IEEE Open Journal of
Signal Processing

Received 14 September 2023; revised 3 December 2023; accepted 5 December 2023. Date of publication 19 December 2023; date of current version 23 January 2024. The review of this article was arranged by Associate Editor Z. Duan.

Digital Object Identifier 10.1109/OJSP.2023.3344379

Group Conversations in Noisy Environments (GiN) – Multimedia Recordings for Location-Aware Speech Enhancement

EMILIE D'OLNE [1] (Student Member, IEEE), ALASTAIR H. MOORE [1], PATRICK A. NAYLOR [1] (Fellow, IEEE), JACOB DONLEY [2], VLADIMIR TOURBABIN [2], AND THOMAS LUNNER[2]

[1]Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K.
[2]Meta Reality Labs Research, Redmond, WA 98052 USA

CORRESPONDING AUTHOR: EMILIE D'OLNE (e-mail: emilie.dolne16@imperial.ac.uk).

This work involved human subjects or animals in its research. The author(s) confirm(s) that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Recent years have seen a growing interest in the use of smart glasses mounted with microphones to solve the cocktail party problem using beamforming techniques or machine learning. Many such approaches could bring substantial advances in hearing aid or Augmented Reality (AR) research. To validate these methods, the EasyCom [Donley et al., 2021] dataset introduced high-quality multi-modal recordings of conversations in noise, including egocentric multi-channel microphone array audio, speech source pose, and headset microphone audio. While providing comprehensive data, EasyCom lacks diversity in the acoustic environments considered and the degree of overlapping speech in conversations. This work therefore presents the Group in Noise (GiN) dataset of over 2 hours of group conversations in noisy environments recorded using binaural microphones and a pair of glasses mounted with 5 microphones. The recordings took place in 3 rooms and contain 6 seated participants as well as a standing facilitator. The data also include close-talking microphone audio and head-pose data for each speaker, an audio channel from a fixed reference microphone, and automatically annotated speaker activity information. A baseline method is used to demonstrate the use of the data for speech enhancement. The dataset is publicly available in d'Olne et al. [2023].

INDEX TERMS Augmented reality (AR), cocktail party, dataset, head-worn array, speech enhancement.

I. INTRODUCTION

The cocktail party problem [3], or the enhancement of speech in noisy environments, has been a focal point of speech processing research for the better part of a century. It finds applications in hearing aids research [4], [5], Automatic Speech Recognition (ASR) [6], [7], or, more recently, AR and Virtual Reality (VR) [8], [9]. The problem can typically be separated into two components: first, a speaker tracking task in which a speaker of interest is identified and localised in space; and second, a speech enhancement task to improve the intelligibility of said speaker of interest. For a long time, researchers have focussed on solving these issues using single microphones or small aperture microphone arrays such as hearing aids. However, recent technological advances as well as commercial trends have opened the door to the use of head-worn devices to solve the cocktail party problem [10], [11], [12]. In particular, the use of smart glasses, or glasses mounted with microphones, has generated a lot of interest and is especially relevant for AR/VR research as the design is analogous to most commercially available systems.

Early work in [13] examined the use of glasses-mounted microphone arrays as a complementary system to conventional hearing aids. It was deemed a robust and simple solution which could reduce the level of ambient diffuse noise. An example of practical beamformer implementation was presented in [14], showcasing the potential of head-worn arrays for interference rejection. More advanced enhancement techniques were evaluated in [15] for the specific cocktail party problem and showed good noise suppression on realistic recordings. However, to truly evaluate the performance of such methods,

TABLE 1. Comparative Table of Existing Egocentric Datasets and GiN

Dataset	Year	# Mics	# Rooms	# Participants	# Hours	Natural speech	Pose
MYRiAD [17]	2023	11-24	2	10-15	76		
CHiME-5&6 [19], [20]	2018/19	32	20	4	60	✓	
Ego4D (Audio-Visual Benchmark) [21]	2022	1-12	100+	1-6	764	✓	
EasyCom [1]	2021	8-10	1	3-5 + 1	5	✓	✓
GiN	**2023**	**15**	**3**	**6 + 1**	**2**	**✓**	**✓**

they must be tested on real cocktail party data recorded using head-worn arrays. Such data must contain a variety of rooms with varying acoustical properties and should accurately represent the behaviour of participants in the scene. Indeed, people in loud environments have a natural tendency to raise their voices and use head and body movements to increase speech intelligibility [16].

Several relevant high-quality databases have been released in recent years, each with their own characteristics. The MYRiAD database [17] contains recordings of live cocktail parties and provides a large number of Room Impulse Responses (RIRs) recorded in two rooms using several microphone configurations. It includes in-ear and behind-the-ear microphones, but not wearable devices. The database described in [18] contains body-related Acoustic Impulse Responses (AIRs) for several devices, including glasses, placed on a human subject. While these AIRs allow for easy simulation of acoustic environments, they also restrict them to static scenes in anechoic rooms, which differ widely from real scenarios. CHiME-5 and CHiME-6 [19], [20] provide recordings of real dinner parties in private homes obtained using binaural microphones and fixed microphone arrays. Such databases better represent natural conversations but lack positional data which are essential for source localisation or traditional spatial-filtering enhancement methods. CHiME-5 and CHiME-6 also do not contain recordings made using wearable arrays such as smart glasses. The Ego4D dataset [21] contains a wide range of egocentric audio-visual recordings of daily activities collected by a large number of participants throughout the world. However, as the focus of the data is mainly on video, most audio recordings are limited to single- or dual-channel and tracking data are not systematically available.

The recent EasyCom database [1] contains egocentric audio-visual data from several group conversation in a single, artificially-created restaurant. It was recorded using AR glasses with 4 microphones and a camera, as well as binaural microphones. Clock-synchronised head-pose information and close-talking speech recordings are provided for all participants in the conversation. Participants in EasyCom were asked to introduce themselves, order food from a menu, play games, and read sentences while immersed in background noise created with loudspeakers. This led participants to raise their voices in order to be heard, as would occur in natural environments. The nature of the tasks, however, limits the amount of overlapping speech in EasyCom. Moreover, the artificial restaurant used in EasyCom was created using acoustic panels to approximate the reverberation time of a typical occupied restaurant [22] which is unlikely to match the acoustic fingerprint of real environments.

