MIRACLE - Microphone Array Impulse Response Dataset for Acoustic Learning

Adam Kujawski[*,1,2] Art J. R. Pelling[†,1,2] Ennes Sarradj[‡,1]

[*]Email: adam.kujawksi@tu-berlin.de, ORCID: 0000-0003-4579-8813
[†]Email: a.pelling@tu-berlin.de, ORCID: 0000-0003-3228-6069
[‡]Email: ennes.sarradj@tu-berlin.de ORCID: 0000-0002-0274-8456
[1] *Department of Engineering Acoustics, TU Berlin, Einsteinufer 25, 10587, Berlin, Germany*
[2] *These authors contributed equally to this work.*

Abstract: This work introduces a large dataset comprising impulse responses of spatially distributed sources within a plane parallel to a planar microphone array. The dataset, named MIRACLE, encompasses 856,128 single-channel impulse responses and includes four different measurement scenarios. Three measurement scenarios were conducted under anechoic conditions. The fourth scenario includes additional specular reflections from a reflective panel. The source positions were obtained by uniformly discretizing a rectangular source plane parallel to the microphone for each scenario. The dataset contains three scenarios with a spatial resolution of 23 mm at two different source-plane-to-array distances, as well as a scenario with a resolution of 5 mm for the shorter distance. In contrast to existing room impulse response datasets, the accuracy of the provided source location labels is assessed and additional metadata, such as the directivity of the loudspeaker used for excitation, is provided. The MIRACLE dataset can be used as a benchmark for data-driven modelling and interpolation methods as well as for various acoustic machine learning tasks, such as source separation, localization, and characterization.

Keywords: impulse response, dataset, microphone array, acoustics

Novelty statement: We provide a large dataset of spatially distributed multichannel impulse response measurements together with a thorough assessment of the source location accuracy.

1. Introduction

A Room Impulse Response (RIR) characterizes the linear time-invariant acoustic propagation between a source and a receiver within a specific acoustic environment. RIRs are crucial for sound field auralization [36] as well as in the realm of room acoustics, where they are used for estimating acoustic properties of a room such as the reverberation time [45].

The emergence of data-driven methods in acoustics [4], particularly deep learning methods, has sparked increasing interest in the availability of rich, high-quality RIR datasets. These datasets play a pivotal role in the training of data-driven (interpolatory) sound field reconstruction methods [13, 17, 21, 26], deep generative models [29, 41], and augmentation methods [6]. In addition, RIR datasets can be flexibly employed in order to synthesize acoustic training data for source localization and characterization [18], sound event detection, and speech separation tasks by convolving arbitrary source signals with RIRs [19, 20, 37]. The same synthesis procedure can be employed for data-driven acoustic parameter estimation problems, such as blind reverberation time estimation [15, 28] and others [7].

While data-driven methods often exhibit superior performance compared to conventional model-based methods, they require large amounts of realistic training data and are sensitive to variations of underlying probability distributions describing the data, also known as *dataset shift* [33]. Experimental data is oftentimes not available or too time-consuming to acquire. Many data-driven methods across various application areas, such as speech enhancement and recognition [16, 25], localization [5, 18], sound field reconstruction [30], room acoustic parameter estimation [10, 49], and acoustical engineering [2, 27, 31], are therefore trained with simulated data, whereby en-

hanced realism helps to improve generalization performance [47, 48]. However, without adaptation to or training with realistic data, the performance of data-driven methods can be significantly impaired [2, 14], which indicates the need for experimentally measured RIR datasets.

Data availability
The dataset presented in this paper can be obtained from

doi:10.14279/depositonce-20106

under the CC BY-NC-SA 4.0 license and authored by Adam Kujawski, Art J. R. Pelling and Ennes Sarradj.

2. Materials and Methods

2.1. Experimental Setup

The experimental setup is illustrated in Fig. 1. Details on the utilized hardware are given in Table 4 in Appendix A.

Microphone Array: The phased microphone setup features a planar microphone array comprising $n_\mathrm{o} = 64$ channels mounted in a 1.5 m × 1.5 m aluminium plate. The microphone arrangement follows Vogel's spiral [44]. The maximum pairwise distance between the array microphones is referred to as the aperture size $d_\mathrm{a} = 1.47\,\mathrm{m}$. The microphone array data was acquired with a multichannel acquisition system (sampling rate: 51.2 kHz).

Sound Source and Excitation Signal: A dynamic 2" cone loudspeaker in a cylindrical 3D-printed enclosure was employed as the sound source. An exponential sine sweep was used as the excitation signal because of its favourable properties with regard to crest factor and rejection of non-linearities [42]. It was designed according to [11,34,35] in the frequency range of the loudspeaker, namely from 100 Hz to 16 kHz. Because the anechoic chamber is nearly free of reflections and has very low noise levels, it was possible to choose a relatively short sweep time of 3 s for the measurement. In order to ensure that the entire system response after excitation is captured, a safety window of 250 ms was added to the recording duration, resulting in $n_s = 166{,}400$ samples per measurement. A loopback of the excitation signal was recorded as a reference signal for post-processing.

Positioning: A high-precision motor-driven 2D positioning system was employed for loudspeaker positioning. The positioning system and the microphone array were manually aligned by using a laser distance meter and a cross-line laser, achieving only minor alignment errors of a few millimetres at worst. The loudspeaker dust cap at the membrane centre was used as reference in the manual alignment. During data post-processing, a spatial offset correction was applied based on a statistical evaluation given in Section 3.3. The corrected positions apply to the acoustical centre of the loudspeaker rather than the center of the membrane.

Environment: All measurements were performed in the anechoic chamber of TU Berlin (room volume $V = 830\,\mathrm{m}^3$, lower cut-off frequency $f_\mathrm{c} = 63\,\mathrm{Hz}$). Neither heating nor air conditioning was active, and the temperature was monitored at the microphone array centre throughout the experiment. A ground plate was placed between the loudspeaker and the microphone array in one of the experimental scenarios to enable a reflective environment. The supporting grid platform and the positioning system were clad with absorptive foam to minimize reflections.

2.2. Experimental Procedure

A customized and fully automated data acquisition procedure was implemented. Before each experiment, the loudspeaker was repeatedly excited with the excitation signal for a duration of 20 minutes (the duration was determined in a dedicated experiment). This warm-up phase accounts for the weakly non-stationary dynamics of the loudspeaker's transfer function, e.g. changes of the properties of the loudspeaker magnet related to internal temperature fluctuations, see [3]. Subsequently, the actual measurement routine was started by positioning the loudspeaker at the desired source location and measuring the room temperature simultaneously. After positioning, two repetitions of background noise measurement (1 s each) and loudspeaker excitation measurements (3 s each) were performed using all n_o microphones at once. Subsequently, the cross-correlation between all n_i recorded channels was evaluated according to the *rule of two* [40]. Based on the measured sweep signals and the noise signal, the rule of two defines a cross-correlation threshold at which a pair of measured sweeps can be regarded free of corruption. In case of any violations, the measurement was repeated automatically.

