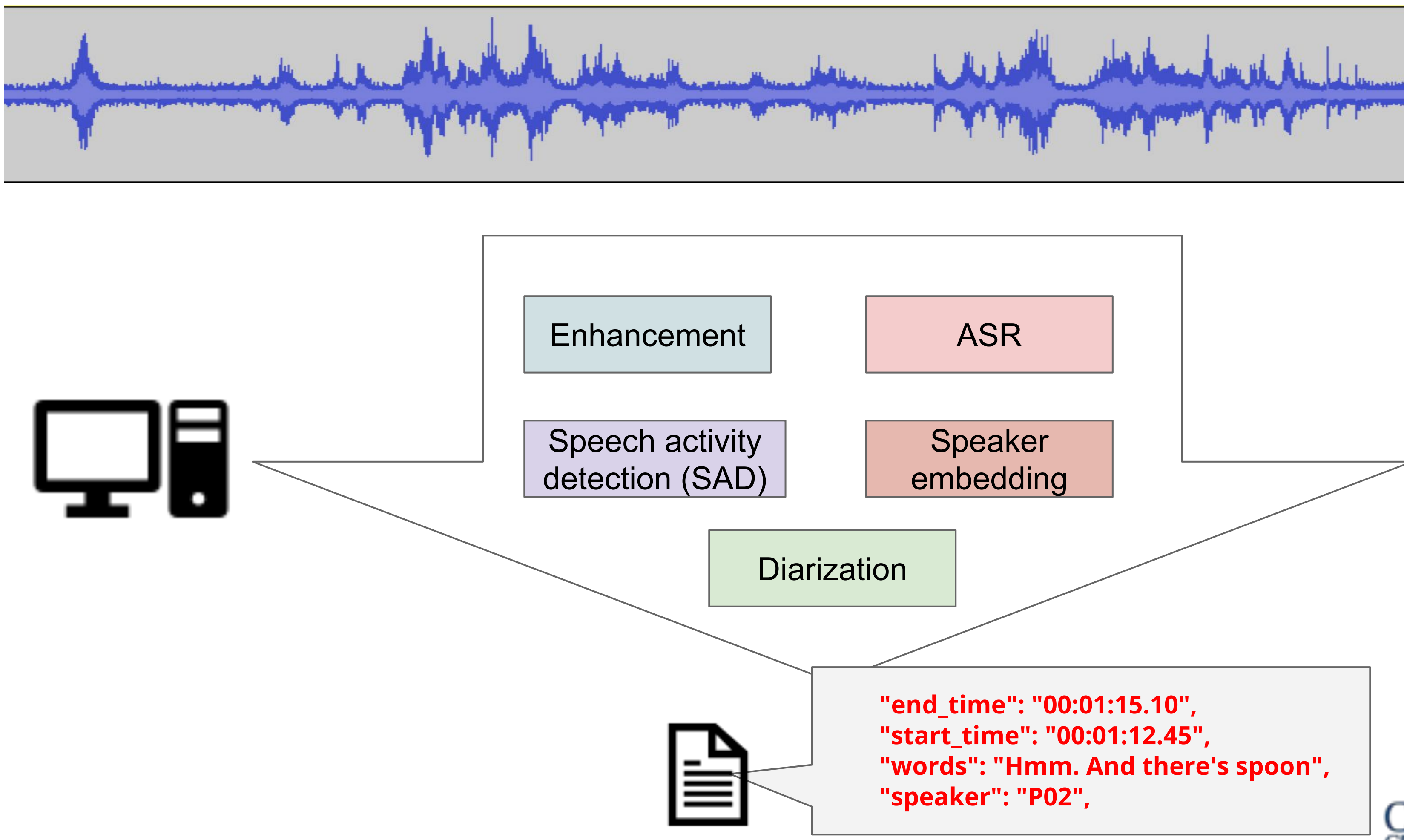
CHiME-6 track 1



CHiME-6 track 2



CHiME-6 track 2



CHiME-6 tracks

The challenge has two tracks:

- **Track 1**: oracle segmentation (equivalent to CHiME-5 multiple array track)
- **Track 2:** no segmentation

Two separate rankings have been produced:

- Ranking A: conventional acoustic model + official language model (`acoustic robustness')
- Ranking B: all other systems

Overview

- Background - From CHiME-1 to CHiME-6
- CHiME-6 data and task
- CHiME-6 baseline systems
- CHiME-6 submissions and results

Policies of the baseline construction

Track 1

- Strong, reproducible (all open source) baseline, but maintain the simplicity

Track 2

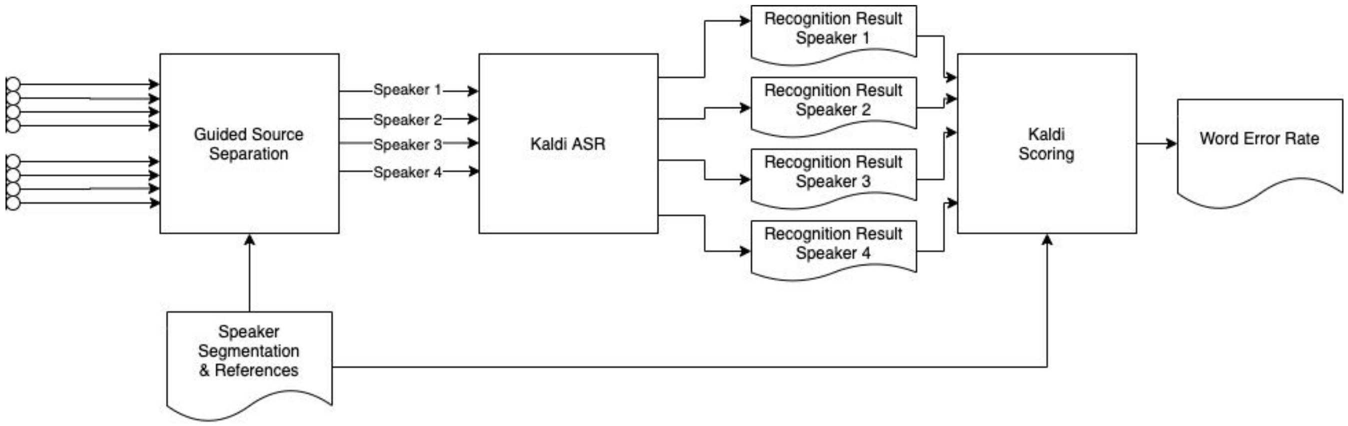
- Integrates speech activity detection (SAD), speaker embedding, and speaker diarization modules in addition to the track 1 system
- All-in-one recipe including training and inference
- This is the first baseline that integrates all multi speaker speech processing in this real scenario

