NOTSOFAR-1 Challenge

Distant Meeting Transcription with a Single Device
Tracks: Single-Channel and Known-Geometry Multi-channel

Speaker: Alon Vinnikov

Authors:
Alon Vinnikov, Amir Ivry, Aviv Hurvitz, Igor Abramovski, Sharon Koubi, Ilya Gurvich, Shai Pe`er, Xiong Xiao, Benjamin Martinez Elizalde, Naoyuki Kanda, Xiaofei Wang, Shalev Shaer, Stav Yagev, Yossi Asher, Sunit Sivasankaran, Yifan Gong, Min Tang, Huaming Wang, Eyal Krupka

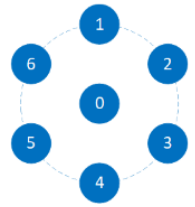
Teams Calling Meeting & Devices and Azure Cognitive Services

Microsoft Corporation

Microsoft

NOTSOFAR-1

- Objective:
  - Generate time-marked, speaker-attributed transcripts (tcpWER)
  - Using a **single device**, aligned with typical conference room setups
- Tracks:
  - **Single-channel**
  - **Multi-channel** with known-geometry
- We provide:
  - New **benchmark dataset** of 280 meetings in the office space covering a wide spectrum of real-world scenarios
  - Realistic simulated **training dataset**, incorporating 15K real RIRs
  - Baseline code: CSS → ASR → Diarization pipeline



Microsoft

Motivation

- Unlock the potential of data-driven methods
  - State-of-the-art pipelines are highly complex
  - Often fail on challenging cases
  - Classical methods are still hard to beat, e.g., GSS (Guided Source Separation) can be more robust than DNNs
  - → We aim to provide **high-quality datasets:**
    - **Training**: close the **train-test gap**
    - **Benchmarking**

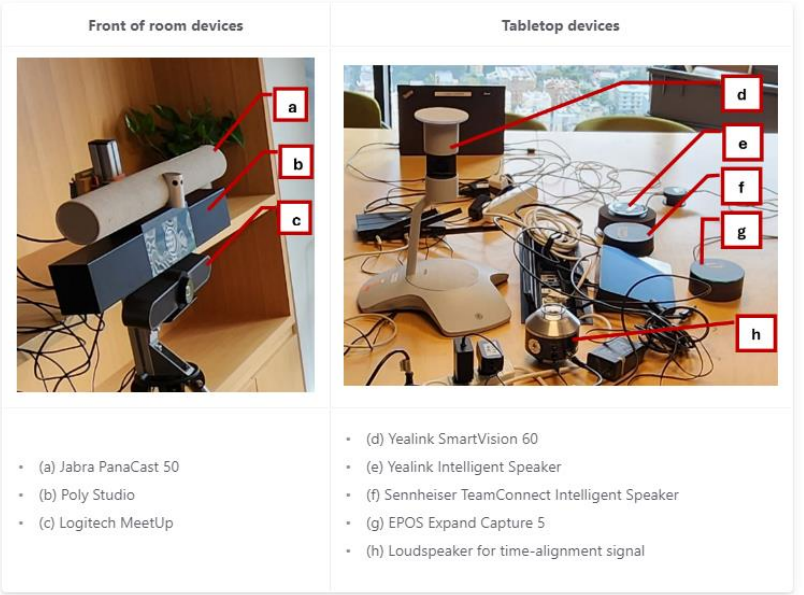
Recorded meeting dataset

- 280 distinct meetings (averaging ~6 minutes)
- 4-8 attendees per meeting
- 30 rooms and 30 unique speakers
- Capturing real-world complexities:
  - Walk and talk
  - Speech near whiteboard
  - Overlap, heated debates
  - Various distances / volumes / noises
- Tagged with meta-data

  (#DebateOverlaps, #TalkNearWhiteboard, #TransientNoise)
- Recorded with a range of commercially available devices
- High quality annotations from head mounted mics:
  - Multi-judge process
  - Bias-free transcription: annotators have **no access to machine-generated transcriptions**

Set	Train.	Dev.	Eval.
Meetings	107	36	137
Avg. Duration	~ 6 minutes		~ 6 minutes
Devices	4 / 5	~3 / 5	~4 / 6
Sessions	428 / 535	106 / 177	548 / 822
Session hours	42 / 53	10 / 17	54 / 82
Rooms	20 rooms in total		10
Speakers	22 speakers in total		10

