The CHiME-8 DASR Task

Generalizable and Array Agnostic Distant Automatic Speech Recognition and Diarization

Samuele Cornell[1] Taejin Park[2] He Huang[2] Christoph Boeddeker[3]
Matthew Wiesner[4] Matthew Maciejewski[4] Xuankai Chang[1]
Paola Garcia[4] Shinji Watanabe[1]

[1]Carnegie Mellon University, USA [2]NVIDIA, USA
[3]Paderborn University, Germany [4]Johns Hopkins University, USA

CHiME CHALLENGE

The CHiME-8 DASR Task

Generalizable and Array Agnostic Distant Automatic Speech Recognition and Diarization

Samuele Cornell[1] Taejin Park[2] He Huang[2] Christoph Boeddeker[3]
Matthew Wiesner[4] Matthew Maciejewski[4] Xuankai Chang[1]
Paola Garcia[4] Shinji Watanabe[1]

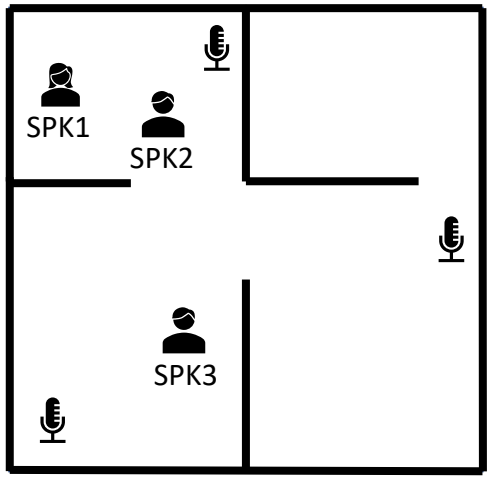
[1]Carnegie Mellon University, USA [2]NVIDIA, USA
[3]Paderborn University, Germany [4]Johns Hopkins University, USA

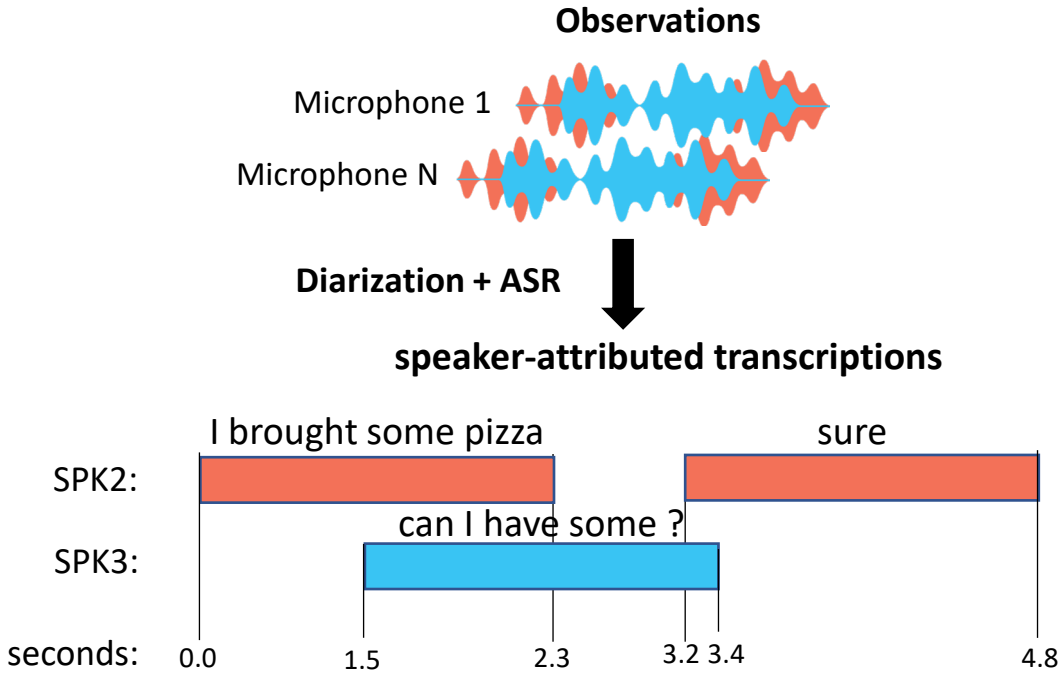
CHiME CHALLENGE

Task Objective

Same end goal as in the past CHiME-6 and CHiME-7 DASR:

- **joint diarization and transcription** of an **unsegmented meeting** scenario

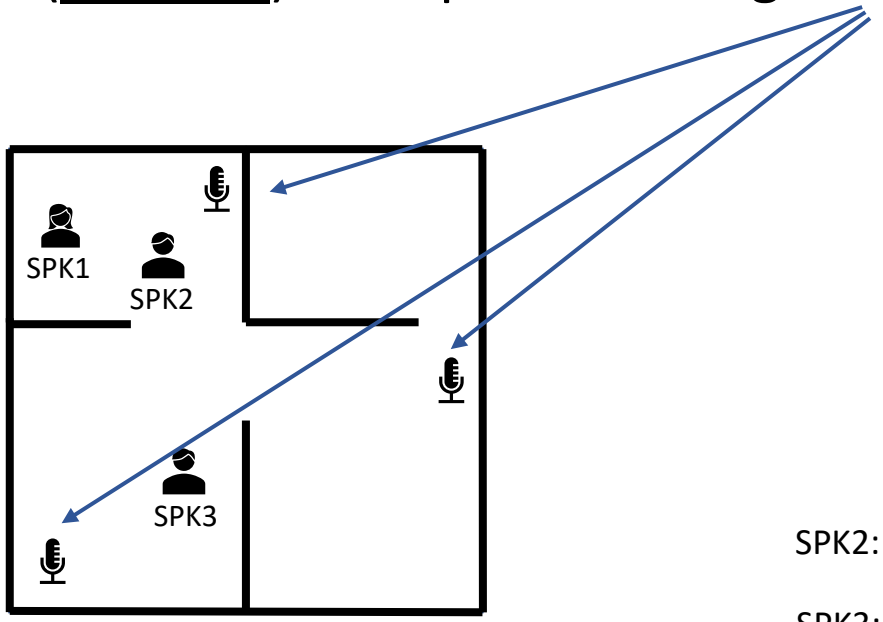


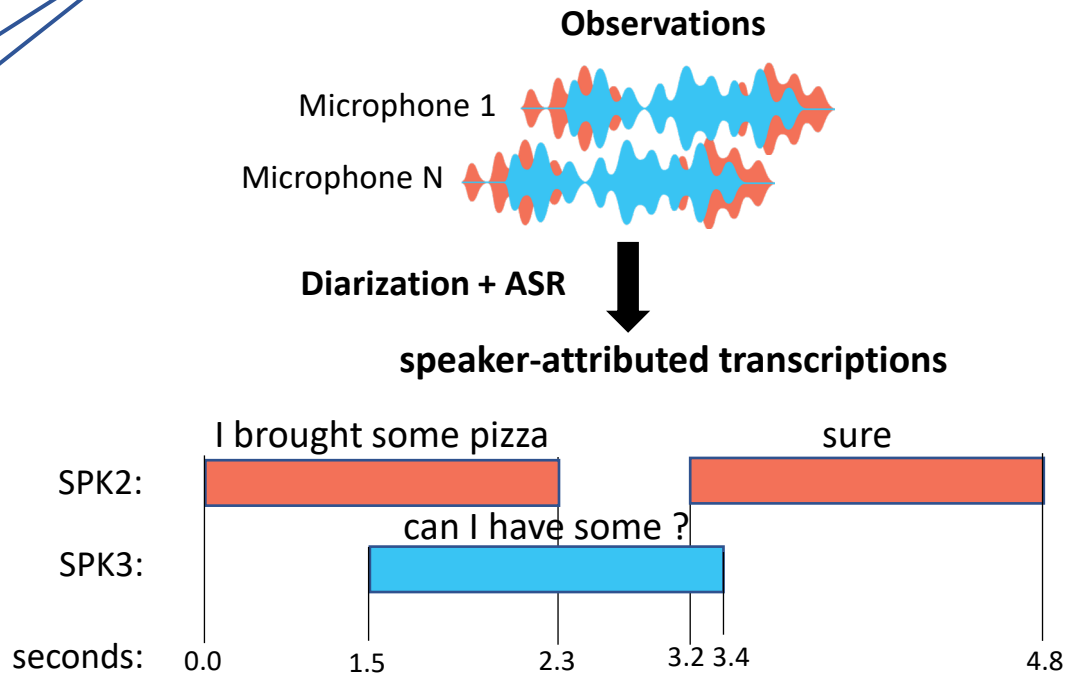


Task Objective

Same end goal as in the past CHiME-6 and CHiME-7 DASR:

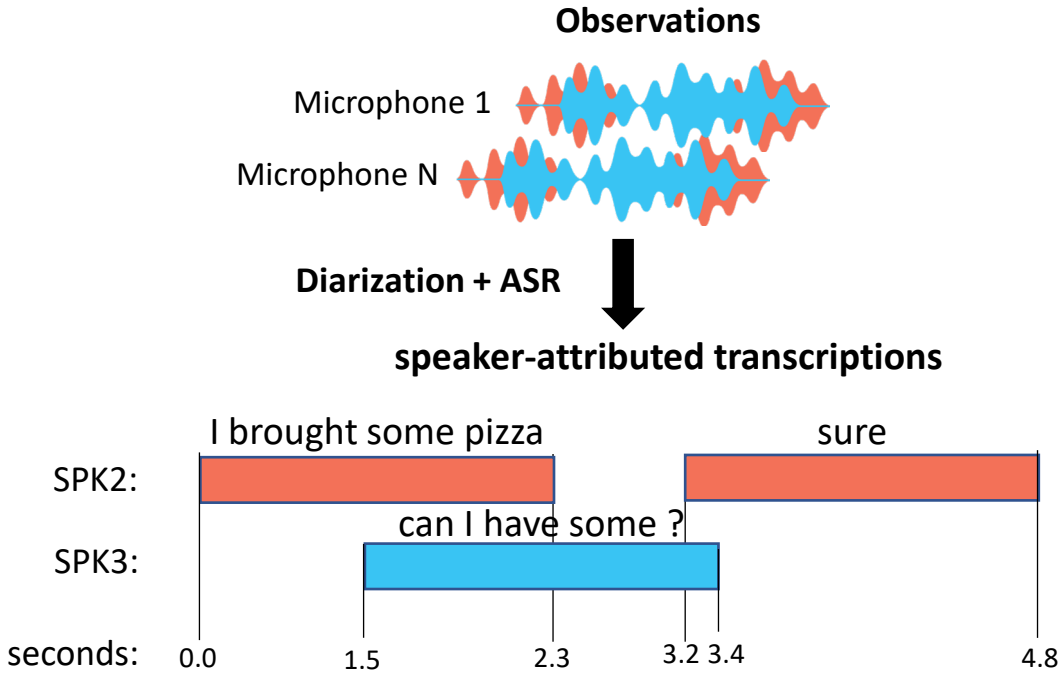
- **joint diarization and transcription** of an **unsegmented meeting** scenario with (possibly) multiple recording devices.





Task Objective

Participants have to produce transcriptions for each speaker with utterance-level segmentation.



Task Objective

Participants have to produce transcriptions for each speaker with utterance-level segmentation.

- The predictions are submitted in the form of a JSON file as depicted here with:
  - start and end time of each utterance.
  - speaker label
  - words uttered
  - Session/meeting id

Hypothesis


```
{
  "end_time": "11.350",
  "start_time": "11.010",
  "words": "so",
  "speaker": "spk1",
 "session_id": "S05" },
{
  "end_time": "14.150",
  "start_time": "12.000",
  "words": "Where is",
  "speaker": "spk2",
  "session_id": "S05"
}
```

Main Goal: Generalization

Foster research towards **robust ASR+diarization, that can generalize to:**

1. arbitrary number of speakers
2. diverse settings (e.g. more formal vs informal style conversation)
3. wide-variety of acoustic scenarios

Main Goal: Generalization

Foster research towards **robust ASR+diarization, that can generalize to:**

1. arbitrary number of speakers
2. diverse settings (e.g. more formal vs informal style conversation)
3. wide-variety of acoustic scenarios
4. different recording devices configurations (incl. ad-hoc array networks and multi-room environments)

Circular array device topology as used in DiPCo

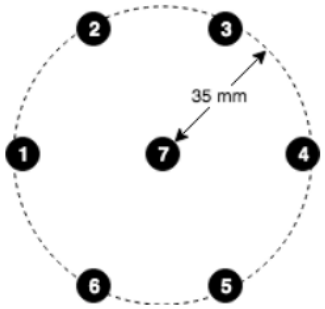
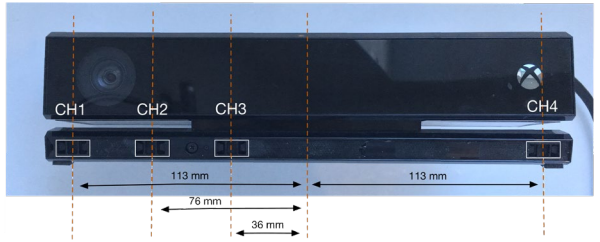
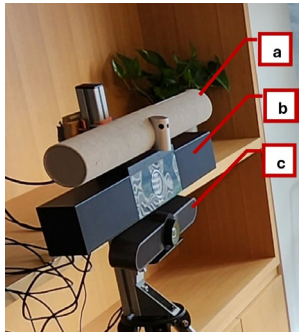


Figure 2: Configuration of the 7-microphone array.