Many thanks to

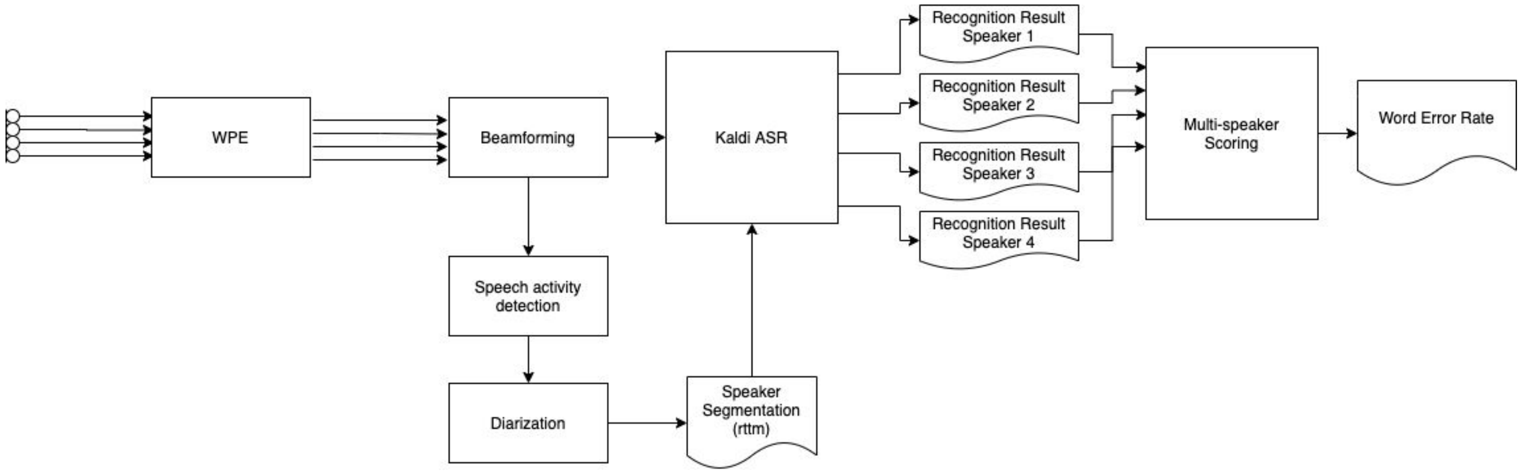
Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, David Snyder, Aswin Shanmugam Subramanian, Jan Trmal, Bar Ben Yair, Christoph Boeddeker, Zhaoheng Ni, Yusuke Fujita, Shota Horiguchi, Naoyuki Kanda, Takuya Yoshioka, Neville Ryant

System overview

Track 1



Track 2



All of them are implemented within a Kaldi recipe

Track 1: Speech enhancement

We used the following open source implementations

- Dereverberation:
  - Nara-WPE: different implementations of "Weighted Prediction Error" for speech dereverberation
- Beamforming
  - Guided Source Separation (GSS) for multiple arrays
    - Uses the context speaker information to extract the target speech from a mixture
    - Reduces the computational cost while keeping the performance (outer mics, reduce #iterations, etc.)
  - BeamformIt
    - We still keep this enhancement option to perform simple weighted delay-and-sum beamforming to the reference array

Track 1: Speech recognition

Kaldi speech recognition toolkit

- Acoustic model: trained with Kinect and worn microphones and augmented data (CHiME noises and simulated RIRs)
  - GMM → TDNN-F
- Language model: 3-gram LM trained with the CHiME-6 transcriptions
- Data cleaning
- Chain model (lattice-free MMI training)
  - Factorized time delay neural network (TDNN-F)
  - I-vector
- Two stage decoding (refine i-vector in the first pass decoding)

Track 1: Baseline performance

	Dev. WER	Eval. WER
CHiME-5 baseline	81.1%	73.3%
CHiME-5 top system (USTC-iFlytek)	45.6%	46.6%
CHiME-6 baseline	**51.8%**	**51.3%**

- Approaching the CHiME-5 top performance with a simple system!

Track 2: Speech enhancement

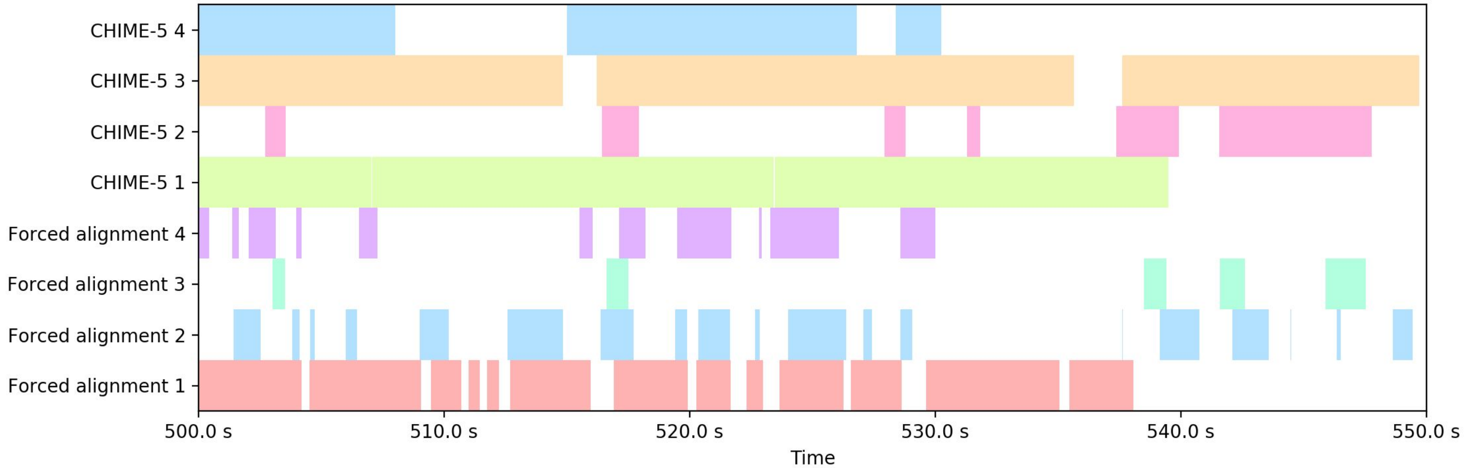
We used the following open source implementations

