NOTSOFAR-1 Challenge: New Datasets, Baseline, and Tasks For Distant Meeting Transcription

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

NOTSOFAR-1 (Natural Office Talkers in Settings Of Far-field Audio Recordings) Challenge

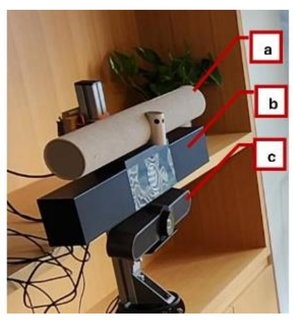
The challenge focuses on distant (far-field) speaker diarization and automatic speech recognition, **using a single recording device**, emphasizing realistic conversational scenarios. It features two tracks: **single-channel (SC)** and, **known-geometry multi-channel (MC).** We are also introducing two new datasets to the community - a **benchmarking dataset** of 280 English meetings and a **matching simulated training dataset** aiming to close train-test gaps observed in prior works.

New Datasets

Recorded Meeting Dataset for Benchmarking

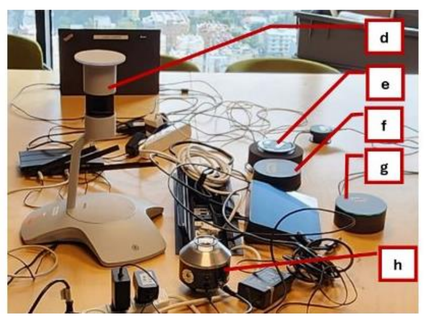
- 280 distinct meetings, averaging ~6 minutes
- 4-8 attendees per meeting
- 30 different rooms and 30 unique speakers
- Capturing real-world complexities:
  - Walk and talk
  - Speech near whiteboard
  - Overlap, heated debates
  - Various distances / volumes / noises
- Recorded with a range of commercial devices
- High quality multi-judge annotations from personal head mounted microphones
- Bias-free transcription: annotators have no access to machine-generated transcriptions
- Tagged with meta-data (e.g., #DebateOverlaps, #TalkNearWhiteboard, #TransientNoise)

Front of room devices

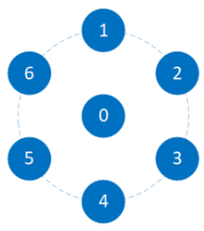


(a) Jabra PanaCast 50
(b) Poly Studio
(c) Logitech MeetUp

Tabletop devices



(a) Yealink SmartVision 60
(b) Yealink Intelligent Speaker
(c) Sennheiser TeamConnect Intelligent Speaker
(d) EPOS Expand Capture 5
(e) Loudspeaker for time-alignment signal



Array geometry of MC devices

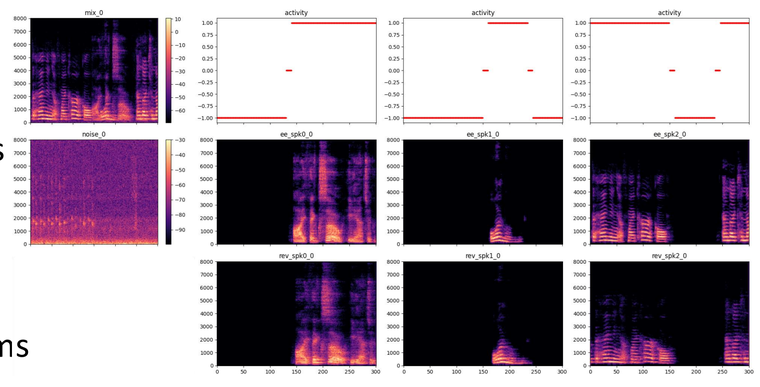
Set	Train.	Dev.	Eval.
Meetings	107	36	137
Avg. Duration	~ 6 minutes		~ 6 minutes
Devices	4 / 5	~3 / 5	~4 / 6
Sessions	428 / 535	106 / 177	548 / 822
Session hours	42 / 53	10 / 17	54 / 82
Rooms	20 rooms in total		10
Speakers	22 speakers in total		10

MC (left) / SC (right)

Simulated Training Dataset (~ 1,000 hours)

Synthesized for real-world generalization:

- Based on 15,000 real room impulse responses
- 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



Mixtures include interruptions, and rapid speaker turns

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

Task and Scientific Goals

Task: To generate time-marked, speaker-attributed transcripts using a single device - as typically found in conference room setups.

Scientific Goals:

1. Unlock the potential of data-driven methods by providing high-quality training and benchmarking datasets
2. We collaborate with the CHiME-8 DASR challenge to offer a fundamental system design comparison:
   Single-channel (NOTSOFAR) vs.
   Geometry-specific multi-channel (NOTSOFAR) vs.
   Geometry-agnostic multi-channel (DASR)

Challenge Tracks

NOTSOFAR-1 features two main tracks, centered on the use of a single device:

1. SC speech transcription, with recordings from various commercial SC devices (not single-microphone)
2. MC speech transcription, with recordings from devices utilizing a known circular geometry

Systems are ranked according to the tcpWER metric to evaluate both speaker diarization and word errors

Baseline System

NOTSOFAR-1 baseline, inspired by Yoshioka et al. [1] and Raj et al. [2], consists of the following three steps:

[1] T. Yoshioka et al., "Advances in online audio-visual meeting transcription," in Proc. 2019 IEEE ASRU Workshop.
[2] D. Raj et al., "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in Proc. 2021 IEEE spoken language technology workshop (SLT).