# HOMULA-RIR: A ROOM IMPULSE RESPONSE DATASET FOR TELECONFERENCING AND SPATIAL AUDIO APPLICATIONS ACQUIRED THROUGH HIGHER-ORDER MICROPHONES AND UNIFORM LINEAR MICROPHONE ARRAYS

*Federico Miotello, Paolo Ostan, Mirco Pezzoli, Luca Comanducci,*
*Alberto Bernardini, Fabio Antonacci, Augusto Sarti*

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy

## ABSTRACT

In this paper, we present HOMULA-RIR, a dataset of room impulse responses (RIRs) acquired using both higher-order microphones (HOMs) and a uniform linear array (ULA), in order to model a remote attendance teleconferencing scenario. Specifically, measurements were performed in a seminar room, where a 64-microphone ULA was used as a multichannel audio acquisition system in the proximity of the speakers, while HOMs were used to model 25 attendees actually present in the seminar room. The HOMs cover a wide area of the room, making the dataset suitable also for applications of virtual acoustics. Through the measurement of the reverberation time and clarity index, and sample applications such as source localization and separation we demonstrate the effectiveness of the HOMULA-RIR dataset.

*Index Terms*— sound field reconstruction, acoustic array processing, acoustic data set, room impulse response

## 1. INTRODUCTION

In recent years, teleconferencing platforms have become part of daily lives of most people, especially after the COVID-19 pandemic. Applications like source separation [1], speech enhancement [2, 3] or echo-canceling [4], dereverberation [5] or audio packet loss concealment [6–8] are customarily used in teleconferencing software. Moreover, with the growing interest within augmented and virtual reality contexts, an increasing number of platforms, including those for teleconferencing, are incorporating spatial audio features. To this end, tasks such as sound field reconstruction [9–11] have gained particular relevance, due to their crucial role in enabling applications like navigable audio. Nevertheless, to accurately assess the performance of these methods in real-world scenarios, it is necessary to test them using data measured in real environments. Additionally, more and more approaches nowadays heavily rely on machine learning or other data-driven algorithms, that thus need large amounts of data for training and validation. For these reasons, several Room Impulse Response (RIR) datasets are present in the literature. In [12] multichannel responses were measured in a room with variable reverberation levels, with the aim of evaluating source separation techniques, while in [13] RIRs in low-reverberation time rooms were measured to test sound zone control and reconstruc-
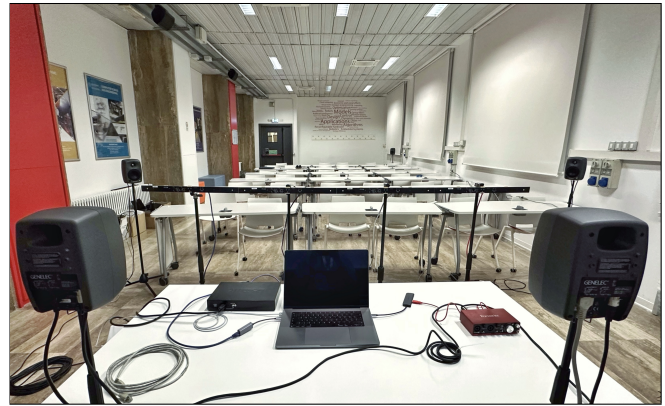


**Fig. 1**: Photograph of "Schiavoni room" taken during the measurement session.

tion methods. The Multi-arraY Room Acoustic Database (MYR-iAD) dataset [14] was created by acquiring measurements using several microphone configurations such as in-ear omnidirectional microphones in a dummy head, circular arrays and behind-the-ear arrays in two recording scenarios. Several RIR datasets acquired through higher-order microphones (HOMs) have also been released. In [15] measurements were acquired in three rooms, each with a static source position and more than one-hundred receivers, using both omnidirectional and B-format microphones. A dataset of six degrees-of-freedom RIRs in controlled and empty rooms, using different reverberation levels was presented in [16], while the Motus dataset [17] was created by acquiring Ambisonic [18] RIRs in a single room and varying the furniture position.