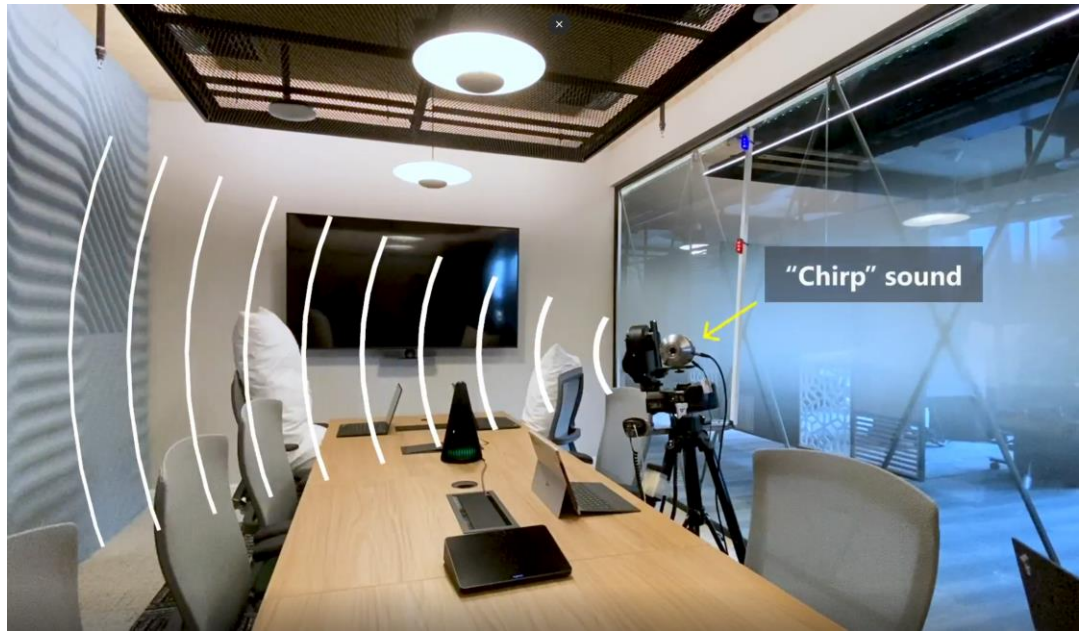
MC (left) / SC (right)

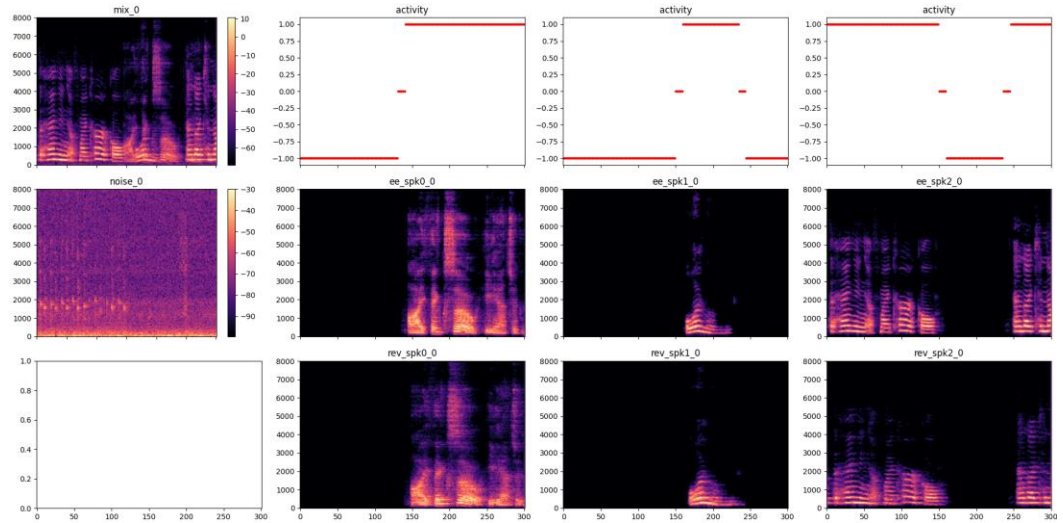


Microsoft

Simulated training dataset

- Synthesized for real-world generalization:
  - Based on 15K real RIRs
  - Realistic overlap patterns
  - Up to 3 simultaneous speakers
  - Transient and stationary noises from real rooms
  - Matched to the known array geometry in the multi-channel track





Microsoft

$$\text{mixture}(t) = \sum_i \left[\text{spk}_i^{\text{direct+early}}(t) + \text{spk}_i^{\text{reverb}}(t) \right] + \text{noise}(t)$$

Tracks

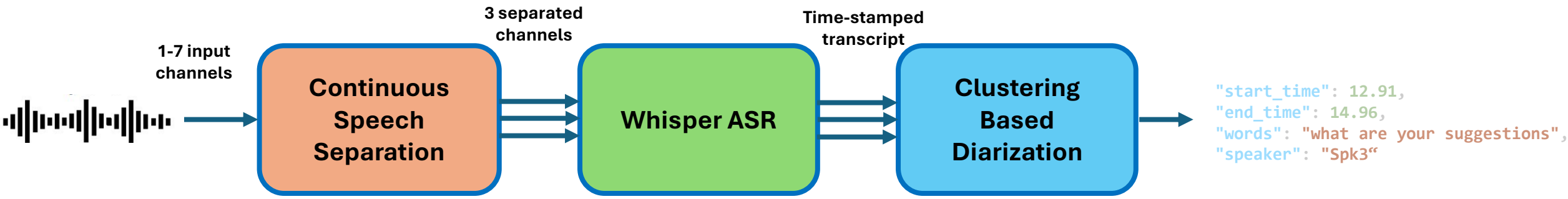
- Two main tracks, centered on the use of a single device:
  - **Single-channel:** a processed stream from commercial devices (not single-microphone)
  - **Multi-channel**: known-geometry, 7 raw data streams
    - To encourage data-driven solutions

