dEchorate: a Calibrated Room Impulse Response Database for Echo-aware Signal Processing

Diego Di Carlo, Pinchas Tandeitnik, Cédric Foy, Nancy Bertin, Antoine Deleforge, Sharon Gannot

RESEARCH

dEchorate: a Calibrated Room Impulse Response Database for Echo-aware Signal Processing

Diego Di Carlo[1*†], Pinchas Tandeitnik[2], Cédric Foy[3], Antoine Deleforge[4], Nancy Bertin[1] and Sharon Gannot[2]

Abstract

This paper presents dEchorate: a new database of measured multichannel Room Impulse Responses (RIRs) including annotations of early echo timings and 3D positions of microphones, real sources and image sources under different wall configurations in a cuboid room. These data provide a tool for benchmarking recent methods in *echo-aware* speech enhancement, room geometry estimation, RIR estimation, acoustic echo retrieval, microphone calibration, echo labeling and reflectors estimation. The database is accompanied with software utilities to easily access, manipulate and visualize the data as well as baseline methods for echo-related tasks.

Keywords: Echo-aware signal processing; Acoustic echoes; Room impulse response; Audio database; Acoustic Echo Retrieval; Spatial Filtering; Room Geometry Estimation; Microphone arrays

1 Introduction

When sound travels from a source to a microphone in a indoor space, it interacts with the environment by being delayed and attenuated due to the distance; and reflected, absorbed and diffracted due to the surfaces. The Room Impulse Response (RIR) represents this phenomenon as a linear and causal time-domain filter. As depicted in Figure 1, RIRs are commonly subdivided into 3 parts: the *direct-path*, corresponding to the line-of-sight propagation; the *early echoes*, stemming from few disjoint reflections on the closest reflectors; and the *late reverberation* comprising the dense accumulation of later reflections and *scattering* effects.

The late reverberation is indicative of the environment size and reverberation time, producing the so-called *listener envelopment*, *i.e.*, the degree of immersion in the sound field [1]. In contrast, the direct path and the early echoes carry precise information on the scene's geometry, such as the position of the source and room surfaces relative to the receiver position [2], and on the surfaces' reflectivity. Such relation is well explained by the Image Source Method (ISM) [3], in which the echoes are associated with the contribution of virtual sound sources lying outside the real room.

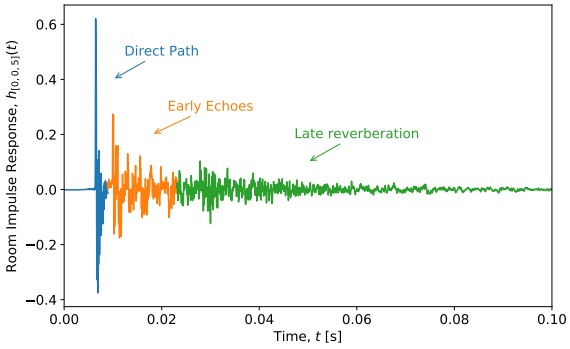

Figure 1 Depiction of a measured room impulse response from the database.

Therefore, one may consider early echoes as "spatialized" copies of the source signal, whose Times of Arrival (TOAs) are related to the source and reflector positions.

Based on this idea, so-called *echo-aware* methods have been introduced a few decades ago, where *matched filters* (or *rake receivers*) are used to constructively sum the sound reflections [4–6] and build beamformers achieving much better sound qualities [7]. These methods have recently regained interested as manifested by the European project SCENIC [8]

*Correspondence: diego.di-carlo@inria.fr
[1]Univ Rennes, Inria, CNRS, IRISA, France
Full list of author information is available at the end of the article
†The first author performed the work while at Bar-Ilan University.

and the UK research project S^3A[1]. Later, a few studies showed that knowing the properties of a few early echoes could boost the performance of typical indoor audio inverse problems such as Speech Enhancement (SE) [9, 10], sound source localization [11–14] and separation [15–18], and speaker verification [19].

Another fervent area of research spanning transversely the audio signal processing field is estimating the room geometry blindly from acoustic signals [20–23]. As recently reviewed by Crocco *et al.* in [22], end-to-end Room Geometry Estimation (RooGE) involves a number of subtasks: RIR estimation, peak picking, microphones calibration, echo labeling and reflectors' position estimation. As interesting applications, these methods have been recently used in active setting (*i.e.*, knowing the transmitted signals) on unmanned aerial vehicles (UAVs, a.k.a. drones) [24, 25] and on mobile-phones [26]. The lowest common denominator of all these tasks is Acoustic Echo Retrieval (AER), that is, estimating the properties of early echoes, such as their TOAs and energies. The former problem is typically referred to as TOA estimation, or Time Difference of Arrival (TDOA) estimation when the direct-path is taken as reference.

As listed in [27] and in [28], a number of recorded RIRs corpora are freely available online, each of them meeting the demands of certain applications. Table 1 summarizes the main characteristics of some of them. One can broadly identify two main classes of echo-aware RIR datasets in the literature: SE/Automatic Speech Recognition (ASR)-oriented datasets, *e.g.* [27, 30, 31], and RooGE-oriented datasets, *e.g.* [21–23]. The former regards acoustic echoes as highly correlated interfering sources coming from close reflectors, such as a table in a meeting room or a near wall. This typically presents a challenge in estimating the correct source's Direction of Arrival (DOA) with further consequences in DOA-based enhancement algorithm, *e.g.*, beamformers. Although this factor is taken into account, such datasets lack proper annotation of these echoes in the RIRs or the absolute position of objects inside the room. The latter group typically features design choices, such as microphones scattered across the room, which are not suitable for SE applications. Indeed, these typically involve compact or ad hoc arrays. The main common drawback of these datasets in that they cannot be easily used for other tasks than the ones which they are designed for.

To bypass the complexity of recording and annotating real RIR datasets, acoustic simulators based on the ISM are extensively used instead [32, 33, 33–35]. While such data are more versatile, simpler and quicker to obtain, they fail to fully capture the complexity and richness of real acoustic environments. Due to this, methods trained, calibrated, or validated on them may fail to generalize to real conditions, as will be shown in this paper. Interestingly, in the context of learning-based blind room volume estimation, the authors of [28] combined multiple real and synthetic RIR datasets in order to find a balance between number of training data and realism.

A good echo-oriented RIR dataset should include a variety of environments (room geometries and surface materials), of microphone placings (close to or away from reflectors, scattered or forming ad-hoc arrays) and, most importantly, precise annotations of the scene's geometry and echo timings in the RIRs. Moreover, in order to be versatile and used in both SE and RooGE applications, geometry and timing annotations should be fully consistent. Such data are difficult to collect since it involves precise measurements of the positions and orientations of all the acoustic emitters, receivers and reflective surfaces inside the environment with dedicated planimetric equipment.

To fill this gap, we present the `dEchorate` dataset: a fully calibrated multichannel RIR database with accurate annotation of the geometry and echo timings in different configurations of a cuboid room with varying wall acoustic profiles. The database currently features 1800 annotated RIRs obtained from 6 arrays of 5 microphones each, 6 sound sources and 11 different acoustic conditions. All the measurements were carried out at the acoustic lab at Bar-Ilan University following a consolidated protocol previously established for the realization of two other multichannel RIRs databases: the BIU's Impulse Response Database [29] gathering RIRs of different reverberation levels sensed by uniform linear arrays (ULAs); and `MIRaGE` [31] providing a set of measurements for a source placed on a dense position grid. The `dEchorate` dataset is designed for AER with linear arrays, and is more generally aimed at analyzing and benchmarking RooGE and echo-aware signal processing methods on real data. In particular, it can be used to assess robustness against the number of reflectors, the reverberation time, additive spatially-diffuse noise and non-ideal frequency and directive characteristics of microphone-source pairs and surfaces in a controlled way. Due to the amount of data and recording conditions, it could also be used to train machine learning models or as a reference to improve RIR simulators. The database is accompanied with a Python toolbox that can be used to process and visualize the data, perform analysis or annotate new datasets.

The remainder of the paper is organized as follows. Section 2 describes the construction and the composition of the dataset, while Section 3 provides an

[1] http://www.s3a-spatialaudio.org/

Table 1 Comparison between some existing RIR databases that account for early acoustic reflections. Receiver positions are indicated in terms of number of microphones per array times number of different positions of the array (∼ stands for partially available information). The read is invited to refer to [27, 28] for more complete list of existing RIR datasets.
†The dataset in [23] is originally intended for RooGE and further extended for (binaural) SE in [18] with a similar setup.
‡These datasets have been recorded in the same room.