EasyCom offers comprehensive high-quality data but is limited in the diversity of scenes considered and the amount of overlapping speech. The SPEAR Challenge [23], [24] overcame this issue by digitally recreating and augmenting the EasyCom dataset using the TASCAR software [25]. The SPEAR data contain more acoustic environments and different conversational setups to EasyCom. While this approach presents a balance between purely simulated data and realistic recordings, it cannot fully replace the acquisition of new conversational data in real rooms.

This work thus aims to extend EasyCom to a wider range of acoustic environments and conversations. The GiN recordings presented herein contain approximately 2 hours of group conversations in three noisy reverberant rooms, as opposed to the single room in EasyCom [1]. Every participant's speech and pose were recorded using close-talking microphones and head trackers mounted on a custom headset. One of the authors took part in the recordings as a facilitator and, contrarily to EasyCom [1], their head pose and speech were also recorded. Furthermore, the data contain egocentric recordings made using binaural microphones and a glasses-mounted microphone array for one of the participants. Diffuse background noise was generated using 10 loudspeakers playing restaurant noise at approximately 75 dB, similarly to EasyCom [1]. Participants were instructed to perform tasks that led to a wide variety of conversational scenarios, with periods of single-speech and period of large amounts of overlapping speech. Table 1 summarises the difference and similarities between GiN and existing egocentric datasets.

The remainder of this paper is structured as follows. Section II describes the dataset creation process, with detailed information regarding the acquisition system, methodology, and post-processing steps applied to the data. Section III provides practical information on the data availability. Section IV investigates speech enhancement as a practical application for the data and presents a baseline beamformer whose performance is then discussed in Section V. Finally, Section VI presents the main conclusions of this work.

II. DATASET DESCRIPTION

This dataset contains clock-synchronised multimedia recordings of conversations in noisy restaurant-like scenarios. It was recorded in three rooms in the Department of Electrical

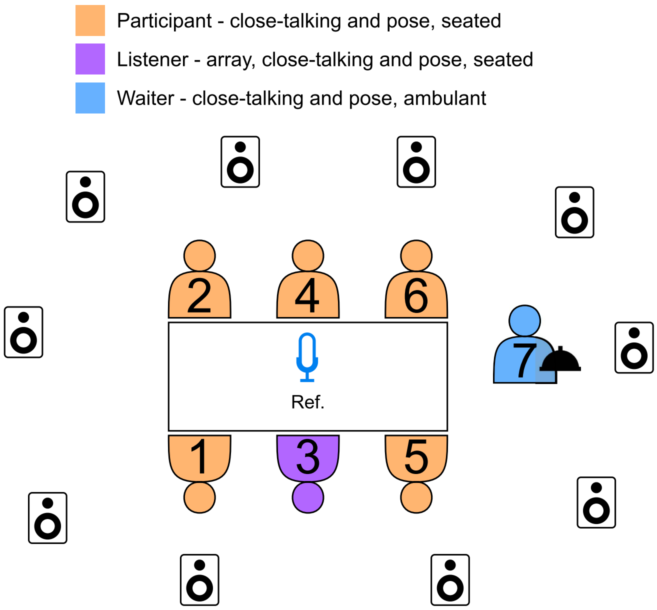
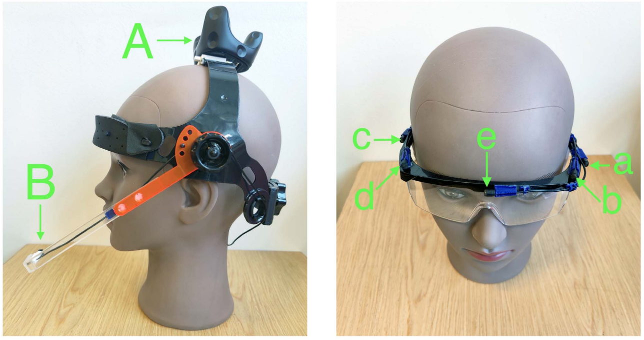

🟧 Participant - close-talking and pose, seated
🟪 Listener - array, close-talking and pose, seated
🟦 Waiter - close-talking and pose, ambulant
2
4
6
Ref.
7
1
3
5

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 coordinate system origin is located behind and to the left of participant 1 (bottom left corner in this view). 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.

TABLE 2. Characteristics of the Rooms in the Dataset[1]

Room	Dimensions (m)	RT60 (ms)	Description
807	$4.8 \times 10.0 \times 2.9$	580	Common eating area
'Faraday'	$5.4 \times 9.9 \times 2.9$	970	Common eating area
403	$9.3 \times 20.4 \times 2.9$	1260	Cafeteria-like room

The reverberation times (RT60) were obtained using the Schroeder method [26] on RIRs measured with pyrirtool[1].

and Electronic Engineering at Imperial College London. In every recording session, 6 participants were seated around a table and held a conversation while noise was played through loudspeakers at a level typically found in restaurants [22], as pictured in Fig. 1. Every participant wore a custom headset equipped with a close-talking microphone and a pose tracker, as shown in Fig. 2(a). In addition, 'Participant 3' wore binaural microphones and a pair of safety glasses mounted with 5 microphones to provide egocentric audio data, as shown in Fig. 2(b). The recordings session were monitored by one of the authors acting as a 'waiter' or 'facilitator', also wearing a custom headset. Each session was recorded by a fixed reference microphone placed in the centre of the table.

This section provides an overview of the data recorded in this work. The methodology used to select rooms and participants as well as the recording sessions details are first described in Section II-A. Then, a detailed description of the data acquisition system is given in Section II-B. Finally, the data post-processing stages are presented in Section II-C.

