THE R3VIVAL DATASET: REPOSITORY OF ROOM RESPONSES AND 360 VIDEOS OF A VARIABLE ACOUSTICS LAB

*Florian Klein**

Ilmenau University of Technology
Electronic Media Technology Group
Ilmenau, Germany

Sebastià V. Amengual Garí

Meta
Reality Labs Research
Redmond, USA

ABSTRACT

This paper presents a dataset of spatial room impulse responses (SRIRs) and 360° stereoscopic video captures of a variable acoustics laboratory. A total of 34 source positions are measured with 8 different acoustic panel configurations, resulting in a total of 272 SRIRs. The source positions are arranged in 30° increments at concentric circles of radius 1.5, 2, and 3 m measured with a directional studio monitor, as well as 4 extra positions at the room corners measured with an omnidirectional source. The receiver is a 7 channel open microphone array optimized for its use with the Spatial Decomposition Method (SDM). The 8 acoustic configurations are achieved by setting a subset of the panels to their absorptive configuration in 5 steps (0%, 25%, 50%, 75%, 100% of the panels), as well as 3 configurations in which entire walls are set to their absorptive configuration (right, right/back, right/back/left). Video captures of the laboratory and a second room are obtained using a 360° stereoscopic camera with a resolution of 4096×2160 pixels, covering the same source/receiver combinations. Furthermore, we present an acoustic analysis of both time-energy and spatio-temporal parameters showcasing the differences in the measured configurations. The dataset, together with spatial analysis and rendering scripts, is publicly released in a GitHub repository[1].

***Index Terms*—** Room Impulse Responses, Stereoscopic video, Dataset, Variable Acoustics

1. INTRODUCTION

The advent of Extended Reality (XR) applications has highlighted the need for low compute solutions to render perceptually accurate acoustic scenes which allow the fusion of virtual and real acoustic sound sources in interactive environments. In Augmented Reality (AR), an excessive acoustic mismatch between the real acoustic space and the acoustics of the virtually rendered sounds result in the room divergence effect, leading to a collapse of externalization of the virtual sounds [1]. However, this effect does not seem to occur with mismatching audiovisual information, as long as no acoustic information about the real environment is provided [2]. This suggests that context and listener expectations that arise from acoustic and visual information about the environment are crucial in the perception of externalized sounds, although the specific processes around these expectations are not fully understood [3]. Additionally, while the room divergence effect has been documented in multiple studies [4, 5], the relationship between the degree of acoustic mismatch and externalization collapse is not yet well established. Moreover, even the detection thresholds of basic room acoustic parameters, such as Reverberation Time (RT), are not currently widely generalized and warrant further research [6].

In this contribution we present a dataset containing both spatial room impulse responses (SRIR) for various acoustic conditions in a variable acoustic laboratory, as well as a collection of stereoscopic 360° videos. The dataset is designed to be especially suitable for the study of audiovisual perception and room acoustics in immersive environments and allows the fine tuning of the acoustic properties of the presented stimuli while maintaining room geometry constant. Additionally, this dataset can be paired with an audiovisual speech corpus[2] to generate arbitrary speech scenes with multiple talkers in various acoustic conditions to be reproduced in Virtual Reality (VR) [7]. To the best of our knowledge, this is the first publicly available dataset of this characteristics, especially given that datasets of variable acoustic rooms are rare.

2. RELATED WORK

Two datasets acquired in the variable acoustics room *Arni* at Aalto Acoustic Labs were released recently. The first dataset [8, 9] contains recordings SRIRs obtained with two spherical microphone arrays (*mh Acoustics Eigenmike em32* and *Zylia ZM-1*) at three source and seven receiver positions. These were measured for five room acoustic configurations

*Florian Klein was employed by Reality Labs Research (Meta) while this work was conducted.