Kinect array device as used in CHiME-6

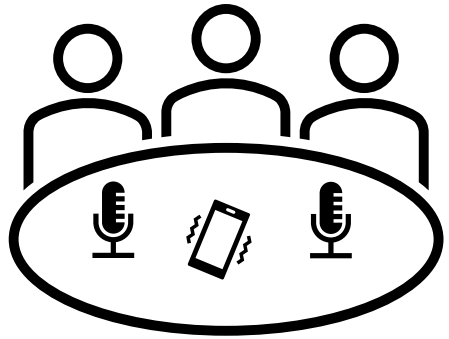


NOTSOFAR1 recording devices

Main Goal: Generalization

Foster research towards **robust ASR+diarization, that can generalize to:**

1. different recording devices configurations (incl. ad-hoc array networks and multi-room environments)
   - Highly interesting and practical problem
     - No need for proprietary devices (more scalable and widely applicable)
       - Multi-device meeting transcription



Main Goal: Generalization

Fills a gap in **current challenges/evaluation benchmarks for meeting transcription**, which **mostly focus on one domain**:

1. AMI, ICSI
2. CHiME-5 & 6, DiPCo
3. Alimeeting
4. In-Car Multi-Channel Automatic Speech Recognition (ICMC-ASR) Challenge
5. Ego4D
6. CHiME-8 NOTSOFAR-1 (Task 2, next presentation)

Main Goal: Generalization

Fills a gap in **current challenges/evaluation benchmarks for meeting transcription**, which **mostly focus on one domain**:

1. AMI, ICSI
2. CHiME-5 & 6, DiPCo
3. Alimeeting
4. In-Car Multi-Channel Automatic Speech Recognition (ICMC-ASR) Challenge
5. Ego4D
6. CHiME-8 NOTSOFAR-1 (Task 2, next presentation)

Similar efforts were done for diarization (e.g. DIHARD, VoxConverse challenges) and non long-form ASR (🤗 Speech Robust Bench)

Datasets

- Two “types” of datasets
  - **4 “Core” datasets (train, dev and eval):**
    - CHiME-6
    - DiPCo
    - Mixer 6 Speech

Datasets

- Two “types” of datasets
  - **4 “Core” datasets (train, dev and eval):**
    - CHiME-6
    - DiPCo
    - Mixer 6 Speech
    - NOTSOFAR1 NEW

DASR and NOTSOFAR1 Tasks

We have a scenario (**NOTSOFAR1 dataset**) in common and agreed together to have **same text normalization and same rules**.

- Every submission to DASR also accounts for a valid submission to the NOTSOFAR1 task.
  - It is one of the four scenarios in CHiME-8 DASR
- Shared scientific goal is to **compare design choices and performance between**:
  - domain specialized systems (NOTSOFAR1 task)
  - generalist systems (DASR task)

DASR and NOTSOFAR1 Tasks

We have a scenario (**NOTSOFAR1 dataset**) in common and agreed together to have **same text normalization and same rules**.

- Every submission to DASR also accounts for a valid submission to the NOTSOFAR1 task.
  - It is one of the four scenarios in CHiME-8 DASR
- Shared scientific goal is to **compare design choices and performance between**:
  - domain specialized systems (NOTSOFAR1 task)
  - generalist systems (DASR task)

Answer might not be obvious, domain-agnostic approaches:

- could generalize better to evaluation set (less biased to the training data)
- their design could allow to leverage more diverse training data
  - E.g. array-agnostic front-end for diarization or separation (e.g. FasNet-TAC, multi-channel EEND-EDA)

Datasets

- Two “types” of datasets
  - **4 “Core” datasets (train, dev and eval):**

Scenario	Train (hh:mm)	Dev (hh:mm)
CHiME-6	40:05	4:27
DiPCo	1:12	1:31
Mixer 6	6:13 (~63 annotated only for one speaker)	8:56
NOTSOFAR-1	14:43	13:25

Datasets

- Two “types” of datasets
  - **4 “Core” datasets (train, dev and eval):**

Scenario	Train (hh:mm)	Dev (hh:mm)
CHiME-6	40:05	4:27
DiPCo	1:12	1:31
Mixer 6	6:13 (~63 annotated only for one speaker)	8:56
NOTSOFAR-1	14:43	13:25

  - **External datasets that participants can use for training and validation**

Datasets

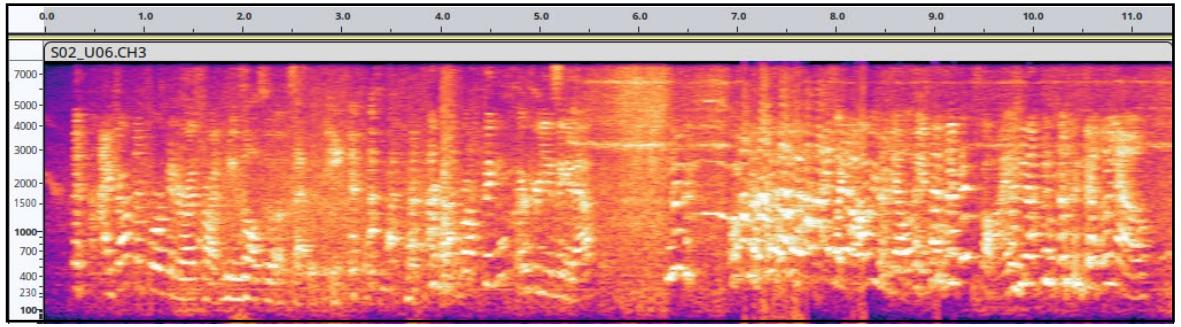
- **External datasets that participants can use for training and validation**
  - Full list available at https://www.chimechallenge.org/current/task1/rules
    - Real meetings: AMI
    - Clean speech datasets: LibriSpeech, WSJ
    - Noise datasets: FSD50k, SINS
    - Speaker verification: VoxCeleb1&2
    - Room impulse responses (RIR): SLR28, MUSAN
    - Synthetic datasets: NOTSOFAR-1 simulated dataset, WHAMR
  - Participants could propose new ones up to 20 March 2024

Core Datasets: Diverse Scenarios

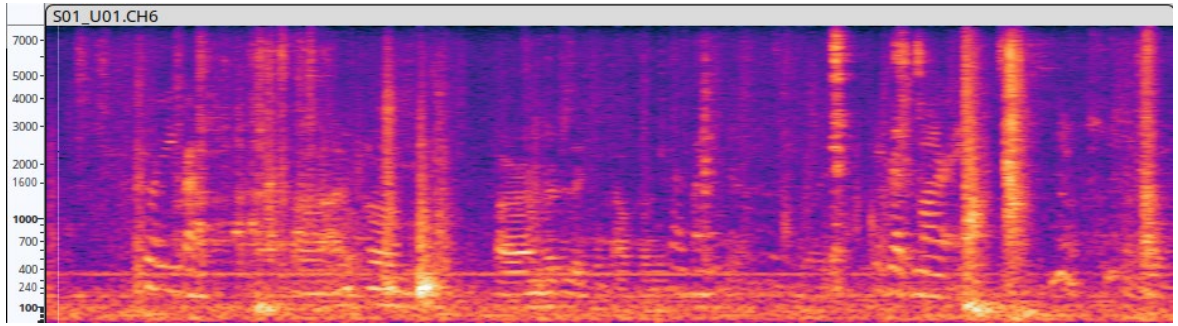
Scenario	Setting	Num. Speakers	Recording Setup	Multi-Room	Meeting Duration
CHiME-6	dinner party	4	6 linear arrays (4 mics each)	Yes	> 2h
DiPCo	dinner party (more formal)	4	5 circular arrays (7 mics each)	No	20-30 mins
Mixer 6 Speech	1-to-1 interview	2	10 heterogeneous devices	No	~15 mins
NOTSOFAR1	office meeting	4-8	1 circular array device (7 mics)	No	~6 mins

Core Datasets: Some Examples

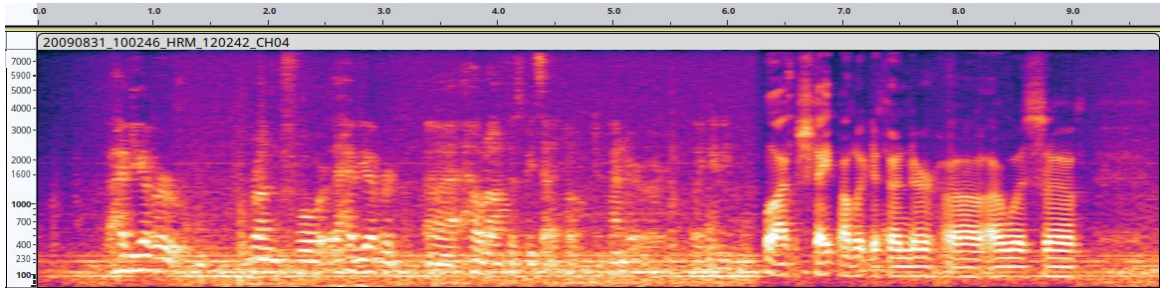
CHiME-6



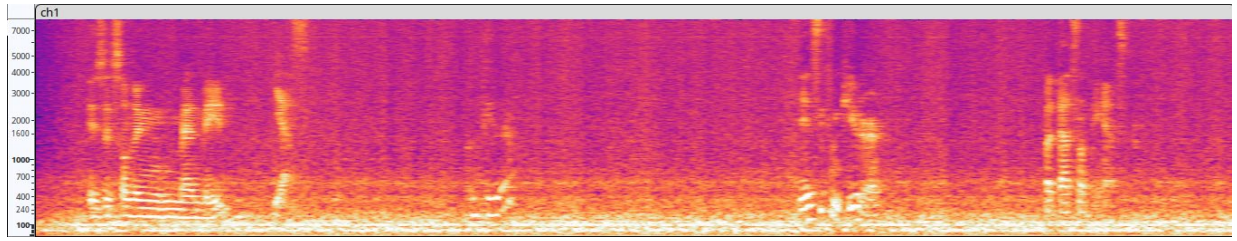
DiPCo



Mixer 6 Speech



NOTSOFAR-1



Figures: Audacity log mel-scaled spectrograms (2048 window size)

Task Rules

Rationale: enforce participants to create just one system for all core scenarios

- **Domain identification is prohibited** (one system must tackle all three core scenarios)
  - Participants could not make any assumption about the microphone configuration used.
    - This would account for domain identification.
  - Systems must estimate the number of speakers automatically (not based on domain)

Task Rules

Rationale: enforce participants to create just one system for all core scenarios

- **Domain identification is prohibited** (one system must tackle all three core scenarios)
  - Participants could not make any assumption about the microphone configuration used.
    - This would account for domain identification.
  - Systems must estimate the number of speakers automatically (not based on domain)

Participants can use a plethora of external pretrained models including:

- Large-scale weakly supervised (e.g. Whisper or OWSM) ASR models
- Large-scale self-supervised models (e.g. WavLM, HuBERT etc).
- Speaker ID Embeddings models (e.g. ECAPA-TDNN)

Full detailed rules, including allowed pre-trained models available at
https://www.chimechallenge.org/current/task1/rules

Challenge Tracks

We proposed two tracks.

In both, participants are tasked to perform ASR and diarization (utterance-level) on meetings from the various core datasets evaluation sets.