Database Name	Annotated			Number of				Key characteristics	Purpose
	Pos.	Echoes	Rooms	RIRs	Rooms	Mic×Pos.	Src		
Dokmanić *et al.* [21]	✓	∼	∼	15	3	5	1	Non shoebox rooms	RooGE
Crocco *et al.* [22]	✓	∼	✓	204	1	17	12	Accurate 3D calibration Many mic and src positions	RooGE
Remaggi *et al.* [23]†	✓	∼	✓	∼1.5k	4	48×2	4-24	Circural dense array Circular placement of sources	RooGE SE
Remaggi *et al.* [18]†	✓	∼	✓	∼1.6k	4	48×2 +2×2	3-24	Circural dense array Binaural Recordings	RooGE SE
`BIU's Database` [29]‡	✓	✗	✗	∼1.8k	3	8×3	26	Linear array with different spacing Circular placement of sources	SE
`BUT-Reverb` [27]	✓	✗	∼	∼1.3k	8	(2-10)×6	3-11	Accurate metadata different device/arrays various rooms	SE/ASR
`VoiceHome` [30]	✓	✗	✗	188	12	8×2	7-9	Various rooms, real homes	SE/ASR
`MIRaGE` [31]‡	✓	✗	✗	371k	3	5×6	25 (+ 4104)	4104 src. pos. in a dense grid different acoustic rooms	SE/ASR
`dEchorate`‡	✓	✓	✓	∼1.8k	11	5×6	6	Accurate echo annotation different surface absorptions	RooGE SE/ASR

Figure 2 Broad-view picture of the acoustic lab at Bar-Ilan university.

overview of the data, studying the variability of typical acoustic parameters. To validate the data, in Section 4 two echo-aware application are presented, one in speech enhancement and one is room geometry estimation. Finally, in Section 5 the paper closes with the conclusions and and offers leads for work.

2 Database Description

2.1 Recording setup

The recording setup is placed in a cuboid room with dimension 6 m × 6 m × 2.4 m. The 6 facets of the room (walls, ceiling, floor) are covered by acoustic panels allowing controllable reverberation time ($\mathtt{RT}_{60}$). We placed 4 directional loudspeakers (direct sources) facing the center of the room and 30 microphones mounted on 6 static linear arrays parallel to the

Table 2 Measurement and recording equipment.

Loudspeakers	(directional, direct) 4× Avanton (directional, indirect) 2× Avanton (omnidirectional) 1× B&G (babble noise) 4× 6301bx Fostex
Microphones	30× AKG CK32
Array	6× nULA (5 mics each, handcrafted)
A/D Converter	ANDIAMO.MC
Indoor Positioning	Marvelmind Starter Set HW v4.9

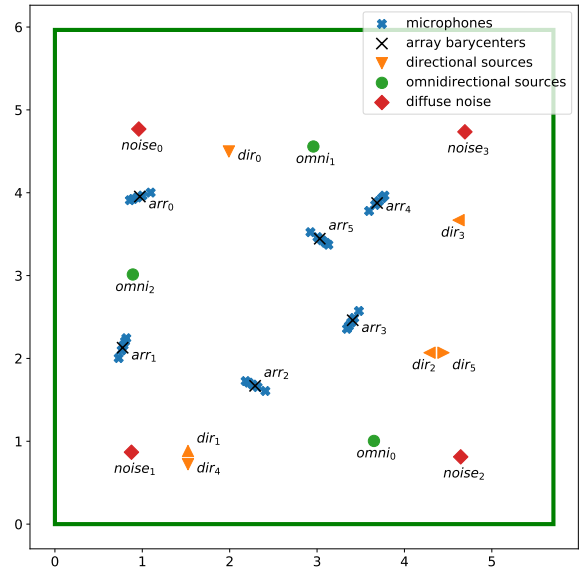

Figure 3 Illustration of the recording setup - top view.

ground. An additional channel is used for the loopback signal, which serves to compute the time of emission and detect errors. Each loudspeaker and each array is positioned close to one of the walls in such a way that the source of the strongest echo can be easily identified. Moreover, their positioning was chosen to cover a wide distribution of source-to-receiver distances, hence, a wide range of Direct-to-Reverberant ratios (DRRs). Further, 2 more loudspeakers were positioned pointing towards the walls (indirect sources). This was done to study the case of early reflections being stronger than the direct-path.

Each linear array consists of 5 microphones with non-uniform inter-microphone spacings of $[4, 5, 7.5, 10]$ cm[2]. Hereinafter we will refer to these elements as non-Uniform Linear Arrays (nULAs).

2.2 Measurements

The main feature of this room is the possibility to change the acoustic profile of each of its facets by flipping double-sided panels with one reflective (made of Formica Laminate sheets) and one absorbing face made of perforated panels filled with rock-wool). A complete list of the materials of the room is available in Section 5. This allows to achieve diverse values of $\mathtt{RT}_{60}$ that range from 0.1 to almost 1 second. In this dataset, the panels of the floor were always kept absorbent.

[2] that is, $[-12.25, -8.25, -3.25, 3.25, 13.25]$ cm w.r.t. the barycenter

Table 3 Surface coding in the dataset: each binary digit indicates if the surface is absrobent (0, ✗) or reflective (1, ✓).

	Surfaces:	Floor	Ceil	West	South	East	North
one-hot	010000	✗	✓	✗	✗	✗	✗
	001000	✗	✗	✓	✗	✗	✗
	000100	✗	✗	✗	✓	✗	✗
	000010	✗	✗	✗	✗	✓	✗
	000001	✗	✗	✗	✗	✗	✓
incremental	000000	✗	✗	✗	✗	✗	✗
	010000	✗	✓	✗	✗	✗	✗
	011000	✗	✓	✓	✗	✗	✗
	011100	✗	✓	✓	✓	✗	✗
	011110	✗	✓	✓	✓	✓	✗
	011111	✗	✓	✓	✓	✓	✓
f.	010001*	✗	✓	✗	✗	✗	✓

Two types of measurement sessions were considered, namely, *one-hot* and *incremental*. For the first type, a single facet was placed in reflective mode while all the others were kept absorbent. For the second type, starting from fully-absorbent mode, facets were progressively switched to reflective one after the other until all but the floor are reflective, as shown in Table 3. The dataset features an extra recording session. For this session, office furnitures (chairs, coat-hanger and a table) were positioned in the room to simulate a typical meeting room with chairs and tables (see Figure 2). Theses recordings may be used to assert the robustness of echo-aware methods in a more realistic scenario

For each room configuration and loudspeaker, three different excitation signals were played and recorded in sequence: chirps, white noise and speech utterances. The former consists in a repetition of 3 Exponentially Swept-frequency Sine (ESS) signals of duration 10 seconds and frequency range from 100 Hz to 14 kHz interspersed with 2 seconds of silence. Such frequency range was chosen to match the characteristics of the loudspeakers. To prevent rapid phase changes and "popping" effects, the signals were linearly faded in and out over 0.2 seconds with a Tuckey taper window.[3] Second, 10 seconds bursts of white noise and 3 anechoic speech utterances from the Wall Street Journal (WSJ) dataset [36] were played in the room. Through all recordings, at least 40 dB of sound dynamic range compared to the room silence was asserted, and a room temperature of $24° \pm 0.5°$C and 80% relative humidity were registered. In these conditions the speed of sounds is $c_{\text{air}} = 346.98$ m/s. In addition, 1 minute of *room tone* (*i.e.*, silence) and 4 minutes of diffuse babble noise were recorded for each session. The latter was

[3] The code to generate the reference signals and to process them is available together with the data. The code is based on the `pyrirtools` Python library

Figure 4 Picture of the acoustic lab. From left to right: the overall setup, one microphone array, the setup with revolved panels.

simulated by transmitting different chunks of the same single-channel babble noise recording from additional loudspeakers facing the four corners of the room.

All microphone signals were synchronously acquired and digitally converted to 48 kHz with 32 bit/sample using the equipment listed in Table 2. The polarity of each microphone was recorded by clapping a book in the middle of the room and their gain is corrected using the room tone.

Finally, RIRs are estimated with the ESS technique [37] where an exponential time-growing frequency sweep is using as probe signal. Then, the RIR is estimated by devolving the microphone signal, implemented as division in the frequency domain (The authors used the same code mentioned in Footnote 3).

2.3 Dataset annotation

2.3.1 RIRs annotation

The objective of this database is to feature annotations in the "geometrical space", namely the microphone, facet and source positions, that are *fully consistent* with annotations in the "signal space", namely the echo timings within the RIRs. This is achieved as follows:

(i) First, the ground-truth positions of the array and source centres are acquired via a Beacon indoor positioning system (`bIPS`). This system consists in 4 stationary bases positioned at the corners of the ceiling and a movable probe used for measurements which can be located within errors of ±2 cm.

(ii) The estimated RIRs are superimposed on synthetic RIRs computed with the Image Source Method (ISM) from the geometry obtained in the previous step. A Python GUI[4] (showed in Figure 5), is used to manually tune a peak finder and label the echoes corresponding to found peaks, that is, annotate their timings and their corresponding image source position and room facet label.