[1] https://github.com/facebookresearch/R3VIVAL

[2] The corpus is no longer accessible from the original link. A mirror source is provided here: *clickable link*

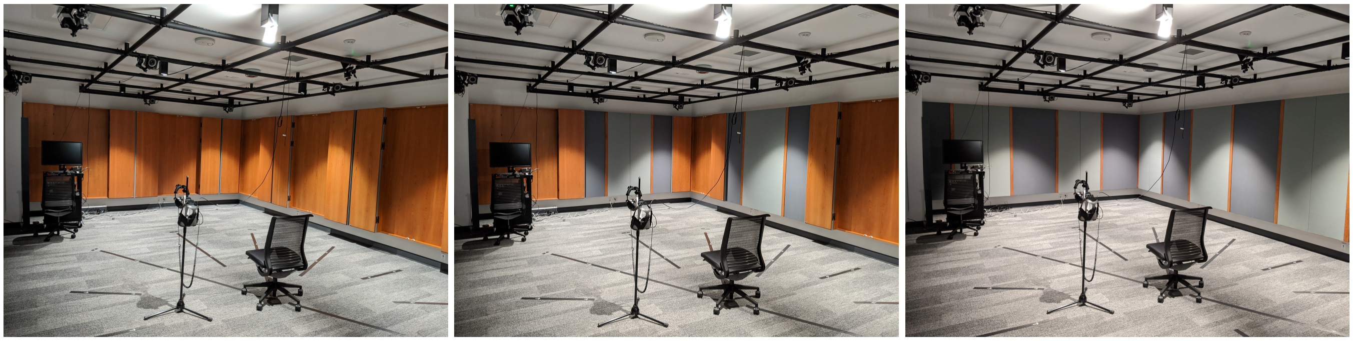


Fig. 1. Variable acoustics laboratory with three acoustic panel configurations: fully reflective (left), 50% absorptive (middle), fully absorptive (right).

(0%, 25%, 50%, 75% and 100% of the absorbers enabled). Reveberation time T_{20} at 1 kHz ranges between 1.22 s to 0.32 s depending on the configuration. The second dataset recorded in the Arni lab [10, 11] contains measurements of 5342 sound absorption configurations measured one source position (omnidirectional source) and 5 receiver positions (omnidirectional single receivers). Another dataset including variable acoustics [12, 13] was recorded in the acoustic lab at Bar-Ilan university. The six faces of the room were covered with acoustic panels which were either reflective or absorptive. This resulted in 11 different acoustic configurations which were measured with both directional and omnidirectional sources, and linear line arrays as receivers.

Two of the above mentioned datasets contain exclusively monaural RIRs. Only the first of the Arni datasets from Aalto University [8, 9], captured with a spherical microphone array (SMA), provides acoustic responses suitable for auralization. However, the dataset does not include any visual captures of the space and it is thus restricted to audio only applications.

Recently, an audiovisual database containing 360° video and Higher-Order Ambisonics recordings of nature and urban environments was released [14]. This dataset was designed for the investigation of Quality of Experience (QoE) in VR applications. Since the scenes are recorded it is not possible to manipulate the acoustic conditions or content.

With the present release we aim at filling a gap in the needs for audiovisual data helpful for the conduction of multiple perceptual tasks in reverberant virtual environments. The dataset that we introduce here allows the generation of arbitrary audiovisual scenes while providing a high degree of fine tuning of the acoustic properties of the presented sounds.

3. VARIABLE ACOUSTICS LABORATORY

The variable acoustics laboratory is a shoebox room with dimensions of 9.7 × 5.5 × 2.7 m (L × W × H) and is equipped with movable panels on its perimeter, covering most of the wall surface of the room. Two different wall materials with different absorption properties can be exposed by opening and closing panels, corresponding to reflective and absorptive configurations, respectively (see figure 1). In its most absorptive configuration (all panels closed), the average reverberation time (RT) of the room is $\mathrm{T}_{30} \approx 0.5$ s at 500 Hz, while in its most reflective configuration the average $\mathrm{T}_{30} \approx 0.75$ s at 500 Hz. Thus, specific panel configurations allow for a finely tuned and potentially position dependent acoustic response. The materials of the floor and ceiling are not configurable and are low pile carpet and plaster, respectively.