In this paper, we present HOMULA-RIR, a complementary dataset of RIRs acquired in a real environment using a hybrid setup, with the objective of representing a realistic teleconferencing scenario. Specifically, we have deployed a linear microphone array to simulate the acquisition of the main speaker, e.g., a lecturer, by a teleconferencing audio system; and HOMs densely sampling the listeners position within the room. Measurements were performed in "Schiavoni room" located at Dipartimento di Elettronica, Informazione e Bioingegneria of Politecnico di Milano in Milan, Italy. The seminar room is named after Prof. Emer. Nicola Schiavoni and it is employed for lectures and teleconferences by the staff of the Politecnico di Milano. After the acquisition of the RIRs and a geometric calibration of the arrays based on acoustic measures, we estimate the reverberation time and the clarity of the room. Moreover, in order to validate the HOMULA-RIR dataset, we test
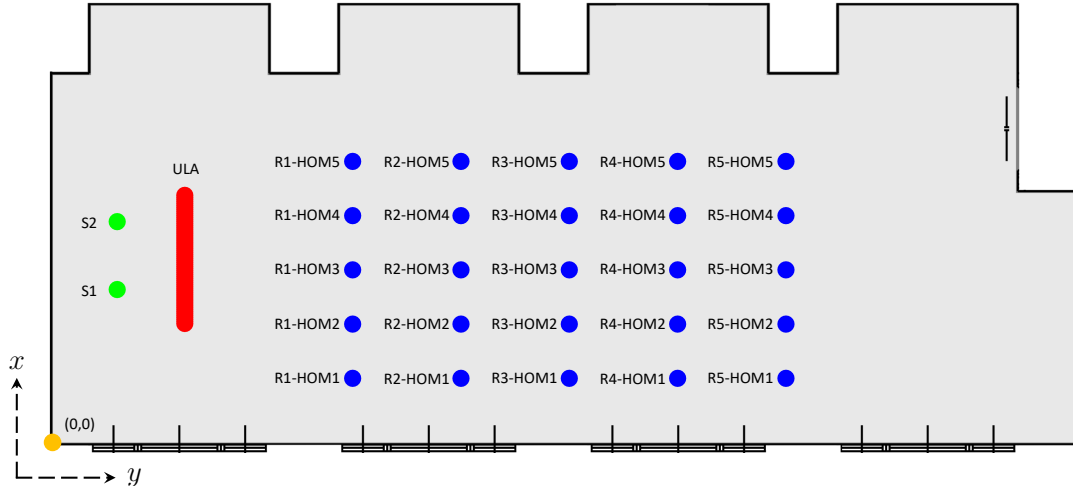
**Fig. 2**: Floor plan of Schiavoni seminar room. Sources are depicted in green, ULAs in red, and HOMs in blue. Spatial measurements are referenced to the orange marker denoting the origin point.

it under two sample applications, namely source localization and source separation, demonstrating its effectiveness. The rest of the paper is organized as follows. In Sec. 2 we describe the dataset in terms of environment and setup, while in Sec. 3 we describe some objective measurements related to the environment conditions. In Sec. 4 we present two sample applications. Finally, in Sec. 5 we draw some conclusions. The dataset is freely available at https://doi.org/10.5281/zenodo.10479726.

## 2. DATASET DESCRIPTION

The proposed dataset is composed of a collection of RIRs, measured in a seminar room covering a wide area thereof. Considering the designated use of the dataset for teleconferencing and spatial audio applications, we chose to employ a hybrid setup by capturing RIRs using both uniform linear arrays (ULAs) and HOMs. More specifically, we opted to position the ULAs in front of the desk, typically where one or more lecturers are located during a presentation, to emulate a teleconferencing system capturing the sound in the proximity of the source. The HOMs, instead, are positioned in correspondence of the listeners' seats, to replicate the listening perspective of the attendees. The RIRs have been recorded using logarithmic sine sweeps ranging from 50 Hz to 22 kHz, each lasting 10 s and sampled at 48 kHz. The room responses have been acquired using Reaper as Digital Audio Workstation and all the audio streams were routed through the Dante™ Controller to a laptop operating a Dante™ Virtual Soundcard.

### 2.1. Room conditions

The room in which we performed the measurements, shown in Figure 1, is typically used for frontal lessons and seminars and thus is furnished with tables and chairs for both lecturers and the audience. To preserve and capture the acoustic properties of the actually used space, we opted to keep the furniture during the collection of the RIRs. The room is 14.52 m long, 5.46 m wide and 3.38 m high. The floormap is shown in Figure 2. It features concrete and tile surfaces throughout, with windows covered by heavy curtains.

### 2.2. Sources setup

We considered two sources located behind the main desk in the room, specifically employing two Genelec 8020D loudspeakers[1], which can be seen in the foreground in Figure 1. The intention is to replicate the scenario of two speakers or lecturers addressing an audience in the room. With respect to the origin indicated in Figure 2, the sources are located at positions $S_1 = [2.28\,\text{m}\ 0.96\,\text{m}\ 1.20\,\text{m}]^T$ and $S_2 = [3.28\,\text{m}\ 0.96\,\text{m}\ 1.20\,\text{m}]^T$, considering the acoustic center indicated in the operating manual of the loudspeakers as measurement point.