(iii) By solving a simple Multi-Dimensional Scaling (MDS) problem [38–40], refined microphone and source positions are computed from echo timings. The non-convexity of the problem is alleviated by using a good initialization (obtained at the previous step), by the high SNR of the measurements and, later, by including additional image sources in the formulation. The prior information about the arrays' structures reduced the number of variables of the problem, leaving the 3D positions of the sources and of the arrays' barycenters in addition to the arrays' tilt on the azimuthal plane.

(iv) By employing a multilateration algorithm [41], where the positions of one microphone per array serve as anchors and the TOAs are converted into distances, it is possible to localize image sources alongside the real sources. This step will be further discussed in Section 4.

Knowing the geometry of the room, in step (i) we were able to initially guess the position of the echoes in the RIR. Then, by iterating through steps (ii), (iii) and (iv), the position of the echoes are refined to be consistent under the ISM.

The final geometrical and signal annotation was chosen as a compromise between the `bIPS` measurements and the `MDS` output. While the former ones are noisy but consistent with the scene's geometry, the latter ones match the TOAs but not necessarily the physical world. In particular, geometrical ambiguities such as global rotation, translation and up-down flips were observed. Instead of manually correcting this error, we modified the original problem from using only the direct path distances (`dMDS`) to considering the image sources' TOA of the ceiling as well in the cost function (`dcMDS`). Table 4 shows numerically the *mismatch* (in cm) between the geometric space (defined by the `bIPS` measurements) and the signal space (the one defined by the echo timings, converted to cm based on the speed of sound). To better quantify it, we introduce here a *Goodness of Match (GoM)* metric: it mea-

[4] This GUI is available in the dataset package.

Table 4 Mismatch between geometric measurements and signal measurements in terms of maximum (Max.), average (Avg.) and standard deviation (Std) of absolute mismatch in centimeters. The goodness of match (GoM) between the signal and geometrical measurements is reported as the fraction of matching echo timings for different thresholds in milliseconds.

	Metrics	bIPS	dMDS	dcMDS
Geom.	Max.	0	6.1	1.07
	Avg.±Std.	0	1.8 ± 1.4	0.39 ± 0.2
Signal	Max.	5.86	1.20	1.86
	Avg.±Std.	1.85 ± 1.5	0.16 ± 0.2	0.41 ± 0.3
Mismatch	GoM (0.5 ms)	97.9%	93.4%	98.1%
	GoM (0.1 ms)	26.6%	44.8%	53.1%
	GoM (0.05 ms)	12.5%	14.4%	30.2%

sures the fraction of (first-order) echo timings annotated in the RIRs matching the annotation produced by the geometry within a threshold. Including the ceiling information, `dcMDS` produces a geometrical configuration which has a small mismatch (0.4 cm on average, 1.86 cm max) in both the signal *and* geometric spaces with a 98.1% matching all the first order echoes within a 0.5 ms threshold (*i.e.*, the position of all the image sources within about 17 cm error). It is worth noting that the `bIPS` measurements produce a significantly less consistent annotation with respect to the signal space.

2.3.2 Other tools for RIRs annotation

Finally, we would like to add that the following tools and techniques were found useful in annotating the echoes.

The "skyline" visualization consists in presenting the intensity of multiple RIRs as an image, such that the wavefronts corresponding to echoes can be highlighted [42]. Let $h_n(l)$ be an RIR from the database, where $l = 0, \dots, L-1$ denotes sample index and $n = 0, \dots, N-1$ is an arbitrary indexing of all the microphones for a fixed room configuration. Then, the *skyline* is the visualization of the $L \times N$ matrix $\mathbf{H}$ created by stacking column-wise N normalized *echograms*[5], that is

$$\mathbf{H}_{l,n} = \mid h_n(l) \mid \, / \max \mid h_n(l) \mid, \tag{1}$$

where $\mid \cdot \mid$ denotes the absolute value.

Figure 6 shows an example of skyline for 120 RIRs corresponding to 4 directional sources, 30 microphones and the most reflective room configuration, stacked horizontally, preserving the order of microphones within the arrays. One can notice several clusters of 5 adjacent bins of similar color (intensity) corresponding to the arrivals at the 5 sensors of each nULA. Thanks to the usage of linear arrays, this visualization allowed us to identify both TOAs and their labeling.

Direct path deconvolution/equalization was used to compensate for the frequency response of the source loudspeaker and microphone [20, 43]. In particular, the direct path of the RIR was manually isolated and used as an equalization filter to enhance early reflections from their superimposition before proceed with peak picking. Each RIR was equalized with its respective direct path. As depicted in Figure 5, in some cases this process was required for correctly identifying the underlying TOAs' peaks.

Different facet configurations for the same geometry influenced the peaks' predominance in the RIR, hence facilitating its echo annotation. An example of RIRs corresponding to 2 different facet configurations is shown in Figure 5: the reader can notice how the peak predominance changes for the different configurations.

An automatic peak finder was used on equalized echograms $\bar{\eta}_n(l)$ to provide an initial guess on the peak positions. In this work, peaks are found using the Python library `peakutils` whose parameters were manually tuned.

2.4 Limitations of current annotation

As stated in [44], we want to emphasize that annotating the correct TOAs of echoes and even the direct path in "clean" real RIRs is far from straightforward. The peaks can be blurred out by the loudspeaker characteristics or the concurrency of multiple reflections. Nevertheless, as showed in Table 4, the proposed annotation was found to be sufficiently consistent both in the geometric and in the echo/signal space. Thus, no further refinement was done. This database can be used as a first basis to develop better AER methods which could be used to iteratively improve the annotation, for instance including 2$^{\text{nd}}$ order reflections.

2.5 The `dEchorate` package

The dataset comes with both data and code to parse and process it. The data are presented in 2 modalities: the `raw` data, that is, the collection of recorded wave files, are organized in folders and can be retrieved by querying a simple database table; the `processed` data, which comprise the estimated RIRs and the geometrical and signal annotations, are organized in tensors directly importable in Matlab or Python (*e.g.* all the RIRs are stored in a tensor of dimension $L \times I \times J \times D$, respectively corresponding to the RIR length in samples, the number of microphones, of sources and of

[5] The echogram is defined either as the absolute value or as the squared value of the RIR.

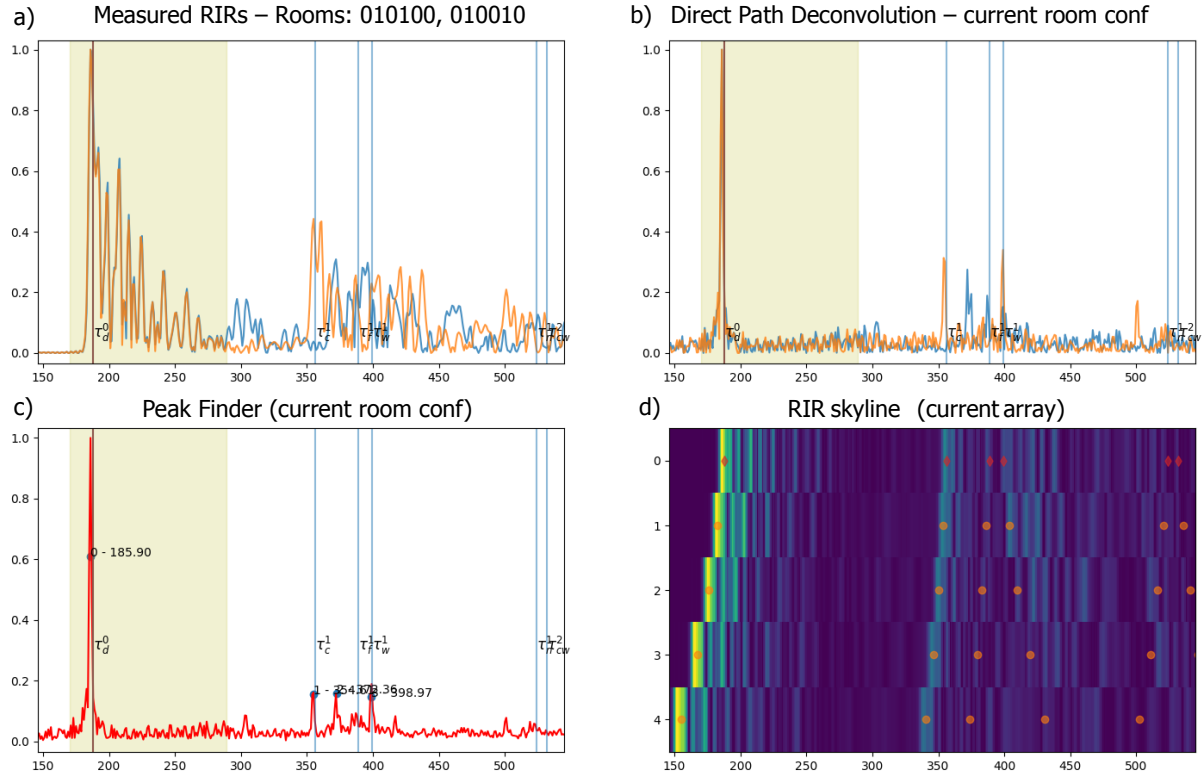

Figure 5 Detail of the GUI used to manually annotate the RIRs. For a given source and a microphone in an nULA, a) and b) each shows 2 RIRs for 2 different room configurations (blue and orange) before and after the direct path deconvolution. c) shows the results of the peak finder for one of the deconvolved RIRs, and d) is a detail on the RIR skyline (See Figure 6) on the corresponding nULA, transposed to match the time axis.

room configurations).