These tracks differ only by the allowed external models:

1. **Constrained LM track**
    1. Participants can use any resource among the training material (incl. external datasets) for LM training.
2. **Unconstrained LM track**
    1. Participants can ALSO use external large language models (LLMs) e.g. (Llama 2, OlMo, TinyLlama)

Ranking Metric

Systems are ranked according to **time-constrained minimum permutation word error rate (tcpWER) as proposed in MeetEval.**

- Evaluates both recognition accuracy, speaker attribution and segmentation

Ranking Metric

Systems are ranked according to **time-constrained minimum permutation word error rate (tcpWER) as proposed in MeetEval.**

- Evaluates both recognition accuracy, speaker attribution and segmentation
- Does not require forced-alignment
  - Uses character-based pseudo alignment
    - Word duration based on character count plus utterance boundaries (available)
  - Allows for a collar (we use 5 seconds)

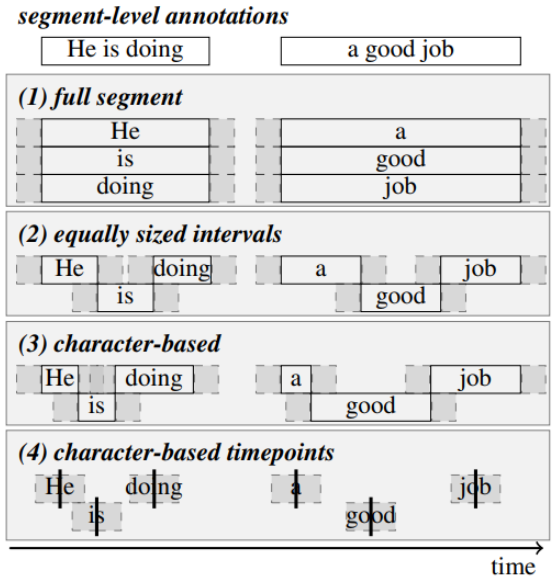


Figure 2: *Visualization of the different pseudo-word-level annotation strategies. The collar is visualized as gray boxes and kept short for better visualization. The character-based annotation strategy correlates best with the actual pronunciation time.*

Image from: von Neumann, Thilo, et al. "Meeteval: A toolkit for computation of word error rates for meeting transcription systems." *CHiME Workshop* 2023.

Ranking Metric

Since we care about domain generalization, we use the **tcpWER macro-average across all the 4 core scenarios as the final metric**

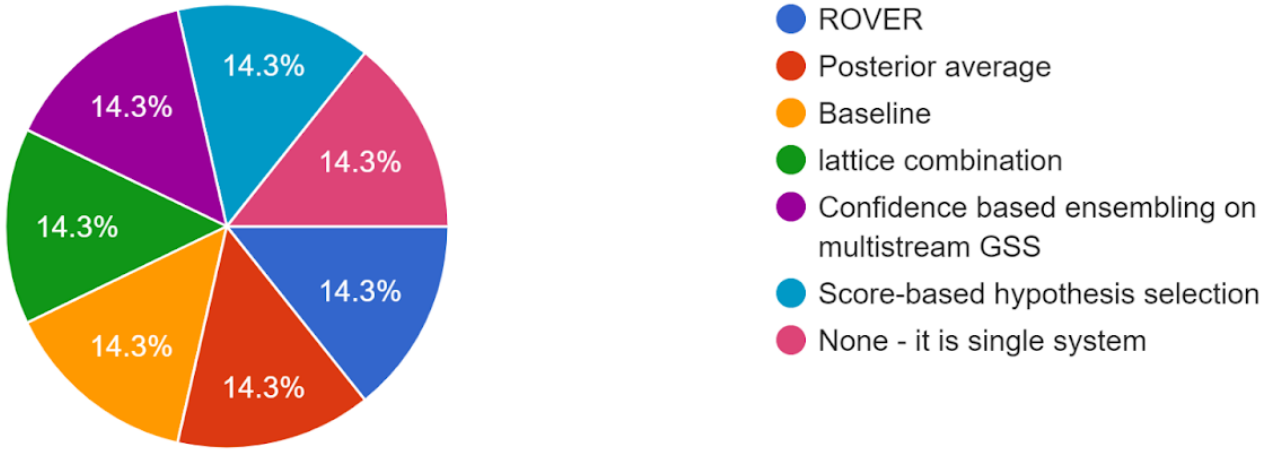
- Teams are ranked based on the best out of 3 submissions (on eval) for each track
  - 3 submissions to allow to explore different strategies including less computationally heavy ones.

Bonus

<u>Last year</u> most participants relied on ensemble methods to boost the performance:

Ensembling (multiple choice, add yours if you want).

7 responses



Bonus

<u>This year</u> we have a jury special award for the most efficient and innovative system.

Bonus

This year we have a jury special award for the most efficient and innovative system.

- The jury will be nominated by the CHiME Committee (so not most of us DASR organizers).
- Systems will be ranked using their description paper according to:
  1. Practicality/efficiency
  2. Innovation/originality
  3. Effectiveness

Bonus

Examples from past CHiME challenges include:

- Guided Source Separation (GSS) (CHiME-5) [1]
- Target speaker VAD (TS-VAD) (CHiME-6) [2]
- BLSTM supported GEV beamformer (CHiME-3) [3]

[1] Boeddeker, Christoph, et al. "Front-end processing for the CHiME-5 dinner party scenario." *CHiME5 Workshop* 2018.
[2] Medennikov, Ivan, et al. "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario." Interspeech. 2020
[3] Heymann, Jahn, et al. "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge." ASRU, 2015.

Baseline Systems

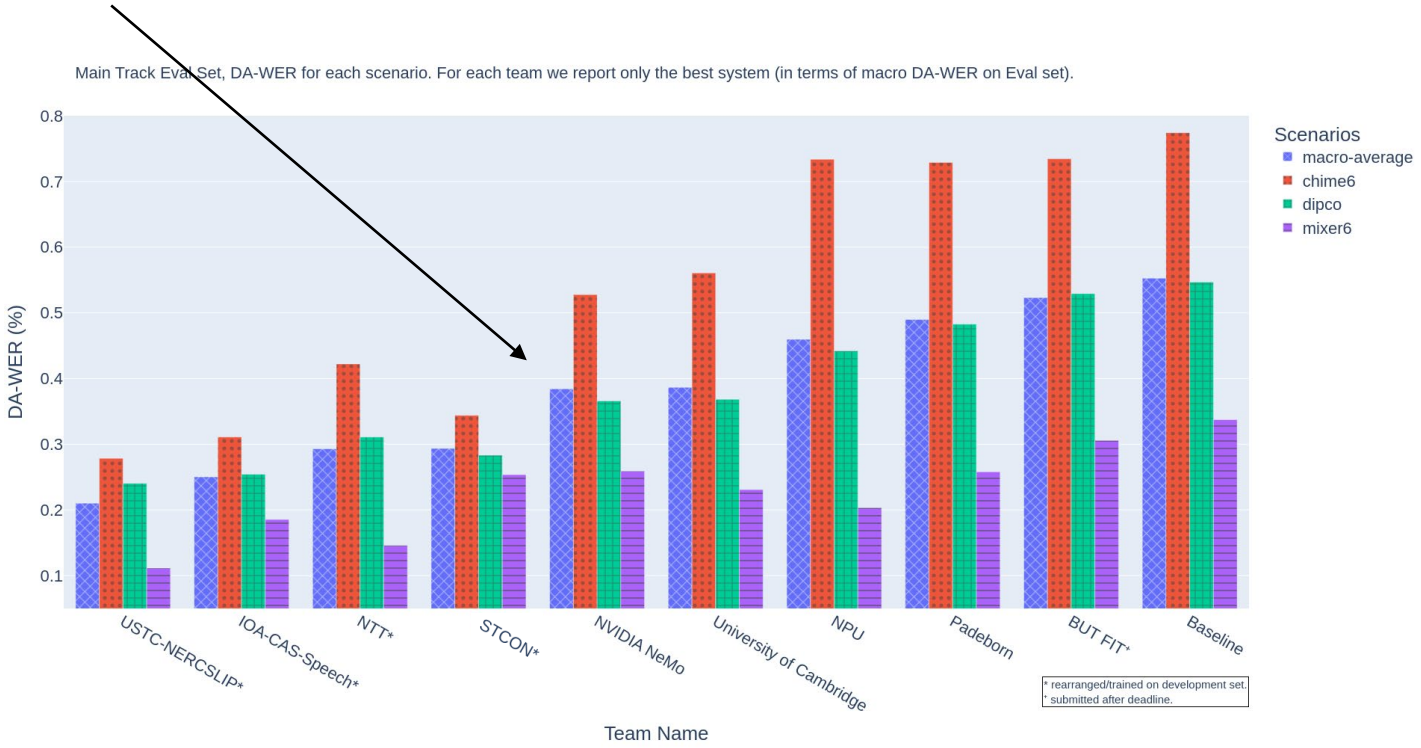
We have **two baseline systems**:

1. ESPNet: https://github.com/espnet/espnet/tree/master/egs2/chime8_task1
   - Updated last year baseline

Baseline Systems

We have **two baseline systems**:

1. ESPNet: https://github.com/espnet/espnet/tree/master/egs2/chime8_task1
   - Updated last year baseline
2. NVIDIA NeMo: https://github.com/chimechallenge/C8DASR-Baseline-NeMo
   - From NVIDIA NeMo team last year submission



Baseline Systems

We have **two baseline systems**:

1. ESPNet: https://github.com/espnet/espnet/tree/master/egs2/chime8_task1
   - Updated last year baseline
2. NVIDIA NeMo: https://github.com/chimechallenge/C8DASR-Baseline-NeMo
   - From NVIDIA NeMo team last year submission

In addition, USTC-NERCSLIP open sourced their extremely effective NSD-MS2S diarization system

- Key component that allowed to rank first in last year CHiME-7 DASR challenge.
- Available at https://github.com/liyunlongaaa/NSD-MS2S

Data Preparation & Download

We also provide a ***chime-utils*** toolkit to allow for **easy data preparation and downloading** as well as scoring:

https://github.com/chimechallenge/chime-utils

- Hopefully its usefulness will extend beyond this challenge
  - Automatic download and convert CHiME-6, DiPCo and NOTSOFAR-1 to have same structure for easy parsing
- It also supports data preparation recipes for DASR for several toolkits including ESPNet, K2, Kaldi, NeMo

Baseline Systems

Both consists in an array topology agnostic meeting transcription pipeline consisting of:

- Multi-channel diarization
- Target speaker separation
  - Envelope variance based channel selection
  - Guided source separation (GSS)
- Monaural ASR

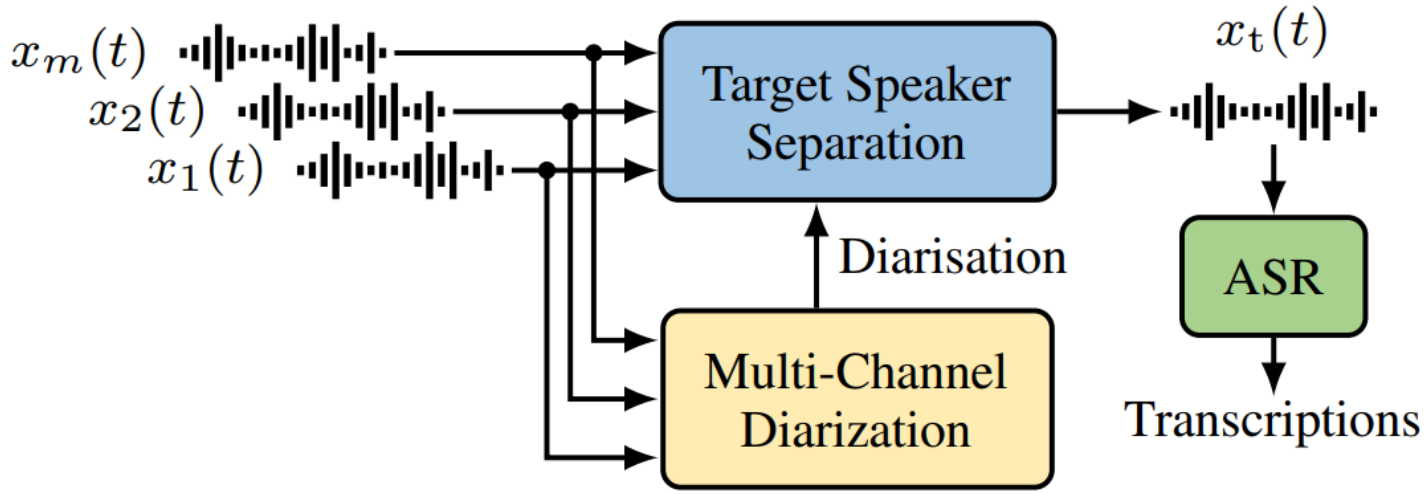


Figure 1: *ESPNet and NeMo baseline systems basic overview.*

Baseline Systems

Both consists in an array topology agnostic meeting transcription pipeline consisting of:

- **Multi-channel diarization**
- Target speaker separation
  - Envelope variance based channel selection
  - Guided source separation (GSS)
- Monaural ASR

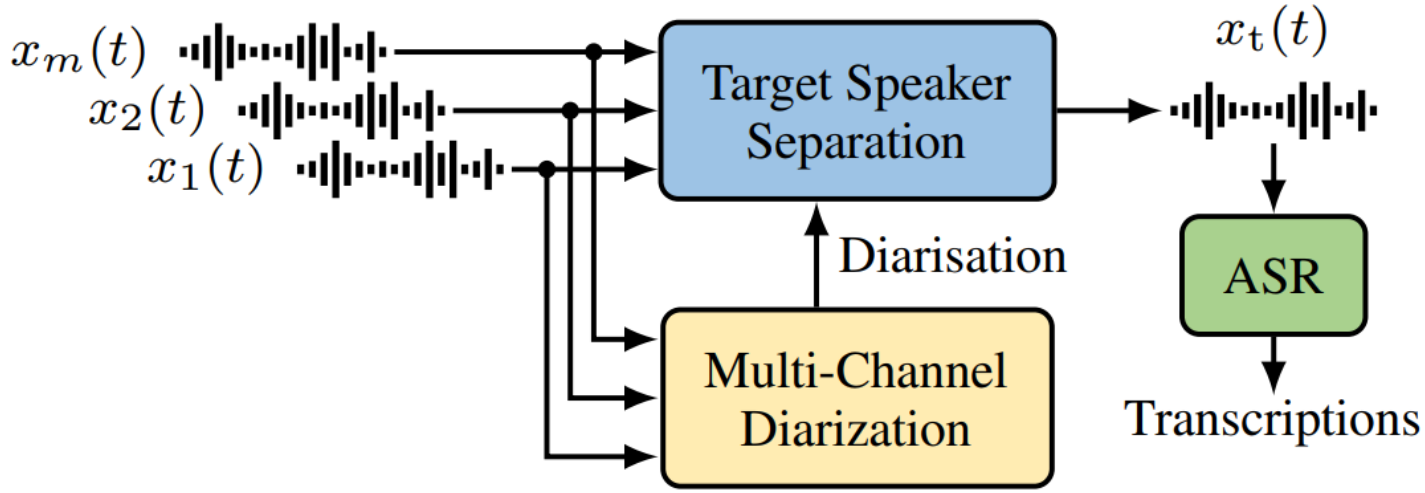
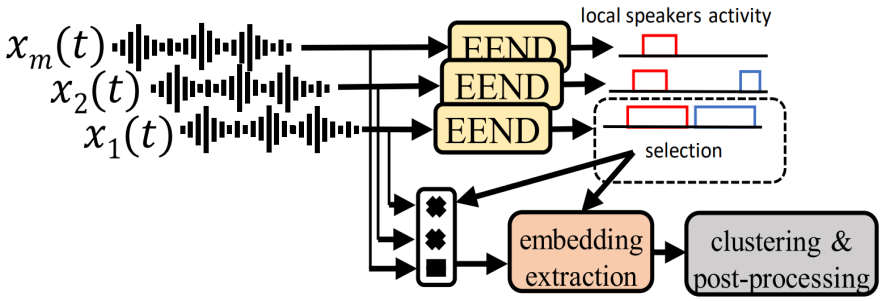


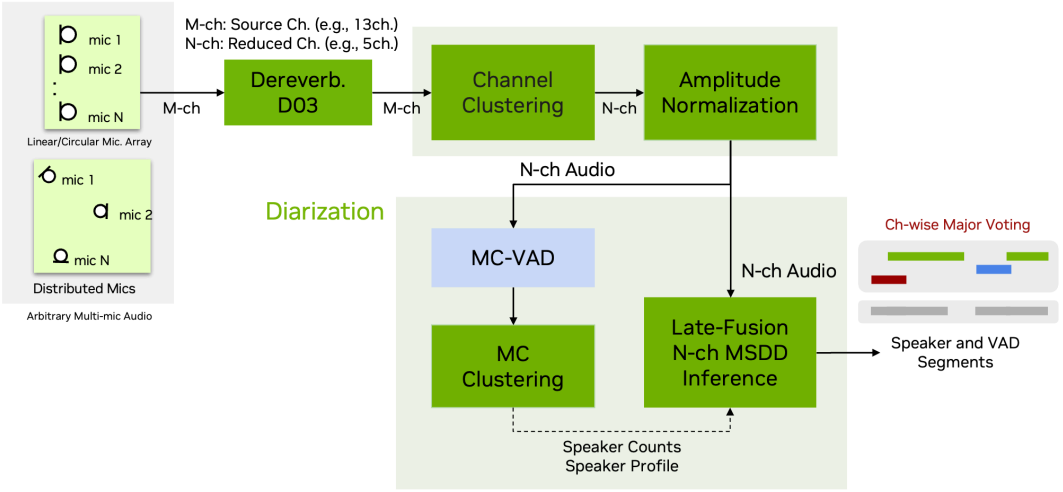
Figure 1: *ESPNet and NeMo baseline systems basic overview.*

Baseline Systems

Multi-channel diarization component

- ESPNet
  - Pyannote diarization pipeline extended to multiple channels
- NeMo
  - MIMO WPE Dereverberation
  - Microphone channel clustering
  - VAD & microphone channel ensembling
  - speaker embedding extraction and clustering
  - Multi scale diarization decoder (TS-VAD like model)





Challenge Participation

Much lower participation compared to last year CHiME-7 **(from 9 down to 3 teams)**

- Participants split between the three tasks this year which were highly related
  - Overall CHiME participation was up (+4 compared to CHiME-7)

Challenge Participation

Much lower participation compared to last year CHiME-7 **(from 9 down to 3 teams)**

- Participants split between the three tasks this year which were highly related
  - Overall CHiME participation was up (+4 compared to CHiME-7)

1. **Constrained LM track**
   - 3 submissions: STCON, NTT and a team which was anonymized due to poor performance
   - **Quality not quantity ?** STCON and NTT have submitted remarkable capable systems

Challenge Participation

Much lower participation compared to last year CHiME-7 **(from 9 down to 3 teams)**

- Participants split between the three tasks this year which were highly related
  - Overall CHiME participation was up (+4 compared to CHiME-7)

1. **Constrained LM track**
   - 3 submissions: STCON, NTT and a team which was anonymized due to poor performance
   - **Quality not quantity ?** STCON and NTT have submitted remarkable capable systems

Challenge Participation

Much lower participation compared to last year CHiME-7 **(from 9 down to 3 teams)**

- Participants split between the three tasks this year which were highly related
  - Overall CHiME participation was up (+4 compared to CHiME-7)

1. **Constrained LM track**
   - 3 submissions: STCON, NTT and a team which was anonymized due to poor performance
   - **Quality not quantity ?** STCON and NTT have submitted remarkable capable systems

2. **Unconstrained LM track**
   - Only STCON submitted a system
     - However, improvement was marginal w.r.t. Constrained LM

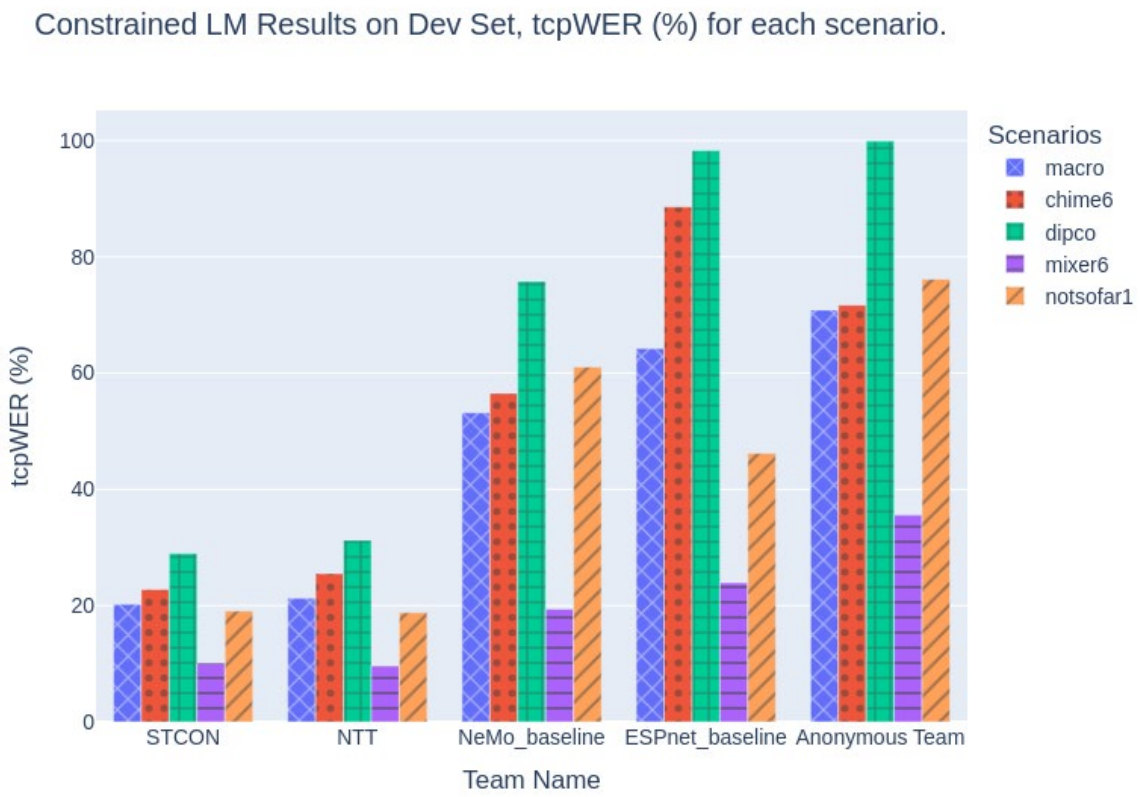
Challenge Participation

3. Jury Award

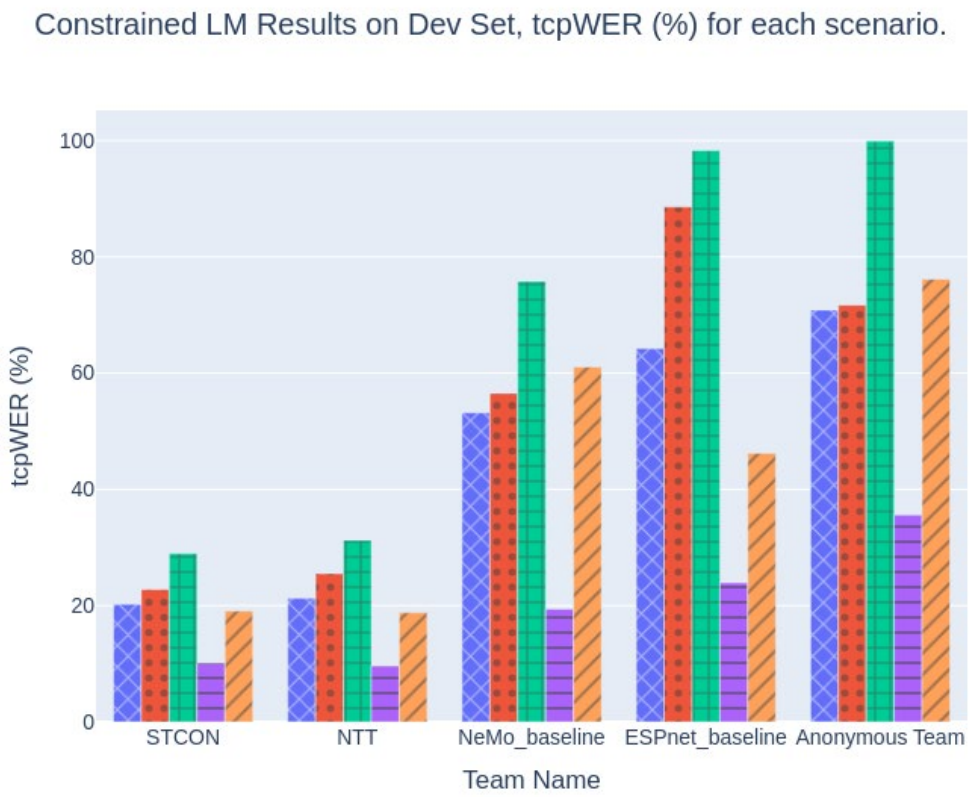
- Considered only together with NOTSOFAR-1 (not enough participants)

