The GiN dataset

Group conversations in Noisy environments (GiN) – Multimedia recordings for location-aware speech enhancement

Emilie d'Olne[1], Jacob Donley[2], Alastair H. Moore[1], Vladimir Tourbabin[2], Patrick A. Naylor[1], Thomas Lunner[2]

Summary

- Over 2 h of group conversations in 3 rooms
- 7 channels of close-talking audio
- Head-pose data for every participant
- Audio from a 7-channel head-mounted array

Motivation

There has been a growing interest in the use of head-worn microphone arrays to solve the cocktail party problem for augmented reality research. However, there is a lack of relevant data for validation of algorithms. The EasyCom dataset was the first to introduce ego-centric multi-modal recordings of conversations in noise [1]. The SPEAR Challenge promoted this area of research by adapting EasyCom, and called in its conclusion for the collection of a wider variety of data [2, 3].

Dataset description

In each session, 6 participants sat around a table and held a conversation while restaurant noise was played through loudspeakers at $\sim$ 75 dB. Every participant wore a headset with a close-talking microphone and a pose tracker. 'Participant 3' wore binaural microphones and safety glasses mounted with 5 microphones. The recordings session were monitored by a 'waiter', also wearing a headset. Each session was recorded by a fixed reference microphone.

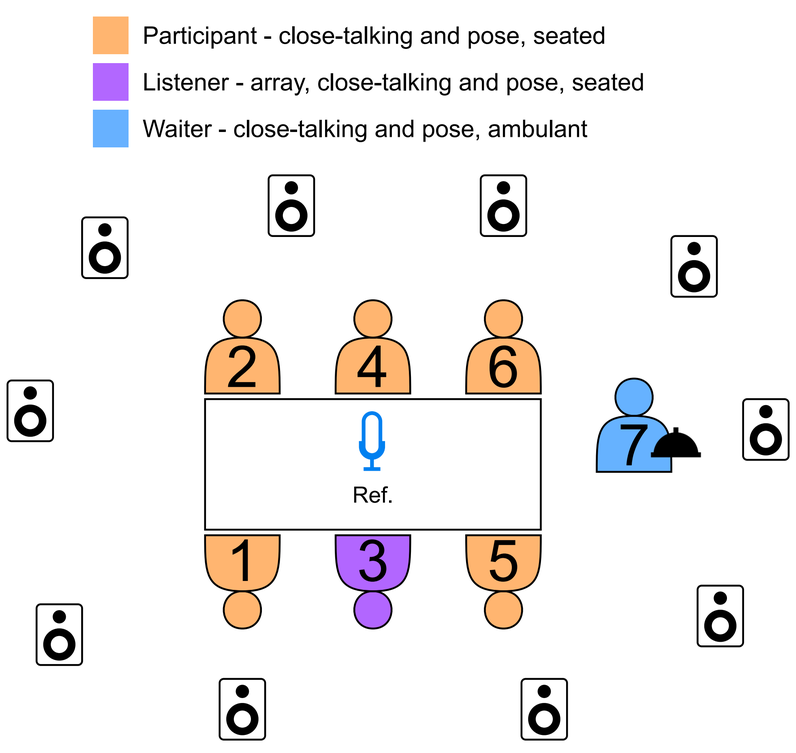
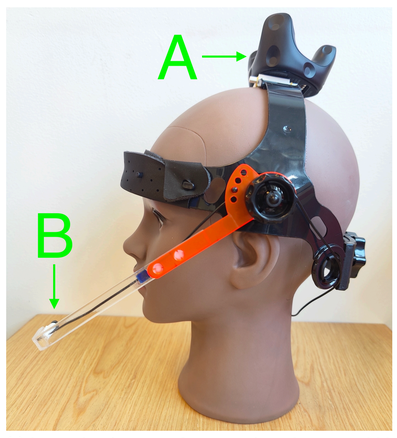
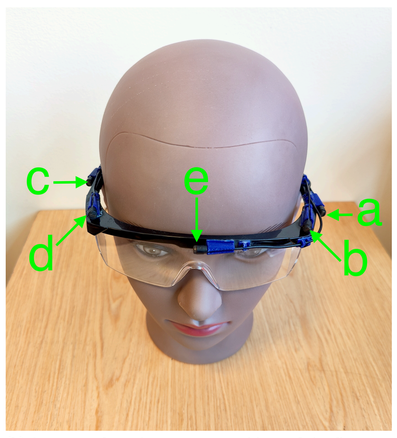


Figure 1: Diagram of the recording setup, not to scale. The 6 participants are shown with their respective speaker ID and the 'waiter' as number 7. Participant 3 is the 'listener', equipped with binaural microphones and custom glasses. The approximate loudspeaker locations are shown. The reference microphone is placed approximately at the centre of the table.



(a) Custom headset; A: HTC VIVE Tracker 3.0, B: DPA 4060 microphone



(b) Custom safety glasses; a-e: positions of the five DPA 4060 microphones

Figure 2: Custom equipment used for data acquisition. Note that the mannequin head in the pictures is smaller than a typical human head.