Together with the data a Python package is available on the same website. This includes wrappers, GUI, examples as well as the code to reproduce this study. In particular, all the scripts used for estimating the RIRs and annotating them are available and can be used to further improve and enrich the annotation or as baselines for future works.

3 Analysing the Data

In this section we will illustrate some characteristics of the collected data in term of acoustic descriptors, namely the $\mathtt{RT}_{60}$, the DRR and the Direct-to-Early Ratio (DER). While the former two are classical acoustic descriptors used to evaluate SE and ASR technologies [45], the latter is less common and used in strongly echoic situations [46, 47].

3.1 Reverberation Time

The $\mathtt{RT}_{60}$ is the time required for the sound level in a room to decrease by 60 dB after the source is turned off, thus, it measures reverberation level. This value is one the most common acoustic descriptor for room acoustics. Besides, as reverberation affects detrimentally the performances of speech processing technologies, the robustness against $\mathtt{RT}_{60}$ has become a common evaluation metric in SE and ASR.

Table 5 reports estimated $\mathtt{RT}_{60}(b)$ values per octave band $b \in \{500, 1000, 2000, 4000\}$ (Hz) for each of the room in the dataset. These values were estimated using the Schroeder's integration methods [48–50] in each octave band. For the octave bands centred at 125 Hz and 250 Hz, the measured RIRs did not exhibit sufficient power for a reliable estimation. This observation found confirmation in the frequency response provided by the loudspeakers' manufacturer, which decays exponentially from 300 Hz downwards.

Ideally, for the $\mathtt{RT}_{60}$ to be reliably estimated, the Schroeder curve, *i.e.* the log of the square-integrated, octave-band-passed RIR, would need to feature a linear decay for 60 dB of dynamic range, which would occur in an ideal diffuse sound regime. However, such range is never observable in practice, due to the presence of noise and possible non-diffuse effects. Hence,

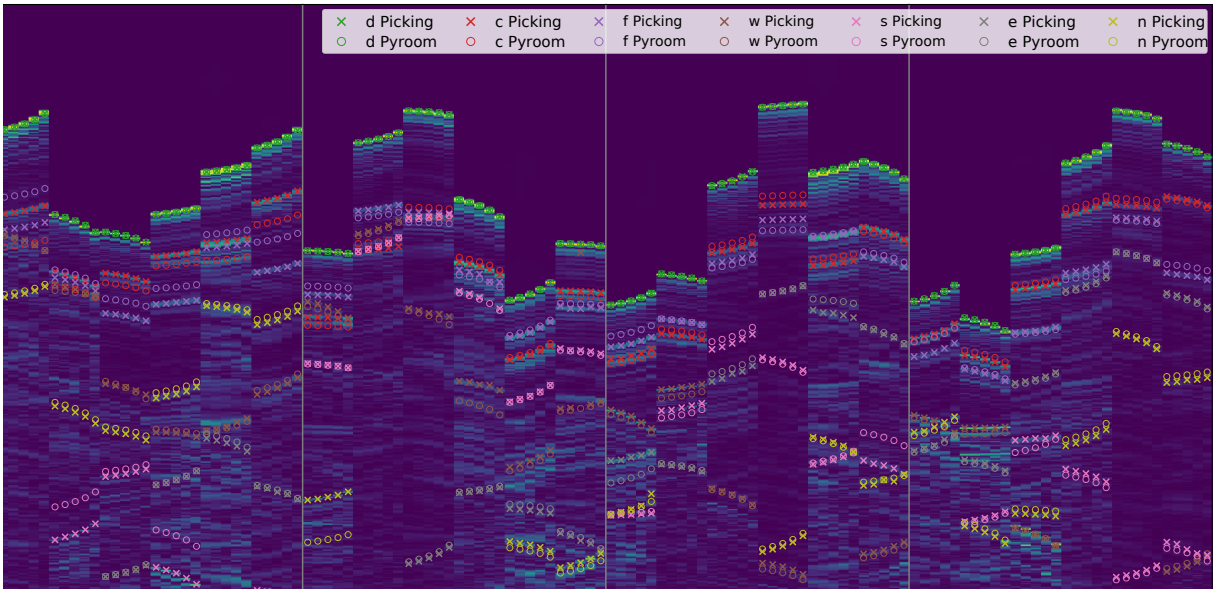

Figure 6 The RIR skyline annotated with observed peaks (×) together with their geometrically-expected position (○) computed with the Pyroomacoustic acoustic simulator. As specified in the legend, markers of different colors are used to indicate the room facets responsible for the reflection: direct path (`d`), ceiling (`c`), floor (`f`), west wall (`w`), . . . , north wall (`n`).

a common technique is to compute, *e.g.*, the $\texttt{RT}_{10}$ on the range $[-5, -15]$ dB of the Schroeder curve and to extrapolate the $\texttt{RT}_{60}$ by multiplying it by 6. We visually inspected all the RIRs of the dataset corresponding to directional sources 1, 2 and 3, *i.e.*, 90 RIRs in each of the 10 rooms. Then, two sets were created. Set $\mathcal{A}$ features all the Schroeder curves featuring linear log-energy decays allowing for reliable $\texttt{RT}_{10}$ estimates. Set $\mathcal{B}$ contains all the other curves. In practice, 49% of the 3600 Schroeder curves were placed in the set $\mathcal{B}$. These mostly correspond to the challenging measurement conditions purposefully included in our dataset, *i.e.*, strong early echoes, loudspeakers facing towards reflectors or receivers close to reflectors. Finally, the $\texttt{RT}_{60}$ value of each room and octave band was calculated from the median of $\texttt{RT}_{10}$ corresponding to Schroeder curves in $\mathcal{A}$ only.

As can be seen in Table 5, obtained reverberation values are consistent with the room progressions described in section 2. Considering the 1000 Hz octave band, the $\texttt{RT}_{60}$ ranges from 0.14 s for the fully absorbent room (`000000`) to 0.73 s for the most reflective room (`011111`). When only one surfaces is reflective the $\texttt{RT}_{60}$ values remains around 0.19 s.

3.2 Direct To Early and Reverberant Ratio

In order to characterize an acoustic environment, it is common to provide the ratio between the energy of the direct and the indirect propagation paths. In particular, one can compute the so-called DRR directly from a measured RIR $h(l)$ [45] as

$$\texttt{DRR} = 10 \log_{10} \frac{\sum_{l \in \mathcal{D}} h^2(l)}{\sum_{l \in \mathcal{R}} h^2 l)} \quad [\text{dB}], \tag{2}$$

where $\mathcal{D}$ denotes the time support comprising the direct propagation path (set to ± 120 samples around its time of arrival, blue part in Figure 1), and $\mathcal{R}$ comprises the remainder of the RIR, including both echoes and late reverberation (orange and green parts in Figure 1).

Similarly, the DER defines the ratio between the energy of the direct path and the early echoes only, that is,

$$\texttt{DER} = 10 \log_{10} \frac{\sum_{l \in \mathcal{D}} h^2(l)}{\sum_{l \in \mathcal{E}} h^2(l)} \quad [\text{dB}], \tag{3}$$

where $\mathcal{E}$ is the time support of the early echoes only (green part in Figure 1).

Differently from the $\texttt{RT}_{60}$ which mainly describes the diffuse regime, both DER and DRR are highly dependent on the position of the source and receiver in the room. Therefore, for each room, wide ranges of these parameters were registered. For the loudspeakers facing the microphones, the DER ranges from 2 dB to 6 dB in one-hot room configurations and from -2 dB to 6 dB in the most reverberant rooms. Similarly, the DRR has a similar trend featuring lower values, such as -2 dB in one-hot rooms and down to -7.5 dB for

Table 5 Reverberation time per octave bands $\mathtt{RT}_{60}(b)$ calculated in the 10 room configurations. For each coefficient, the number of corresponding Schroeder curves in $\mathcal{A}$ used to compute the median estimate is given in parentheses.