Following the main measurement campaign, an additional measurement was conducted in the anechoic chamber to obtain the angle-dependent frequency response of the loudspeaker at discrete azimuth angles at a resolution of $\Delta\theta = 2.5°$. A microphone was placed at a distance of 0.5 m from the loudspeaker centre. The latter was mounted on a motor-driven dispersion measurement turntable. A photograph of the measurement setup can be found in Fig. 2. The same excitation signal and processing parameters as in the previous measurement campaign were used to determine the loudspeaker impulse response. Due to the cylindrical enclosure enclosing the loudspeaker, rotational symmetry around the z-axis can be assumed.

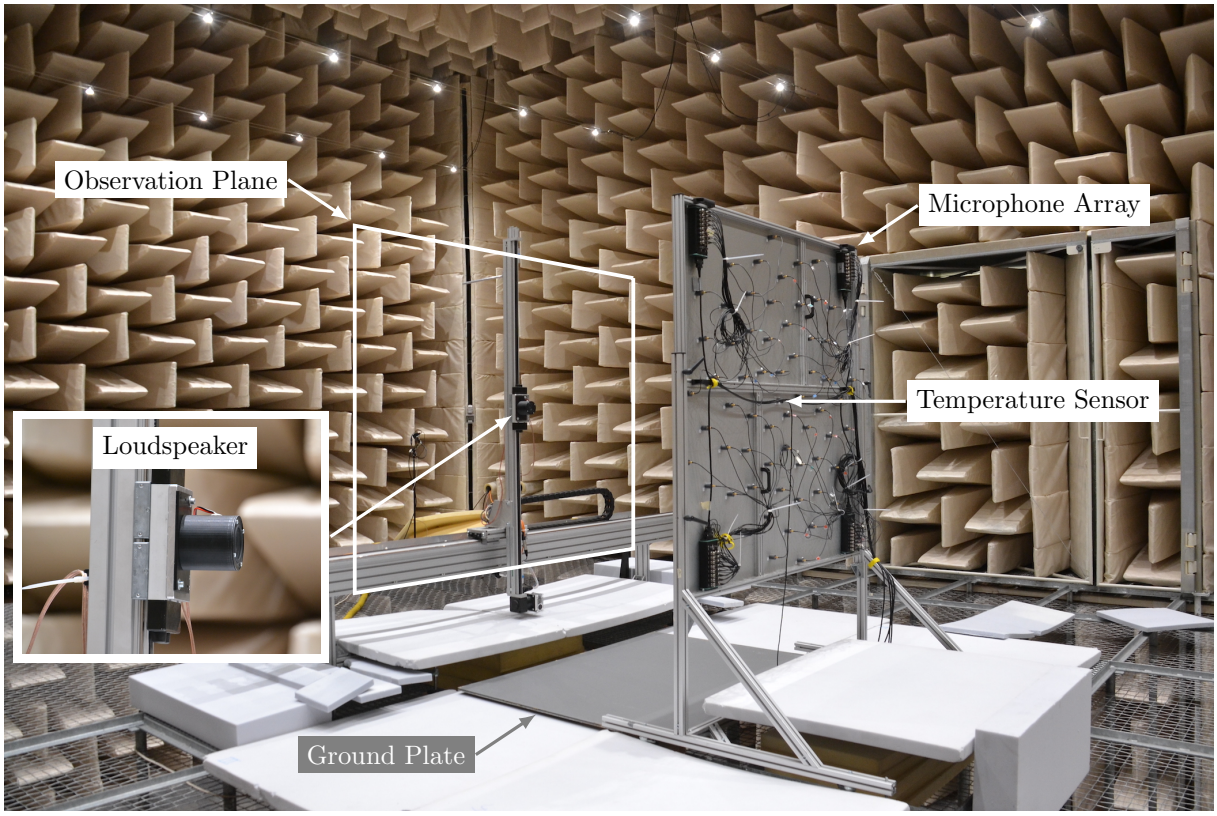

Figure 1: Experimental setup for the main experiment (R2) with reflective ground plate.

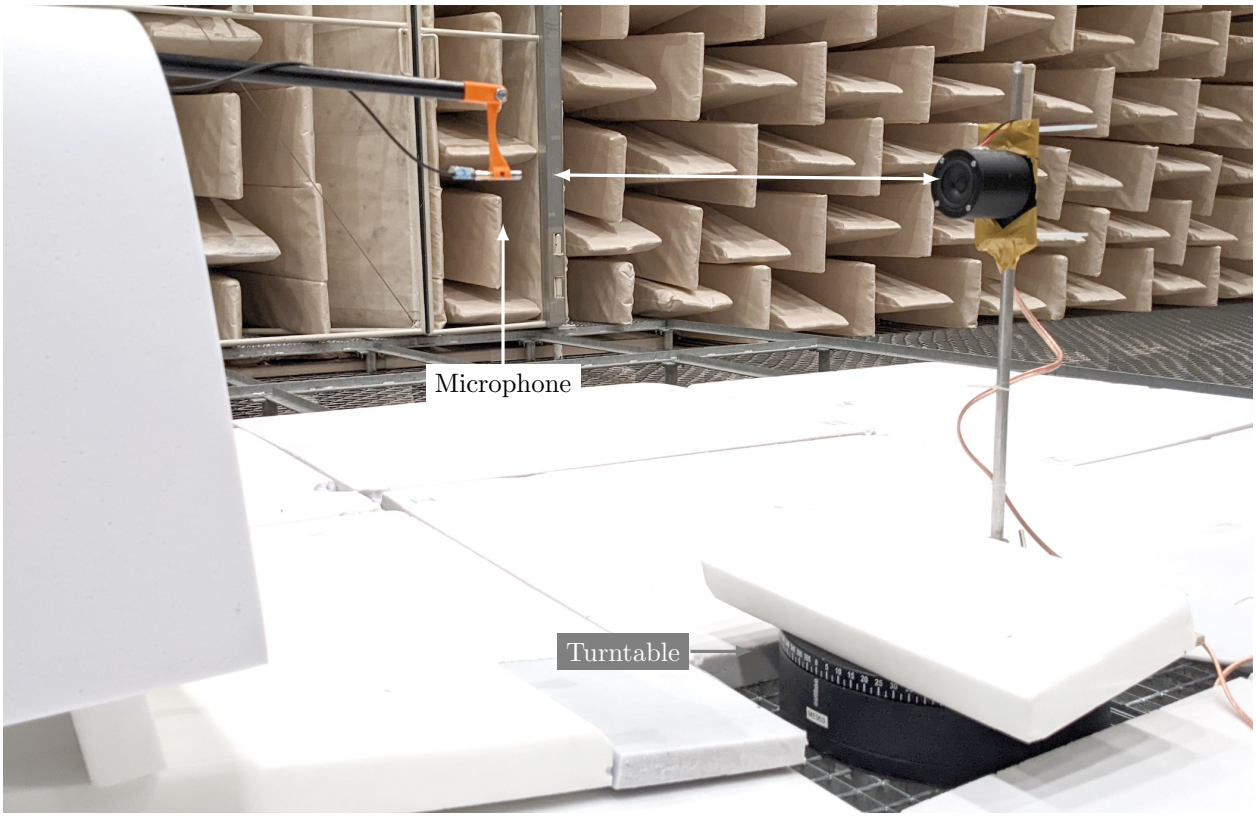

Figure 2: Experimental setup for the directivity measurement.