- Dereverberation:
  - Nara-WPE: different implementations of "Weighted Prediction Error" for speech dereverberation
- Beamforming
  - BeamformIt
    - We still keep this enhancement option to perform simple weighted delay-and-sum beamforming to the reference array

- Note that we did not include GSS due to the risk of degradation in GSS performance using estimated diarization information

Track 2: Speaker segmentation (RTTM) refinement

- CHiME-6 utterance boundaries sometimes included long pauses within a sentence, bad for diarization
- We created a new reference RTTM by force aligning the transcripts with the binaural recordings using the HMM-GMM system
- Utterances were then sequences of words separated by less than 300ms silence or noise
- Raised at the CHiME challenge forum



Track 2: Speech activity detection

Kaldi speech recognition toolkit

- Generate speech activity labels from the HMM-GMM system
- 5-layer TDNN with statistics pooling
- **Only use the U06 array** for simplicity

	Dev			Eval.		
	Missed speech	False alarm	Total error	Missed speech	False alarm	Total error
Original RTTM	2.5%	0.8%	3.3%	4.1%	1.8%	5.9%
Force-aligned RTTM	1.9%	0.7%	2.6%	4.3%	1.5%	5.8%

Track 2: Speaker diarization

Kaldi speech recognition toolkit

- **x-vector** neural diarization model is trained with **VoxCeleb**
- Probabilistic linear discriminant analysis (**PLDA**) model [44] is trained on **CHiME-6**
- Agglomerative hierarchical clustering (**AHC**) is performed
- The number of speakers is given (=4)

	Dev.		Eval.	
	DER	JER	DER	JER
Original RTTM	61.6%	69.8%	62.0%	71.4%
Force-aligned RTTM	63.4%	70.8%	68.2%	72.5%

Track 2: Speech recognition

- Same as track 1

Track 2: Evaluation metrics

Speaker diarization

- diarization error rate (DER)
- Jaccard error rate (JER)
- Both are computed by using *dscore* (official DIHARD scoring tool)

Speech recognition

- Concatenated minimum-permutation word error rate (**cpWER**).
  a. *Concatenate* all utterances of each speaker for both reference and hypothesis files.
  b. Compute the WER between the reference and *all possible speaker permutations* of the hypothesis.
  c. Pick the lowest WER among them
- cpWER includes the diarization error and we used it as an official metric for our ranking

Track 2: Baseline performance

- CHiME-6 Track 1 and 2 baseline ASR results with BeamformIt-based and GSS-based speech enhancement.
- We used the same acoustic and language models for both tracks.

	Enhancement	Segmentation	Dev. WER	Eval. WER
Track 1	GSS	Oracle	51.8%	51.3%
Track 2	BeamformIt	Diarization	84.3% (cpWER)	77.9% (cpWER)

- Significant degradation due to the diarization errors (Challenge!!!)

Overview

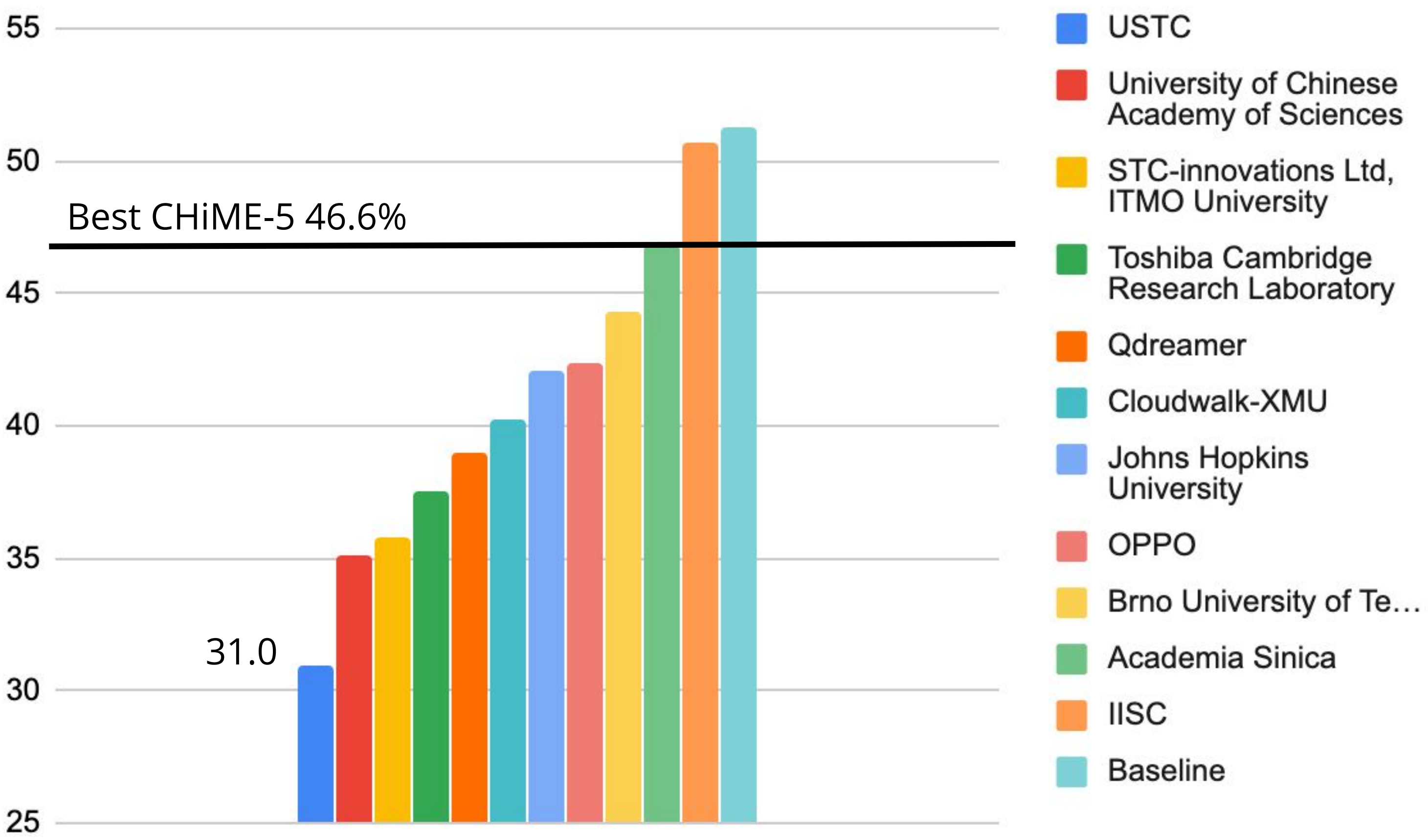
- Background - From CHiME-1 to CHiME-6
- CHiME-6 data and task
- CHiME-6 baseline systems
- CHiME-6 submissions and results