4. DATASET

4.1. Acoustic Measurements

Eight different panel configurations were measured. By switching a portion of the panels to the acoustically absorptive side, 5 configurations (0%, 25%, 50%, 75%, 100% of absorption) were created. For intermediate configurations, each wall was modified in the same way. Fig 1 shows three panel configurations. In addition, 3 configurations were measured in which entire walls are set to their absorptive configuration (right, right/back, right/back/left), while the remaining were kept reflective. These configurations lead to a non-uniform distribution of absorptive surfaces.

For each panel configuration, the same source-receiver combinations were measured. Twelve speaker positions at 1.5 m and 2 m as well as six speakers at 3 m were measured around a microphone array (see Fig. 2). For these measurements directional speakers (Genelec 8320) were pointed towards the array. In addition, four positions (BG 1 to BG4) are measured with an omnidirectional sound source (B&K 4295). The microphone array is a custom array of 7 microphones based on an open sphere design with a diameter of 10 cm. A center microphone (Earthworks M50) is surrounded by 6 miniature microphones (DPA 4060), a suitable design for spatial acoustic analysis and rendering with SDM [15, 16]. All measurements were conducted at a height of 1.3 m and using an 20 s sine sweep in the range of 50 Hz - 20 kHz at a sampling frequency of 192 kHz.

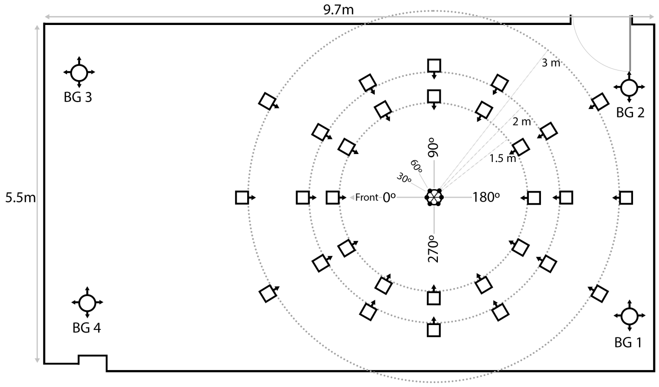
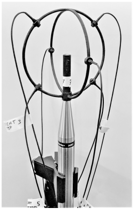


Fig. 2. Measurement plan: Directional speakers placed at 12 positions with 1.5 m distance, 12 positions with 2 m distance and 6 positions with 3 m distance to the receiver. Four omnidirectional sound sources in the corners of the room.



Channel	Position	Azimuth	Elevation
1	Front	0	0
2	Back	180	0
3	Right Top	90	45
4	Left Top	90	45
5	Right Bottom	270	-45
6	Left Bottom	270	-45
7	Center	N/A	N/A

Fig. 3. Open microphone array used in the measurements (left) and positions of each microphone (right).

4.2. Acoustic Analysis

Figure 4 shows the frequency dependent reverberation times for each panel configuration. For all configurations, a similar frequency dependency can be observed, showing the expected reduction of RT for more absorptive settings. Below 250 Hz the panels only have minimal effects on the reverberation.

Figure 5 shows the spatiotemporal analysis (performed via the SDM Toolbox [15]) of the acoustic energy for one loudspeaker and two exemplary panel configurations. By comparing the top plot (0% absorptive) with the bottom plot (100% absorptive) we can clearly observe a decrease of the acoustic energy, starting as early as 20 ms after the arrival of the direct sound. Additionally, more absorptive settings show an obvious reduction of the lateral energy. However, the impact on first reflections appears to be negligible (see strong specular reflections in the time range 5 to 1000 ms). Matlab scripts and the analysis for all loudspeaker positions are provided in the online repository.