Table 1: MIRACLE experimental scenarios.

Scenario	Anechoic	n_i	n_o	$d_x = d_y$	$\Delta d_x = \Delta d_y$	d_z
A1	✓	$64^2 = 4{,}096$	64	146.7 cm	23.3 mm	73.4 cm
D1	✓	$33^2 = 1{,}089$	64	16.0 cm	5.0 mm	73.4 cm
A2	✓	$64^2 = 4{,}096$	64	146.7 cm	23.3 mm	146.7 cm
R2	✗	$64^2 = 4{,}096$	64	146.7 cm	23.3 mm	146.7 cm

2.3. Post-processing

Several post-processing steps were performed to obtain a good estimate of the system impulse response from the measurements. Firstly, the (loopback) excitation and microphone signals were averaged across the two measurement repetitions to obtain a single averaged excitation signal $\tilde{u}_{i,j} \in \mathbb{R}^{n_s}$ and averaged microphone signal $\tilde{y}_{i,j} \in \mathbb{R}^{n_s}$ at the i-th source to the j-th receiver location, respectively. According to that, all signals were resampled to a sampling rate of $f_\mathrm{d} = 32\,\mathrm{kHz}$ since the loudspeaker transmission capability and excitation sweep have an upper frequency limit of 16 kHz. We applied the polyphase method for resampling (see [1] for details).

Deconvolution

In the following, let $n_\mathrm{d} = 104{,}000$ denote the number of samples after resampling. An estimate of the frequency response was obtained by dividing the Discrete Fourier Transform (DFT) of the averaged and downsampled measurement signals $Y_{i,j} = \mathsf{DFT}(\tilde{y}_{i,j}) \in \mathbb{C}^{n_\mathrm{d}}$ by the corresponding DFT of the averaged and resampled loopback excitation signals $U_{i,j} = \mathsf{DFT}(\tilde{u}_{i,j}) \in \mathbb{C}^{n_\mathrm{d}}$, i.e.

$$H_{i,j}\left(e^{\imath\omega_k}\right) = Y_{i,j}\left(e^{\imath\omega_k}\right) U_{i,j}^{-1}\left(e^{\imath\omega_k}\right) \in \mathbb{C},$$

for the angular frequency $\omega_k = 2\pi k/n_\mathrm{d}$ with $k \in [-n_\mathrm{d}/2, n_\mathrm{d}/2] \subset \mathbb{Z}$. The inverse spectra $U_{i,j}^{-1} \in \mathbb{C}^{n_\mathrm{d}}$ were obtained by regularized inversion [12, 22–24, 35, 38]

$$U_{i,j}^{-1}\left(e^{\imath\omega_k}\right) = \frac{U_{i,j}^{*}\left(e^{\imath\omega_k}\right)}{U_{i,j}^{*}\left(e^{\imath\omega_k}\right) U_{i,j}\left(e^{\imath\omega_k}\right) + M\lambda\left(e^{\imath\omega_k}\right)},$$

where $M = \max_{k\in\{1\,\ldots,\,n_\mathrm{d}\}}\{|U_{i,j}\left(e^{\imath\omega_k}\right)|^2\} = 1$. Regularization is necessary to avoid instabilities in the deconvolved frequency response that arise from persistently exciting only over a limited frequency range. Practical considerations for choosing the regularization term in acoustic applications can be found in [35]. The regularization term $\lambda \in \mathbb{R}^{n_\mathrm{d}}$ was chosen as

$$\lambda\left(e^{\imath\omega_k}\right) = \begin{cases} 1 & \text{for } |\omega_k| \in [0,\, \omega_\mathrm{fade}] \\ \frac{1+\cos\left(\frac{\omega_\mathrm{fade}-|\omega_k|}{\omega_\mathrm{fade}-\omega_\mathrm{cut}}\right)}{2} & \text{for } |\omega_k| \in [\omega_\mathrm{fade},\, \omega_\mathrm{cut}] \\ 0 & \text{for } |\omega_k| \in [\omega_\mathrm{cut},\, \pi] \end{cases}$$

such that the regularization term $\lambda\left(e^{\imath\omega_k}\right)$ is equal to 0 above the cutoff frequency

$$\omega_\mathrm{cut} = 2\pi\frac{100\,\mathrm{Hz}}{f_\mathrm{d}}$$

which is chosen according to the lower limit of the loudspeaker's frequency range of 100 Hz and equal to 1 below $\omega_\mathrm{fade} = \frac{\omega_\mathrm{cut}}{\sqrt{2}}$. A cross-fade based on a Hann window (raised-cosine) is used to smoothly transition in between. The estimate of the frequency response $H_{i,j}$ was then transformed back to the time domain to finally obtain the impulse response

$$h_{i,j} = \mathsf{DFT}^{-1}\left(H_{i,j}\right).$$

Truncation

The calculated impulse responses were subsequently truncated in order to contain the size of the final dataset. For user convenience, the impulse responses of all measurement scenarios were truncated identically. For this, the minimum cumulative energy $e \in \mathbb{R}^{n_\mathrm{d}}$ given by

$$e(t) = \min_{i\in n_\mathrm{i}, j\in n_\mathrm{o}} \sum_{\tau=1}^{t} |h_{i,j}(\tau)|^2, \quad t \in \{1,\,\ldots,\,n_\mathrm{d}\},$$

was calculated for each scenario. The truncation index n_t was chosen to be the smallest power of two that is larger than the time index for which 0.1 % of the energy is truncated at worst, namely

$$n_t = 1{,}024 \geq \tilde{t} = \operatorname*{arg\,max}_{t\in\{1,\,\ldots,\,n_\mathrm{d}\}} \left\{e(t) \leq 0.999\,\|e\|_\infty\right\}.$$

3. Results and Discussion

3.1. Impulse Responses

A total of four different experimental scenarios were realized, which are summarized in Table 1. The acquisition time for each of the large-scale scenarios A1, A2, and R2 was about 20 hours. The total number of single-channel impulse responses across all scenarios is 856, 128. The scenarios differ regarding the environment as well as the spatial dimension ($d_y = d_x$), sampling resolution ($\Delta d_y = \Delta d_x$), and distance d_z of the source plane. The two large anechoic scenarios A1 and A2 each include 4, 096 measured source positions on an equidistantly spaced 64×64

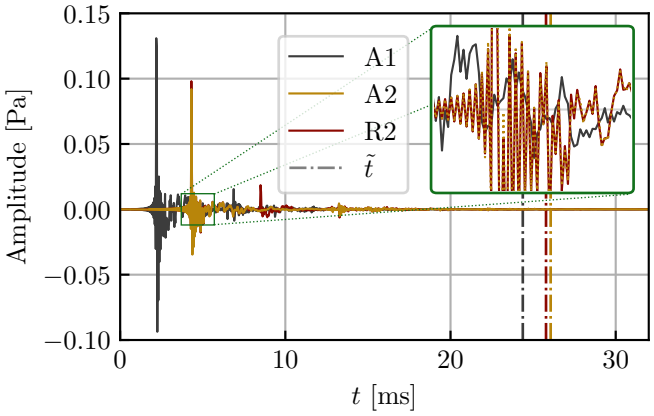

Figure 3: Measured impulse responses for the scenarios A1, A2, and R2 and the centremost locations in the source and receiver plane. The dash-dotted vertical lines indicate the truncation index $\tilde{t}$.