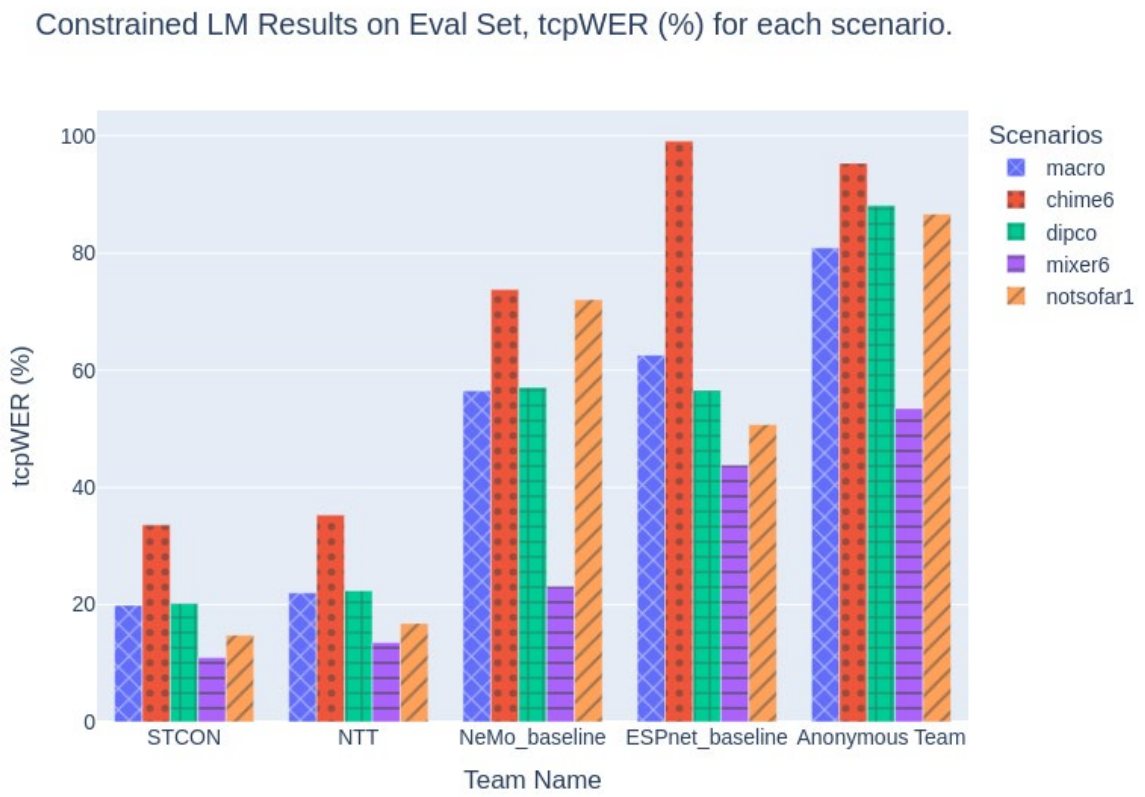
👏 NTT team however made significant efforts in producing also a more practical system and report the real time factor (RTF).