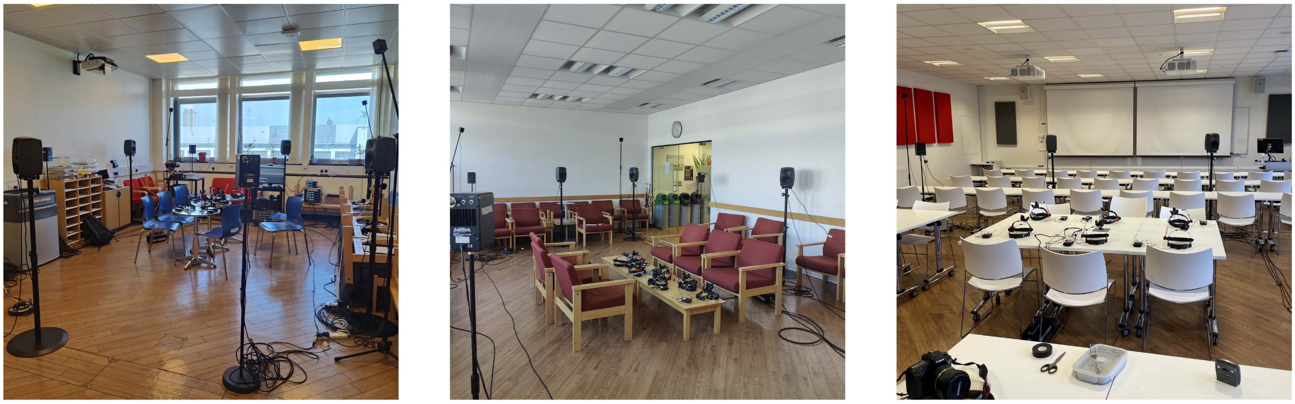
A. METHODOLOGY AND RECORDING PROTOCOL

This section describes the room and participant selection process as well as the recording protocol.

1) ROOM SELECTION AND DESCRIPTION

Three rooms in the Department of Electrical and Electronic Engineering at Imperial College London were selected to carry out the recordings. The rooms are pictured in Fig. 3 and have characteristics summarised in Table 2. The first two rooms, *Room 807* and *Room 'Faraday'* are popular breakout areas, commonly used for eating lunch, and as such were deemed suitable to emulate a restaurant scenario. The third room, *Room 403*, is a large, reverberant teaching space which was selected for its resemblance to a typical cafeteria. Prior to each session, the rooms were prepared to allow for easy access and movement around the recording area. Furniture was kept in the room but rearranged as to allow for a large table to fit comfortably in the approximate centre of the space. The tracking system described in Section II-B1 was set-up with each of the 4 Valve Index Base Stations placed in a corner of the room at a height of approximately 2 m and facing downwards at a small angle in order to fully capture the scene. The Valve Index Head-mounted Display (HMD) was placed on a mannequin head in the lower left corner of the room at a height of 103 cm and serves as the origin of the tracking coordinate system. The 10 loudspeakers were placed in a circle around the table with a radius constrained by the geometry of the rooms and with random orientations; the loudspeaker positions are provided in the database. Finally, a reference microphone was placed near the centre of the table. Fig. 1 shows a schematic view of the arrangement of participants and loudspeakers.

[1] [Online]. Available: https://github.com/maj4e/pyrirtool



(**a**) Room 807

(**b**) Room 'Faraday'

(**c**) Room 403

FIGURE 3. Rooms at Imperial College London included in the database.

2) PARTICIPANT SELECTION AND INSTRUCTIONS

Participants were recruited using convenience sampling on the Imperial College London campus and through word-of-mouth. In total, 32 participants took part in the recordings, with a mixture of native and non-native English speakers and with some participants knowing each other beforehand. Whenever possible, two sessions were recorded in the same room with different groups of participants, giving 12 unique speakers per room. Some participants took part in multiple recordings, each time in a different room. Before every session, participants were asked to review an information sheet detailing the objectives of the recordings and were given the opportunity to ask questions. They were prompted to act as naturally as possible during the recordings. Each participant was fitted with a custom headset, containing a close-talking microphone and an HTC VIVE Tracker. The corresponding author was the waiter in all sessions and also wore a custom headset, such that their speech and motion is also included in the database. In each session, one participant was picked to act as the listener and was additionally fitted with the binaural microphones and the glasses-mounted microphone array.

3) SESSION DETAILS

Every session began with a calibration phase, recorded in quiet. Each participant was first asked to read the sentence '*Polly Piper picked a peck of pickled peppers*' to detect potential audio clipping during plosives. The positions of the close-talking microphones were adjusted whenever necessary. Then, participants were asked to move their heads according to a predetermined pattern (forwards, left, right, forwards, up, down, forwards) to identify the axes of rotation of the participants' heads and confirm the correct operation of the tracking system in post-processing. The calibration sessions are included in the database.

The data recording session then started, coinciding with noise playback from the loudspeakers. Each recording session was comprised of four distinct tasks, labelled 1–4. Instructions for each task were delivered verbally by the waiter as part of the recording and participants were able to ask questions. These waiter interaction phases are labelled *Task 0*. In *Task 1*, participants were asked to read in turn a snippet from *Alice in Wonderland* [27]. Silent participants were asked to listen in a natural fashion without forced head orientation. *Task 2* was a group exercise in which the participants were asked to suggest and collectively agree on a suitable holiday destination. Each participant was given a secret constraint, e.g. '*The country must start with the letter A*', to inform their choices but were asked to not divulge the constraint itself to other participants. In *Task 3*, the participants played the game '*I spy*', in pairs with the person opposite them, such that the recording contained 3 simultaneous conversations. The participants were asked to refrain from using the words '*yes*' and '*no*' to avoid single word answers and maintain a balanced interaction. Finally, in *Task 4*, the participants were asked to play the card game '*Go Fish*' as a group until the end of the recording.

Each task was designed to last 5 minutes but the nature of the tasks means that individual sessions exhibit variations. The recordings ended after approximately 24 minutes.

B. DATA ACQUISITION METHODS

The data acquisition system used in these recordings is described in this section and summarised by the diagram in Fig. 4. The main challenge consisted in the synchronisation of three components: the tracking system, the audio recording system, and the audio playback system. This was achieved using the LSL software [28], [29]. Moreover, a custom headset, shown in Fig. 2(a), was designed to facilitate the wireless acquisition of head tracking and close-talking speech data.