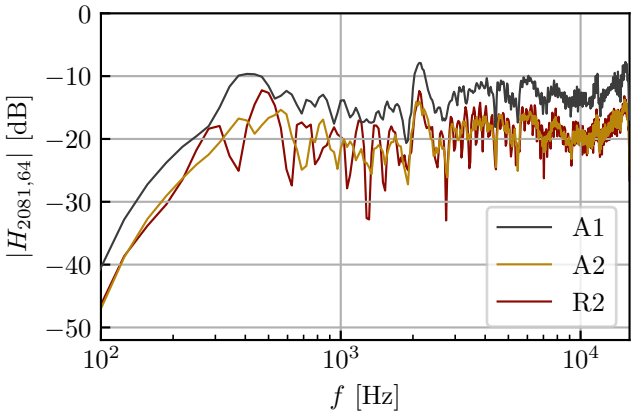

Figure 4: Magnitude of the frequency response of the measured transfer functions for the scenarios A1, A2, and R2 at the centremost locations in the source and receiver plane.

grid at different source-plane distances d_z. In addition, a densely-sampled scenario D1 was acquired on a smaller 33×33 grid with a spacing of only 5 mm. Scenario R2 is based on the same geometric setup as scenario A2, but an aluminium plate on the floor introduces a specular reflection. Fig. 3 and Fig. 4 exemplarily show the measured impulse response and its magnitude spectrum for a single source-receiver combination for scenarios A1, A2, and R2, respectively. It can be readily verified that the doubling of the distance to the source is also reflected in a doubling of the delay shift and an attenuation of the magnitude spectrum by approximately −6 dB. Furthermore, the specular reflection for scenario R2 manifests in a prominent second peak in the impulse response and comb filtering in its magnitude spectrum.

The mean and standard deviation of temperature and the speed of sound for each of the scenarios are given in Table 2. The speed of sound has been calculated according to [8, 9][1]. It reveals that the temperature and the speed of sound are almost identical across all scenarios with an absolute difference of $\Delta\mu < 1\,°\mathrm{C}$ and $\Delta\mu \leq 0.6\,\frac{m}{s}$, respectively, which is expected due to the fairly constant environmental conditions inside the anechoic chamber.

3.2. Loudspeaker Directivity

Fig. 5 shows the directivity D and the directivity index DI of the loudspeaker measured with a dispersion measurement turntable in the azimuthal plane. In this work, the directivity is defined as the ratio between the measured squared sound pressure $p_{\mathrm{RMS}}(\theta, f)$ at an angle θ and the maximum among all angles, i.e.

$$\mathrm{D}(\theta, f) = 10 \log_{10} \left(\frac{p_{\mathrm{RMS}}(\theta, f)}{\max_{\phi \in [0, 2\pi]} p_{\mathrm{RMS}}(\phi, f)} \right),$$

The directivity index under the assumption of rotational symmetry is expressed as

$$\mathrm{DI}(f) = 10 \log_{10} \left(\frac{4\pi p^2_{\mathrm{RMS}}(0, f)}{2\pi \int_0^{\pi} p^2_{\mathrm{RMS}}(\phi, f) \sin(\phi)\, \mathrm{d}\phi} \right),$$

where $p^2_{\mathrm{RMS}}(0, f)$ represents the squared sound pressure in front of the speaker.

It is seen that the loudspeaker exhibits a radiation pattern similar to a monopole until an upper frequency of 2 kHz. Above this frequency, the directivity index increases. Still, the directivity observed by the microphone array is close to a monopole at relevant radiation angles, i.e. $\theta \leq \theta_{\max} = 67.3°$, as indicated by the dashed line in Fig. 5.

Table 2: Mean μ and standard deviation σ of the temperature and speed of sound for each experiment.

Scenario	Temperature [°C]		Speed of Sound [m s^{-1}]	
A1	$\mu = 21.6$	$\sigma = 0.12$	$\mu = 344.8$	$\sigma = 0.07$
D1	$\mu = 21.8$	$\sigma = 0.01$	$\mu = 345.0$	$\sigma = 0.01$
A2	$\mu = 22.3$	$\sigma = 0.05$	$\mu = 345.3$	$\sigma = 0.03$
R2	$\mu = 22.5$	$\sigma = 0.02$	$\mu = 345.4$	$\sigma = 0.01$

3.3. Positional Validation

Several uncertainty factors affected the spatial alignment precision regarding the microphone array centre and the

[1] An atmospheric pressure of 101.325 kPa and a carbon dioxide mole fraction of 0.0004 was used. A generic value of 38% was used for the relative humidity approximating the humidity conditions throughout the experiments

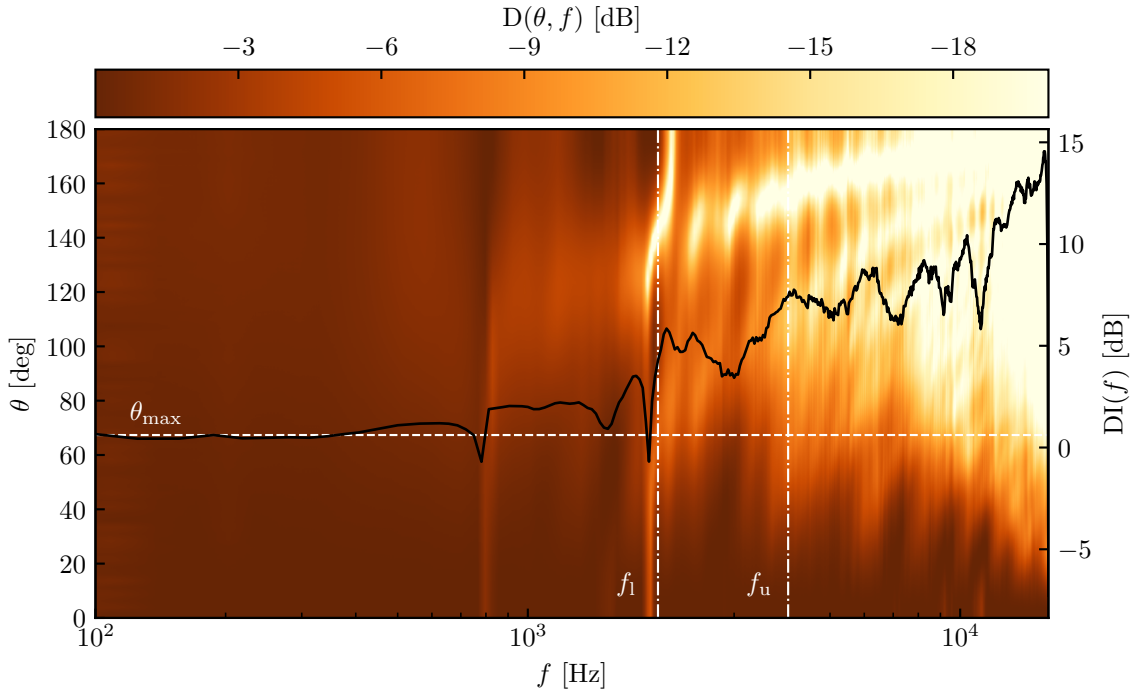

Figure 5: Directivity D and directivity index DI of the loudspeaker. The maximum opening angle across all experiments is denoted by $\theta_{\max}$.

centre of the observation area. These factors include measurement uncertainties with regard to the utilized cross-line laser and distance meter as well as mechanical backlash, which occurred primarily with horizontal changes of direction. Therefore, a systematic spatial offset within the range of a few millimetres can be assumed.