Challenge Results



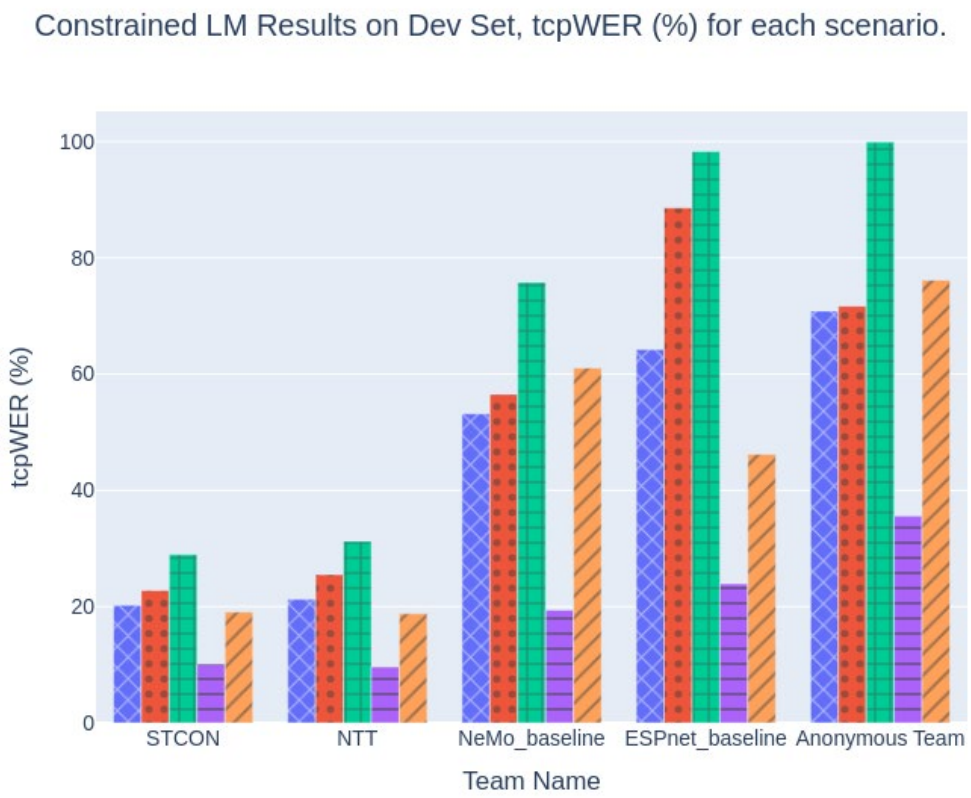
Challenge Results

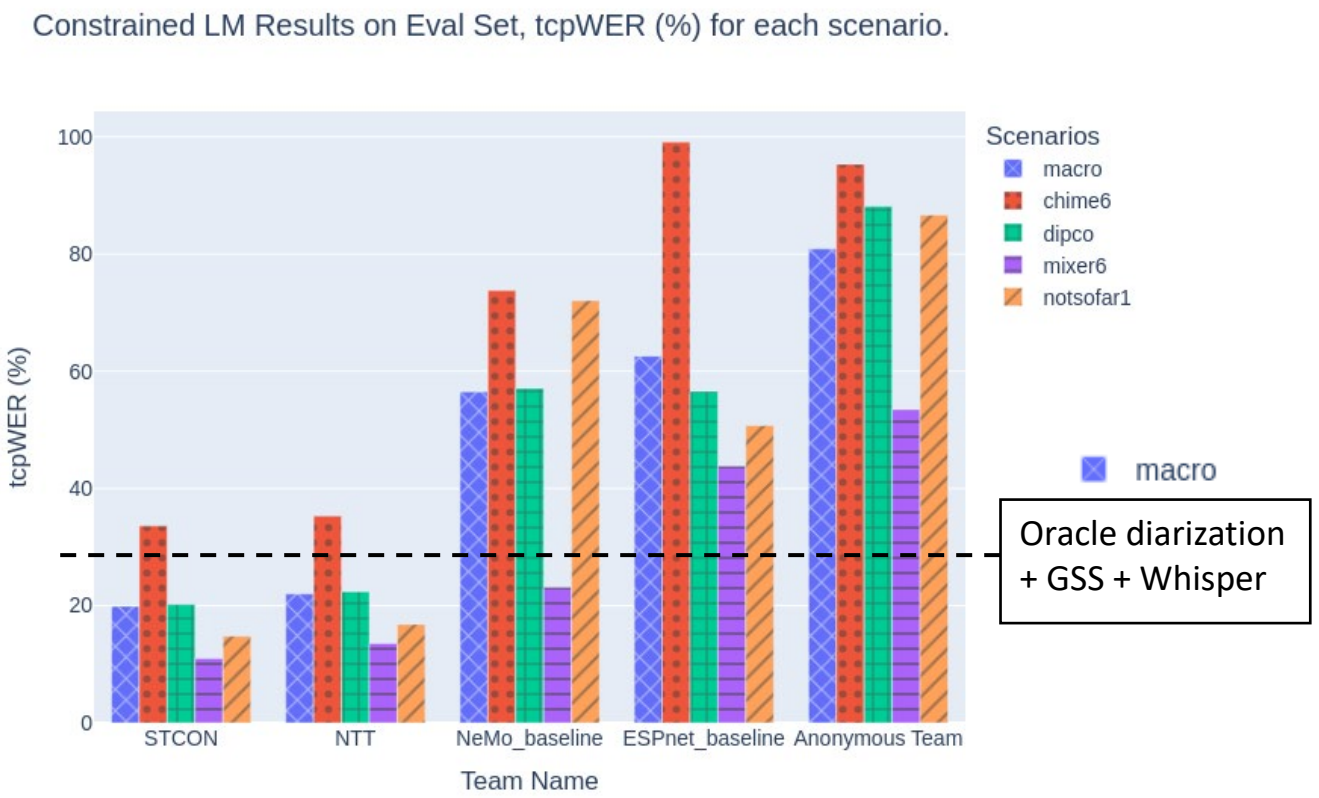




🏆 Congrats to STCON

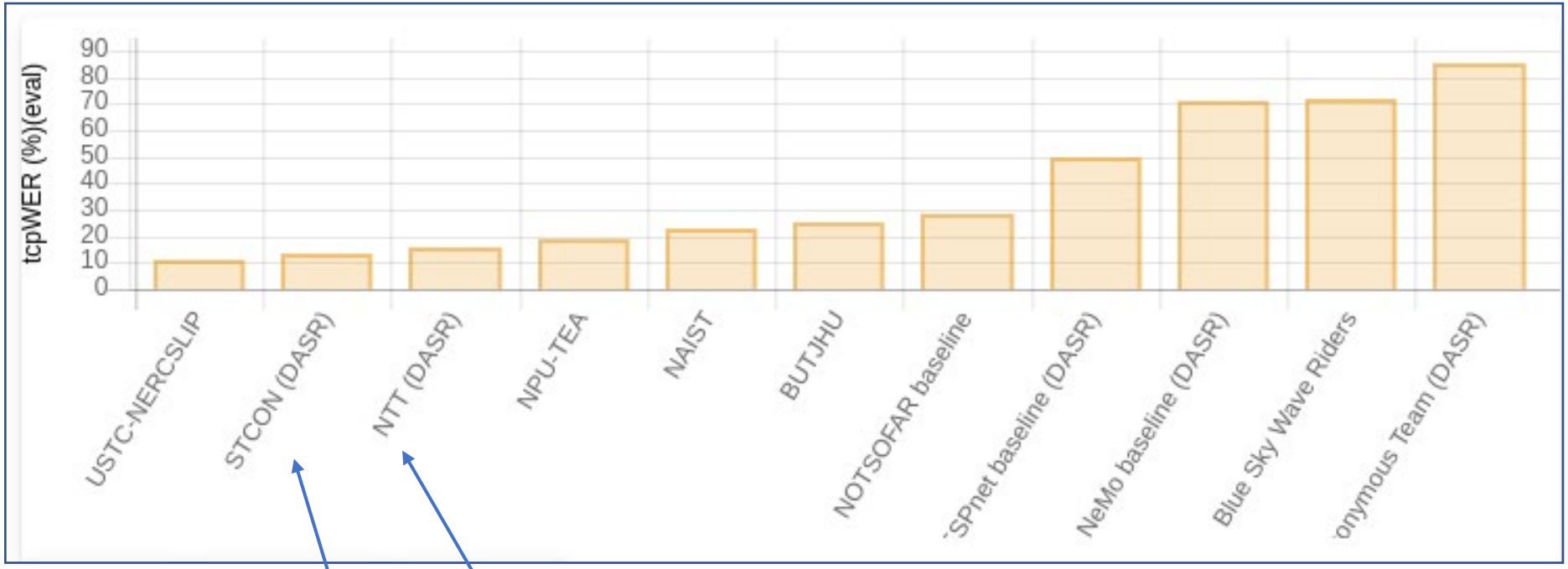
Challenge Results





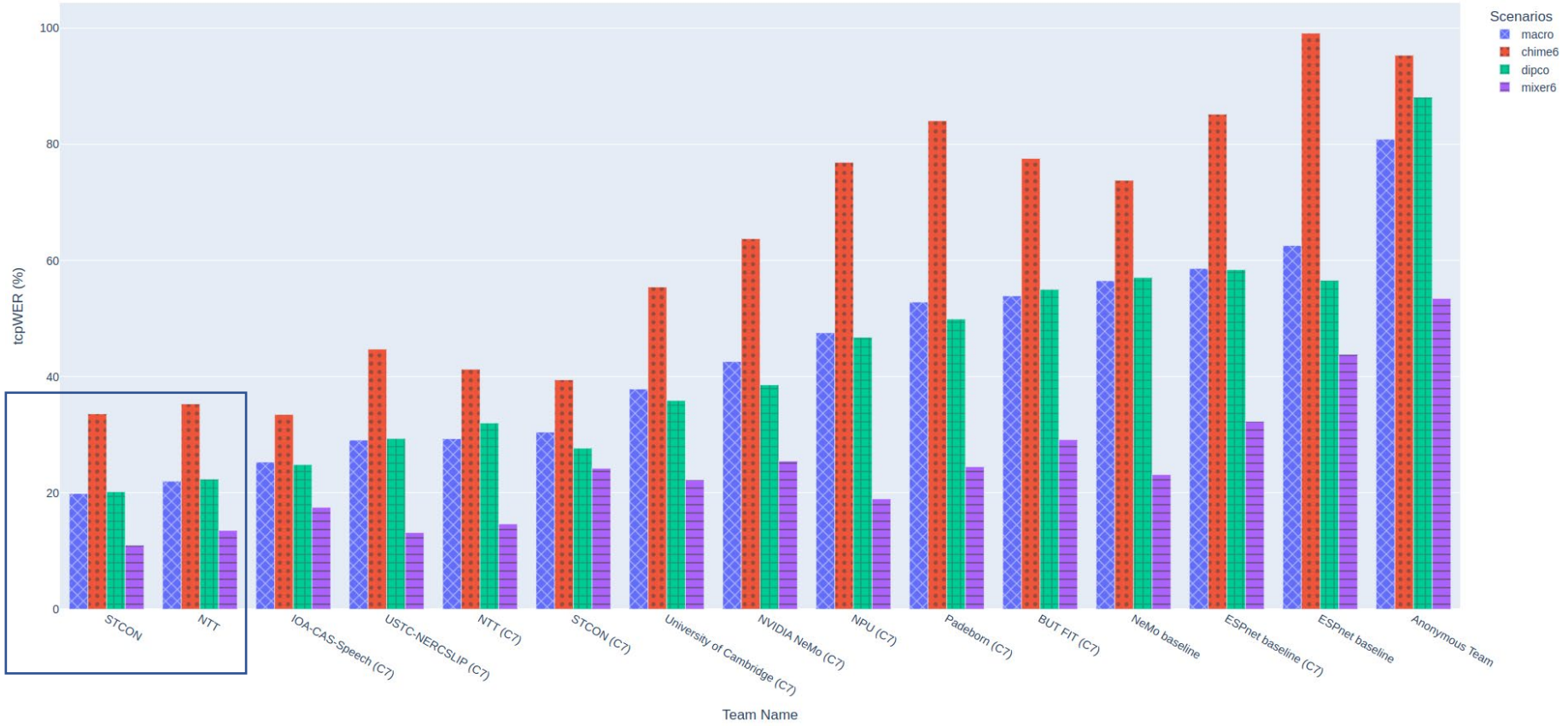
🏆 Congrats to STCON

NOTSOFAR-1 Task 2 Results



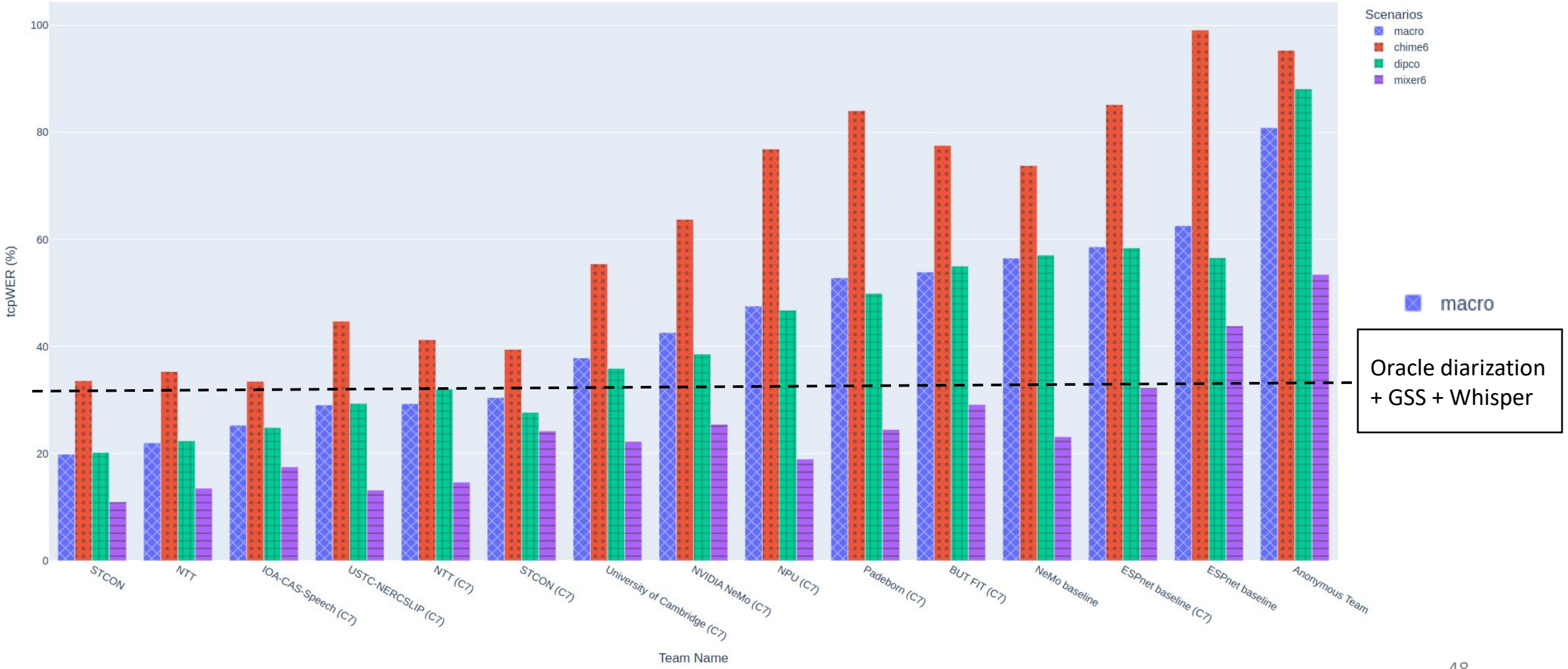
✨ Remarkably, STCON and NTT systems place also 2° and 3° in the NOTSOFAR-1 Task 2 challenge, despite being array and domain agnostic

C8+C7 Results



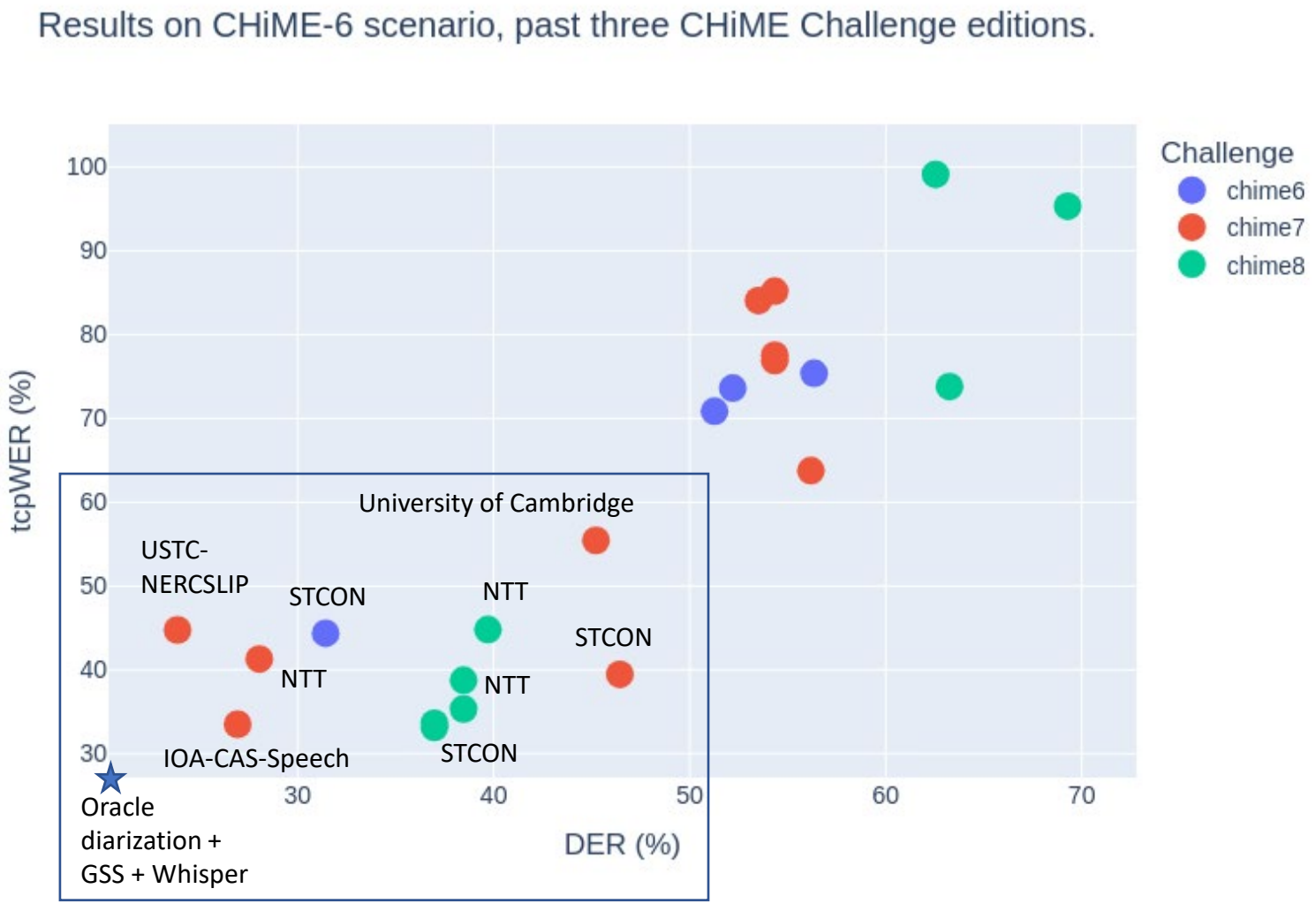
✨ STCON and NTT systems are able to push the performance further on CHiME-6, DiPCo and Mixer 6 despite having to deal also with the highly different NOTSOFAR-1 scenario