1) TRACKING SYSTEM

The head orientations and positions of every speaker in the recordings were tracked using HTC VIVE Trackers 3.0

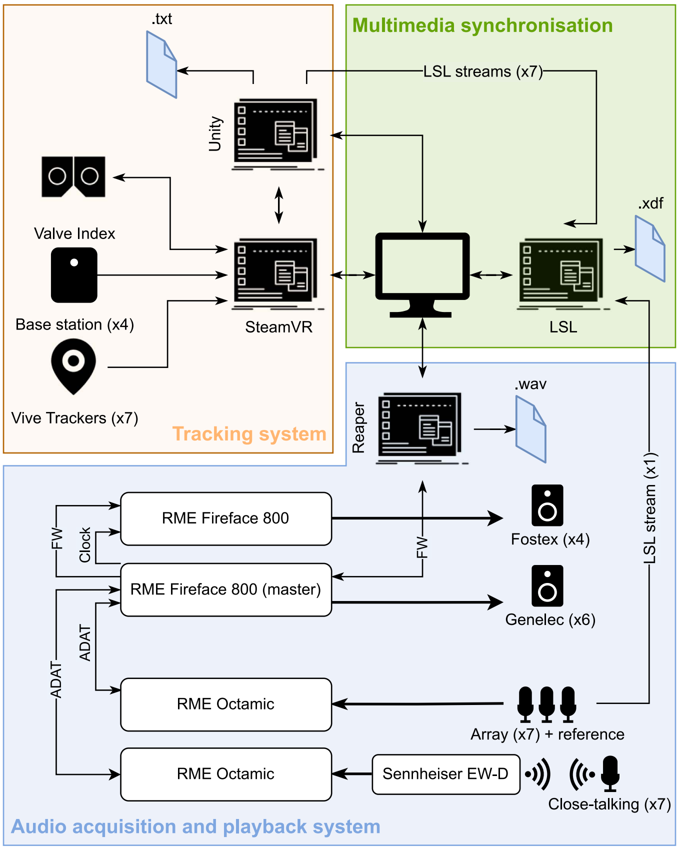


FIGURE 4. Data acquisition system diagram. The data were captured using a PC running Windows 10. Pose data were stored in .txt files, microphone recordings were stored in .wav files, and synchronisation data were stored in a .xdf file following the Lab Streaming Layer (LSL) convention.

mounted on custom headsets. The VIVE Trackers are designed to track body-movement for VR and are accurate to within a few centimetres [30]. The trackers were paired with a Valve Index HMD and 4 Valve Index Base Stations through the Steam VR system. The HMD was used to fix the origin of the tracking coordinate system, and data was acquired from the trackers at a rate of 90 Hz using a Unity script adapted from [31]. Compared to 'outside-in' camera-based tracking systems, the apparatus is sufficiently portable that recordings could feasibly be made in several rooms. Compared to 'inside-out' systems, using VIVE Trackers enabled head pose tracking data to be obtained for all participants without anyone having to wear an HMD and without risk of occlusion. A few samples of head orientation or position were determined to be unreliable during the recordings, either due to sensor occlusion or connectivity issues with the Steam VR system. These samples were discarded and appropriately labelled as missing data points.

2) AUDIO CAPTURE AND PLAYBACK

In each session, a total of 15 audio channels were recorded using the Reaper software [32] at a sampling rate of 48 kHz. All microphones signals were recorded using two RME OctaMic preamplifiers, clock-synchronised through a primary RME FireFace 800 Firewire Audio Interface. The custom headset microphones for the 6 participants and the waiter were DPA 4060 miniature omni-directional condenser microphones. To allow for natural participant movement, these were recorded wirelessly using seven Sennheiser EW-D receivers and body-packs. In each session, 'Participant 3' acted as the listener and was fitted with DPA 4560 binaural microphones and custom glasses mounted with 5 DPA 4060 microphones, shown in Fig. 2(b). Finally, a single DPA 4060 microphone was mounted on a 10 cm stand in the centre of the table to act as a static reference microphone.

Loudspeaker playback was also managed by the Reaper software. A single-channel recording of a natural restaurant environment was used to generate 10 channels of uncorrelated noise. This noise was played through 10 loudspeakers connected to the balanced outputs of a secondary RME FireFace 800 Firewire audio interface, synchronised with the primary interface also used for audio capture. The 10 loudspeakers consisted of 6 Genelec 8030 and 4 Fostex 6301B, and the level of the noise was designed to approximate 75 dB to match the level typically found in restaurants [22].

3) MULTIMEDIA SYNCHRONISATION

The synchronisation of audio recordings and pose data was achieved using the LSL software [28], [29]. LSL is an open-source platform designed for synchronised multimedia signal acquisition. The `LabRecorder`[2] module was used to capture independent streams from the audio capture and tracking systems. These can be exposed over an IP network but, in this case, they were all generated from the same PC. Each LSL stream includes both the stream's own timestamps, obtained from the local clock, and timestamps generated from the recorder's clock. The `xdf-Matlab`[3] utility was used to obtain synchronised multi-modal data. For these recordings, the `LSL4Unity`[4] plug-in created a stream with the positional data from the tracking system and the `AudioCapture`[5] software was used to capture the static reference microphone channel. The multichannel audio recorded in Reaper was subsequently time-aligned using `sigalign` from Voicebox [33].

C. POST-PROCESSING AND ADDITIONAL DATA

The head-pose data were obtained using Unity and therefore obeyed a left-handed, y-up coordinate system. To facilitate the use of the data for audio processing tasks, the pose data were converted to the right-handed, z-up system found in TASCAR [25] using Voicebox [33].