Submission statistics

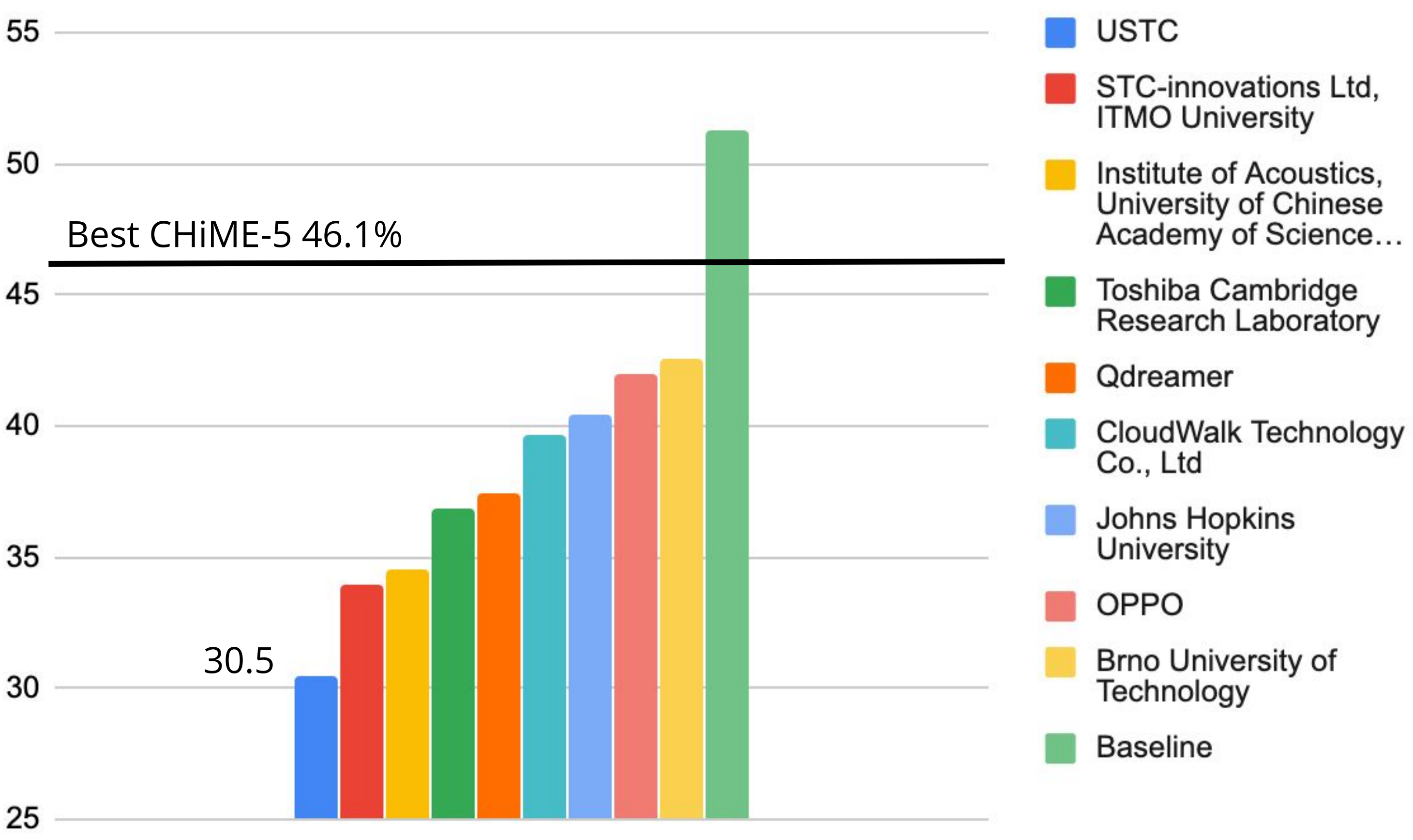
- In total, **34 submissions** by **13 papers**
  - CHiME-5: 35 submissions by 20 papers
  - Track 1-A: 11, Track 1-B: 9, Track 2-A: 9, Track 2-B: 5
- In total, 111 authors, 8.5 authors per paper
- Academia 10 vs. Industry 6
- Asia 9, Europe 4, North America 2

- We have several new participants (Welcome!)