Due to the anechoic environment and the use of a large-scale microphone array enabling an excellent spatial resolution, Conventional Frequency Domain Beamforming [32] serves as an appropriate method to obtain an estimate of the actual source location. The large number of acoustic cases also permits a statistical approach to determine the spatial offset for a measurement scenario and to quantify the uncertainty regarding the source position information.

Beamforming

Let $\omega_k = 2\pi k/n_\mathrm{d}$ with $k \in [-n_\mathrm{d}/2, n_\mathrm{d}/2] \subset \mathbb{Z}$ and let

$$H(e^{\imath\omega_k}) = \begin{bmatrix} H_{i,1}(e^{\imath\omega_k}) & \dots & H_{i,n_\mathrm{o}}(e^{\imath\omega_k}) \end{bmatrix} \in \mathbb{C}^{n_\mathrm{o}}$$

denote the transfer function measurements from the i-th source at location x_s for $i \in \{1, \dots, n_\mathrm{i}\}$ to each of the n_o microphones. The cross-spectral matrix induced by a sound source with unit strength is then given by

$$C(\omega_k) = H(e^{\imath\omega_k})H(e^{\imath\omega_k})^* \in \mathbb{C}^{n_\mathrm{o}\times n_\mathrm{o}}.$$

The beamforming result for an assumed source location $x_\mathrm{s} \in \mathbb{R}^3$ is then given by the square of the C-weighted norm of the steering vector $a(x_\mathrm{s}, \omega_k) \in \mathbb{C}^{n_\mathrm{o}}$, i.e.

$$b(x_\mathrm{s}, \omega_k) = \|a(x_\mathrm{s}, \omega_k)\|^2_{C(\omega_k)} = a(x_\mathrm{s}, \omega_k)^* C(\omega_k) a(x_\mathrm{s}, \omega_k).$$

Many formulations of the steering vector can be found in the literature. The formulations I and IV in [43] result in a coincidence of the beamformer's steered response power maximum and the actual source location for a single monopole source radiating under free-field conditions. In this work, formulation IV was used, which defines the entries of a via

$$\{a(x_\mathrm{s}, \omega)\}_j = \frac{e^{\imath\omega(r_j - r_0)/c}}{r_j \sqrt{n_\mathrm{o} \sum_{k=1}^{n_\mathrm{o}} r_k^{-2}}},$$

where $r_j = \|x_\mathrm{s} - x_j\|_2$ is the distance between the assumed source location x_s and the j-th microphone location x_j, and $r_0 = \|x_\mathrm{s} - x_0\|_2$ is the distance between x_s and the reference position, in this case the origin of the coordinate system.

Validation of each measured source position commenced with the spatial discretization of a neighbourhood around the assumed source position. A 201×201 equidistantly spaced focus-grid with a resolution of $\Delta x = 0.5\,\mathrm{mm}$ was employed. The beamforming map was computed on the discretized region for every frequency in the range

$$\Omega = \left[2\pi \frac{f_\mathrm{l}}{f_\mathrm{d}}, 2\pi \frac{f_\mathrm{u}}{f_\mathrm{d}}\right]$$

which was chosen such that the lower frequency limit $f_\mathrm{l} = 2\,\mathrm{kHz}$ enabled a sufficiently large spatial resolution in the resulting beamforming map, and the upper frequency limit $f_\mathrm{u} = 4\,\mathrm{kHz}$ ensures that the wavelength is larger than the loudspeaker diameter. The latter is important to ensure that the loudspeaker has a radiation pattern close to a monopole at relevant radiation angles in order to meet the monopole assumption needed for the steering vector formulation. As indicated by the dashed line in Fig. 5, the radiation angle from the loudspeaker to any microphone in the array is bounded by $\theta_\mathrm{max} = 67.3°$. The global spatial maximum is then determined by

$$\hat{x}_i = \arg\max_{x_s} \sum_{\omega\in\Omega} \hat{b}(x_s, \omega),$$

where $\hat{b}(x_s, \omega)$ denotes the amplitude normalized beamforming result

$$\hat{b}(x_s, \omega) = \frac{b(x_s, \omega)}{b(\hat{x}_s, \omega)},$$

with $b(\hat{x}_s, \omega)$ being the beamformer's maximum output among all source locations x_s at a given frequency ω. The evaluation was conducted for different distances within a range of up to $\pm 12\,\mathrm{mm}$ around the assumed source distance with a sampling interval of $\Delta z = 1\,\mathrm{mm}$ to account for a potential mismatch of the source plane distance. Finally, the positional offset between the beamformer's prediction and the assumed source position is determined by $\Delta x_i = \hat{x}_i - x_i$.

Statistical Evaluation

The systematic positional offset between the centre of the observation area and the microphone array in the horizontal and vertical direction can be statistically determined by using the estimates $\Delta x_i \in \mathbb{R}^2$ for each individual measured source position. Thereby, each estimated positional deviation Δx_i can be seen as a realization of the jointly distributed random variables R_x, R_y with the joint Probability Density Function (PDF) $f_{R_x,R_y}(\Delta x_i)$. It is assumed that the individual positional offset estimations Δx_i are symmetrically distributed around the true positional offset due to the approximate symmetry of the microphone array and observation plane around the origin. Then, the true positional offset corresponds to the deviation associated with the greatest probability. A simple method to determine the joint PDF of jointly distributed random variables based on a finite set of samples is the kernel density estimation [39], denoted by

$$\hat{f}_{R_x,R_y}(\Delta x_i) = \frac{1}{N}\sum_{n=1}^{N} K_h(\Delta x_i - \Delta x_i^{(n)}),$$

where N refers to the sample size and K_h is the so-called kernel. A bivariate Gaussian kernel with bandwidth h was used, where h was chosen according to the *Silverman's rule of thumb* [46].

Offset Correction

The correction procedure's first step was determining the distance Δz between the loudspeaker and the microphone array plane for the experiments $\{A1, D1\}$ and $\{A2\}$. The joint PDF was estimated individually for each evaluated distance Δz. Note that source cases from experiment R2 were excluded from the statistical evaluation since the ground plate reflections would introduce an additional disruptive factor in the positional estimation. It is assumed that the true distance minimizes the variance among any direction associated with $\hat{f}_{R_x,R_y}(\Delta x_i)$, i.e. the spectral norm of the covariance matrix $\Sigma_{\Delta x_i}(\Delta z)$ is minimized, such that

$$\arg\min_{\Delta z} ||\Sigma_{\Delta x_i}(\Delta z)||_2.$$

Fig. 6 shows the joint PDF with the smallest spectral norm for the experiments $\{A1, D1\}$ and $\{A2\}$. Based on the joint PDF corresponding to the optimal distance correction Δz, the true positional offset in vertical and horizontal direction is determined from the maximum of the corresponding marginal distributions depicted in Fig. 7. Table 3 shows the positional offset correction values for each of the experiments.