### 2.3. EStick setup

As far as the ULAs are concerned, we used four co-linear EStick V3, made by Eventide Inc. in collaboration with Politecnico di Milano [19]. ESticks are modular ULAs, and each unit consists of 16 omnidirectional MEMS microphones with a spacing of 3 cm between them. One advantage of the system is its versatility in rapidly deploying various linear and planar array configurations, accommodating up to 64 microphone elements. For our particular setup, visible in Figure 1, we opted for a linear geometry to achieve a 64-microphone-long array, positioned in front of the main desk, at the same height $z = 1.20\,\text{m}$ of the sources. The microphone signals are accessible through the integrated Dante™ connectivity, using a single CAT6 cable for each array, which provides both power and synchronization. Even if the position of each capsule is known a priori, we have implemented a self-calibration procedure to localize the ULA within the room, as described in Section 3.1.

### 2.4. Spatial Mic setup

For the acquisition of the RIRs through HOMs, we adopted the Spatial Mic Dante by Voyage Audio[2]. The Spatial Mic is based on the geometry presented in [20], and it features 8 prepolarized condenser capsules, allowing a higher-order Ambisonics [18] encoding up to the 2nd order. Similarly to the ESticks, the Spatial Mic features integrated Dante™ connectivity, using a single CAT6 cable for each

---

[1]https://www.genelec.com/8020d
[2]https://voyage.audio/spatialmic

2

unit that provides both power and synchronization. In the considered setup, we positioned the HOMs in correspondence of the audience seats, at the same height $z = 1.20$ m of the sources, covering 25 different locations divided in 5 rows. With five devices at our disposal, we conducted the measurements five times, relocating the microphones while maintaining precise consistency in the surrounding environment and position of the sources. The position of each capsule has been estimated through acoustic measures computed as described in Section 3.1.

## 2.5. Data format

The released RIRs have been recorded at a sample rate of fs $=$ 48 kHz and truncated to a duration of 1 s. They are provided as multichannel `wav` files, saved at 32 bit per sample. RIRs of individual arrays are saved as separate files, following the naming convention: `rir-`**`source`**`-`**`array`**`.wav`. Here, **`source`** can be either `S1` or `S2`, depending on the considered source, and **`array`** is an acronym representing a specific microphone array, as depicted in Figure 2. The term **`array`** can take on either `ULA` for the ESticks measures, or a pair **`row`**`-`**`HOM`** for the Spatial Mics measures. Specifically, **`row`** $= \{$`R1, R2, R3, R4, R5`$\}$ designates the row where a particular Spatial Mic is positioned, and **`HOM`** $= \{$`HOM1, HOM2, HOM3, HOM4, HOM5`$\}$ denotes a specific array within each row. The positions of each capsule in every array are released as `csv` files, adopting the naming convention `pos-`**`array`**`.csv`, where **`array`** is the same acronym denoting a specific microphone array. Additionally, the positions of the two sources are reported in the file `pos-sources.csv`.

## 3. EVALUATION OF MEASUREMENTS

### 3.1. Calibration

A calibration process is carried out to determine the absolute placement within the room of each capsule of the microphone arrays. This allows us to provide acoustically precise positions, which are particularly relevant for many potential applications of the data set. To infer each microphone's location, the calibration process uses four additional loudspeakers (Genelec 8020D), whose positions are known and which can be seen in the background in Figure 1.

First, we measured the RIRs between each capsule of both the ESticks and the Spatial Mics, and the additional sources placed around the area in which the microphones are located. The acquisition of the RIRs follows the same procedure detailed in Section 2. From each RIR, by identifying the first peak associated to the direct path, we compute the Time of Arrival (ToA), which is the time that it takes for a pressure wave to travel from a source to one of the microphone capsules. Then, to estimate the unknown positions of each capsule $\mathbf{r}_m$, we solve an optimization least-squares problem in 2D as

$$(\mathbf{r}_m, d) = \arg\min_{(\mathbf{r}_m, d)} \sum_{m,l} [\|\mathbf{r}_m - \mathbf{r}_l\| - (\tau_{m,l} \cdot c - d)]^2, \quad (1)$$

where $c$ is the speed of sound, $\mathbf{r}_l$ is the position of loudspeaker $l$, $\tau_{m,l}$ is the ToA between $\mathbf{r}_m$ and $\mathbf{r}_l$ and $d$ is an estimate of the delay caused by the acquisition system latency, expressed in meters.