[2] [Online]. Available: https://github.com/labstreaminglayer/App-LabRecorder

[3] [Online]. Available: https://github.com/xdf-modules/xdf-Matlab

[4] [Online]. Available: https://github.com/labstreaminglayer/LSL4Unity

[5] [Online]. Available: https://github.com/labstreaminglayer/App-AudioCapture

Additionally, Voice Activity Detection (VAD) labels were obtained for the close-talking microphones using Silero [34]. Segments of voice activity were identified for each of the 7 channels independently, such that the ith close-talking channel $x_i[n]$ is associated with a voice activity mask $V_i[n]$, with n the time index. However, the VAD in a given channel sometimes identified voice activity from nearby speakers, e.g. voice activity from Speaker 2 was detected in the close-talking channel associated with Speaker 1. This can be explained by the close physical proximity between participants and the frame-based approach used in Silero which ignores that such cross-talk occurs at a much lower energy than speech from the principal speaker. Therefore, a cross-channel energy thresholding approach was used to refine the VAD. For every close-talking channel $x_i[n]$ and for every segment of voice activity $\mathbf{l}$ detected in channel $x_i[n]$, the segment was discarded if channel $x_i[n]$ contained considerably less energy than any other channel, or, equivalently

$$V_i[\mathbf{l}] = \begin{cases} \mathbf{1} & \text{if } E\{x_i[\mathbf{l}]\} > \alpha\, E\{x_j[\mathbf{l}]\} \;\; \forall j \\ \mathbf{0} & \text{otherwise,} \end{cases} \tag{1}$$

with $0 \leq \alpha \leq 1$, $E\{x_i[\mathbf{l}]\} = \sum_{\mathbf{l}} |x_i[\mathbf{l}]|^2$ the energy in the ith channel over the segment $\mathbf{l}$, and α a tuning parameter which was empirically set to 0.2. A small α leads to more cross-talk being admitted in the the voice activity mask while a large α may lead to segments being discarded when multiple speakers are simultaneously active. The time regions associated with each of tasks presented in II-A were manually identified and included as labels in VAD segments.

The directional characteristics of the microphone array were measured in an anechoic room using the head-related transfer measurement system of the Audio Experience Design team at Imperial College London, described in detail in [35]. Briefly, the AIRs were measured from an arc of loudspeakers to the glasses array and binaural microphones whilst being worn by the second author. The vertical spacing of the speakers is 10° for angles within 30° of the horizontal plane and 15° elsewhere. The horizontal density of measurements is determined by rotation of the subject, which was stepped in 5° increments. Two versions of the measured impulse responses are included in the database, differing only in how tightly cropped they are. Tighter cropping has fewer room reflections and was used for the beamformer design described below. However, some applications may benefit from the longer responses.

III. DATASET AVAILABILITY

Imperial College London has released the GiN database in [2]. It contains 2 hours of synchronised microphone and pose data as well as automatically generated VAD information. To facilitate the use of the data, sessions were divided into 1 minute segments similar to EasyCom [1]. Segments containing personal information regarding the participants were removed, leading to some files lasting less than 1 minute. The total size of the database is approximately 14GB with 2803 files. The pose data is given in the right-hand z-up coordinate system, as used in TASCAR [25]. The position of loudspeakers and AIRs for the glasses array are also provided. Tools to visualise the data and modify the VAD are provided.[6]

IV. APPLICATION: SPEECH ENHANCEMENT

The dataset presented in this work was recorded primarily for the purposes of evaluating multichannel speech enhancement algorithms. To confirm the use of the presented data for such purposes, a baseline beamforming algorithm is introduced and applied to the 7-channel head-worn array, composed of binaural and glasses-mounted microphones. Given the methodology described in Section II, the listener may choose to attend to any of the conversation participants at any point during the recordings. Furthermore, the considerable level of restaurant noise in the scenes requires a baseline method capable of achieving high noise reduction. Therefore, a Minimum Variance Distortionless Response (MVDR) beamformer assuming a stationary cylindrically isotropic noise field is proposed as the baseline enhancement method. A cylindrically isotropic noise field assumption is appropriate here as rooms typically have more reflective walls than floors or ceilings [36]. This is a similar baseline to the one presented in EasyCom [1]. Letting $\mathbf{x}^{\mathrm{A}}(k, \ell) = [x_1^{\mathrm{A}}(k, \ell), x_2^{\mathrm{A}}(k, \ell), \ldots, x_7^{\mathrm{A}}(k, \ell)]^T$ be the 7-channel head-worn array expressed in the Short-time Fourier Transform (STFT) domain with k and ℓ the frequency- and time-frame indices, the enhanced single-channel beamformer output $y(k, \ell)$ can be obtained using

$$y(k, \ell) = \mathbf{w}^H(k, \ell)\mathbf{x}^{\mathrm{A}}(k, \ell), \tag{2}$$

with $\mathbf{w}(k, \ell) \in \mathbb{C}^{7\times 1}$ the beamformer weights and $(\cdot)^H$ denoting the Hermitian transpose. Under the baseline assumption of a stationary cylindrically-isotropic noise field, the MVDR beamformer weights are defined by

$$\mathbf{w}(k, \ell) = \frac{\mathbf{R}^{-1}(k)\mathbf{d}(\phi(\ell), \theta(\ell), k)}{\mathbf{d}^H(\phi(\ell), \theta(\ell), k)\mathbf{R}^{-1}(k)\mathbf{d}(\phi(\ell), \theta(\ell), k)}, \tag{3}$$

where $\mathbf{R}(k) \in \mathbb{C}^{7\times 7}$ is the Noise Covariance Matrix (NCM) for a stationary cylindrically isotropic noise field, and $\mathbf{d}(\phi(\ell), \theta(\ell), k) \in \mathbb{C}^{7\times 1}$ are the Acoustic Transfer Functions (ATFs) from a look direction, defined by the azimuth and elevation angles $\phi(\ell)$ and $\theta(\ell)$, to the 7-channel head-worn array. The NCM assumes that the noise arrives at the array with equal power from all azimuthal directions at 0 elevation, and can therefore be obtained in theory as [37]

$$\mathbf{R}(k) = \frac{1}{2\pi}\int_0^{2\pi} \mathbf{d}(\phi, 0, k)\mathbf{d}^H(\phi, 0, k)d\phi. \tag{4}$$

In practice, however, only a discrete number of ATFs, $\mathbf{d}_{\mathrm{ATF}}(\phi_q, 0^\circ, k)$ are available from a discrete set of Q azimuth angles, ϕ_q, at elevation $\theta = 0^\circ$, as described in Section II. The