	Room 1 000000	Room 2 011000	Room 3 011100	Room 4 011110	Room 5 011111	Room 6 001000	Room 7 000100	Room 8 000010	Room 9 000001	Room 10 010001*
500 Hz	0.18 (11)	0.40 (7)	0.46 (20)	0.60 (51)	0.75 (48)	0.22 (8)	0.21 (5)	0.21 (8)	0.22 (7)	0.37 (12)
1000 Hz	0.14 (62)	0.33 (83)	0.34 (86)	0.56 (89)	0.73 (90)	0.19 (79)	0.19 (74)	0.18 (69)	0.19 (70)	0.26 (72)
2000 Hz	0.16 (65)	0.25 (81)	0.30 (86)	0.48 (82)	0.68 (88)	0.18 (74)	0.20 (64)	0.18 (66)	0.18 (67)	0.24 (69)
4000 Hz	0.22 (15)	0.25 (17)	0.37 (22)	0.55 (16)	0.81 (29)	0.22 (17)	0.23 (12)	0.26 (14)	0.24 (18)	0.28 (14)

the most reverberant ones. A complete annotation of these metrics is available in the database.

4 Using the Data

The dEchorate database is now used to investigate the performance of state-of-the-art methods on two echo-aware acoustic signal processing applications on both synthetic and measured data, namely, spatial filtering and room geometry estimation.

4.1 Application: Echo-aware Beamforming

Let I microphones acquire to a single static point sound source, contaminated by noise sources. In the short-time Fourier transform (STFT) domain, we stack the I complex-valued microphone observations at frequency f and time t into a vector $\mathbf{x}(f,t) \in \mathbb{C}^I$. Let us denote $s(f,t) \in \mathbb{C}$ and $\mathbf{n}(f,t) \in \mathbb{C}^I$ the source signal and the noise signals at microphones, which are assumed to be statistically independent. By denoting $\mathbf{h} \in \mathbb{C}^I$ the Fourier transforms of the RIRs, the observed microphone signals in the STFT domain can be expressed a follows:

$$\mathbf{x}(f,t) = \mathbf{h}(f)s(f,t) + \mathbf{n}(f,t). \tag{4}$$

Here, the STFT windows are assumed long enough so that the discrete convolution-to-multiplication approximation holds well.

Beamforming is one of the most widely used techniques for enhancing multichannel microphone recordings. The literature on this topic spans several decades of array processing and a recent review can be found in [51]. In the frequency domain, the goal of beamforming is to estimate a set of coefficients $\mathbf{w}(f) \in \mathbb{C}^I$ that are applied to $\mathbf{x}(f,t)$, such that $s(f,t) \approx \mathbf{w}^H\mathbf{x}(f,t)$. Hereinafter, we will consider only the *distortionless* beamformers aiming at retrieving the clean target speech signal, as it is generated at the source position.

As mentioned throughout the paper, the knowledge of early echoes is expected to boost spatial filtering performances. However, estimating these elements is difficult in practice. To quantify this, we compare *echo-agnostic* and *echo-aware* beamformers. In order to study their empirical potential, we will evaluate their performance using both synthetic and measured data, as available in the presented dataset.

Echo-agnostic beamformers do not need any echo-estimation step: they either ignore their contributions, as in the direct-path delay-and-sum beamformer (DS) [52], or they consider coupling filters between pairs of microphones, called Relative Transfer Functions (ReTFs) [7]. Note that contrary to RIRs, there exist efficient methods to estimate ReTFs from multichannel recordings of unknown sources (see [51, Section VI.B] for a review). The ReTFs can then be naturally incorporated in powerful beamforming algorithms achieving speech dereverberation and noise reduction in static [53] and dynamic scenarios [54]. In this work, ReTFs are estimated using Generalized Eigenvector Decomposition (GEVD) method [55], using the approach illustanted in [56].

Echo-aware beamformers fall in the category of *rake receivers*, borrowing the idea from telecommunication where an antenna *rakes* (*i.e.*, combines) coherent signals arriving from different propagation paths [4–6]. To this end, they typically consider that for each RIR i, the delays and frequency-independent attenuation coefficients of R early echoes are known, denoted here as $\tau_i^{(r)}$ and $\alpha_i^{(r)}$. In the frequency domain, this translates into the following:

$$\mathbf{h}(f) = \left[\sum_{r=0}^{R-1} \alpha_i^{(r)} \exp\left(2\pi j f \tau_i^{(r)}\right)\right]_i, \tag{5}$$

where $r = 0, \dots, R-1$ denotes the reflection order,

Recently, these methods have been used for noise and interferer suppression in [9, 57] and for noise and reverberation reduction in [10, 58]. The main limitation of these works is that echo properties, or alternatively the position of image sources, must be known *a priori*. Hereafter, we will assume these properties known by using the annotations of the dEchorate dataset, as described in Section 2.3. In particular, we will assume that the RIRs follow the echo model (5) with $R = 4$, corresponding to the 4 strongest echoes. Knowing the echo delays, the associated attenuation coefficients are retrieved from the RIRs using a simple maximum-likelihood approach, as in [59, Eq. 10].

We evaluate the performance of both types of beamformers on the task of noise and late reverberation

suppression. Different Minimum Variance Distortionless Response (MVDR) beamformers are considered, assuming either spatially white noise (*i.e.*, classical `DS` design), diffuse noise (*i.e.*, the Capon filter) or diffuse noise *plus* the late reverberation [60]. In the latter case, the late reverberation statistics are modeled by a spatial coherence matrix [61] weighted by the late reverberation power, which is estimated using the procedure described in [60].

Overall, the different RIR models considered are direct propagation (`DP`, *i.e.*, ignoring echoes), multipath propagation (`Rake`, *i.e.*, using 4 known early echoes) [9, 10] or the full reverberant propagation (`ReTF`) [7, 56].Table 6 summarizes the considered beamformers designs.

Table 6 Summary of the considered beamformers. "n." and "lr." are used as short-hand for noise and late reverberation. (*) denotes echo-aware beamformers.

Acronym	Steering Vectors	Noise Model
DS [52]	Direct Path AOA	Spatially white n.
MVDR-DP [52]	Direct Path AOA	Diffuse n.
MVDR-ReTF [7]	ReTF	Diffuse n.
MVDR-Rake* [9]	4 Echoes/chan.	Diffuse n.
MVDR-DP-Late [10]	Direct Path AOA	Spat.ly white n.+lr.
MVDR-ReTF-Late [56]	ReTF	Diffuse n. + lr.
MVDR-Rake-Late* [10]	4 Echoes/chan.	Diffuse n. + lr.

Performances of the different designs are compared on the task of enhancing a target speech signal in a 5-channel mixture using the nULAs in the `dEchorate` dataset. They are tested in scenarios featuring high reverberation and diffuse babble noise, appropriately scaled to pre-defined signal-to-noise ratios $\texttt{SNR} \in \{0, 10, 20\}$. Using the `dEchorate` data, we consider the room configuration `011111` ($\texttt{RT}_{60} \approx 600$ ms) and all possible combinations of (target, array) positions. Both real and corresponding synthetic RIRs are used, which are then convolved with anechoic utterances from the WSJ corpus [36] and corrupted by recorded diffuse babble noise. The synthetic RIRs are computed with the Python library `pyroomacoustics` [62], based purely on the ISM. Hence, on synthetic RIRs, the known echo timings perfectly match the components in their early part (no model mismatch).

The evaluation is conducted similarly to the one in [10] where the following metrics are considered:

- the Signal-to-Noise plus Reverberation Ratio improvement (iSNRR) in dB, computed as the difference between the input `SNRR` at the reference microphone and the `SNRR` at the filter output;
- the Speech-to-Reverberation energy Modulation Ratio improvement (iSRMR) [63] to measure dereverberation;
- the Perceptual Evaluation of Speech Quality improvement (iPESQ) score [64] to assess the perceptual quality of the signal and indirectly the amount of artifacts.

Implementations of the SRMR and Perceptual Evalution of Speech Quality (PESQ) metrics are available in the Python library `speechmetrics`. Both the Signal-to-Noise plus Reverberation Ratio (SNRR) and the PESQ are relative metrics, meaning they require a target reference signal. Here we consider the clean target signal as the dry source signal convolved with the early part of the RIR (up to R-th echo) of the reference (first) microphone. On the one hand, this choice numerically penalizes both direct-path-based and ReTF-based beamformers, which respectively aim at extracting the direct-path signal and the full reverberant signal in the reference microphone. On the other hand, considering only the direct path or the full reverberant signal would be equally unfair for the other beamformers. Moreover, including early echoes in the target signal is perceptually motivated since they are known to contribute to speech intelligibility [65].