Table 3: Positional correction values for each experiment.

Scenarios	Δx[mm]	Δy[mm]	Δz[mm]
A1, D1	−4.6 mm	1.4 mm	4.0 mm
A2, R2	−5.2 mm	−0.4 mm	6.0 mm

With the correction offset applied, one can conclude that the positional uncertainties regarding the true source positions are in the order of a few millimetres. Given the 2.5 and 97.5 percentiles of the marginal distributions, the positional uncertainty is in the range of $[-3.6\,\mathrm{mm}, 3.4\,\mathrm{mm}]$ in x-direction and $[-2.1\,\mathrm{mm}, 3.5\,\mathrm{mm}]$ in y-direction for the experiments $\{A1, D1\}$. Regarding the experiments $\{A2, R2\}$, the positional uncertainty is in the range of $[-4.9\,\mathrm{mm}, 1.4\,\mathrm{mm}]$ in x-direction and $[-2.6\,\mathrm{mm}, 3.7\,\mathrm{mm}]$ in y-direction.

Abbreviations

DFT Discrete Fourier Transform

PDF Probability Density Function

RIR Room Impulse Response

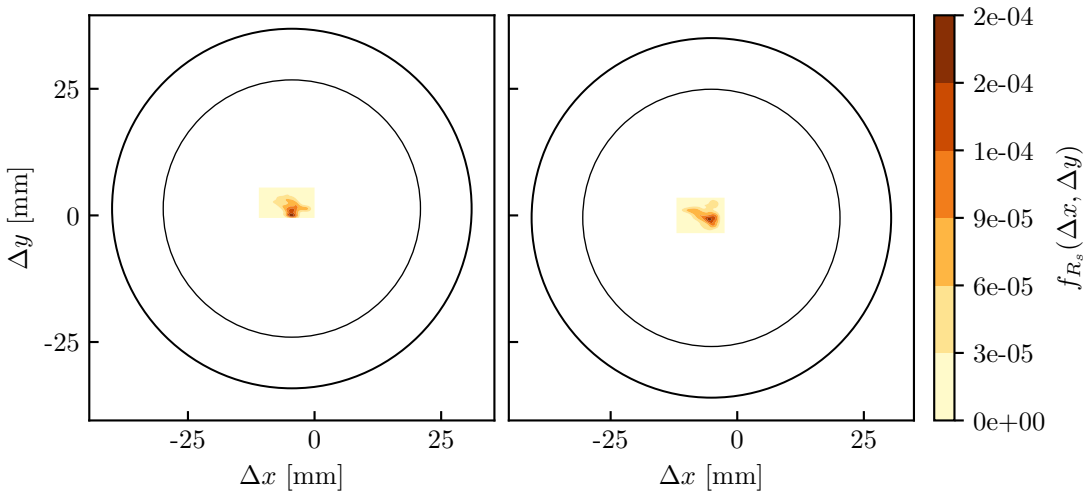

Figure 6: Estimated joint PDF of the positional deviations between the beamforming results and the assumed source positions. The inner black circle corresponds to the outer rim of the loudspeaker and the outer black circle indicates the outer rim of the enclosure box (left: Experiments {A1, D1}, right: Experiment A2).

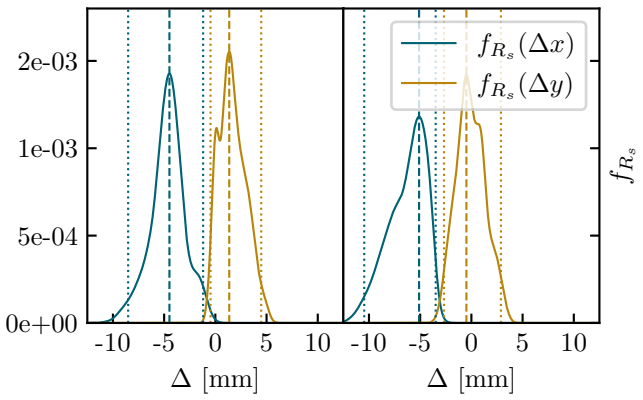

Figure 7: Marginal distribution functions characterizing the positional offset between the microphone array and the observation plane (left: Experiments {A1, D1}, right: Experiment A2). The dashed line indicates the positional offset corresponding to the maximum of the corresponding PDF. The dotted lines indicate the 2.5% and 97.5% percentiles.

Acknowledgments

The authors also thank Arya Prasetya, Serdar Gareayaghi, Can Kurt Kayser and Roman Tschakert for their help with the experimental measurements and Fabian Brinkmann for valuable insights into sweep synthesis and experiment design.

The authors thankfully acknowledge the support of this research by Deutsche Forschungsgemeinschaft through projects 439144410 and 504367810.

References

[1] *Scipy v1.11.4 manual*, https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.resample_poly.html (accessed 2023-12-18).

[2] E. J. Arcondoulis, Q. Li, S. Wei, Y. Liu, and P. Xu, *Experimental validation and performance analysis of deep learning acoustic source imaging methods*, in 28th AIAA/CEAS Aeroacoustics Conference, Southampton, UK, 6 2022, https://doi.org/10.2514/6.2022-2852.

[3] L. L. Beranek and T. J. Mellow, *Acoustics: Sound Fields and Transducers*, Academic Press, and imprint of Elsevier, Amsterdam, first edition ed., 2012.

[4] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C.-A. Deledalle, *Machine learning in acoustics: Theory and applications*, The Journal of the Acoustical Society of America, 146 (2019), pp. 3590–3628, https://doi.org/10.1121/1.5133944.

[5] G. Bologni, R. Heusdens, and J. Martinez, *Acoustic reflectors localization from stereo recordings using neural networks*, in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 1–5, https://doi.org/10.1109/ICASSP39728.2021.9414473.

[6] N. J. Bryan, *Impulse Response Data Augmentation and Deep Neural Networks for Blind Room Acoustic Parameter Estimation*, in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 5 2020, IEEE, pp. 1–5, https://doi.org/10.1109/ICASSP40776.2020.9052970.

[7] M. Cobos, J. Ahrens, K. Kowalczyk, and A. Politis, *An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction*, EURASIP Journal on Audio, Speech, and

Music Processing, 2022 (2022), p. 10, https://doi.org/10.1186/s13636-022-00242-x.

[8] O. Cramer, *The variation of the specific heat ratio and the speed of sound in air with temperature, pressure, humidity, and CO2 concentration*, The Journal of the Acoustical Society of America, 93 (1993), pp. 2510–2516, https://doi.org/10.1121/1.405827.