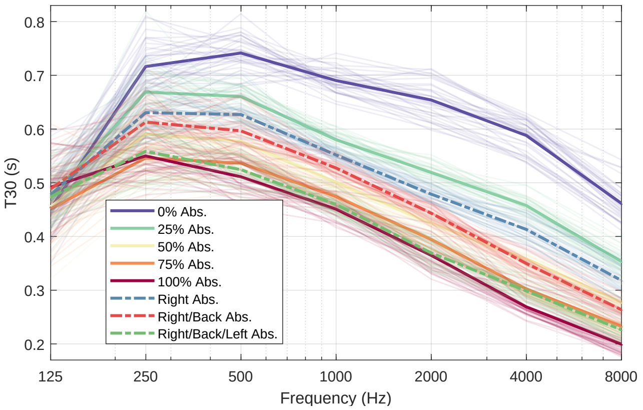


Fig. 4. Estimated reverberation time (T30) for all measured positions (thin lines) and average reverberation time in each panel configuration (thick lines).

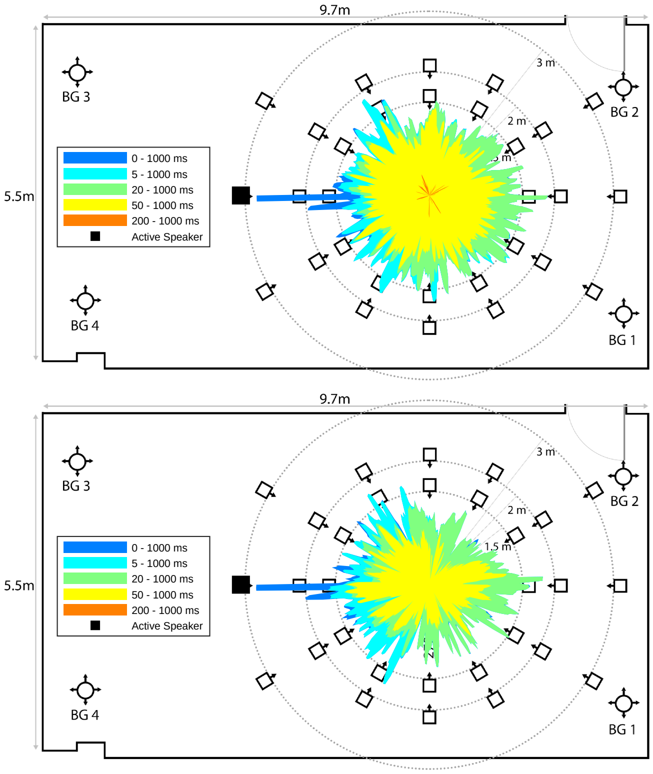


Fig. 5. Spatio-temporal acoustic analysis for 2 panel configurations: 0% absorptive (top), 100% absorptive (bottom).

4.3. Video captures

Stereoscopic 360° video captures are completed in two rooms: the variable acoustics laboratory and a second shoebox room (*Room A*) of similar dimensions. A second room is added to allow the synthesis of audio-visually divergent scenarios, in which the acoustics of the variable acoustics room are reproduced along with the visuals of *Room A*.

The visual arrangement in both rooms is identical, and follows the same structure as the acoustical measurements – 3 concentric rings of loudspeakers placed at intervals of 30°. Each of the videos has a duration of 4 s and contains 3 simultaneously captured loudspeakers. Besides the loudspeakers,

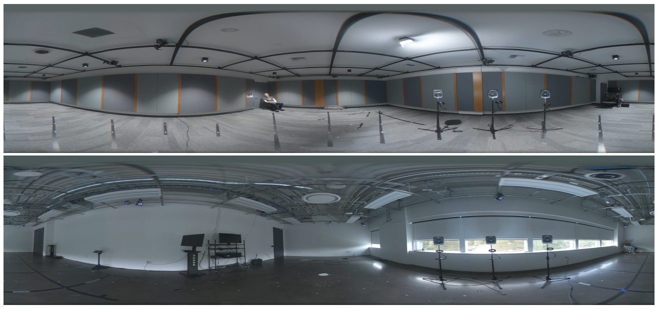


Fig. 6. Panoramic images of the rooms for which the dataset contains stereoscopic 360 videos – Variable Acoustics Lab (top) and Room A (bottom).

only the camera operator is present in the room and maintains a static position. Additionally, the lighting is kept constant in all the videos. Hence, the videos can be easily looped and combined to produce any potential loudspeaker configuration for the synthesis of arbitrary audiovisual scenes. A screenshot of a video for each of the rooms is shown in Fig. 6.