[6] [Online]. Available: https://github.com/ImperialCollegeLondon/sap-ic-gin

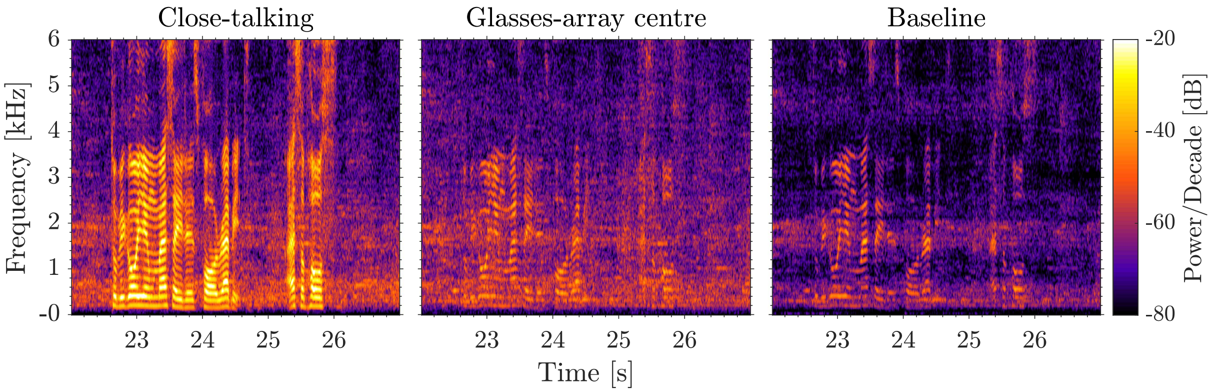


FIGURE 5. Spectrograms for a sample sentence from Participant 4, as measured by their close-talking microphone, the right-ear binaural microphone, and as enhanced using the baseline beamformer.

NCM must therefore be approximated using

$$\hat{\mathbf{R}}(k) = \frac{1}{Q}\sum_{q} \mathbf{d}_{\mathrm{ATF}}(\phi_q, 0^\circ, k)\mathbf{d}^H_{\mathrm{ATF}}(\phi_q, 0^\circ, k). \tag{5}$$

Note from (3) that the baseline beamformer weights are fully described by a stationary NCM and an ATF whose direction is determined by time-varying azimuth and elevation angles, $\phi(\ell)$ and $\theta(\ell)$. These angles are calculated using the head-pose information available in the database as the relative angle between a speaker of interest and the head-worn array, taking into account that the head-worn array is itself rotated. The measured ATF with the angle closest to the calculated relative angle is then used to steer the beamformer.

The baseline method is applied to the head-worn array recordings frame-by-frame using a STFT with 16 ms frames and 50% overlap. Any missing pose data are set to retain the value held in the previous head-tracking frame, and the samples are linearly interpolated to match the audio sampling rate of 48 kHz. The speaker of interest is chosen to be Participant 4, i.e. the speaker sitting directly across the table from the listener. As a qualitative example of the baseline performance, Fig. 5 shows three example spectrograms obtained for a sample sentence from Participant 4. The first spectrogram was obtained using Participant 4's close-talking microphone and exhibits clear speech characteristics. However, a non-negligible level of background noise can also be seen, e.g. for $n > 26$ s. The signal measured at the centre of the glasses array shows that whereas Participant 4's speech is still visible, it is at a lower level relative to the background noise. Finally, the spectrogram of the baseline enhancement shows that the energy of the background noise is reduced while preserving the energy of Participant 4's speech.

V. BASELINE RESULTS AND DISCUSSION

To objectively evaluate the performance of the baseline method, intrusive measures pertaining to signal-to-noise ratio, speech intelligibility and speech quality, were computed.

It should be noted that enhancement performance strongly depends on the selected target. For example, focusing on the 'waiter' is likely to be challenging, as the target is far away from the glasses-mounted array and is moving across the room. It will also depend on which other participants are producing competing speech. In all tasks, if Participant 4 is talking then they are a legitimate 'target' for Participant 3, the array wearer, to be listening to. Thus, in this evaluation Participant 4 was selected to be the speaker of interest.

Participant 4's close-talking microphone was used as the clean reference signal for intrusive metrics. As shown in Fig. 5, this signal contains a non-negligible amount of background noise, and is susceptible to capture cross-talk from other participants. While it cannot strictly be considered a clean reference, it captures Participant 4's speech at a level close to the highest achievable SNR in real scenarios. The microphone boom in Fig. 2(a) was specifically designed to adjust to different participants and capture their speech reliably. The issue of leaky reference is well-known in recordings of real scenarios and is also present in EasyCom [1]. It can be mitigated by using loudspeaker playback instead of live speech in recordings, at the cost of naturalness of participant interactions. The reference could also be denoised using various techniques such as in the SPEAR challenge [23], [24] but this may introduce unwanted distortions in the reference signal. Therefore, in this work, in order to ensure that results accurately represent the baseline enhancement of the target speech, metrics were only computed during segments of voice-activity as determined using VAD labels described in Section II.

The noise reduction performance of the baseline was evaluated using segmental SNR (SegSNR) [33]. The SegSNR computes the power ratio of the desired speech signal to all other undesired signals, e.g. background noise or competing speakers, during segments where speech is active. A higher value of SegSNR indicates a higher level of target speech relative to noise. To measure target speech intelligibility, the Short-time Objective Intelligibility Measure (STOI) [38] was