C8+C7 Results



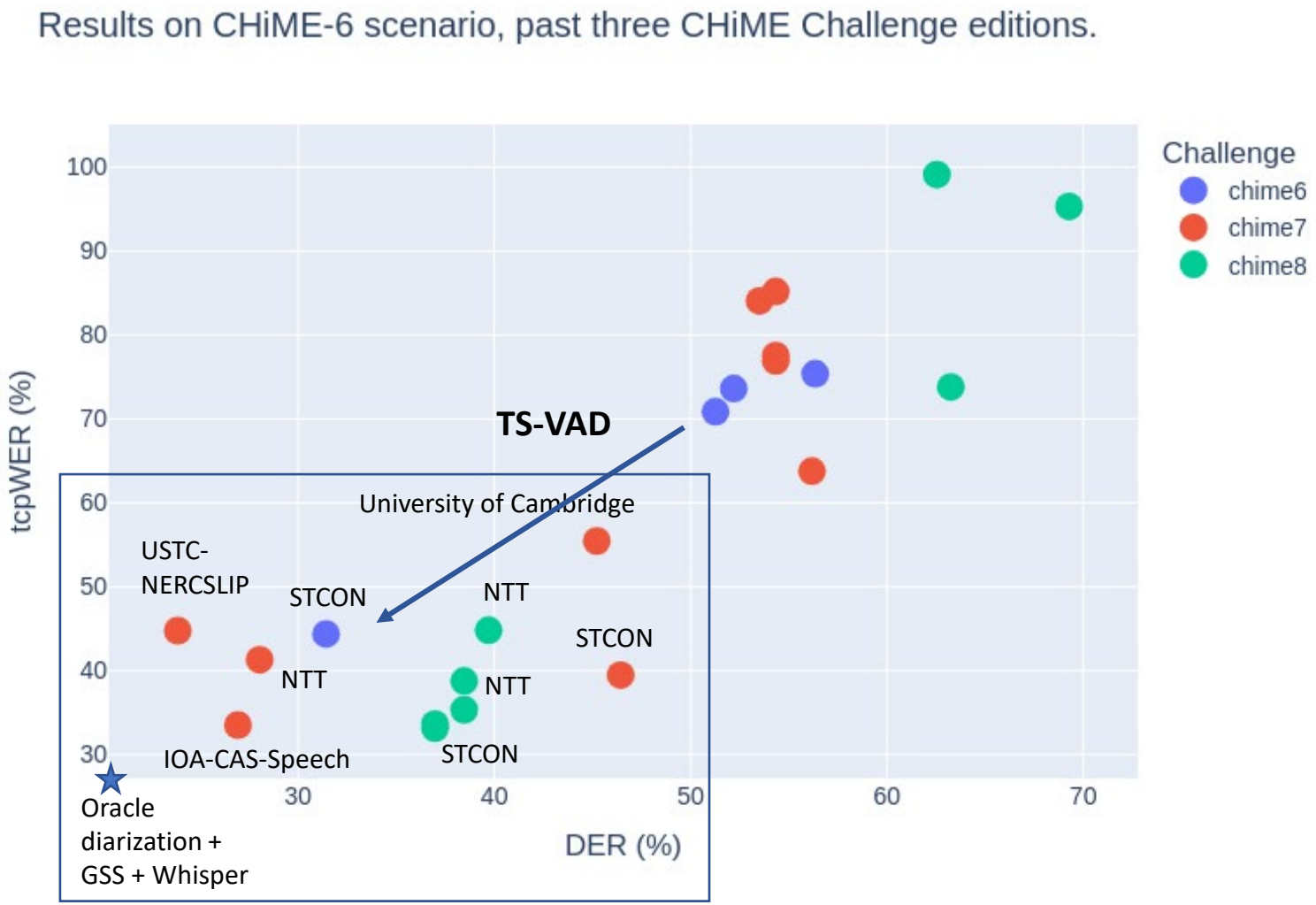
4.5 years of CHiME-6

4.5 years of CHiME-6

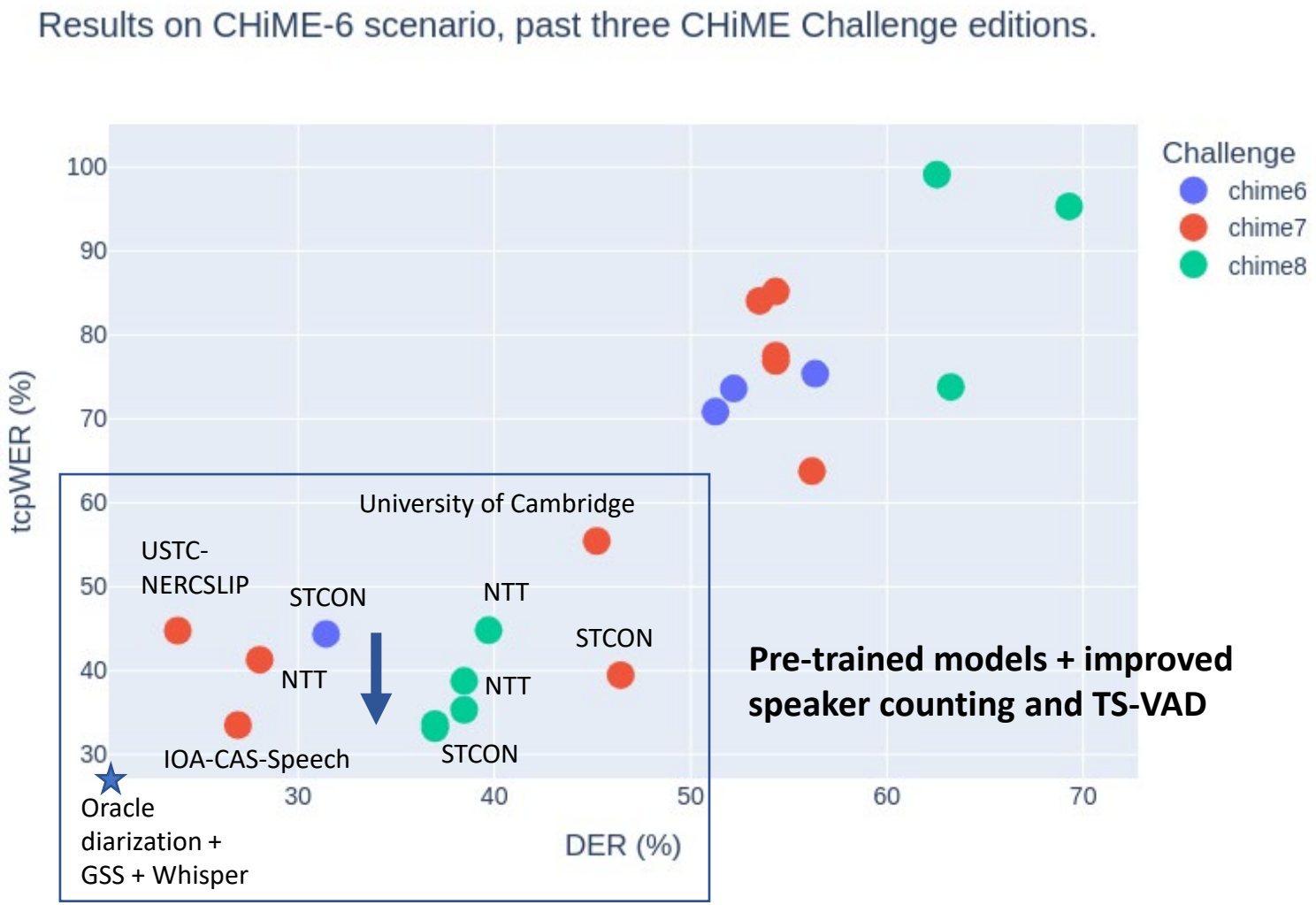


- DER (%) is computed w.r.t. JSON annotation and with 250ms collar

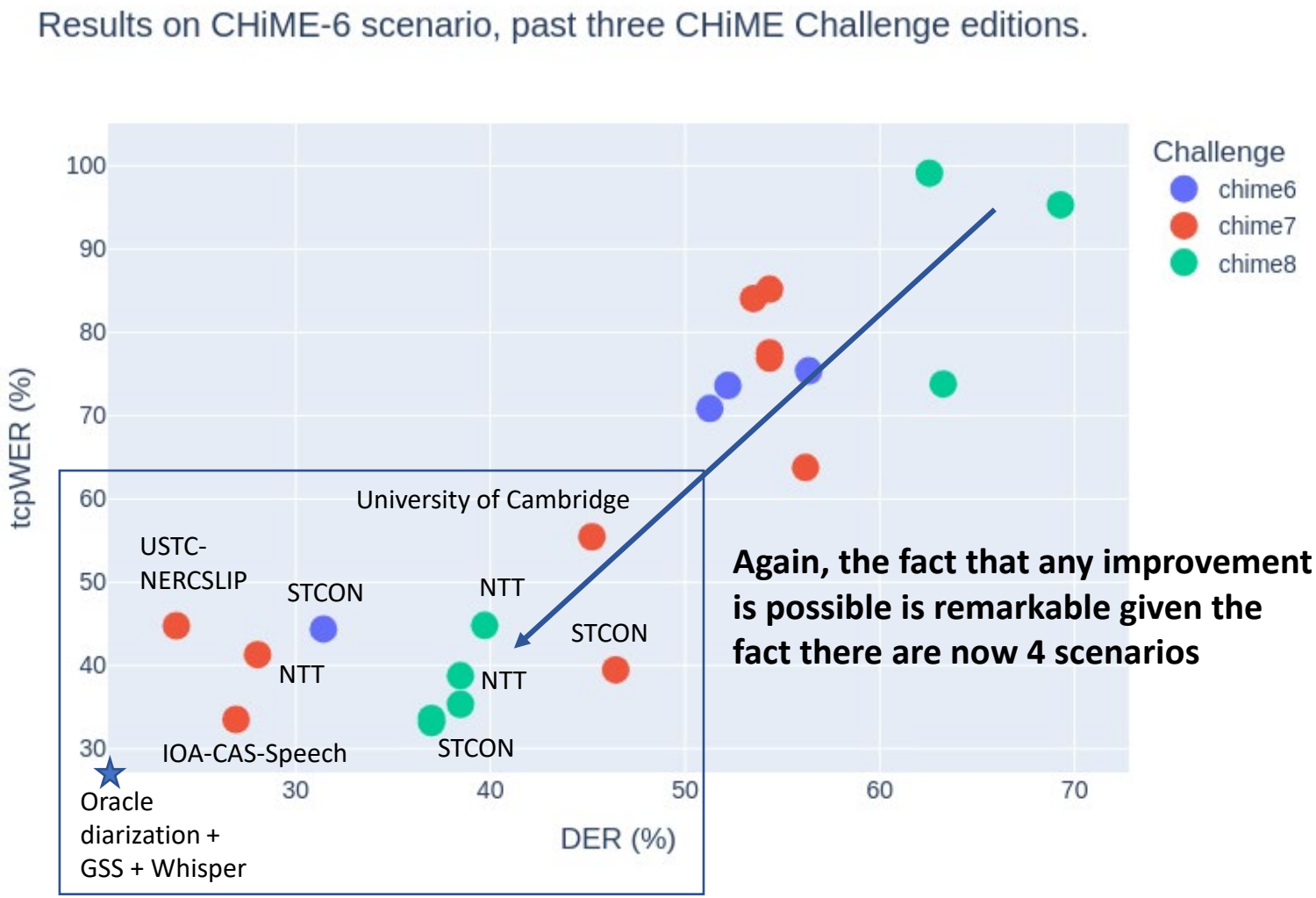
4.5 years of CHiME-6



4.5 years of CHiME-6



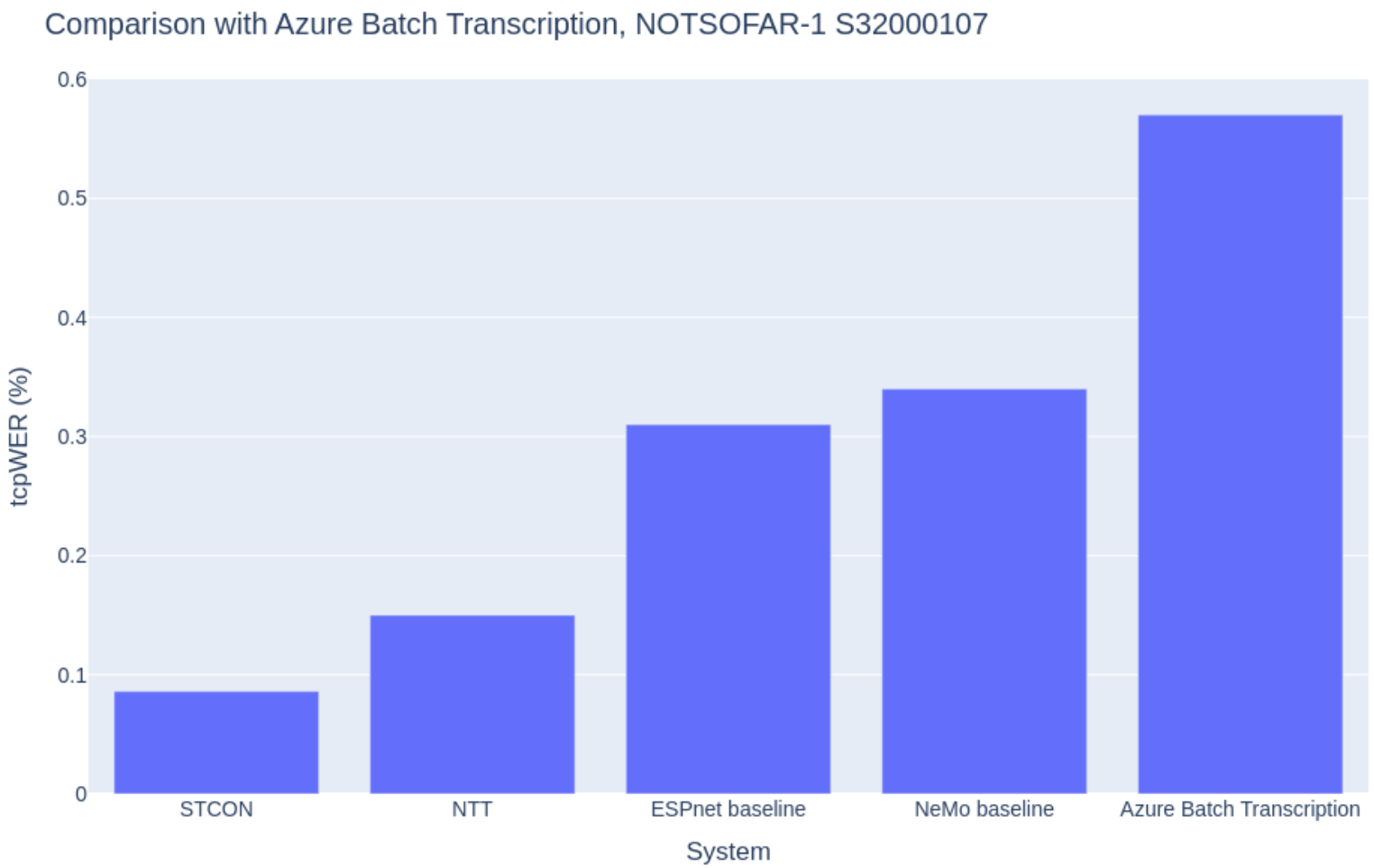
4.5 years of CHiME-6



Again, the fact that any improvement is possible is remarkable given the fact there are now 4 scenarios