- Collaborate with CHiME-8 DASR task to offer a fundamental system design comparison:
  - Single-channel (NOTSOFAR)

  vs.
  - Geometry-specific multi-channel (NOTSOFAR)

  vs.
  - Geometry-agnostic multi-channel (DASR)

Baseline system



Participation

Team Name	Affiliation
UWB	University of West Bohemia
Fano Labs	Fano Labs
Blue Sky Wave Riders	Blue Sky
USTC-NERCSLIP	University of Science and Technology of China
ToTaTo	Tallinn University of Technology
NJU-AALab	Nanjing University
NPU-TEA	Tencent
NAIST	Nara Institute of Science and Technology
BUT/JHU	Brno University of Technology

Official Results: Eval

Single-channel

RANK	TEAM NAME	SYSTEM TAG	TCPWER (%) (EVAL)	TCPWER (%) (DEV)	TCORCWER (%) (EVAL)	TCORCWER (%) (DEV)
1	**USTC-NERCSLIP**	**sys2**	**22.2 %**	**24.6 %**	**17.7 %**	**19.9 %**
2	NPU-TEA	sys4	30.0 %	28.6 %	25.7 %	25.1 %
3	NJU-AALab	sys1	33.5 %	36.4 %	30.4 %	33.1 %
4	NAIST	sys2	36.6 %	41.0 %	33.3 %	36.2 %
5	**BUTJHU**	**sys2**	**40.1 %**	**35.7 %**	**27.2 %**	**25.4 %**
6	ToTaTo	sys4	41.2 %	39.8 %	29.2 %	28.5 %
7	NOTSOFAR baseline	sys1	41.4 %	45.8 %	35.5 %	38.6 %
8	Fano Labs	sys3	43.1 %	40.7 %	40.0 %	38.0 %
9	UWB	sys1	45.8 %	41.0 %	39.1 %	32.6 %
10	Blue Sky Wave Riders	sys2	74.1 %		37.0 %	

Multi-channel

RANK	TEAM NAME	SYSTEM TAG	TCPWER (%) (EVAL)	TCPWER (%) (DEV)	TCORCWER (%) (EVAL)	TCORCWER (%) (DEV)
1	**USTC-NERCSLIP**	**sys1**	**10.8 %**	**14.7 %**	**9.5 %**	**12.7 %**
2	NPU-TEA	sys3	18.7 %	21.4 %	17.0 %	19.4 %
3	NAIST	sys2	23.0 %	28.9 %	21.1 %	26.1 %
4	**BUTJHU**	**sys2**	**24.9 %**	**29.6 %**	**15.6 %**	**18.7 %**
5	NOTSOFAR baseline	sys1	28.3 %	31.6 %	24.6 %	26.6 %
6	Blue Sky Wave Riders	sys2	71.9 %	72.5 %	34.4 %	