The videos were captured using a Live Planet VR camera system. The camera is a stereoscopic 360° camera with 16 Sony IMX 326 sensors and performs stitching on device. The resolution of the videos is 4096 × 2160 at a framerate of 30 fps. The only post-processing applied to the videos was applying a Gaussian blur to the face of the camera operator.

5. APPLICATIONS

The dataset is especially suited for perceptual studies, a plausible Binaural Room Impulse Responses (BRIRs) can easily be rendered using publicly available tools [16]. Matlab scripts for the generation of BRIRs are provided in the online repository. Additionally, we also provide scripts that utilize *ffmpeg* for the generation of arbitrary multi-talker synthetic scenes by combining the data from our database with the audiovisual speech corpus from [7] (see Fig. 7).

Research in room acoustics perceptual thresholds has been conducted either with acoustic simulations or using recordings from different spaces. However, this poses serious limitations. Simplified simulations might often suffer from plausibility issues, and the use of recordings from different rooms does not allow the independent modification of single acoustical dimensions. Alternatively, variable acoustic rooms with granular adjustment of their acoustics provide important advantages, since acoustics can be modified while keeping the room geometry and the spatial properties of the acoustics constant. Our dataset could be especially suitable for the research on the audibility of early reflections, just noticeable differences of reverberation, or distance perception. An exemplary study recently used this dataset derived perceptual thresholds of reverberation time [6]. Participants had to perform a triangle test comparing three sound sources placed at different positions in the room. One sound source was recorded under different acoustic conditions than the other two, and the task was to detect the differing source. For the most critical signal in the test (castanets), a reverberation time difference of 8% was detectable by the participants, while the difference was 15% across all conditions (sound signals: castanets, speech; sound source distances: 1.5 m, 2 m, 3 m).

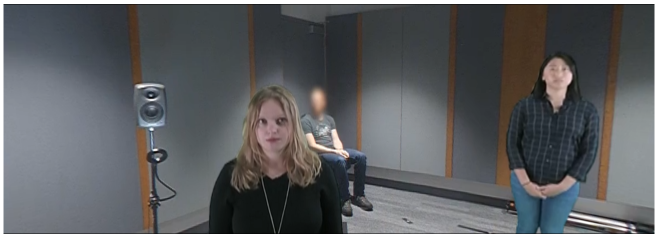


Fig. 7. Example image of the combination of the video recordings of two speakers from [7] overlaid on a capture from our dataset.

Room acoustic divergence is a well documented phenomenon that is however not well understood. Previous studies investigated the effects of reverberation time mismatch [4], adaptation to room divergent scenarios [5], or audiovisual divergence [2]. This dataset provides the flexibility to prototype a variety of relevant scenarios in which the listener could be exposed to divergent audiovisual scenarios or to scenes containing sources with different acoustical properties.

Audiovisual Spatial Congruence using the example of the ventriloquist effect explains the bimodal integration of an audiovisual stimulus in which the visual and acoustic locations of the source are slightly different [17]. The current dataset, paired with the audiovisual speech corpus from [7] allows the creation of synthetic arbitrary scenes with talkers at multiple positions (see Fig. 7). Systematically modifying the acoustic and visual location of the talkers could lead to a generalized characterization of the ventriloquist effect. This could in turn result in important gains for perceptually motivated audio in XR, for instance, by determining the context dependent needs for individualized Head-Related Transfer Functions (HRTF).

6. CONCLUSION

In this contribution we presented an open source dataset of SRIR of a variable acoustics lab and 360 stereoscopic captures of the lab and a second room. In the public repository we provide routines for the generation of binaural acoustic renderings and synthetic audiovisual scenes, enabling multiple applications related to perceptual research in immersive environments.