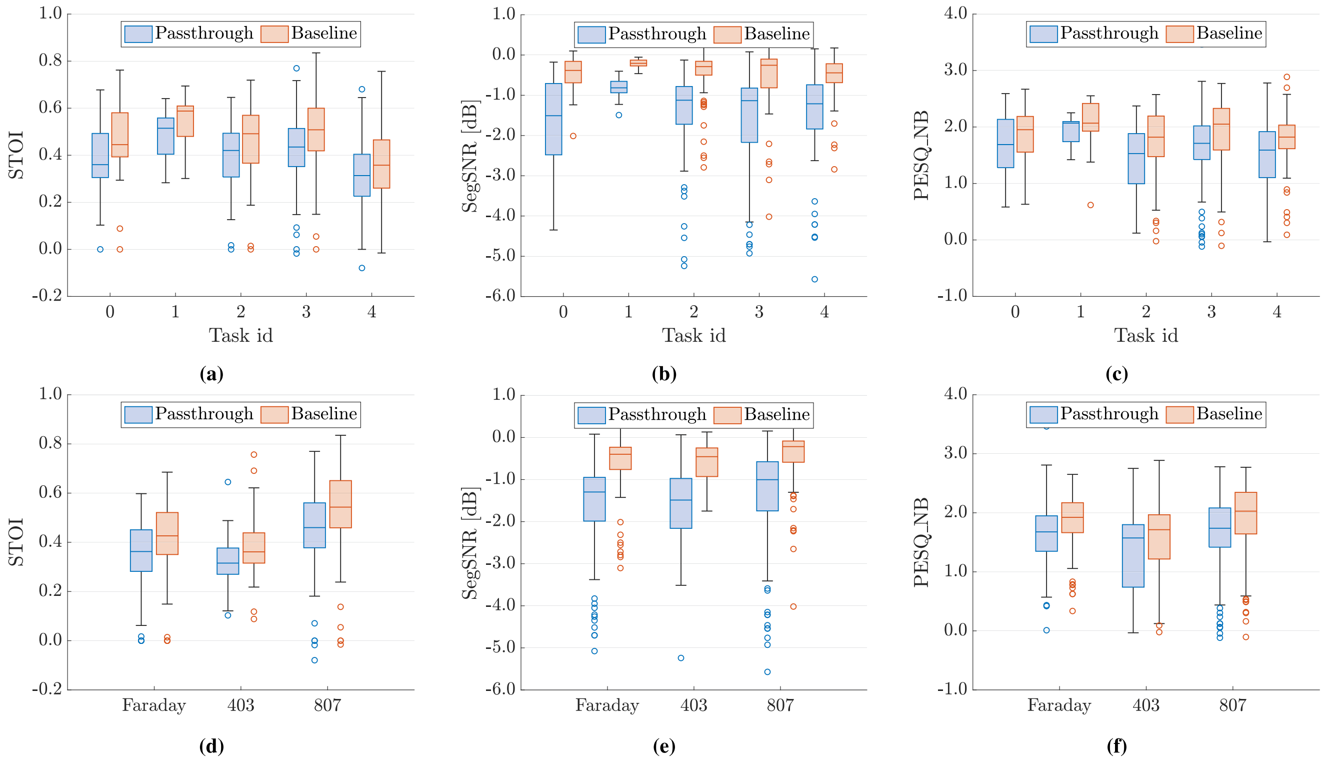


FIGURE 6. **Selected intrusive metrics for unprocessed ('Passthrough') and processed ('Baseline') speech. Metrics are (a), (d) STOI, (b), (d) Segmental SNR and (c), (f) narrowband PESQ. Top row (a), (b), (c) shows dependence on Task, bottom row (d), (e), (f) shows dependence on recording room.**

computed. STOI is calculated over short signal frames and yields values between 0 and 1 which highly correlate to perceived speech intelligibility. A higher value of STOI relates to a higher predicted intelligibility. Finally, speech quality was measured using the narrowband Perceptual Evaluation of Speech Quality (PESQ) [39]. PESQ is an industry standard designed to assess the quality of speech for telephone communication. It leads to scores between −0.5 and 4.5 and higher scores correspond to higher signal quality.

Fig. 6 shows the intrusive metrics computed as a function of task (top row) and room (bottom row) for both a passthrough signal and the baseline beamformer. The passthrough was selected as the right-ear binaural microphone. The passthrough metrics in blue boxes give a measure of the difficulty of listening without any enhancement. Comparing tasks, all metrics are higher for *Task 1*, which is consistent with the comparative simplicity of the scene, with one person reading aloud whilst others listens. It also exhibits the smallest variability in results, indicating that the noise field is approximately stationary. This also relates to the nature of the task that limited the number of competing sources and the frequency of listener head movements. Comparing rooms, metrics are slightly higher for '807', which is consistent with it being the least reverberant of the rooms. The variability of passthrough results across rooms further demonstrates the need to evaluate enhancement methods on real data recorded in multiple acoustic environments. Beamforming improves all metrics for all tasks and rooms as shown in red boxes. Informal listening confirmed that the baseline beamformer achieves a noticeable level of noise reduction.

The considerable level of background noise, the varying number and level of competing sources, and the highly-dynamic nature of the recorded scenes make speech enhancement on this data very challenging. Non-intrusive metrics may be considered to circumvent the leaky reference issue [40]. Many steps could be taken to improve the baseline beamformer, including but not limited to: head-rotation aware adaptive noise-covariance estimation [41], steering direction smoothing to limit sharp changes in beamformer weights, or binaural beamforming to gain benefits from binaural unmasking. Other methods of enhancement could also be explored, such as subspace-based methods [42] or deep neural networks. Readers are also referred to the SPEAR Challenge [23], [24] for more approaches. Finally, this dataset can be exploited for other applications including sound-source localisation and tracking, source counting or speaker diarization. Additionally, using the close-talking microphone signal to aid labelling, the array data could be used for single- or multi-channel ASR in adverse environments.

VI. CONCLUSION

This work presented the GiN database of multimedia recordings for location-aware speech enhancement. It contains 2 hours of group conversations in restaurant noise in 3 different rooms. It was acquired using 7 close-talking microphones, a pair of binaural microphones, a 5-channel array mounted on a pair of glasses, and a reference microphone. Additionally, head position and orientation data are available for every participant in the group conversation. The dataset was successfully used to evaluate the efficacy of a baseline MVDR beamformer for speech enhancement using the binaural microphones and glasses-mounted array.

ACKNOWLEDGMENT

The authors would like to thank the AXD team, especially Aidan Hogg, for their assistance with making the AIR measurements. They would also like to thank Phil Jones for mechanical engineering support and all of the volunteers who participated in these recordings.