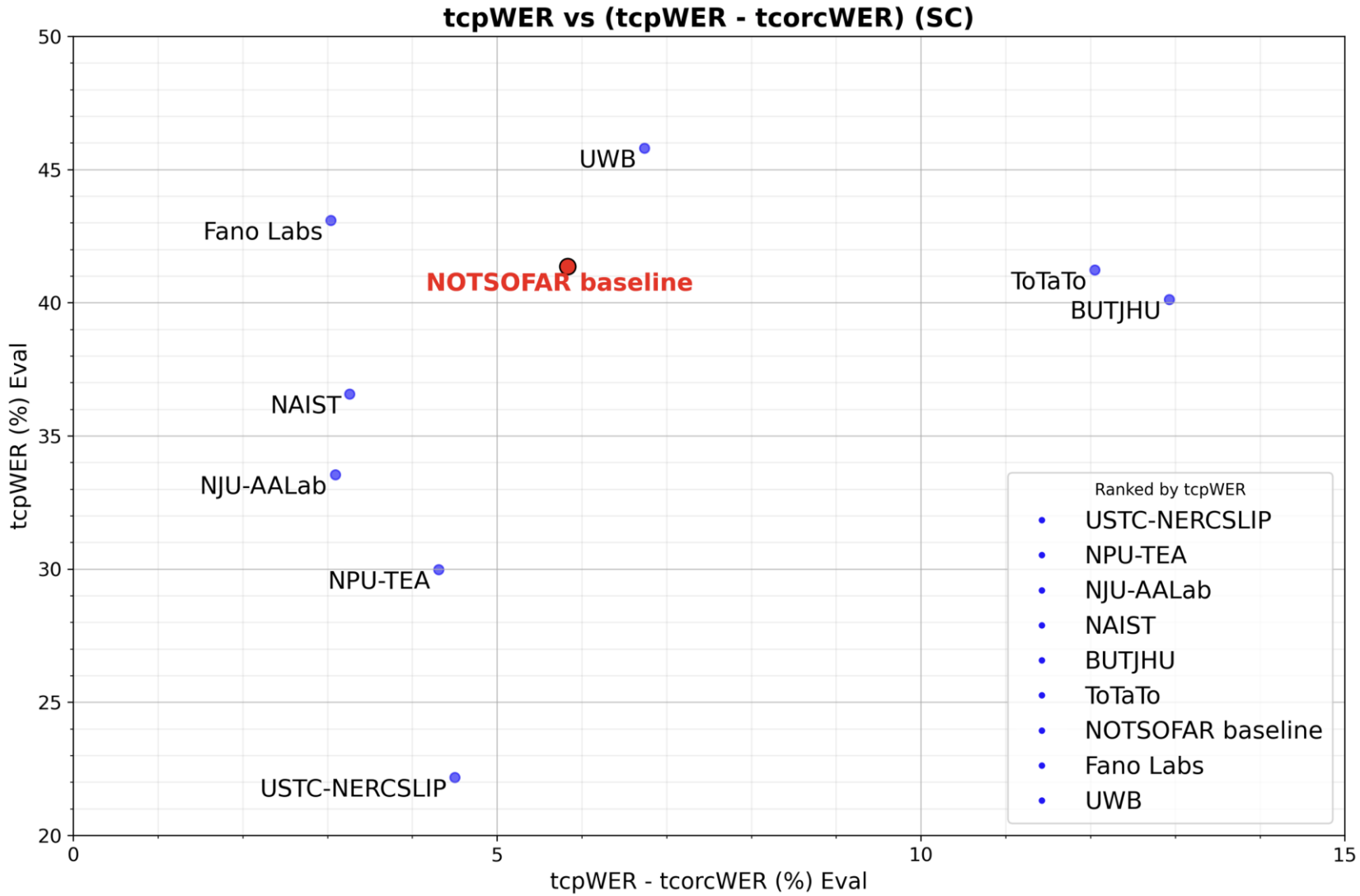
🏅 **The Jury Award goes to: BUT/JHU**

Multi-channel Supplementary Results: DASR included

Team Name	System Tag	tcpWER (%) (eval)	tcpWER (%) (dev)	tcorcWER (%) (eval)	tcorcWER (%) (dev)	
1	USTC-NERCSLIP	sys1	10.8 %	14.7 %	9.5 %	12.7 %
2	STCON (DASR)	sys1	13.8 %	19.5 %	11.6 %	16.2 %
3	NTT (DASR)	sys3	15.9 %	19.4 %	12.6 %	15.2 %
4	NPU-TEA	sys3	18.7 %	21.4 %	17.0 %	19.4 %
5	NAIST	sys2	23.0 %	28.9 %	21.1 %	26.1 %
6	BUTJHU	sys2	24.9 %	29.6 %	15.6 %	18.7 %
7	NOTSOFAR baseline	sys1	28.3 %	31.6 %	24.6 %	26.6 %
8	ESPnet_baseline (DASR)	sys1	49.7 %	48.1 %	40.2 %	36.3 %
9	NeMo_baseline (DASR)	sys1	70.8 %	61.3 %	59.9 %	47.9 %
10	Blue Sky Wave Riders	sys2	71.9 %	72.5 %	34.4 %	31.6 %
11	Anonymous Team (DASR)	sys1	85.3 %	76.0 %	59.2 %	54.5 %