### 3.2. Reverberation time

In order to characterize the acoustic properties of "Schiavoni room", we estimate the reverberation time T60 of the environment. In par-
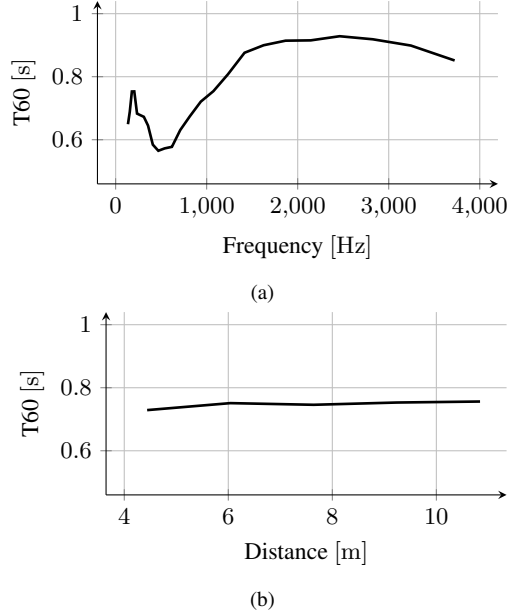


(a)



(b)

**Fig. 3**: Reverberation time as a function of (a) frequency and as a function of (b) distance from the sources.

ticular, we compute the average room T60 as the mean T60 measured from each RIR using Schroeder method, in the implementation provided by the `pyroomacoustics` package [21]. In particular, we considered a 30 dB decay in the energy decay curve to actually estimate the T60. Figure 3a shows the variation of T60 across third-octave frequency bands, spanning from 125 Hz to 4000 Hz. Figure 3b, instead, depicts the relationship between T60 and the absolute $y$ position in the room, providing insights into reverberation time dependence on the distance from the sources. As expected, the reverberation time exhibits variability among different frequency bands, ranging from a minimum of 0.56 s to a maximum of 0.93 s, with a mean value of 0.74 s. In contrast, there is no discernible dependence on distance, indicating that the reverberation time remains consistent throughout different locations within the room.

### 3.3. Clarity index

The clarity index, as defined in ISO-3382-1 standard, serves as an estimation of perceived clarity within a room and depends on the energy ratio between the early and late parts of the RIR [22]. Specifically, we calculated both C50 and C80 clarity indices, using the implementation available in the `python-acoustics` package. Figure 4a displays the variation of both metrics across third-octave frequency bands, spanning from 125 Hz to 4000 Hz. Figure 4b, instead, presents the relationship between clarity index and the absolute $y$ position in the room, enabling the exploration of its dependence on the distance from the sources. Except for the lower frequencies, clarity remains relatively consistent across different frequency bands, converging towards average values of C50 $= 0.5$ dB and C80 $= 5.4$ dB. Conversely, when considering the dependence on distance, C50 exhibits a decline from 2.5 dB at the front row of the room to $-6.8$ dB at the back of the room, while C80 exhibits a decline from 7.0 dB to 2.4 dB.
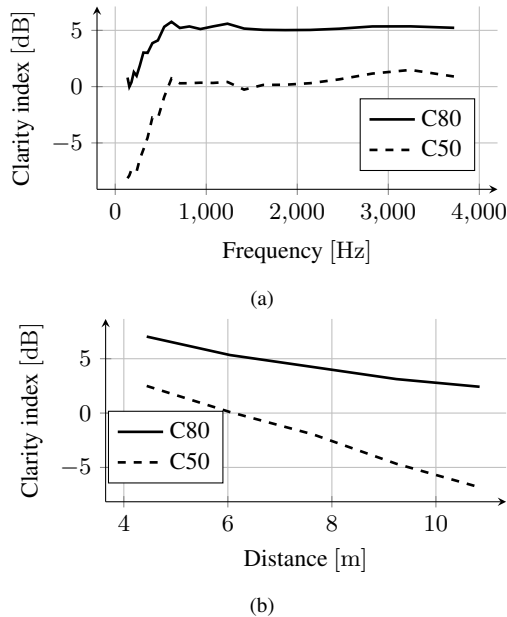
(a)



(b)

**Fig. 4**: Clarity index as a function of (a) frequency and as a function of (b) distance from the sources.

## 4. SAMPLE APPLICATION

In order to validate the HOMULA-RIR dataset, we conducted various tests targeting two classical applications: blind source separation and source localization. Specifically, blind source separation was employed to show the use of ULA signals, while source localization was used to validate the HOMs signals.