Rooms description

Sessions were recorded in 3 rooms of the Electrical and Electronic Engineering Department at Imperial College London.

Table 1: Characteristics of the rooms in the dataset.

Room	Size (m)	RT60 (ms)	Description
807	$4.8 \times 10.0 \times 2.9$	580	Eating area
'Faraday'	$5.4 \times 9.9 \times 2.9$	970	Eating area
403	$9.3 \times 20.4 \times 2.9$	1260	Cafeteria

Session details

Sessions were divided into 4 tasks of approximately 5 min each and containing different levels of overlapping speech.

1. Reading task, one speaker at a time
2. Group task, 6 speakers, potential overlap
3. Pair task, 3 simultaneous conversations between pairs, almost continuous overlap
4. Group task, 6 speakers with clear turn taking, limited overlap

Additional information

The head pose ($f_s = 90$ Hz) and microphone data ($f_s = 48$ kHz) were clock-synchronised using the Lab Streaming Layer [4].

Automatic voice activity labels are also provided in the dataset, based on Silero [5].

Anechoic acoustic impulse responses are provided for the head-mounted array.

Use for speech enhancement

The dataset was validated for speech enhancement using a fixed MVDR beamformer applied to the head-mounted array.

Signal model

Expressing the array in the STFT domain as $\mathbf{x}^{\mathsf{A}}(k,\ell) \in \mathbb{C}^{7\times 1}$, the beamformer output is

$$y(k,\ell) = \mathbf{w}^H(k,\ell)\mathbf{x}^{\mathsf{A}}(k,\ell),$$

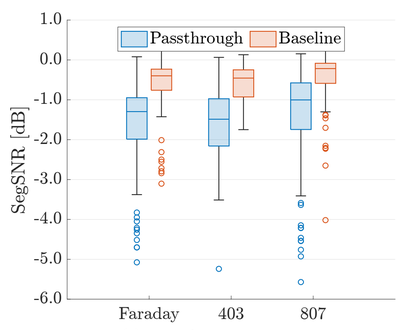
with

$$\mathbf{w}(k,\ell) = \frac{\mathbf{R}^{-1}(k)\mathbf{d}(\phi(\ell),\theta(\ell),k)}{\mathbf{d}^H(\phi(\ell),\theta(\ell),k)\mathbf{R}^{-1}(k)\mathbf{d}(\phi(\ell),\theta(\ell),k)},$$

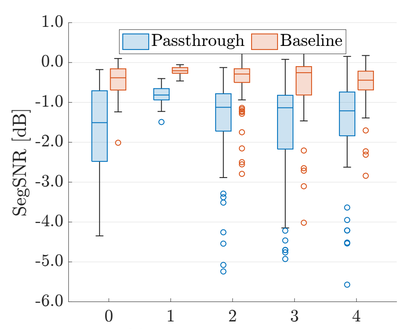
- $\mathbf{R}(k)$ is the NCM for a stationary cylindrically isotropic noise field
- $\mathbf{d}(\phi(\ell),\theta(\ell),k)$ is obtained using head-pose data and the anechoic AIRs

Enhancement results

'Participant 4' was selected as target speaker and metrics were computed only in segments of voice activity determined by the VAD labels.



(a) Dependence on room



(b) Dependence on task

Figure 3: Selected intrusive metrics for unprocessed ('Passthrough') and processed ('Baseline') speech.

Results across rooms and tasks show that

- The baseline beamformer improves metrics over the passthrough in all cases
- Room '807' has overall higher metrics and therefore is a comparatively easier environment, which is consistent with it being the least reverberant room
- Similarly, Task 1, the reading task, is easier to listen to than other tasks, which is consistent with the absence of overlapping speech

Availability

The dataset is made publicly available by Imperial College London, together with useful functions for speech enhancement.

Affiliations

[1] Electrical and Electronic Engineering Department, Imperial College London
[2] Meta Reality Labs Research, Redmond, USA

References

[1] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, "EasyCom: An Augmented Reality Dataset to Support Algorithms for Easy Communication in Noisy Environments," arXiv:2107.04174 [cs, eess], Oct. 2021.

[2] P. Guiraud, S. Hafezi, P. A. Naylor, A. H. Moore, J. Donley, V. Tourbabin, and T. Lunner, "An introduction to the speech enhancement for augmented reality (SPEAR) challenge," in Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC), 2022, pp. 1–5.

[3] V. Tourbabin, P. Guiraud, S. Hafezi, P. A. Naylor, A. H. Moore, J. Donley, and T. Lunner, "The SPEAR challenge - Review of results," in Proc Forum Acusticum, Sep. 2023.

[4] C. Kothe, "Lab Streaming Layer (LSL) - A software framework for synchronizing a large array of data collection and stimulation devices," 2014. [Online]. Available: https://github.com/sccn/labstreaminglayer/

[5] Silero, "Silero VAD: Pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier," 2021. [Online]. Available: https://github.com/snakers4/silero-vad

Imperial College London

emilie.dolne16@imperial.ac.uk