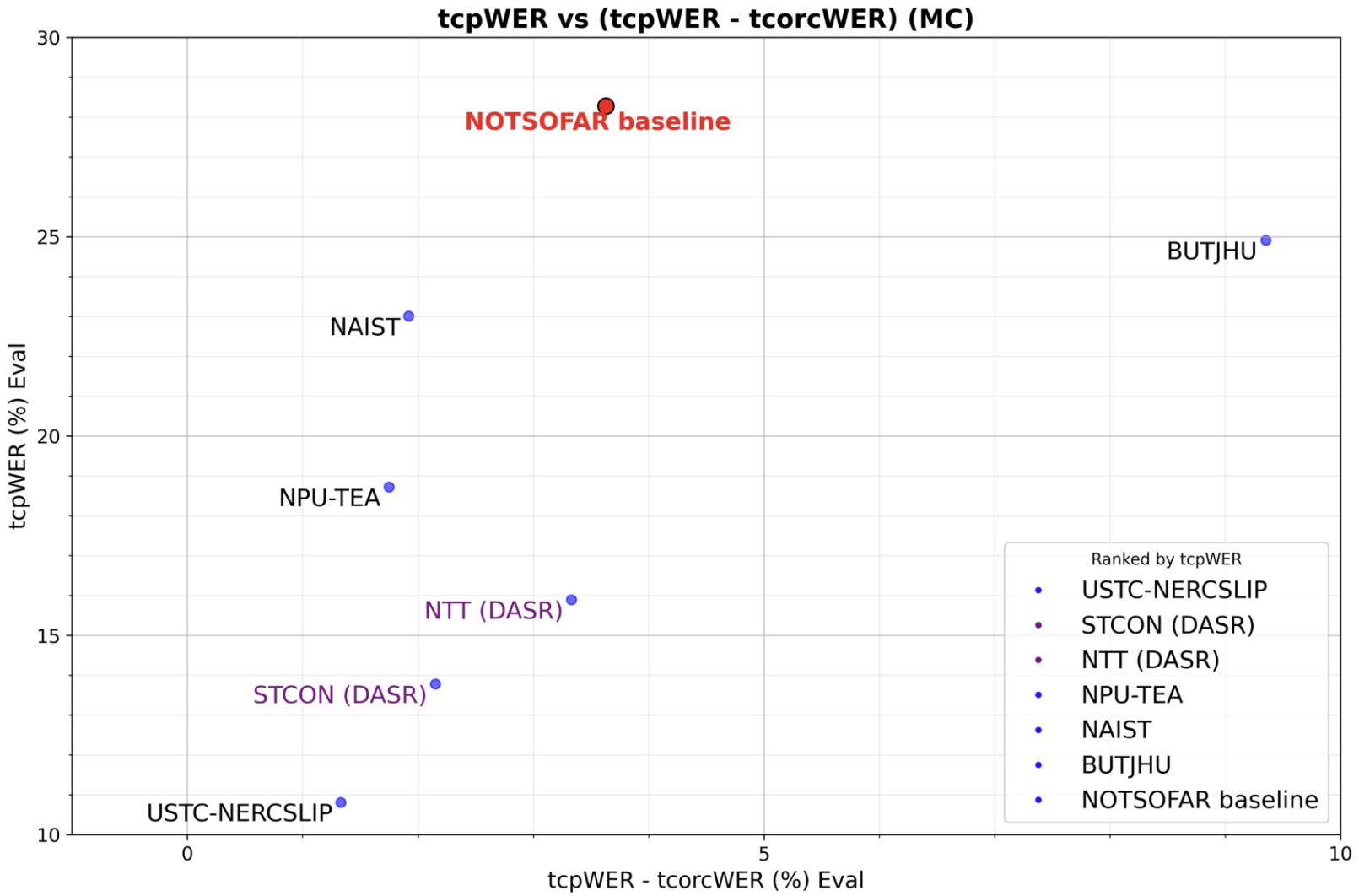
Single-channel Track Results: Eval

The (tcpWER-tcpWER) difference isolates speaker ID errors

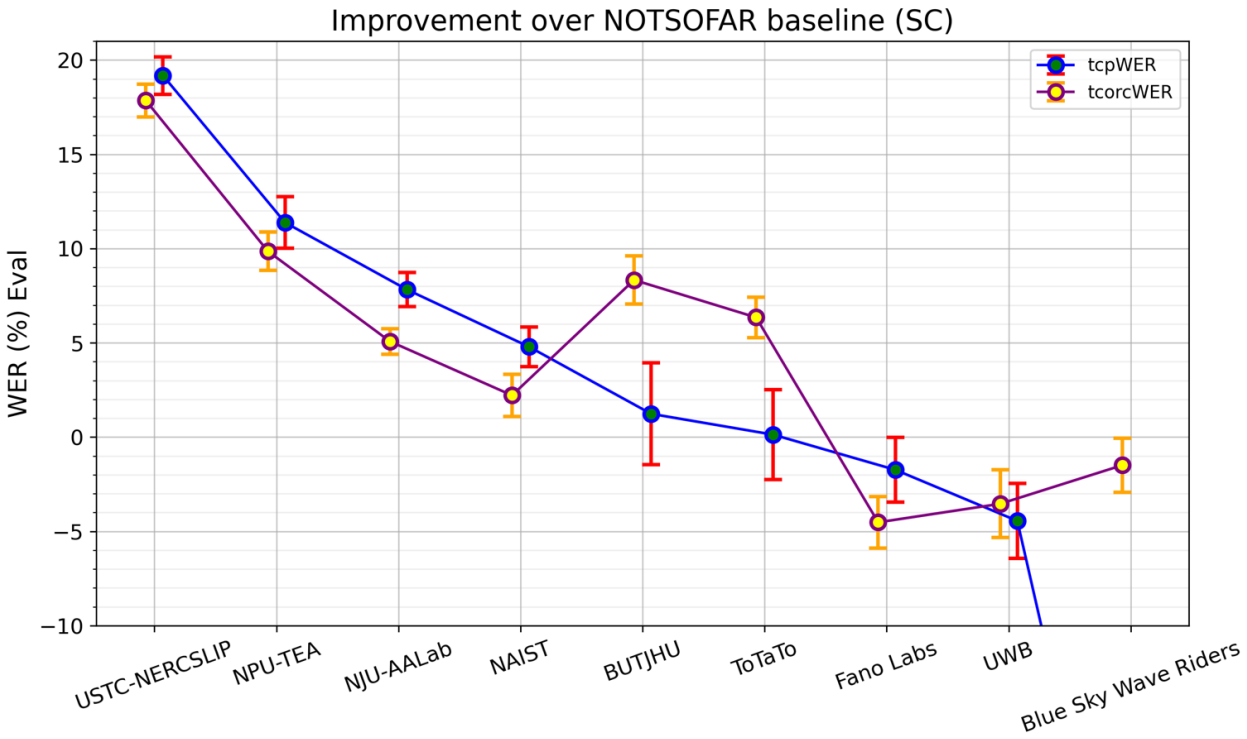


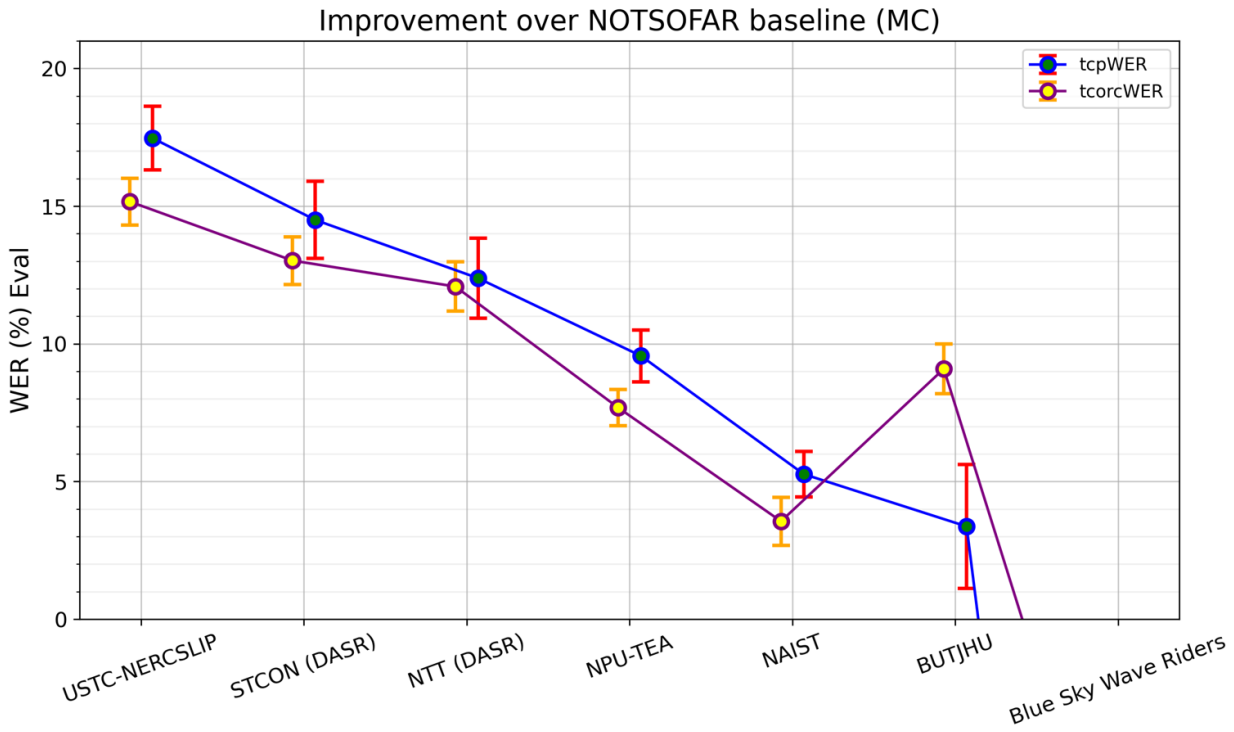
Multi-channel Track Results: Eval

The (tcpWER-tcpWER) difference isolates speaker ID errors