### 4.1. Blind source separation

The task of blind audio source separation involves extracting multiple unknown audio signals (sources) by processing their combined mixture. This process is referred to as *blind* because the algorithm has access only to the mixed signals and lacks information about the individual source signals. To achieve this, we exploit the Ray-Space-Based Multichannel Nonnegative Matrix Factorization algorithm (RS-MNMF), originally proposed in [23]. The algorithm leverages the Ray Space Transform [24], to project the microphone signals acquired by ULAs (the ESticks) into the Ray Space domain. In such domain, the position of sources is encoded directly into the magnitude of the ray-space-transformed signals. This results in an effective use of the spatial information present in the mixture and encoded in the ray space data, allowing for a direct application of the conventional multichannel NMF algorithm [23].

Results are computed in terms of three classic blind source separation metrics, namely Source to Distortion Ratio (SDR), Source to Interferences Ratio (SIR) and Sources to Artifacts Ratio (SAR) [25]. The tests were conducted considering $3\,\mathrm{s}$ long speech signals, and the average value of these metrics over all the microphone signals is taken into account. It can be noticed that the results remain consistent for both examined sources, $S_1$ and $S_2$, indicating that the separation can be successfully performed considering either sources and independently from the locations. Additionally, the obtained values align with those presented in [23] using a similar setup.
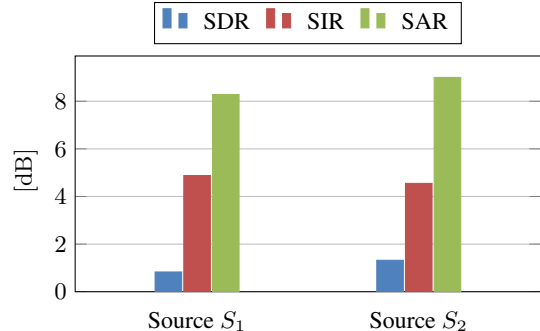


**Fig. 5**: Blind source separation results: SDR, SIR and SAR metrics for both considered sources.

### 4.2. Source localization

In the context of acoustic signal processing, source localization is the task of identifying the direction of arrival (DOA) of a sound emitted by a source from a multichannel acquisition. In this work, we leverage on SHD-LRA [26], a DOA estimation approach that exploits low-rank signal approximations in the spherical harmonic domain. In particular, the algorithm makes use of the spherical harmonics representation in order to estimate the direction-dependent components that characterize the source position by means of low-rank decomposition of the expansion coefficients [26].

As in [26], results are evaluated in terms of two performance metrics: the probability of detection (PD) and the root mean squared error (RMSE) of the DOA. In particular, the former is computed as the percentage of DOA estimates below an absolute DOA error of $10°$. Using the microphone signals captured by the Spatial Mics and considering the acoustically calibrated positions as ground truth, a probability of detection of $79\,\%$ can be achieved. This aligns with the results presented in [26] when dealing with reverberant room conditions. Also the RMSE values are consistent with those in [26] computed for non-ideal conditions, yielding an azimuth RMSE of $3.33°$ and an elevation RMSE of $6.10°$.

## 5. CONCLUSION

We presented HOMULA-RIR, a dataset of RIRs measured in a seminar room, acquired through the use of both ULAs and HOMs. This diverse and versatile configuration guarantees suitability for a broad range of application scenarios in the context of telecommunications, teleconferencing, and spatial audio. We provide precise measurements, together with acoustically calibrated microphone positions. Various analyses, including the measurement of reverberation time and clarity, have been conducted to offer a comprehensive understanding of the acoustic characteristics within the environment. To validate the usability of the dataset, we performed tests with two classic applications: blind source separation and source localization. Results prove the effectiveness of the dataset in these contexts, showcasing its potential for the intended applications.

## 6. ACKNOWLEDGEMENTS

# 7. REFERENCES

[1] M. Olivieri, L. Comanducci, M. Pezzoli, D. Balsarri, L. Menescardi, M. Buccoli, S. Pecorino, A. Grosso, F. Antonacci, and A. Sarti, "Real-time multichannel speech separation and enhancement using a beamspace-domain-based lightweight cnn," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[2] Y. Hsu, Y. Lee, and M. R. Bai, "Learning-based personal speech enhancement for teleconferencing by exploiting spatial-spectral features," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8787–8791, IEEE, 2022.