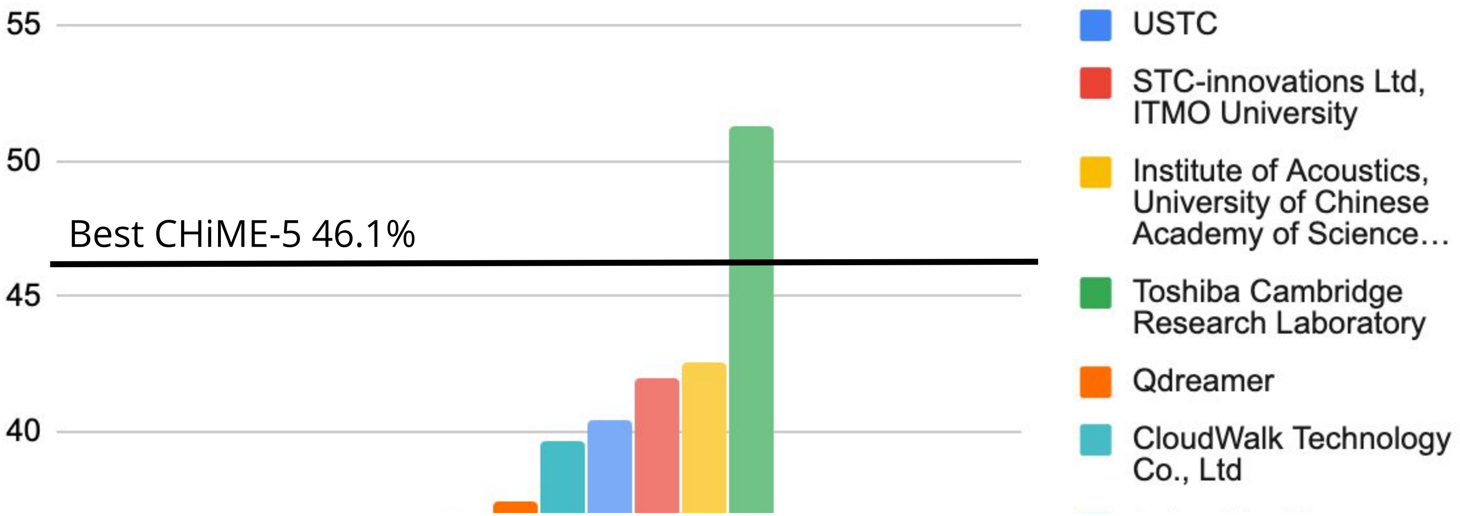
Results: Track 1-A WER

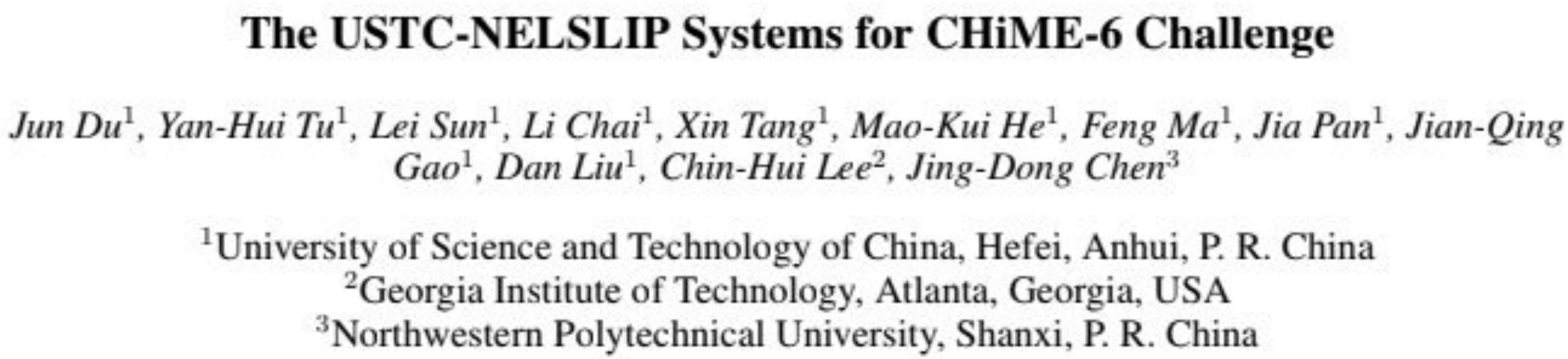


Results: Track 1-B WER

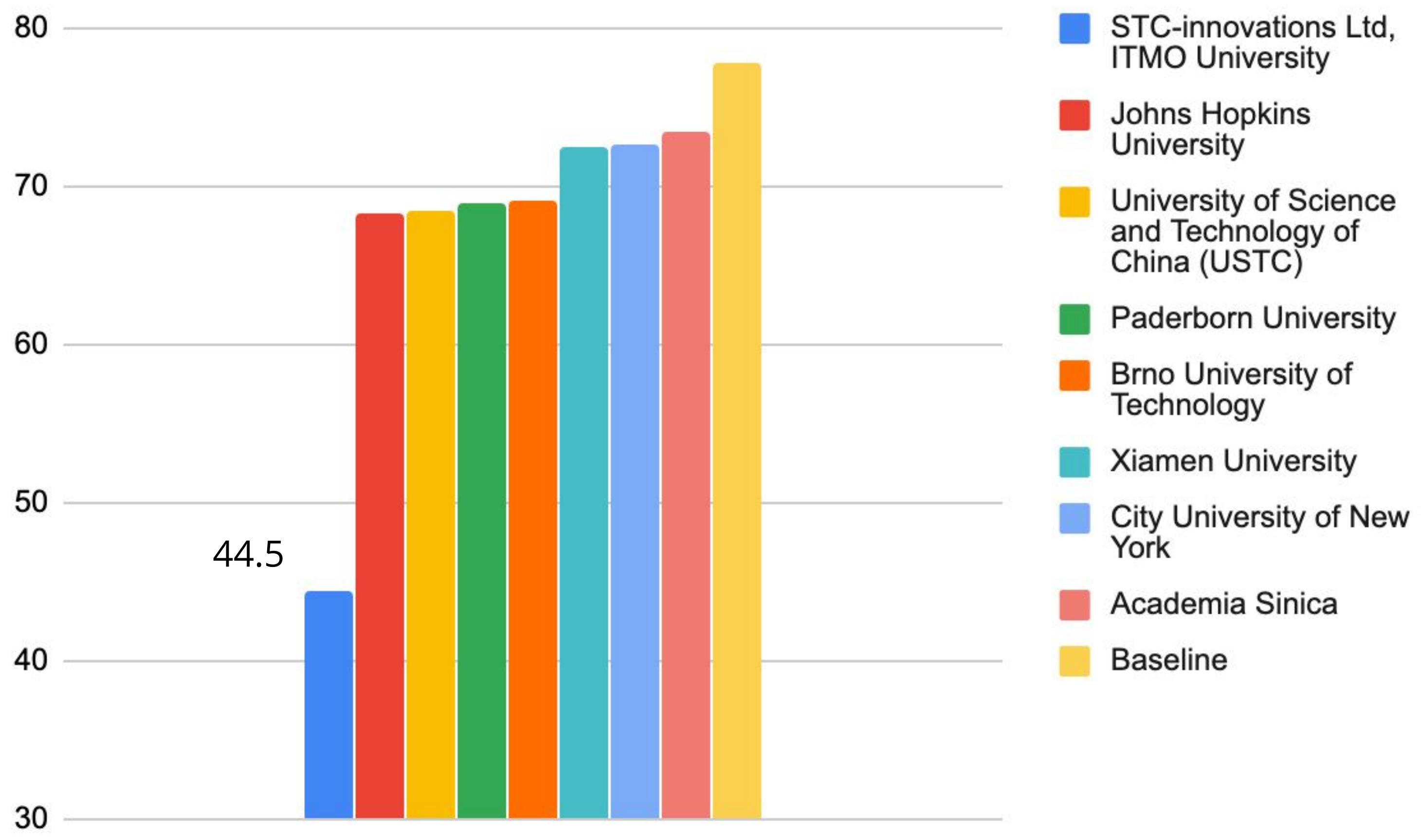


Results: Track 1-B WER

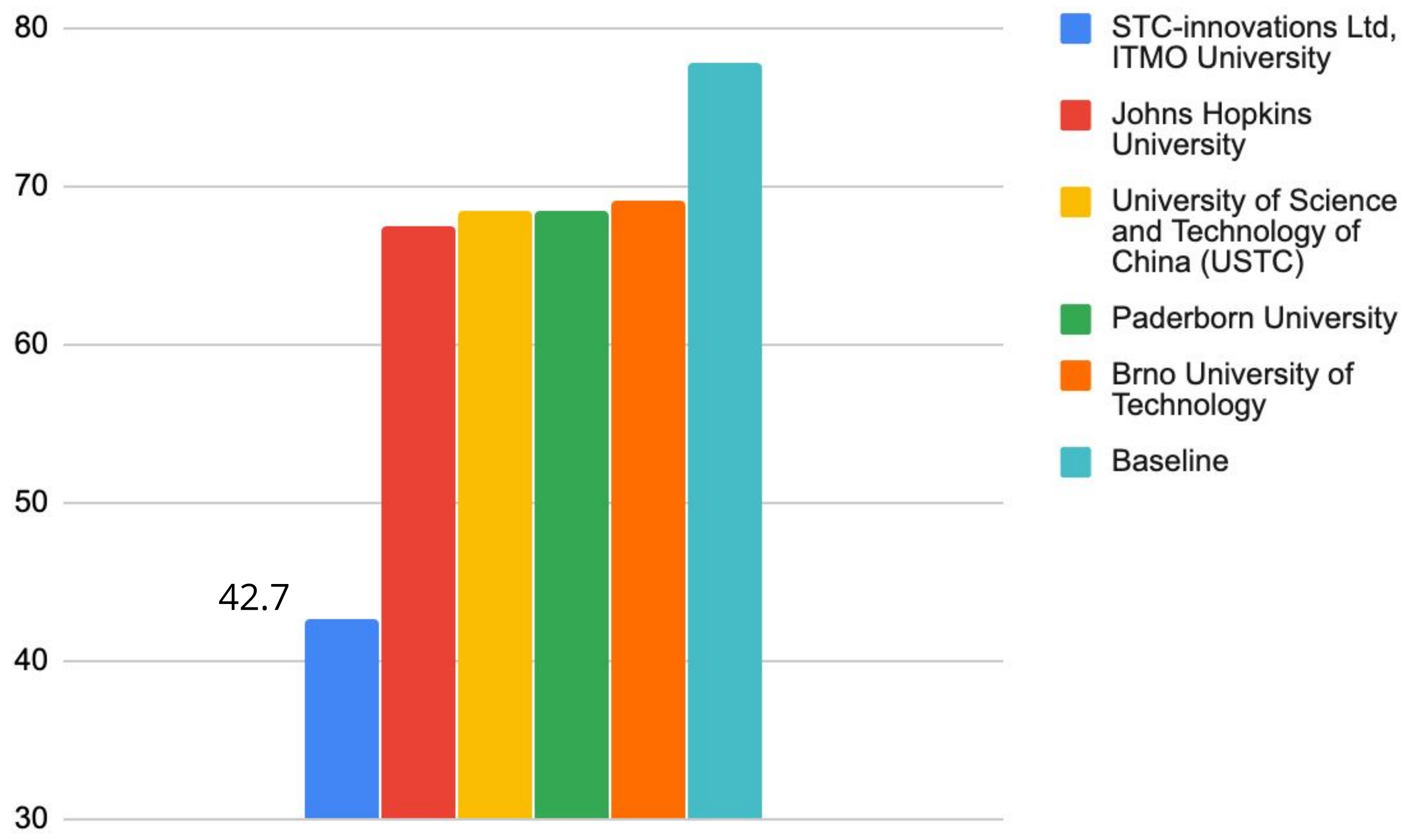




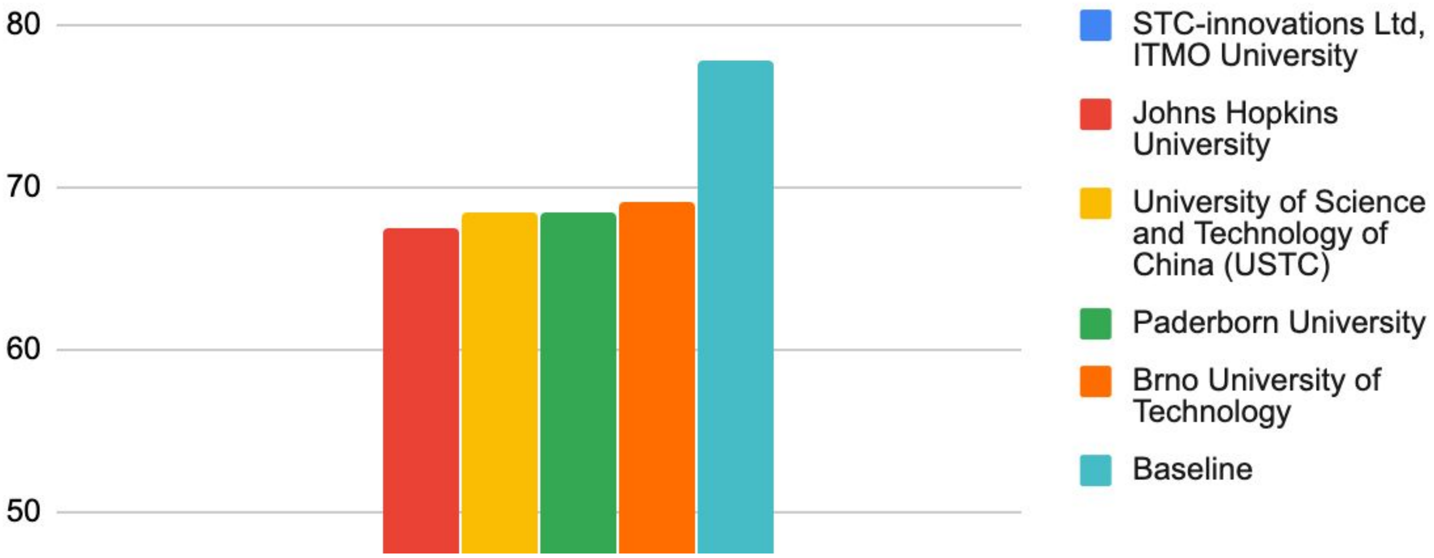
Results: Track 2-A cpWER



Results: Track 2-B cpWER