REFERENCES

[1] J. Donley et al., "EasyCom: An augmented reality dataset to support algorithms for easy communication in noisy environments," Oct. 2021, *arXiv:2107.04174*.

[2] E. d'Olne, A. H. Moore, P. A. Naylor, J. Donley, V. Tourbabin, and T. Lunner, "Group in noise (GiN) data - 2023," doi: 10.14469/hpc/13463.

[3] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, Sep. 1953.

[4] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica United Acustica*, vol. 86, pp. 117–128, 2000.

[5] A. Moore, P. Naylor, and M. Brookes, "Improving robustness of adaptive beamforming for hearing devices," in *Proc. Int. Symp. Auditory Audiological Res.*, 2019, vol. 7, pp. 305–316.

[6] Y.-m. Qian, C. Weng, X.-k. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Front. Inf. Technol. Electron. Eng.*, vol. 19, pp. 40–63, 2018.

[7] B. Shi, W.-N. Hsu, and A. Mohamed, "Robust self-supervised audio-visual speech recognition," 2022, *arXiv.2201.01763*.

[8] J. Fernandez, L. McCormack, P. Hyvärinen, A. Politis, and V. Pulkki, "Enhancing binaural rendering of head-worn microphone arrays through the use of adaptive spatial covariance matching," *J. Acoust. Soc. Amer.*, vol. 151, no. 4, pp. 2624–2635, 2022.

[9] H. Jiang, C. Murdock, and V. K. Ithapu, "Egocentric deep multi-channel audio-visual active speaker localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10544–10552.

[10] R. M. Corey, N. Tsuda, and A. C. Singer, "Acoustic impulse responses for wearable audio devices," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 216–220.

[11] E. d'Olne, A. H. Moore, and P. A. Naylor, "Model-based beamforming for wearable microphone arrays," in *Proc. Eur. Signal Process. Conf.*, 2021, pp. 1105–1109.

[12] K. Sekiguchi, A. A. Nugraha, Y. Du, Y. Bando, M. Fontaine, and K. Yoshii, "Direction-aware adaptive online neural speech enhancement with an augmented reality headset in real noisy conversational environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 9266–9273.

[13] W. Soede, A. J. Berkhout, and F. A. Bilsen, "Development of a directional hearing instrument based on array technology," *J. Acoust. Soc. Amer.*, vol. 94, no. 2, pp. 785–798, 1993.

[14] M. H. Anderson et al., "Towards mobile gaze-directed beamforming: A novel neuro-technology for hearing loss," in *Proc. Annu. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 5806–5809.

[15] S. Hafezi et al., "Subspace hybrid beamforming for head-worn microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[16] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Amer.*, vol. 26, no. 2, pp. 212–215, 1954.

[17] T. Dietzen, R. Ali, M. Taseska, and T. van Waterschoot, "MYRiAD: A multi-array room acoustic database," *EURASIP J. Audio, Speech, Music Process.*, vol. 2023, no. 1, pp. 1–14, 2023.

[18] R. M. Corey and A. C. Singer, "Motion-tolerant beamforming with deformable microphone arrays," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 115–119.

[19] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," Mar. 2018, *arXiv: 1803.10609*.

[20] S. Watanabe et al., "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," 2020, *arXiv:2004.09249*.

[21] K. Grauman et al., "Ego4D: Around the world in 3,000 hours of egocentric video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18995–19012.

[22] J. H. Rindel, "Acoustical capacity as a means of noise control in eating establishments," in *Proc. Baltic-Nord Acoust. Meeeting*, vol. 2429, 2012.

[23] P. Guiraud et al., "An introduction to the speech enhancement for augmented reality (SPEAR) challenge," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2022, pp. 1–5.

[24] V. Tourbabin et al., "The SPEAR challenge - review of results," in *Proc Forum Acusticum*, 2023.

[25] G. Grimm, J. Luberadzka, and V. Hohmann, "A toolbox for rendering virtual acoustic environments in the context of audiology," *Acta Acustica United Acustica*, vol. 105, no. 3, pp. 566–578, 2019.

[26] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *J. Acoust. Soc. Amer.*, vol. 43, no. 4, pp. 829–834, Apr. 1968.

[27] L. Carroll, "Alice's Adventures in Wonderland," *Project Gutenberg*, Jun. 2008. [Online]. Available: https://www.gutenberg.org/ebooks/11

[28] 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/

[29] Q. Wang, Q. Zhang, W. Sun, C. Boulay, K. Kim, and R. L. Barmaki, "A scoping review of the use of lab streaming layer framework in virtual and augmented reality research," in *Proc. IEEE Virtual Reality*, 2023, pp. 1–16.

[30] V. Holzwarth, J. Gisler, C. Hirt, and A. Kunz, "Comparing the accuracy and precision of Steamvr tracking 2.0 and oculus quest 2 in a room scale setup," in *Proc. Int. Conf. Virtual Augmented Reality Simul.*, 2021, pp. 42–46.

[31] E. Badier, "A simple library to use HTC vive tracker devices in unity," 2020. [Online]. Available: https://github.com/ebadier/ViveTrackers

[32] "Reaper - digital audio workstation," Cockos Inc, 2023. [Online]. Available: https://www.reaper.fm/

[33] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," 1997. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[34] 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

[35] I. Engel et al., "The SONICOM HRTF dataset," *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 241–253, 2023.

[36] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 1006–1018, Jun. 2015.

[37] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Amer.*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.

[38] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4214–4217.

[39] Int. Telecommun.Union (ITU-T), "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," (ITU-T), Geneva, Switzerland, Tech. Rep. Recommendation P.862, Nov. 2003.

[40] A. Kumar et al., "Torchaudio-Squim: Reference-less speech quality and intelligibility measures in torchaudio," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[41] A. H. Moore, W. Xue, P. A. Naylor, and M. Brookes, "Noise covariance matrix estimation for rotating microphone arrays," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 3, pp. 519–530, Mar. 2019.

[42] V. W. Neo, C. Evers, and P. A. Naylor, "Enhancement of noisy reverberant speech using polynomial matrix eigenvalue decomposition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3255–3266, 2021.