[3] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, "Personalized speech enhancement: New models and comprehensive evaluation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 356–360, IEEE, 2022.

[4] K. Sridhar, R. Cutler, A. Saabas, T. Parnamaa, M. Loide, H. Gamper, S. Braun, R. Aichner, and S. Srinivasan, "Icassp 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 151–155, IEEE, 2021.

[5] C. Chen, W. Sun, D. Harwath, and K. Grauman, "Learning audio-visual dereverberation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[6] L. Diener, S. Sootla, S. Branets, A. Saabas, R. Aichner, and R. Cutler, "Interspeech 2022 audio deep packet loss concealment challenge," *arXiv preprint arXiv:2204.05222*, 2022.

[7] A. I. Mezza, M. Amerena, A. Bernardini, and A. Sarti, "Hybrid packet loss concealment for real-time networked music applications," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 266–273, 2024.

[8] F. Miotello, M. Pezzoli, L. Comanducci, F. Antonacci, and A. Sarti, "Deep prior-based audio inpainting using multi-resolution harmonic convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[9] M. Pezzoli, F. Borra, F. Antonacci, S. Tubaro, and A. Sarti, "A parametric approach to virtual miking for sources of arbitrary directivity," *IEEE/ACM Trans. Acoust., Speech, Signal Process.*, vol. 28, pp. 2333–2348, 2020.

[10] L. McCormack, A. Politis, R. Gonzalez, T. Lokki, and V. Pulkki, "Parametric ambisonic encoding of arbitrary microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2062–2075, 2022.

[11] X. Karakonstantis and E. Fernandez-Grande, "Generative adversarial networks with physical sound field priors," *The Journal of the Acoustical Society of America*, vol. 154, no. 2, pp. 1226–1238, 2023.

[12] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 313–317, IEEE, 2014.

[13] S. Koyama, T. Nishida, K. Kimura, T. Abe, N. Ueno, and J. Brunnström, "Meshrir: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, IEEE, 2021.

[14] T. Dietzen, R. Ali, M. Taseska, and T. van Waterschoot, "Myriad: a multi-array room acoustic database," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, pp. 1–14, 2023.

[15] R. Stewart and M. Sandler, "Database of omnidirectional and b-format room impulse responses," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 165–168, IEEE, 2010.

[16] J.-W. Choi and F. Zotter, "Six degrees-of-freedom room impulse response dataset measured over a dense loudspeaker grid (6drir-dl)," in *Audio Engineering Society Conference: AES 2023 International Conference on Spatial and Immersive Audio*, Audio Engineering Society, 2023.

[17] G. Götz, S. J. Schlecht, and V. Pulkki, "A dataset of higher-order ambisonic room impulse responses and 3d models measured in a room with varying furniture," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pp. 1–8, IEEE, 2021.

[18] F. Zotter and M. Frank, *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality.* Springer Nature, 2019.

[19] M. Pezzoli, L. Comanducci, J. Waltz, A. Agnello, L. Bondi, A. Canclini, and A. Sarti, "A dante powered modular microphone array system," in *Audio Engineering Society Convention 145*, Audio Engineering Society, 2018.

[20] E. M. Benjamin, "A second-order soundfield microphone with improved polar pattern shape," in *Audio Engineering Society Convention 133*, Audio Engineering Society, 2012.

[21] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 351–355, IEEE, 2018.

[22] P. Götz, C. Tuna, A. Walther, and E. A. Habets, "Online reverberation time and clarity estimation in dynamic acoustic conditions," *The Journal of the Acoustical Society of America*, vol. 153, no. 6, pp. 3532–3542, 2023.

[23] M. Pezzoli, J. J. Carabias-Orti, M. Cobos, F. Antonacci, and A. Sarti, "Ray-space-based multichannel nonnegative matrix factorization for audio source separation," *IEEE Signal Processing Letters*, vol. 28, pp. 369–373, 2021.

[24] L. Bianchi, F. Antonacci, A. Sarti, and S. Tubaro, "The ray space transform: A new framework for wave field processing," *IEEE Transactions on Signal Processing*, vol. 64, no. 21, pp. 5696–5706, 2016.

[25] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[26] M. Cobos, M. Pezzoli, F. Antonacci, and A. Sarti, "Acoustic source localization in the spherical harmonics domain exploiting low-rank approximations," in *Int. Conf. Acoust. Speech Signal Process*, pp. 1–5, IEEE, 2023.