[9] R. S. Davis, *Equation for the determination of the density of moist air (1981/91)*, Metrologia, 29 (1992), p. 67, https://doi.org/10.1088/0026-1394/29/1/008, https://dx.doi.org/10.1088/0026-1394/29/1/008.

[10] S. Dilungana, A. Deleforge, C. Foy, and S. Faisan, *Learning-based estimation of individual absorption profiles from a single room impulse response with known positions of source, sensor and surfaces*, in INTER-NOISE and NOISE-CON Congress and Conference Proceedings, vol. 263, 2021, pp. 5623–5630, https://doi.org/10.3397/IN-2021-3186.

[11] A. Farina, *Simultaneous Measurement of Impulse Response and Distortion with Swept-sine technique*, in 108th AES Convention, Paris, France, 2 2000.

[12] A. Farina, *Advancements in impulse response measurements by sine sweeps*, in 122nd AES Convention, Vienna, Austria, 2007, p. 21.

[13] E. Fernandez-Grande, X. Karakonstantis, D. Caviedes-Nozal, and P. Gerstoft, *Generative models for sound field reconstruction*, The Journal of the Acoustical Society of America, 153 (2023), pp. 1179–1190, https://doi.org/10.1121/10.0016896.

[14] A. Francl and J. McDermott, *Deep neural network models of sound localization reveal how perception is adapted to real-world environments*, Nature Human Behaviour, 6 (2022), pp. 111–133, https://doi.org/10.1101/2020.07.21.214486.

[15] H. Gamper and I. J. Tashev, *Blind reverberation time estimation using a convolutional neural network*, in 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 9 2018, pp. 136–140, https://doi.org/10.1109/IWAENC.2018.8521241.

[16] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, *A consolidated perspective on multimicrophone speech enhancement and source separation*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25 (2017), pp. 692–730, https://doi.org/10.1109/TASLP.2016.2647702.

[17] A. Geldert, N. Meyer-Kahlen, and S. J. Schlecht, *Interpolation of Spatial Room Impulse Responses Using Partial Optimal Transport*, in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, IEEE, pp. 1–5, https://doi.org/10.1109/ICASSP49357.2023.10095452.

[18] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, *A Survey of Sound Source Localization with Deep Learning Methods*, The Journal of the Acoustical Society of America, 152 (2022), pp. 107–151, https://doi.org/10.1121/10.0011809.

[19] E. Guizzo, R. F. Gramaccioni, S. Jamili, C. Marinoni, E. Massaro, C. Medaglia, G. Nachira, L. Nucciarelli, L. Paglialunga, M. Pennese, S. Pepe, E. Rocchi, A. Uncini, and D. Comminiello, *L3DAS21 Challenge: Machine Learning for 3D Audio Signal Processing*, in Proceedings of the International Workshop on Machine Learning for Signal Processing (MLSP), Gold Coast, Australia, 10 2021, IEEE, https://doi.org/10.1109/MLSP52302.2021.9596248.

[20] E. Guizzo, C. Marinoni, M. Pennese, X. Ren, X. Zheng, C. Zhang, B. Masiero, A. Uncini, and D. Comminiello, *L3DAS22 Challenge: Learning 3D Audio Sources in a Real Office Environment*, in Proceedings of the ICASSP, Singapore, Singapore, 5 2022, IEEE, pp. 9186–9190, https://doi.org/10.1109/ICASSP43922.2022.9746872.

[21] Y. Haneda, Y. Kaneda, and N. Kitawaki, *Common-acoustical-pole and residue model and its application to spatial interpolation and extrapolation of a room transfer function*, IEEE Transactions on Speech and Audio Processing, 7 (1999), pp. 709–717, https://doi.org/10.1109/89.799696.

[22] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM Monographs on Mathematical Modeling and Computation, SIAM, Philadelphia, 1998, https://doi.org/10.1137/1.9780898719697.

[23] T. Hironori, K. Ole, N. P. A, and H. Hareo, *Inverse filter of sound reproduction systems using regularization*, IEICE Trans. Fundamentals, A, 80 (1997), pp. 809–820, https://cir.nii.ac.jp/crid/1572824502324497664 (accessed 2024-03-13).

[24] M. Holters, T. Corbach, and U. Zölzer, *Impulse response measurement techniques and their applicability in the real world*, in 12th Int. Conference on Digital Audio Effects (DAFx-09), 2009, https://www.dafx.de/paper-archive/details.php?id=1u-OdqevtbweDYmNY2_kuA (accessed 2024-03-13).

[25] J. Huang and T. Bocklet, *Intel Far-Field Speaker Recognition System for VOiCES Challenge 2019*, in Proc. Interspeech 2019, 2019, pp. 2473–2477, https://doi.org/10.21437/Interspeech.2019-2894.

[26] F. Katzberg, R. Mazur, M. Maass, M. Böhme, and A. Mertins, *Spatial interpolation of room impulse responses using compressed sensing*, in 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 9 2018, pp. 426–430, https://doi.org/10.1109/IWAENC.2018.8521390.

[27] A. Kujawski and E. Sarradj, *Fast grid-free strength mapping of multiple sound sources from microphone array data using a Transformer architecture*, The Journal of the Acoustical Society of America, 152 (2022), pp. 2543–2556, https://doi.org/10.1121/10.0015005.

[28] M. Lee and J.-H. Chang, *Deep neural network based blind estimation of reverberation time based on multichannel microphones*, Acta Acustica united with Acustica, 104 (2018), pp. 486–495, https://doi.org/10.3813/AAA.919191.

[29] S. Lee, H.-S. Choi, and K. Lee, *Yet another generative model for room impulse response estimation*, in 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 10 2023, pp. 1–5, https://doi.org/10.1109/WASPAA58266.2023.10248189.

[30] F. Lluís, P. Martínez-Nuevo, M. Bo Møller, and S. Ewan Shepstone, *Sound field reconstruction in rooms: Inpainting meets super-resolution*, The Journal of the Acoustical Society of America, 148 (2020), pp. 649–659, https://doi.org/10.1121/10.0001687.

[31] T. Lobato, R. Sottek, and M. Vorländer, *Deconvolution with neural grid compression: A method to accurately and quickly process beamforming results*, The Journal of the Acoustical Society of America, 153 (2023), pp. 2073–2089, https://doi.org/10.1121/10.0017792.

[32] R. Merino-Martínez, P. Sijtsma, M. Snellen, T. Ahlefeldt, J. Antoni, C. J. Bahr, D. Blacodon, D. Ernst, A. Finez, S. Funke, T. F. Geyer, S. Haxter, G. Herold, X. Huang, W. M. Humphreys, Q. Leclère, A. Malgoezar, U. Michel, T. Padois, A. Pereira, C. Picard, E. Sarradj, H. Siller, D. G. Simons, and C. Spehr, *A review of acoustic imaging methods using phased microphone arrays*, CEAS Aeronautical Journal, 10 (2019), pp. 197–230, https://doi.org/10.1007/s13272-019-00383-4.

[33] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, *A unifying view on dataset shift in classification*, Pattern Recognition, 45 (2012), pp. 521–530, https://doi.org/10.1016/j.patcog.2011.06.019.