Results: Track 2-B cpWER



The STC System for the CHiME-6 Challenge

Ivan Medennikov[1,2], *Maxim Korenevsky*[1], *Tatiana Prisyach*[1], *Yuri Khokhlov*[1], *Mariya Korenevskaya*[1], *Ivan Sorokin*[1], *Tatiana Timofeeva*[1], *Anton Mitrofanov*[1], *Andrei Andrusenko*[1,2], *Ivan Podluzhny*[1], *Aleksandr Laptev*[1,2], *Aleksei Romanenko*[1,2]

[1]STC-innovations Ltd, [2]ITMO University, Saint Petersburg, Russia

Technology summary

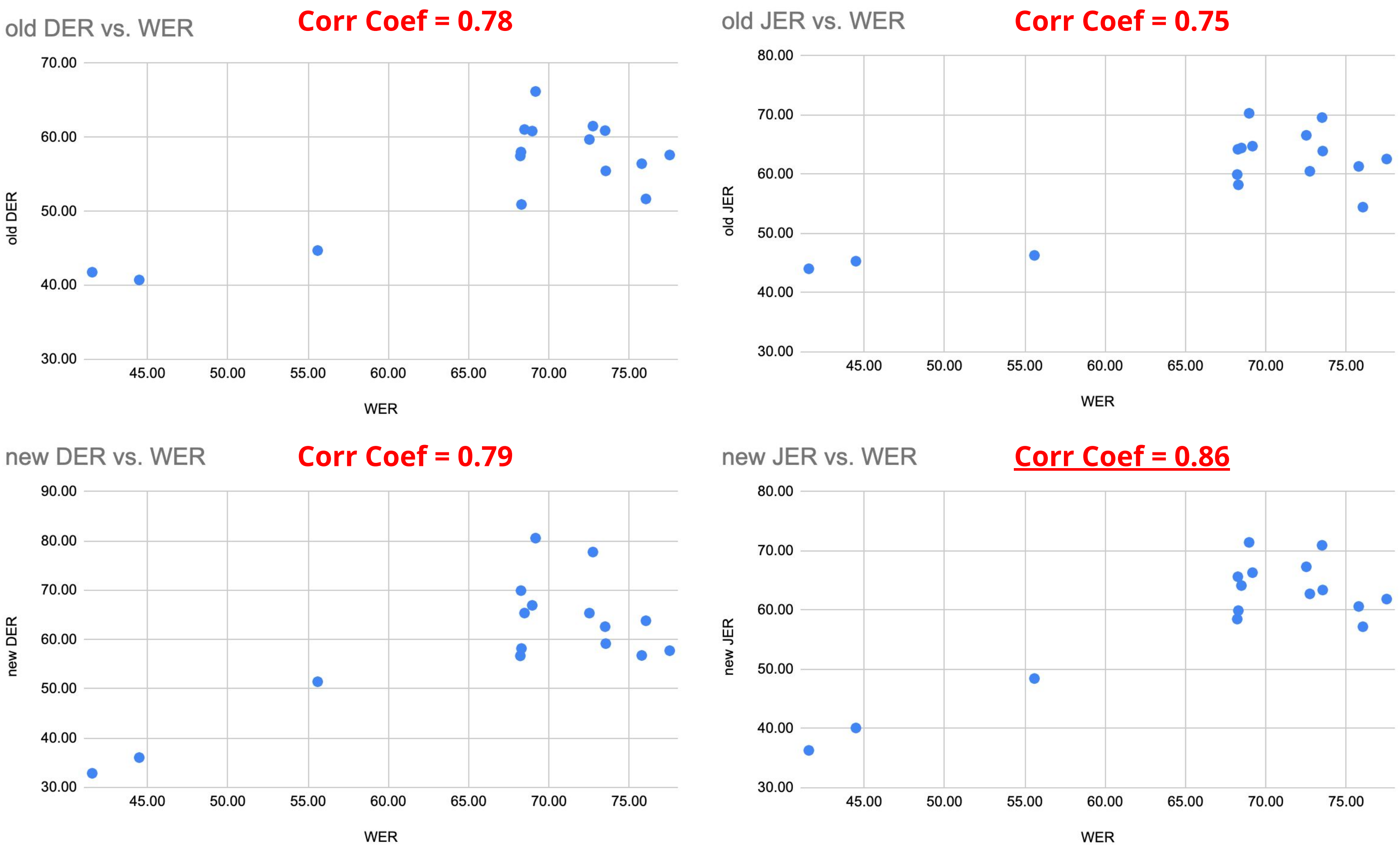
Track 1

- Speech enhancement: Guided source separation, speech separation, iterative method, full use of multiple arrays
- Data augmentation: mixing the enhanced signals
- Acoustic model: combine CNN