7. REFERENCES

[1] Stephan Werner, Florian Klein, Thomas Mayenfels, and Karlheinz Brandenburg, “A summary on acoustic room divergence and its effect on externalization of auditory events,” in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6.

[2] Juan C Gil-Carvajal, Jens Cubick, Sébastien Santurette, and Torsten Dau, “Spatial hearing with incongruent visual or auditory room cues,” *Scientific reports*, vol. 6, no. 1, pp. 1–10, 2016.

[3] Virginia Best, Robert Baumgartner, Mathieu Lavandier, Piotr Majdak, and Norbert Kopčo, “Sound externalization: A review of recent research,” *Trends in Hearing*, vol. 24, pp. 2331216520948390, 2020.

[4] Sebastia V Amengual Gari, Henrik G Hassager, Florian Klein, Johannes M Arend, and Philip W Robinson, “Towards determining thresholds for room divergence: A pilot study on perceived externalization,” in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. IEEE, 2021, pp. 1–7.

[5] Florian Klein, Stephan Werner, and Thomas Mayenfels, “Influences of training on externalization of binaural synthesis in situations of room divergence,” *Journal of the Audio Engineering Society*, vol. 65, no. 3, pp. 178–187, 2017.

[6] Florian Klein, Sebastia V. Amengual Gari, Johannes M. Arend, and Philip W. Robinson, “Towards determining thresholds for room divergence: A pilot study on detection thresholds,” in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2021, pp. 1–7.

[7] Lindsey R Kishline, Scott W Colburn, and Philip W Robinson, “A multimedia speech corpus for audio visual research in virtual reality (l),” *The Journal of the Acoustical Society of America*, vol. 148, no. 2, pp. 492–495, 2020.

[8] Thomas McKenzie, Leo McCormack, and Christoph Hold, “Dataset of spatial room impulse responses in a variable acoustics room for six degrees-of-freedom rendering and analysis,” *arXiv*, 2021.

[9] Thomas McKenzie, Leo McCormack, and Christoph Hold, “Dataset of Spatial Room Impulse Responses in a Variable Acoustics Room for Six Degrees-of-Freedom Rendering and Analysis,” Zenodo, Nov. 2021.

[10] Karolina Prawda, Sebastian J. Schlecht, and Vesa Välimäki, “Calibrating the sabine and eyring formulas,” *The Journal of the Acoustical Society of America*, vol. 152, no. 2, pp. 1158–1169, 2022.

[11] Prawda Karolina, Sebastian J. Schlecht, and Vesa Välimäki, “Dataset of impulse responses from variable acoustics room Arni at Aalto Acoustic Labs,” Zenodo, Aug. 2022.

[12] Diego Di Carlo, Pinchas Tandeitnik, Cédric Foy, Antoine Deleforge, Nancy Bertin, and Sharon Gannot, ,” *arXiv*, 2021.

[13] Diego Di Carlo, Pinchas Tandeitnik, Cedric Foy, Nancy Bertin, Antoine Deleforge, and Sharon Gannot, “The dechorate dataset,” Zenodo, Mar. 2021.

[14] Thomas Robotham, Ashutosh Singla, Olli S. Rummukainen, Alexander Raake, and Emanuël A. P. Habets, “Audiovisual database with 360° video and higher-order ambisonics audio for perception, cognition, behavior, and qoe evaluation research,” in *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*, 2022, pp. 1–6.

[15] Sakari Tervo, Jukka Pätynen, Antti Kuusinen, and Tapio Lokki, “Spatial decomposition method for room impulse responses,” *Journal of the Audio Engineering Society*, vol. 61, no. 1/2, pp. 17–28, 2013.

[16] Sebastià V Amengual Garí, Johannes M Arend, Paul T Calamia, and Philip W Robinson, “Optimizations of the spatial decomposition method for binaural reproduction,” *Journal of the Audio Engineering Society*, vol. 68, no. 12, pp. 959–976, 2021.

[17] David Alais and David Burr, “The ventriloquist effect results from near-optimal bimodal integration,” *Current biology*, vol. 14, no. 3, pp. 257–262, 2004.