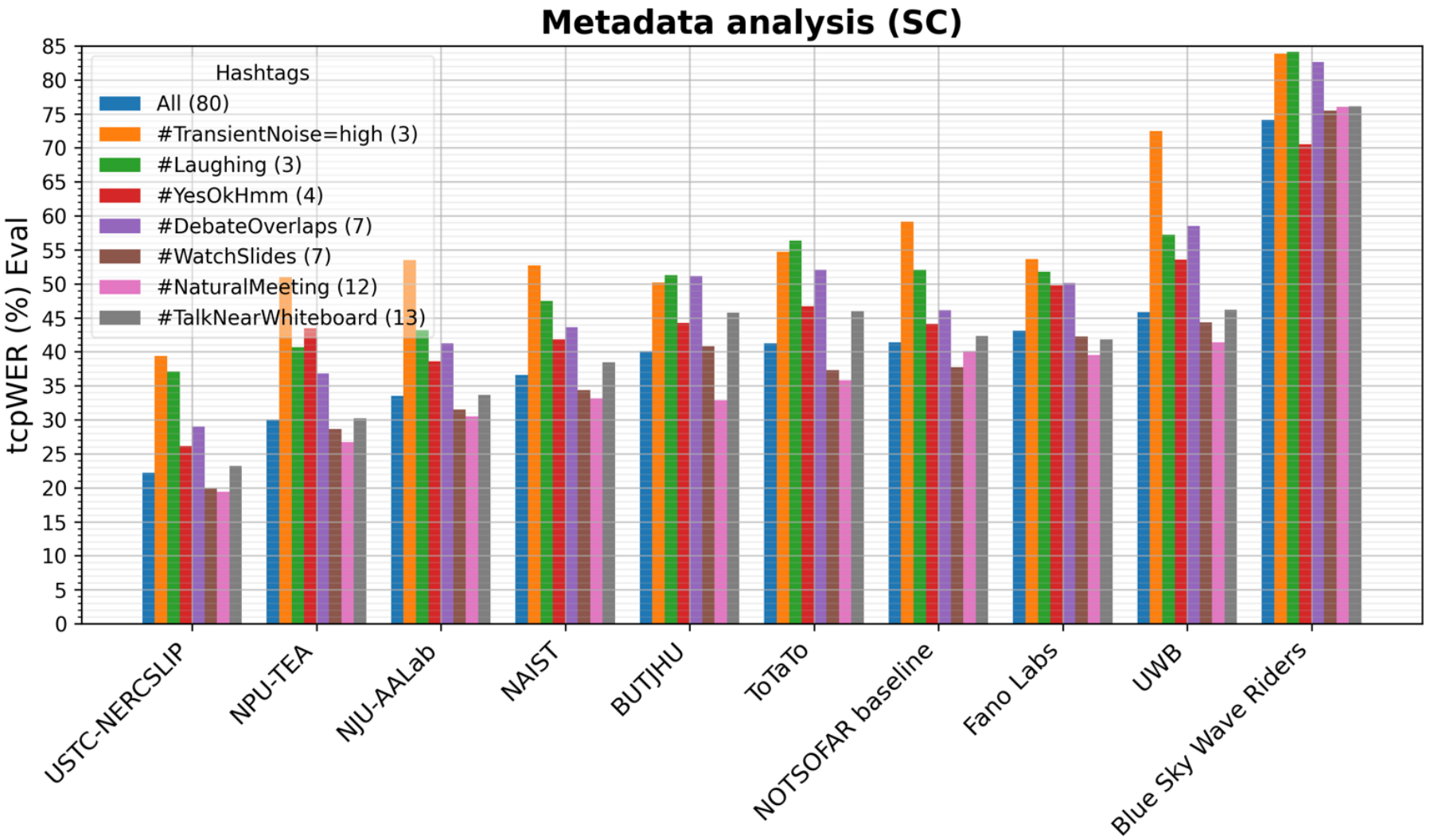


Confidence Intervals - Improvement Over Baseline

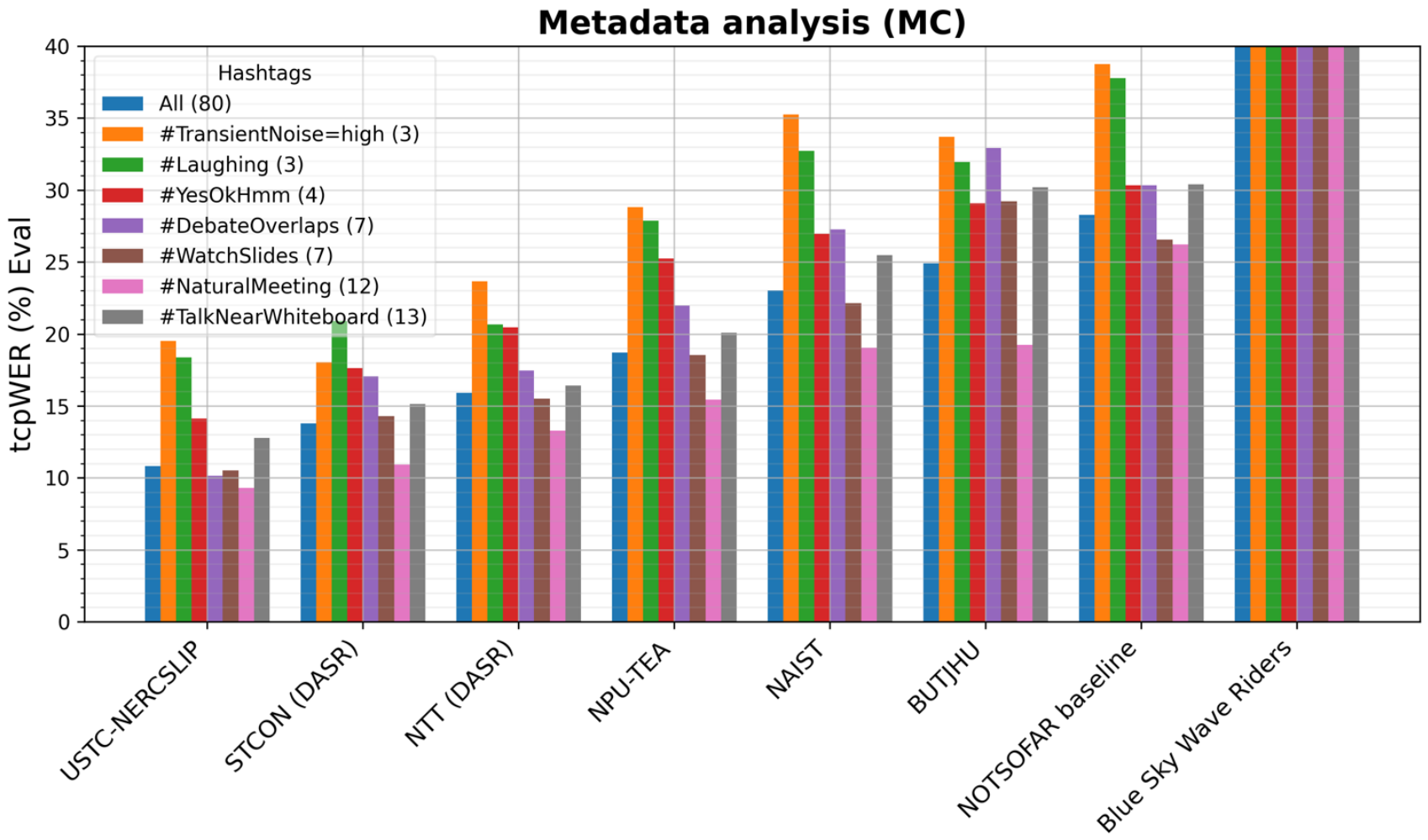




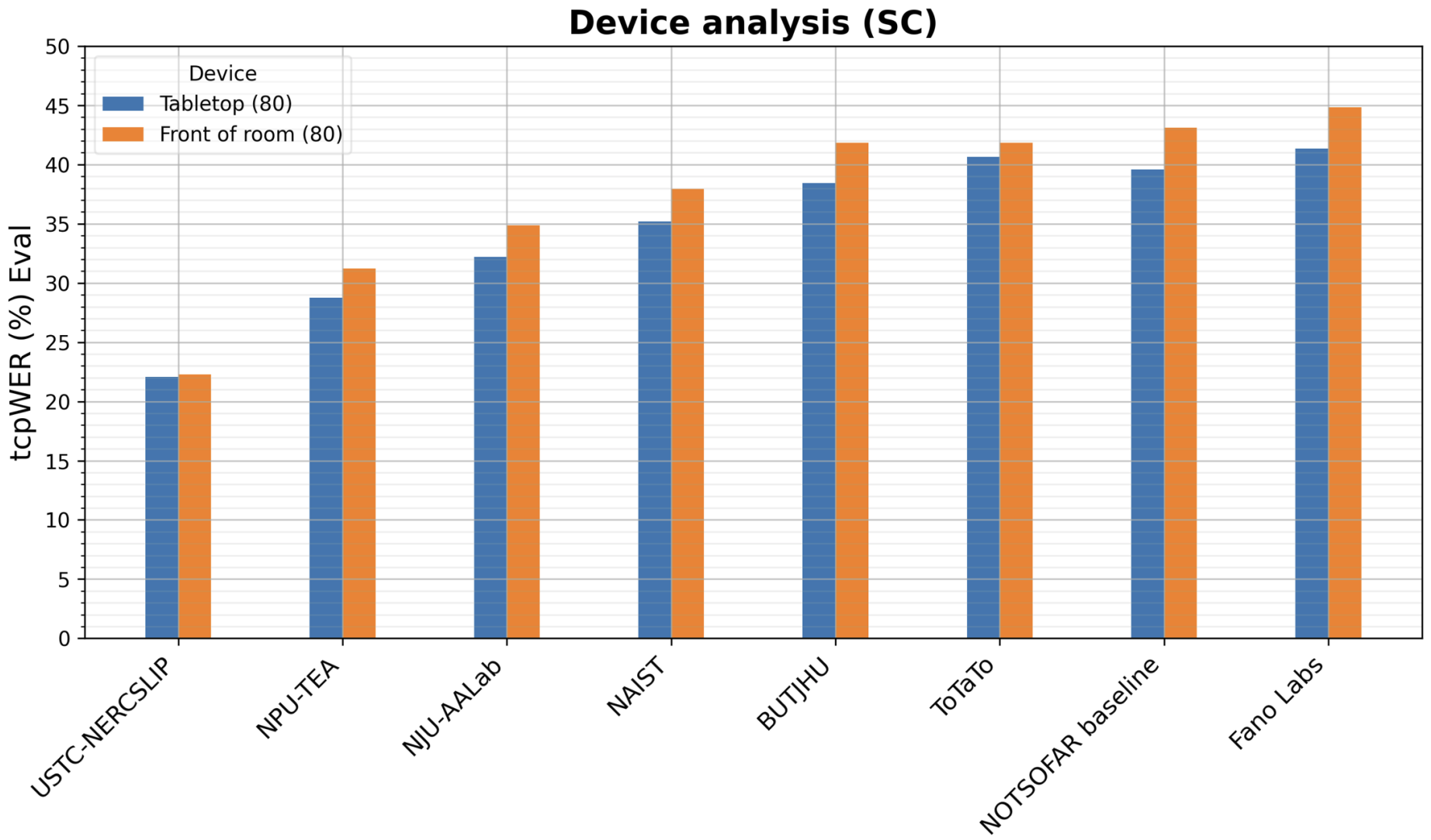
Metadata analysis



Metadata analysis



Single-channel: Front-of-room vs Tabletop



NOTSOFAR datasets for open research

- Available under the Creative Commons Attribution 4.0 International License (CC BY 4.0)
- Latest release: Train/Dev/Eval-small/Eval-full
- Comparison to the subsets used during the NOTSOFAR-1 Challenge:
  - Cleaned up data
  - Larger eval set (eval-full) now available
  - **Dev-set-2** removed due to legal constraints; **Dev-set-1** will be the development set for the open dataset
  - Upcoming **upgrades**: Train/Dev annotation quality.

Thank you!

Any questions?

The NOTSOFAR team

Find us at:
https://github.com/microsoft/NOTSOFAR1-CHALLENGE

Microsoft

- NOTSOFAR - Natural Office Talkers in Settings Of Far-field Audio Recordings