[34] S. Müller and P. Massarani, *Transfer-function measurement with sweeps*, Journal of the Audio Engineering Society, 49 (2001), pp. 443–471, https://www.aes.org/e-lib/browse.cfm?elib=10189 (accessed 2023-12-19).

[35] M. Müller-Trapet, *On the practical application of the impulse response measurement method with swept-sine signals in building acoustics*, The Journal of the Acoustical Society of America, 148 (2020), pp. 1864–1878, https://doi.org/10.1121/10.0001916.

[36] K. Müller and F. Zotter, *Auralization based on multi-perspective ambisonic room impulse responses*, Acta Acustica, 4 (2020), 25, https://doi.org/10.1051/aacus/2020024.

[37] K. Nagatomo, M. Yasuda, K. Yatabe, S. Saito, and Y. Oikawa, *Wearable Seld Dataset: Dataset For Sound Event Localization And Detection Using Wearable Devices Around Head*, in Proceedings of the ICASSP, Singapore, Singapore, 5 2022, IEEE, pp. 156–160, https://doi.org/10.1109/ICASSP43922.2022.9746544.

[38] S. G. Norcross, M. Bouchard, and G. A. Soulodre, *Inverse filtering design using a minimal-phase target function from regularization*, in Audio Engineering Society Convention 121, San Francisco, CA, USA, Oct. 2006, Audio Engineering Society, https://www.aes.org/e-lib/browse.cfm?elib=13763 (accessed 2024-03-13).

[39] E. Parzen, *On estimation of a probability density function and mode*, The Annals of Mathematical Statistics, 33 (1962), pp. 1065–1076.

[40] K. Prawda, S. J. Schlecht, and V. Välimäki, *Robust selection of clean swept-sine measurements in non-stationary noise*, The Journal of the Acoustical Society of America, 151 (2022), pp. 2117–2126, https://doi.org/10.1121/10.0009915.

[41] A. Ratnarajah, Z. Tang, R. Aralikatti, and D. Manocha, *Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes*, in Proceedings of the 30th ACM International Conference on Multimedia, Lisboa Portugal, 10 2022, Association for Computing Machinery, New York, NY, United States, pp. 924–933, https://doi.org/10.1145/3503161.3548253.

[42] M. Rébillat, R. Hennequin, É. Corteel, and B. F. Katz, *Identification of cascade of hammerstein models for the description of nonlinearities in vibrating devices*, Journal of Sound and Vibration, 330 (2011), pp. 1018–1038, https://doi.org/10.1016/j.jsv.2010.09.012.

[43] E. Sarradj, *Three-dimensional acoustic source mapping with different beamforming steering vector formulations*, Advances in Acoustics and Vibration, (2012), 292695, https://doi.org/10.1155/2012/292695.

[44] E. Sarradj, *A Generic Approach To Synthesize Optimal Array Microphone Arrangements*, in 6th Berlin Beamforming Conference, Berlin, Germany, 2 2016, Gesellschaft zur Förderung angewandter Informatik (GFaI), pp. 1–12.

[45] M. R. Schroeder, *New Method of Measuring Reverberation Time*, The Journal of the Acoustical Society of America, 37 (2005), pp. 409–412, https://doi.org/10.1121/1.1909343.

[46] B. W. Silverman, *Density estimation for statistics and data analysis*, Chapman & Hall/CRC monographs on statistics and applied probability, Chapman and Hall, London, 1986, https://cds.cern.ch/record/1070306.

[47] P. Srivastava, *Realism in virtually supervised learning for acoustic room characterization and sound source localization*, theses, Université de Lorraine, 2023, https://theses.hal.science/tel-04313405.

[48] P. Srivastava, A. Deleforge, A. Politis, and E. Vincent, *How to (Virtually) Train Your Speaker Localizer*, in Proc. INTERSPEECH 2023, Dublin, Ireland, 8 2023, ISCA, pp. 1204–1208, https://doi.org/10.21437/Interspeech.2023-1065.

[49] W. Yu and W. B. Kleijn, *Room acoustical parameter estimation from room impulse responses using deep neural networks*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29 (2021), pp. 436–447, https://doi.org/10.1109/TASLP.2020.3043115.

A. Experiment Equipment

Table 4 lists the hardware devices that were used in the experiments. The calibration of the temperature sensor was performed after the measurement campaign using a reference sensor with a temperature accuracy of $\pm 0.1\,°C$.

Table 4: Utilized hardware devices.

Device	Manufacturer	Type	Usage
Microphones	GRAS	40PL-1 Short CCP	Sound pressure acquisition
Temperature Sensor	OMNI SENSORS	OT60-B ($\pm 0.8\,°C$)	Temperature acquisition
Acquisition System	SINUS	Typhoon	Data acquisition
Stepper Motor	Stepperonline	NEMA23	Axes positioning
Motor Control Unit	OpenBuilds	Blackbox X32	Control loudspeaker position
Amplifier	Klein & Hummel	Monoblock MB 80	Loudspeaker amplification
Turntable	Outline	ET2	Directivity measurement
Laser distance meter	PeakTech	2800A	Positional alignment
Cross line laser	Bosch	PCL20	Positional alignment

B. File Structure

The files `A1.h5`, `A2.h5` and `R2.h5` have a size of about 1.07 GB and `D1.h5` has a size of about 302.3 MB. Their contents are organized as follows:

<Dataset>

- `data`
  - `impulse_response` — `float32` array of shape $(n_\mathrm{i}, n_\mathrm{o}, n_t)$ - measured impulse responses
  - `location`
    - `receiver` — `float64` array of shape $(n_\mathrm{o}, 3)$ - microphone locations
    - `source` — `float64` array of shape $(n_\mathrm{o}, 3)$ - corrected source locations
    - `source_raw` — `float64` array of shape $(n_\mathrm{o}, 3)$ - uncorrected source locations
- `metadata`
  - `c0` — `float32` array of shape $(n_\mathrm{i},)$ - speed of sound
  - `temperature` — `float32` array of shape $(n_\mathrm{i},)$ - ambient temperature
  - `sampling_rate` — `int64` - sampling rate

We also supply the file `loudspeaker.h5` with a size of about 468 KB which contains the directivity measurements of the loudspeaker. Its contents are organzine as follows:

<Dataset>

- `data`
  - `angle` — `float32` array of shape $(73,)$ - measurement angles
  - `impulse_response` — `float32` array of shape $(73, n_t)$ - measured impulse responses
- `metadata`
  - `directivity` — `float32` array of shape $(73, 513)$ - directivity D
  - `directivity_index` — `float64` array of shape $(513,)$ - directivity index DI
  - `fftfreq` — `float64` array of shape $(513,)$ - corresponding frequencies
  - `sampling_rate` — `int64` - sampling rate

C. Loading the Files

Listing 1: Python code snippet for loading the data.

```python
from h5py import File

with File('A1').with_suffix('.h5'), 'r') as f:
    ir = f['data']['impulse_response'][()]
    fs = f['metadata']['sampling_rate'][()]
```

Listing 2: Matlab code snippet for loading the data.

```
ir = h5read('A1.h5', '/data/impulse_response')
fs = h5read('A1.h5', '/metadata/sampling_rate')
```