CHiME vs Current Prod Systems

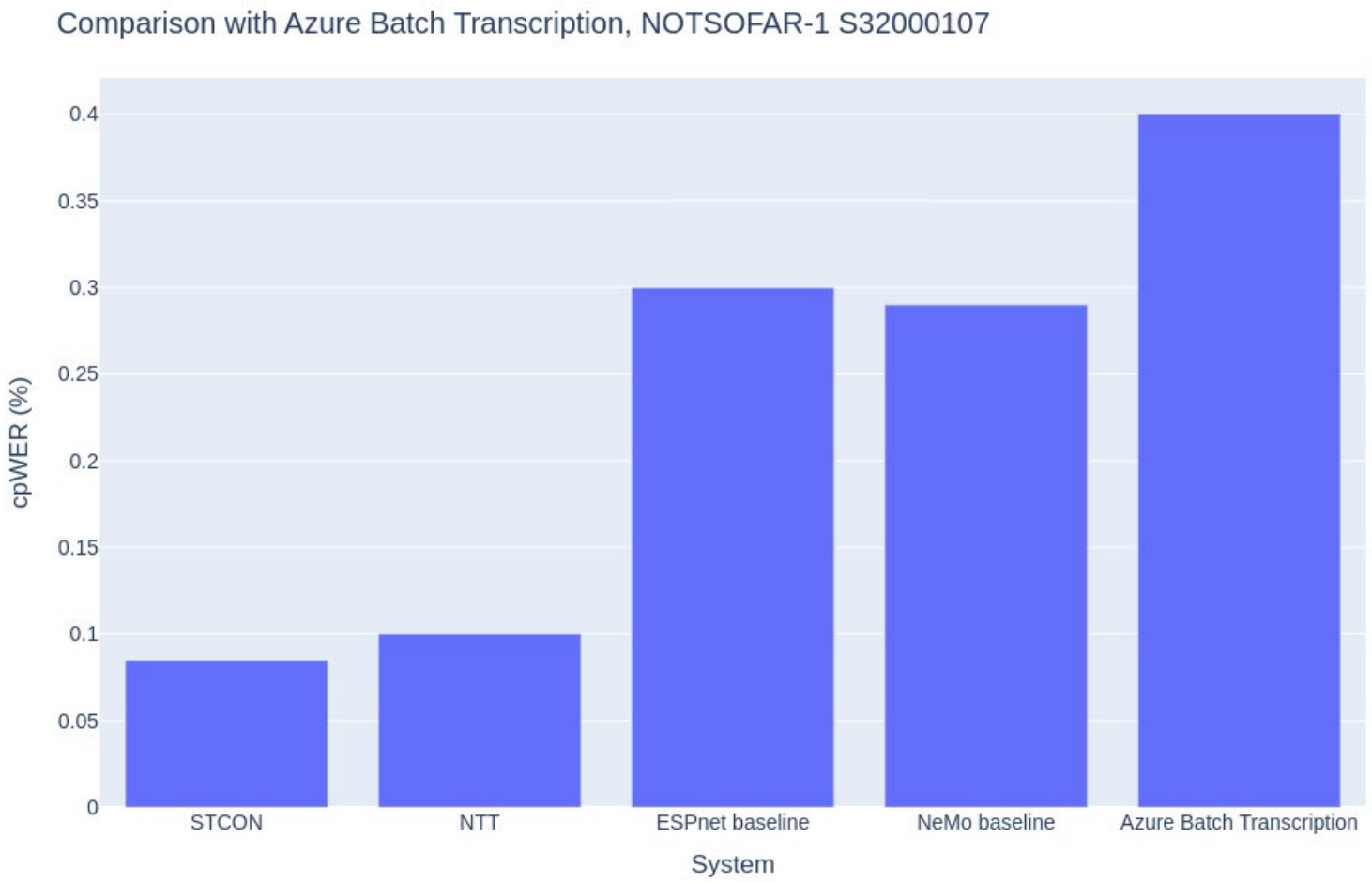
CHiME vs Current Prod Systems



NOTE: Azure results are single channel as diarization appears to be not supported for multi-channel

- Random single session from NOTSOFAR-1 eval set
  - Easiest scenario for single-channel systems

CHiME vs Current Prod Systems



Gap with baselines is less pronounced for cpWER

My Takeaways from Task 1 & 2

My Takeaways from Task 1 & 2

1. **Guided Source Separation (GSS) still reigns supreme for front-end processing**
    - STCON, NTT and USTC (NOTSOFAR-1 Task) use it as the main separation component

My Takeaways from Task 1 & 2

1. **Guided Source Separation (GSS) still reigns supreme for front-end processing**
    - STCON, NTT and USTC (NOTSOFAR-1 Task) use it as the main separation component

(Target) speech separation with real world data is hard, even when reasonably matched synthetic data is available and array geometry is known (NOTSOFAR-1 scenario)

My Takeaways from Task 1 & 2

1. **Guided Source Separation (GSS) still reigns supreme for front-end processing**
    - STCON, NTT and USTC (NOTSOFAR-1 Task) use it as the main separation component

(Target) speech separation with real world data is hard, even when reasonably matched synthetic data is available and array geometry is known (NOTSOFAR-1 scenario)

- <u>No team except STCON performed frontend + ASR E2E fine-tuning</u>
    - This model however works best when used for GSS refinement (G-TSep).
    - STCON also tried to use continuous source separation (CSS) but failed to achieve good results.

My Takeaways from Task 1 & 2

1. **Guided Source Separation (GSS) still reigns supreme for front-end processing**
    - STCON, NTT and USTC (NOTSOFAR-1 Task) use it as the main separation component

(Target) speech separation with real world data is hard, even when reasonably matched synthetic data is available and array geometry is known (NOTSOFAR-1 scenario)

- <u>No team except STCON performed frontend + ASR E2E fine-tuning</u>
    - This model however works best when used for GSS refinement (G-TSep).
    - STCON also tried to use continuous source separation (CSS) but failed to achieve good results.

- <u>NTT team proposes some improvements over the baseline channel selection + GSS pipeline</u>
    - Brouhuaha estimated C50 speech clarity index based channel selection
    - Spatial-prediction MWF instead of the MIMO MVDR

My Takeaways from Task 1 & 2

2. **For diarization, all top teams use TS-VAD techniques**
    - USTC (NOTSOFAR-1 Task), STCON and NTT **all used NSD-MS2S** [1]

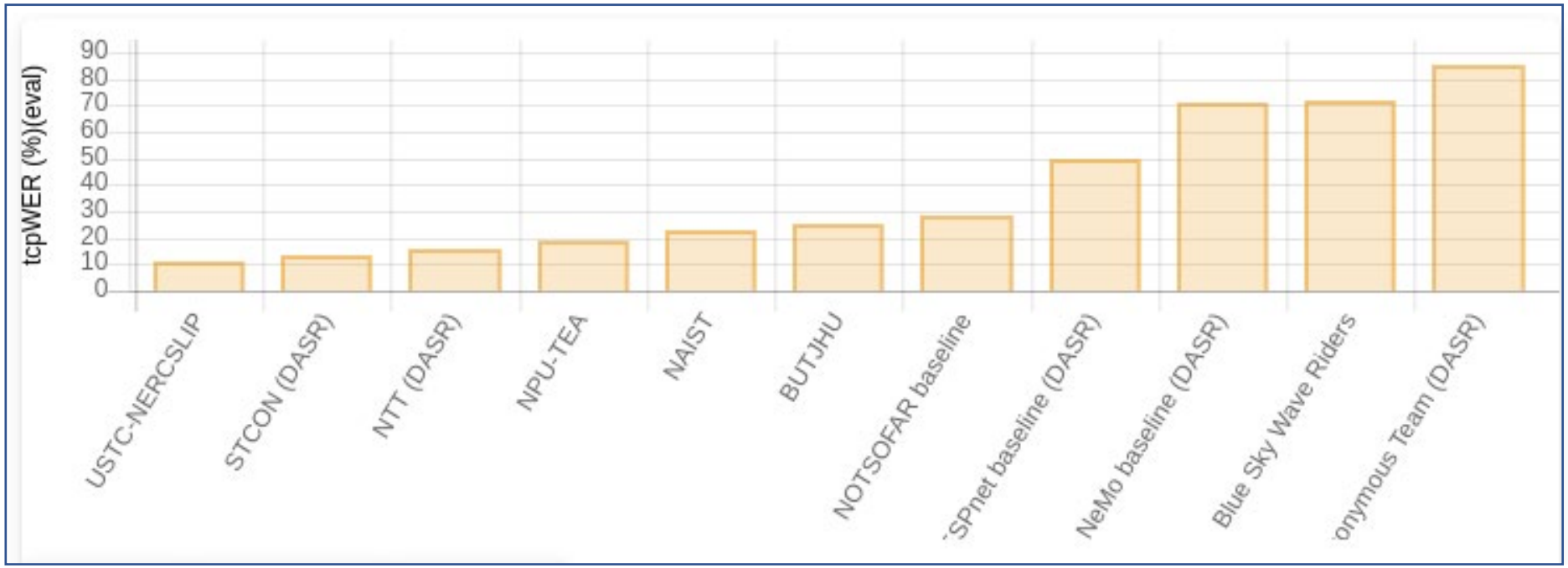
Accurate speaker counting for TS-VAD initialization is crucial

- STCON: Wav2vec 2.0 speaker ID embeddings AED + ECAPA-TDNN
- NTT: multi-channel speaker counting

[1] Yang, Gaobin, et al. "Neural Speaker Diarization Using Memory-Aware Multi-Speaker Embedding with Sequence-to-Sequence Architecture." ICASSP, 2024.

My Takeaways from Task 1 & 2

3. **Array/Domain-agnostic approaches (DASR) are competitive with domain specific ones (NOTSOFAR-1)**
   - STCON and NTT systems achieve 2° and 3° place in NOTSOFAR-1 multi-channel track
   - USTC-NERCSLIP NOTSOFAR-1 Task 2 winning system is heavily based on their CHiME-7 submission



Limitations & Future Work

1. **Generalization to unseen/unknown domains**
   - This is not really addressed, participants knew the domains in advance
   - We need to collect new data for this purpose (expensive)

Limitations & Future Work

1. **Generalization to unseen/unknown domains**
   - This is not really addressed, participants knew the domains in advance
   - We need to collect new data for this purpose (expensive)
2. **The entry bar is (still) very high**
   - Building a SotA multi-channel ASR+diarization pipeline is difficult for small teams
     - Requires decent amount of computational & human resources
     - Baselines are still difficult to experiment with for students

Thank you and thanks to all participants

Q&A also at the poster session or email me: samuele.cornell@ieee.org

If you are interested in CHiME challenges and workshop, consider joining:

CHiME Slack

CHiME Mailing List

CHiME CHALLENGE