Track 2

- Target-Speaker Voice Activity Detection (TS-VAD) largely solve the overlap problems in speaker diarization!

ASR vs. diarization metrics



Conclusion

Conclusion

Take home message from Track 1:

- Top system ~30%
- Most systems outperformed the best 2018 system (reproducible)

Steadily improve the performance in this really challenging environments

Conclusion

Take home message from Track 2:

- Totally 8 research groups could build their systems
- The top system is filling out the gap comes from the oracle segmentation

We established a method to tackle multispeaker unsegmented recordings!

Thanks a lot!

We are publishing our baseline efforts in arXiv (https://arxiv.org/abs/2004.09249) and this workshop proceedings

CHiME-6 Challenge:
Tackling Multispeaker Speech Recognition for Unsegmented Recordings

[1]*Shinji Watanabe*, [2]*Michael Mandel*, [3]*Jon Barker*, [4]*Emmanuel Vincent*
[1]*Ashish Arora*, [1]*Xuankai Chang*, [1]*Sanjeev Khudanpur*, [1]*Vimal Manohar*, [1]*Daniel Povey*, [1]*Desh Raj*,
[1]*David Snyder*, [1]*Aswin Shanmugam Subramanian*, [1]*Jan Trmal*, [1]*Bar Ben Yair*, [5]*Christoph Boeddeker*,
[2]*Zhaoheng Ni*, [6]*Yusuke Fujita*, [6]*Shota Horiguchi*, [7]*Naoyuki Kanda*, [7]*Takuya Yoshioka*, [8]*Neville Ryant*

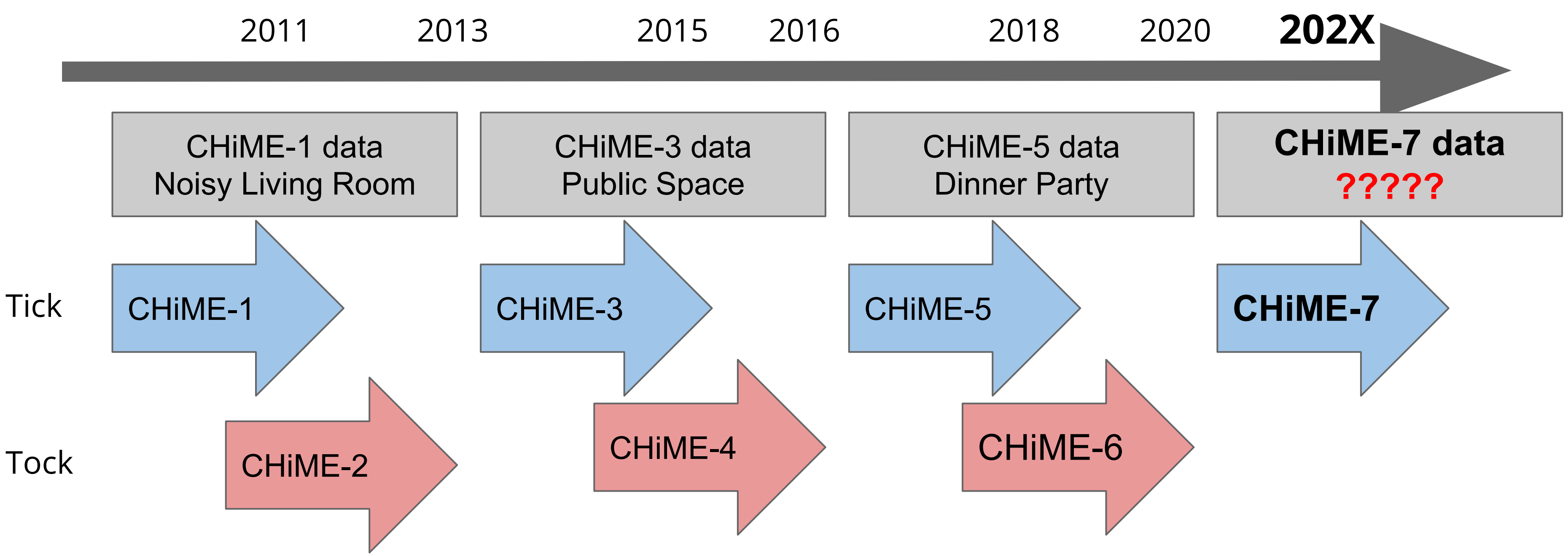
[1]Johns Hopkins University, USA, [2]The City University of New York, USA, [3]University of Sheffield, UK, [4]Inria, France, [5]Paderborn University, Germany, [6]Hitachi, Ltd., Japan, [7]Microsoft, USA, [8]Linguistic Data Consortium, USA

We welcome your input!

- Other (more challenging or less challenging but more organized) scenarios?
- More data?
- Multimodal (audio and video)?
- Multilingual?
- Dynamic environment (e.g., wearable or robot)?
- Simplified systems?
- Online vs. offline?
- Other tasks (keyword search)?

Questionnaires based on Google form
URL: https://forms.gle/jgXaxFEcqSN7dQau7

Conclusion



Now we should move on to the next stage

- Other (more challenging) scenarios?
- More data or more techniques to further establish the scenario?
- Multi-modal?
- Dynamic environment (e.g., robot)
- Online or off line?