Numerical results are reported in Figure 7. On synthetic data, as expected, one can see that the more information is used, the better performances are. Including late reverberation statistics considerably boosts performance in all cases. Both the ReTFs-based and the echo-aware beamformers significantly outperform the simple designs based on direct path only. While the two designs perform comparably in terms of iSNRR and iPESQ, the former has a slight edge over the latter in terms of median iSRMR. A possible explanation is that GEVD methods tend to consider the stronger and more stable components of the ReTFs, which in the considered scenarios may identify with the earlier portion of the RIRs. Moreover, since it is not constrained by a fixed echo model, the ReTFs can capture more information, *e.g.*, frequency-dependent attenuation coefficients. Finally, one should consider the compacity of the model (5) with respect to the ReTF model in terms of the number of parameters to be estimated. In fact, when considering 4 echoes, only 8 parameters per channel are needed, as opposed to several hundreds for the ReTF (ideally, as many as the number of frequency bins per channel).

When it comes to measured RIRs, however, the trends are different. Here, the errors in echo timings due to calibration mismatch and the richness of real acoustic propagation lead to a drop in performance for echo-aware methods, both in terms of means and variances. This is clearest when considering the iPESQ metric, which also accounts for artifacts. the echo-agnostic beamformer considering late reverberation `MVDR-ReTF-Late` outperforms the other methods, maintaining the trend exhibited on simulated data. Finally, conversely to the `MVDR-ReTF-Late`,

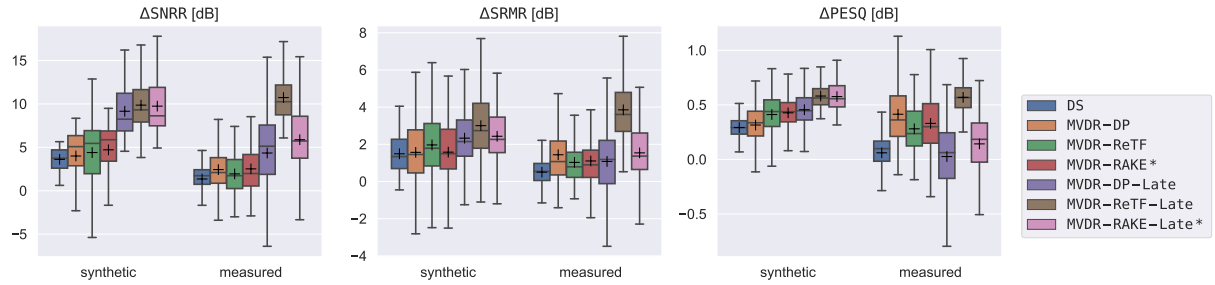

Figure 7 Boxplot showing the comparison of different echo-agnostic and echo-aware (*) beamformers for the room configuration `011111` ($\texttt{RT}_{60} \approx 600$ ms) on measured and synthetic data for all combinations of source-array positions in the `dEchorate` dataset. Mean values is indicated as +, while whiskers indicates extreme values.

the `MVDR-Rake-Late` yields a significant portion of negative performances. As already observed in [10], this is probably due to the tiny annotation mismatches in echo timings as well as the fact that their frequency-dependent strengths, induced by reflective surfaces, are not modeled in rake beamformers. This suggests that in order to be applicable to real conditions, future work in echo-aware beamforming should include finer blind estimates of early echo properties from signals, as investigated in, *e.g.*, [35, 66].

4.2 Application: Room Geometry Estimation

The shape of a convex room can be estimated knowing the positions of first-order image sources. Several methods have been proposed which take into account different levels of prior information and noise (see [23, 67] for a review). When the echoes' TOA and their labeling are known for 4 non-coplanar microphones, one can perform this task using simple geometrical reasoning as in [21]. In details, the 3D coordinates of each image source can be retrieved solving a multilateration problem [68], namely the extension of the trilateration problem to 3D space, where the goal is to estimate the relative position of an object based on the measurement of it distance with respect to anchor points. Finally, the position and orientation of each room facet can be easily derived from the ISM equations as the plane bisecting the line joining the real source position and the position of its corresponding image (see Figure 8)

In `dEchorate`, the annotation of all the first order echo timings are available, as well as correspondences between echoes and room facets. This information can be used directly as input for the above-mentioned multilateration algorithm. We illustrate the validity of these annotations by employing the RooGE technique in [21] (with know labels) based on them.

Table 7 shows the results of the estimation of the room facets position in terms of Distance Error (DE)

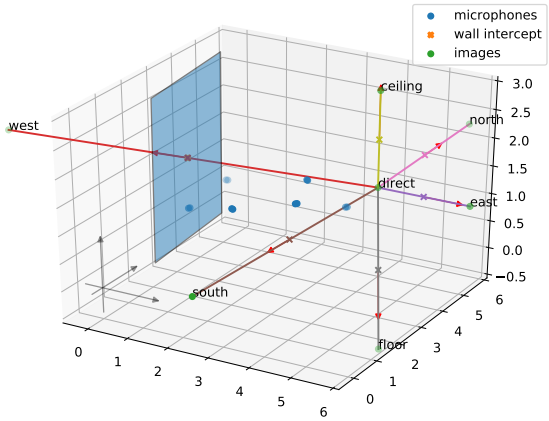

Figure 8 Images source estimation and reflector estimation for one of the sound sources in the dataset.

(in centimeters) and surface orientation error, (dubbed here Angular Error (AE), in degrees) using a single source and all 30 microphones, namely the 6 arrays. Room facets are estimated using each of the sources #1 to #4 as a probe. Despite a few outliers, the majority of facets are estimated correctly in terms of their placement and orientation with respect to the coordinate system computed in Section 2.3. For instance, using source #4, all 6 surfaces were localized with 1.49 cm DE on average and their inclinations with 1.3° AE on average. Apart from the outliers, these results are in line with the ones reported by Dokmanić *et al.*in the work [21] using a setup of 5 microphones listening to 1 sound source.

Furthermore, one can use all the 4 sources to estimate the room geometry as suggested in [22]. By doing so, the entire room geometry estimation results in 1.15 cm DE and 2.6° AE on average.

The small errors are due to a concurrency of multiple factors, such as tiny offsets in the annotations and

Table 7 Distance Error (DE) in centimeters and Angular Error (AE) in degrees between ground truth and estimated room facets using each of the sound sources (#1 to #4) as a probe. For each wall, bold font is used for the source yielding the best DE and AE, while italic highlights outliers when present.

source id	1		2		3		4	
wall	DE	AE	DE	AE	DE	AE	DE	AE
west	0.74	8.99°	4.59	8.32°	5.89	5.75°	**0.05**	**2.40°**
east	**0.81**	**0.08°**	0.9	0.50°	*69.51*	*55.70°*	0.31	0.21°
south	3.94	16.08°	**0.18**	1.77°	*14.37*	*18.55°*	0.82	**1.65°**
north	1.34	0.76°	1.40	8.94°	**0.63**	**0.17°**	2.08	1.38°
floor	**5.19**	**1.76°**	7.27	2.66°	7.11	2.02°	5.22	1.90°
ceiling	1.16	0.28°	0.67	0.76°	**0.24**	1.16°	0.48	**0.26°**

the ideal shoebox approximation. In the real recording room, some gaps were present between revolving panels in the room facet. In addition, it is possible that for some (image source, receiver) pairs the far-field assumption is not verified, causing inaccuracies when inverting the ISM. The 2 outliers for source #3 are due to a wrong annotation caused by the source directivity which induced an echo mislabeling. When a wall is right behind a source, the energy of the related 1st reflection is very small and might not appear in the RIRs. This happened for the eastern wall and a second order image was taken instead. Finally, the contribution of multiple reflections arriving at the same time can result in large late spikes in estimated RIRs. This effect is particularly amplified when the microphone and loudspeakers exhibit long impulse responses. As a consequence, some spikes can be miss-classified. This happened for the southern-wall where again a second-order image was taken instead. Note that such echo mislabelings can either be corrected manually or using Euclidean distance matrix criteria as proposed in [21]. Overall, this experiment illustrates well the interesting challenge of estimating and exploiting acoustic echoes in RIRs when typical sources and receivers with imperfect characteristics are used.

ec

5 Conclusions and Perspectives

This paper introduced a new database of room impulse responses featuring accurate annotation of early echo timings that are consistent with source, microphone and room facet positions. These data can be used to test methods in the room geometry estimation pipeline and in echo-aware audio signal processing. In particular, robustness of these methods can be validated against different levels of $\mathtt{RT}_{60}$, SNR, surface reflectivity, proximity, or early echo density.

This dataset paves the way to a number of interesting future research directions. By making this dataset freely available to the audio signal processing community, we hope to foster research in AER and echo-aware signal processing in order to improve the performance of existing methods on real data. Moreover, the dataset could be updated by including more robust annotations derived from more advanced algorithms for calibration and AER.

In addition, the data analysis conducted in this work brings the attention to exploring the impact of mismatch between simulated and real RIRs on audio signal processing methods. Finally, by using the pairs of simulated vs. real RIRs available in the dataset, it should be possible to develop techniques to convert one to the other, using style transfer or domain adaptation techniques, thus opening the way to new types of learning-based acoustic simulators.

Appendix

Room materials

Table 8 Materials covering the acoustic laboratory in Bar-Ilan University.

Surface	Mode	Material
Floor	absorbent	Hairy carpet
Ceiling	absorbent	Glass wool mats covered with porous tin
Ceiling	reflective	Formica (20 mm thick)
Walls	absorbent	Glass wool mats covered with porous tin
Walls	reflective	Panels: Formica (20 mm thick) Wall: Plaster

Acknowledgements
Luca Remaggi, Marco Crocco, Alessio Del Bue, and Robin Scheibler are thanked for help during experimental design.

Availability of data and materials
The database is publicly available at zenodo and github.

Abbreviations

AE	Angular Error
AER	Acoustic Echo Retrieval
ASR	Automatic Speech Recognition
DE	Distance Error
DER	Direct-to-Early Ratio
DRR	Direct-to-Reverberant ratio
DOA	Direction of Arrival
ESS	Exponentially Swept-frequency Sine
GEVD	Generalized Eigenvector Decomposition
GoM	Goodness of Match
MDS	Multi-Dimensional Scaling
MVDR	Minimum Variance Distortionless Response
nULA	non-Uniform Linear Array
PESQ	Perceptual Evalution of Speech Quality
RIR	Room Impulse Response

ReTF Relative Transfer Function
TOA Time of Arrival
TDOA Time Difference of Arrival
ISM Image Source Method
SE Speech Enhancement
SNRR Signal-to-Noise plus Reverberation Ratio
iPESQ Perceptual Evaluation of Speech Quality improvement
iSNRR Signal-to-Noise plus Reverberation Ratio improvement
iSRMR Speech-to-Reverberation energy Modulation Ratio improvement
RooGE Room Geometry Estimation
WSJ Wall Street Journal

Author details

[1]Univ Rennes, Inria, CNRS, IRISA, France. [2]Faculty of Engineering, Bar-Ilan University, Ramat-Gan, 5290002, Israel. [3]UMRAE, Cerema, Univ. Gustave Eiffel, Ifsttar, Strasbourg, 67035, France. [4]Université de Lorraine, Inria, CNRS, LORIA, F-54000 Nancy, France.

References

1. Griesinger, D.: The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces. Acta Acustica united with Acustica **83**(4), 721–731 (1997)
2. Kuttruff, H.: Room Acoustics. Crc Press, ??? (2016)
3. Allen, J.B., Berkley, D.A.: Image method for efficiently simulating small-room acoustics. Journal of the Acoustical Society of America **65**(4), 943–950 (1979). doi:10.1121/1.382599
4. Flanagan, J.L., Surendran, A.C., Jan, E.-E.: Spatially selective sound capture for speech and audio processing. Speech Communication **13**(1-2), 207–222 (1993)
5. Jan, E.E., Svaizer, P., Flanagan, J.L.: Matched-filter processing of microphone array for spatial volume selectivity. In: IEEE International Symposium on Circuits and Systems, vol. 2, pp. 1460–1463 (1995). doi:10.1109/iscas.1995.521409. IEEE
6. Affes, S., Grenier, Y.: A signal subspace tracking algorithm for microphone array processing of speech. IEEE Transactions on Speech and Audio Processing **5**(5), 425–437 (1997). doi:10.1109/89.622565
7. Gannot, S., Burshtein, D., Weinstein, E.: Signal enhancement using beamforming and non-stationarity with applications to speech. IEEE Transactions on Signal Processing **49**(8), 1614–1626 (2001). doi:10.1109/78.934132
8. Annibale, P., Antonacci, F., Bestagini, P., Brutti, A., Canclini, A., Cristoforetti, L., Habets, E., Kellermann, W., Kowalczyk, K., Lombard, A., Mabande, E., Markovic, D., Naylor, P., Omologo, M., Rabenstein, R., Sarti, A., Svaizer, P., Thomas, M.: The SCENIC project: Environment-aware sound sensing and rendering. Procedia Computer Science **7**, 150–152 (2011). doi:10.1016/j.procs.2011.09.039
9. Dokmanić, I., Scheibler, R., Vetterli, M.: Raking the Cocktail Party. IEEE Journal on Selected Topics in Signal Processing **9**(5), 825–836 (2015). doi:10.1109/JSTSP.2015.2415761
10. Kowalczyk, K.: Raking early reflection signals for late reverberation and noise reduction. The Journal of the Acoustical Society of America **145**(3), 257–263 (2019). doi:10.1121/1.5095535
11. Ribeiro, F., Ba, D., Zhang, C., Florêncio, D.: Turning enemies into friends: Using reflections to improve Sound Source Localization. In: IEEE International Conference on Multimedia and Expo (ICME), pp. 731–736 (2010). doi:10.1109/ICME.2010.5583886
12. Salvati, D., Drioli, C., Foresti, G.L.: Sound source and microphone localization from acoustic impulse responses. IEEE Signal Processing Letters **23**(10), 1459–1463 (2016)
13. Di Carlo, D., Deleforge, A., Bertin, N.: Mirage: 2D Source Localization Using Microphone Pair Augmentation with Echoes. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2019-May, pp. 775–779 (2019). doi:10.1109/ICASSP.2019.8683534
14. Daniel, J., Kitić, S.: Time domain velocity vector for retracing the multipath propagation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 421–425 (2020). IEEE
15. Asaei, A., Golbabaee, M., Bourlard, H., Cevher, V.: Structured sparsity models for reverberant speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **22**(3), 620–633 (2014)
16. Leglaive, S., Badeau, R., Richard, G.: Multichannel Audio Source Separation with Probabilistic Reverberation Priors. IEEE/ACM Transactions on Audio Speech and Language Processing **24**(12), 2453–2465 (2016). doi:10.1109/TASLP.2016.2614140
17. Scheibler, R., Di Carlos, D., Deleforge, A., Dokmanic, I.: Separake: Source Separation with a Little Help from Echoes. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2018-April, pp. 6897–6901 (2018). doi:10.1109/ICASSP.2018.8461345. http://arxiv.org/abs/1711.06805
18. Remaggi, L., Jackson, P.J., Wang, W.: Modeling the comb filter effect and interaural coherence for binaural source separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **27**(12), 2263–2277 (2019)
19. Al-Karawi, K.A., Mohammed, D.Y.: Early reflection detection using autocorrelation to improve robustness of speaker verification in reverberant conditions. International Journal of Speech Technology **22**(4), 1077–1084 (2019)
20. Antonacci, F., Filos, J., Thomas, M.R., Habets, E.A., Sarti, A., Naylor, P.A., Tubaro, S.: Inference of room geometry from acoustic impulse responses. IEEE Transactions on Audio, Speech, and Language Processing **20**(10), 2683–2695 (2012)
21. Dokmanić, I., Parhizkar, R., Walther, A., Lu, Y.M., Vetterli, M.: Acoustic echoes reveal room shape. Proceedings of the National Academy of Sciences of the United States of America **110**(30), 12186–12191 (2013). doi:10.1073/pnas.1221464110
22. Crocco, M., Trucco, A., Del Bue, A.: Uncalibrated 3D room geometry estimation from sound impulse responses. Journal of the Franklin Institute **354**(18), 8678–8709 (2017). doi:10.1016/j.jfranklin.2017.10.024
23. Remaggi, L., Jackson, P.J.B., Coleman, P., Wang, W.: Acoustic Reflector Localization: Novel Image Source Reversion and Direct Localization Methods. IEEE/ACM Transactions on Audio Speech and Language Processing **25**(2), 296–309 (2017). doi:10.1109/TASLP.2016.2633802
24. Jensen, J.R., Saqib, U., Gannot, S.: An em method for multichannel toa and doa estimation of acoustic echoes. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 120–124 (2019). IEEE
25. Boutin, M., Kemper, G.: A drone can hear the shape of a room. SIAM Journal on Applied Algebra and Geometry **4**(1), 123–140 (2020)
26. Shih, O., Rowe, A.: Can a phone hear the shape of a room? In: ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), pp. 277–288 (2019). IEEE
27. Szoke, I., Skacel, M., Mosner, L., Paliesek, J., Cernocky, J.H.: Building and Evaluation of a Real Room Impulse Response Dataset. IEEE Journal on Selected Topics in Signal Processing **13**(4), 863–876 (2019). doi:10.1109/JSTSP.2019.2917582
28. Genovese, A.F., Gamper, H., Pulkki, V., Raghuvanshi, N., Tashev, I.J.: Blind Room Volume Estimation from Single-channel Noisy Speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2019-May, pp. 231–235 (2019). doi:10.1109/ICASSP.2019.8682951
29. Hadad, E., Heese, F., Vary, P., Gannot, S.: Multichannel audio database in various acoustic environments. In: 2014 14th International Workshop on Acoustic Signal Enhancement, IWAENC 2014, pp. 313–317 (2014). doi:10.1109/IWAENC.2014.6954309
30. Bertin, N., Camberlein, E., Lebarbenchon, R., Vincent, E., Sivasankaran, S., Illina, I., Bimbot, F.: VoiceHome-2, an extended corpus for multichannel speech processing in real homes. Speech Communication **106**, 68–78 (2019). doi:10.1016/j.specom.2018.11.002
31. Čmejla, J., Kounovskỳ, T., Gannot, S., Koldovskỳ, Z., Tandeitnik, P.: Mirage: Multichannel database of room impulse responses measured on high-resolution cube-shaped grid. In: European Signal Processing Conference (EUSIPCO), pp. 56–60 (2021). IEEE
32. Gaultier, C., Kataria, S., Deleforge, A.: VAST: The virtual acoustic space traveler dataset. In: Lecture Notes in Computer Science, vol. 10169 LNCS, pp. 68–79 (2017). doi:10.1007/978-3-319-53547-07
33. Kim, C., Misra, A., Chin, K., Hughes, T., Narayanan, A., Sainath, T.N., Bacchiani, M.: Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech

recognition in google home. In: Proc. Interspeech 2017, pp. 379–383 (2017). doi:10.21437/Interspeech.2017-1510. http://dx.doi.org/10.21437/Interspeech.2017-1510
34. Perotin, L., Serizel, R., Vincent, E., Guerin, A.: CRNN-Based Multiple DoA Estimation Using Acoustic Intensity Features for Ambisonics Recordings. IEEE Journal on Selected Topics in Signal Processing **13**(1), 22–33 (2019). doi:10.1109/JSTSP.2019.2900164
35. Di Carlo, D., Elvira, C., Deleforge, A., Bertin, N., Gribonval, R.: Blaster: An Off-Grid Method for Blind and Regularized Acoustic Echoes Retrieval. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 156–160 (2020). doi:10.1109/icassp40776.2020.9054647
36. Paul, D.B., Baker, J.M.: The design for the wall street journal-based csr corpus. In: Proceedings of the Workshop on Speech and Natural Language, pp. 357–362 (1992). Association for Computational Linguistics
37. Farina, A.: Advancements in impulse response measurements by sine sweeps. 122nd Audio Engineering Society Convention **3**, 1626–1646 (2007)
38. Dokmanić, I., Ranieri, J., Vetterli, M.: Relax and unfold: Microphone localization with Euclidean distance matrices. In: European Signal Processing Conference, (EUSIPCO), pp. 265–269 (2015). doi:10.1109/EUSIPCO.2015.7362386
39. Crocco, M., Del Bue, A.: Estimation of TDOA for room reflections by iterative weighted l1 constraint. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2016-May, pp. 3201–3205. IEEE, ??? (2016). doi:10.1109/ICASSP.2016.7472268
40. Plinge, A., Jacob, F., Haeb-Umbach, R., Fink, G.A.: Acoustic microphone geometry calibration. IEEE Signal Processing Magazine (July), 14–28 (2016)
41. Beck, A., Stoica, P., Li, J.: Exact and approximate solutions of source localization problems. IEEE Transactions on Signal Processing **56**(5), 1770–1778 (2008)
42. Baba, Y.E., Walther, A., Habets, E.A.P.: 3D room geometry inference based on room impulse response stacks. IEEE/ACM Transactions on Audio Speech and Language Processing **26**(5), 857–872 (2018). doi:10.1109/TASLP.2017.2784298
43. Eaton, J., Gaubitch, N.D., Moore, A.H., Naylor, P.A.: Estimation of Room Acoustic Parameters: The ACE Challenge. IEEE/ACM Transactions on Audio Speech and Language Processing **24**, 1681–1693 (2016). doi:10.1109/TASLP.2016.2577502
44. Defrance, G., Daudet, L., Polack, J.-D.: Finding the onset of a room impulse response: Straightforward? The Journal of the Acoustical Society of America **124**(4), 248–254 (2008). doi:10.1121/1.2960935
45. Eaton, J., Gaubitch, N.D., Moore, A.H., Naylor, P.A.: The ace challenge—corpus description and performance evaluation. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 1–5 (2015). IEEE
46. Eargle, J.M.: Characteristics of performance and recording spaces. In: Handbook of Recording Engineering, pp. 57–65. Springer, ??? (1996)
47. Naylor, P.A., Gaubitch, N.D.: Speech Dereverberation. Springer, ??? (2010)
48. Schroeder, M.R.: New method of measuring reverberation time. The Journal of the Acoustical Society of America **37**(6), 1187–1188 (1965)
49. Chu, W.T.: Comparison of reverberation measurements using schroeder's impulse method and decay-curve averaging method. The Journal of the Acoustical Society of America **63**(5), 1444–1450 (1978)
50. Xiang, N.: Evaluation of reverberation times using a nonlinear regression approach. The Journal of the Acoustical Society of America **98**(4), 2112–2121 (1995)
51. Gannot, S., Vincent, E., Markovich-Golan, S., Ozerov, A.: A consolidated perspective on multi-microphone speech enhancement and source separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**(4), 692–730 (2017)
52. Van Trees, H.L.: Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory. John Wiley & Sons, ??? (2004)
53. Schwartz, O., Gannot, S., Habets, E.A.: Multi-microphone speech dereverberation and noise reduction using relative early transfer functions. IEEE/ACM Transactions on Audio, Speech, and Language Processing **23**(2), 240–251 (2014)
54. Kodrasi, I., Doclo, S.: Evd-based multi-channel dereverberation of a moving speaker using different retf estimation methods. In: 2017 Hands-free Speech Communications and Microphone Arrays (HSCMA), pp. 116–120 (2017). IEEE
55. Markovich, S., Gannot, S., Cohen, I.: Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. IEEE Transactions on Audio, Speech, and Language Processing **17**(6), 1071–1086 (2009)
56. Markovich-Golan, S., Gannot, S., Kellermann, W.: Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function. In: 2018 26th European Signal Processing Conference (EUSIPCO), pp. 2499–2503 (2018). IEEE
57. Scheibler, R., Dokmanić, I., Vetterli, M.: Raking echoes in the time domain. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 554–558 (2015). IEEE
58. Javed, H.A., Moore, A.H., Naylor, P.A.: Spherical microphone array acoustic rake receivers. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 111–115 (2016). IEEE
59. Condat, L., Hirabayashi, A.: Cadzow denoising upgraded: A new projection method for the recovery of dirac pulses from noisy linear measurements. Sampling Theory in Signal and Image Processing **14**(1), 17–47 (2015)
60. Schwartz, O., Gannot, S., Habets, E.A.: Joint estimation of late reverberant and speech power spectral densities in noisy environments using frobenius norm. In: 2016 24th European Signal Processing Conference (EUSIPCO), pp. 1123–1127 (2016). IEEE
61. Kuster, M.: Objective sound field analysis based on the coherence estimated from two microphone signals. The Journal of the Acoustical Society of America **131**(4), 3284–3284 (2012)
62. Scheibler, R., Bezzam, E., Dokmanić, I.: Pyroomacoustics: A Python package for audio room simulations and array processing algorithms. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Calgary, CA (2018). accepted
63. Falk, T.H., Zheng, C., Chan, W.-Y.: A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Transactions on Audio, Speech, and Language Processing **18**(7), 1766–1774 (2010)
64. Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In: IEEE International Conference on Acoustics, Speech, and Signal (ICASSP), vol. 2, pp. 749–752 (2001). IEEE
65. Bradley, J.S., Sato, H., Picard, M.: On the importance of early reflections for speech in rooms. The Journal of the Acoustical Society of America (JASA) **113**(6), 3233–3244 (2003)
66. Tukuljac, H.P., Deleforge, A., Gribonval, R.: Mulan: A blind and off-grid method for multichannel echo retrieval. In: NeurIPS 2018-Thirty-second Conference on Neural Information Processing Systems, pp. 1–11 (2018)
67. Crocco, M., Trucco, A., Del Bue, A.: Room Reflector Estimation from Sound by Greedy Iterative Approach. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2018-April, pp. 6877–6881 (2018). doi:10.1109/ICASSP.2018.8461640
68. Beck, A., Stoica, P., Li, J.: Exact and approximate solutions of source localization problems. IEEE Transactions on Signal Processing **56**(5), 1770–1778 (2008). doi:10.1109/TSP.2007.909342