








Auditory-visual scenes for hearing research

Steven van de Par^{1,a,*} , Stephan D. Ewert^{1,a} , Lubos Hladek², Christoph Kirsch¹, Julia Schütze¹, Josep Llorca-Boff³ , Giso Grimm¹ , Maartje M.E. Hendrikse^{1,4} , Birger Kollmeier¹, and Bernhard U. Seeber² 

¹ Carl-von-Ossietzky Universität, Oldenburg, Dept. Medical Physics and Acoustics, Cluster of Excellence “Hearing4all”, Carl-von-Ossietzky-Str. 9–11, 26129 Oldenburg, Germany

² Technical University of Munich, Audio Information Processing, Department of Electrical and Computer Engineering, Theresienstr. 90, 80333 München, Germany

³ RWTH Aachen University, Institute for Hearing Technology and Acoustics, Kopernikusstr. 5, 52074 Aachen, Germany

⁴ Erasmus University Medical Center, Rotterdam, Department of Otorhinolaryngology and Head and Neck Surgery, Burgemeester Oudlaan 50, 3062 PA Rotterdam, Netherlands

Received 25 October 2021, Accepted 27 July 2022

Abstract – While experimentation with synthetic stimuli in abstracted listening situations has a long standing and successful history in hearing research, an increased interest exists on closing the remaining gap towards real-life listening by replicating situations with high ecological validity in the lab. This is important for understanding the underlying auditory mechanisms and their relevance in real-life situations as well as for developing and evaluating increasingly sophisticated algorithms for hearing assistance. A range of ‘classical’ stimuli and paradigms have evolved to de-facto standards in psychoacoustics, which are simplistic and can be easily reproduced across laboratories. While they ideally allow for across laboratory comparisons and reproducible research, they, however, lack the acoustic stimulus complexity and the availability of visual information as observed in everyday life communication and listening situations. This contribution aims to provide and establish an extendable set of complex auditory-visual scenes for hearing research that allow for ecologically valid testing in realistic scenes while also supporting reproducibility and comparability of scientific results. Three virtual environments are provided (underground station, pub, living room), consisting of a detailed visual model, an acoustic geometry model with acoustic surface properties as well as a set of acoustic measurements in the respective real-world environments. The current data set enables i) audio-visual research in a reproducible set of environments, ii) comparison of room acoustic simulation methods with “ground truth” acoustic measurements, iii) a condensation point for future extensions and contributions for developments towards standardized test cases for ecologically valid hearing research in complex scenes.

Keywords: Complex acoustic environments, Speech intelligibility, Room acoustics, Ecological validity

1 Introduction

Speech is an important mode of communication in a wide range of daily-life situations. In real life, however, speech communication is often challenged by a number of complicating factors, such as the presence of ambient, interfering noise, a dynamically changing acoustic environment, and the presence of reverberation (e.g. [1]). In addition, speech communication may be impaired due to individual difficulties in understanding speech caused by hearing impairment (e.g., [2]) or due to unfamiliarity with the specific language (e.g., [3]).

In better understanding the effect of these factors on persons with a reduced speech recognition ability (e.g., hearing impaired listeners), much prior research has focused

on conditions with relatively simple, artificial stimuli. For example, balanced speech corpora (e.g., rhyme tests or matrix sentence tests) have been employed in conditions with one or two interfering sources like speech shaped noise played over a few loudspeakers or headphones. A particular advantage of such artificial stimuli is that stimulus properties are well defined (see, e.g., [4]) and dedicated strategies can be employed to investigate how a certain stimulus property affects speech intelligibility. This helps understanding the mechanisms underlying speech processing in humans, and allows developing models that predict speech intelligibility (e.g., [5–10]). Moreover, stimuli and experimental methods as well as model predictions can be reproduced and compared across different labs, which generally contributes to scientific development.

However, the ecological validity of the described classical “lab-based” speech and psychoacoustical experiments using such artificial stimuli has been questioned with regard

*Corresponding author: Steven.van.de.Par@uni-oldenburg.de, stephan.ewert@uni-oldenburg.de, seeber@tum.de

^aThese authors share first authorship.

to real-world outcomes (e.g., [11–14]). For noise reduction schemes and beamforming algorithms in hearing aids, it has indeed been shown that there is a discrepancy between laboratory results obtained with simple speech intelligibility measurements and real-life results [15, 16]. To close the gap towards real-life listening, tests need to consider additional factors that may affect speech perception in everyday listening situations (for a review see [17]). Keidser et al. [17] provided a comprehensive set of such factors which were grouped in five methodological dimensions: “Sources of stimuli, Environment, Context of participation, Task, and Individual”.

Beyond speech intelligibility, virtual acoustic and audio–visual environments can help us to better understand the ability to locate sound sources in reverberant environments (e.g., [18]), the mechanisms of auditory scene analysis, and the mechanisms of shared attention between acoustic and visual sources. Like for speech research, the complexity of the acoustic scene and the audio–visual configuration will affect performance. In addition, virtual audio–visual environments can help to better assess the quality of sound reproduction achieved with ear-level devices, or room acoustics of a to-be-built concert hall.

Real-life acoustical *Environments* are typically more complex than classical “lab-based” speech intelligibility tests using speech in (stationary) noise. Such complex acoustic environments (e.g., [19]) typically contain multiple, diverse, spatially distributed interfering *Sources of stimuli* such as speech, music, and diffuse background noise. Moreover, reverberation is typical to occur in enclosed spaces. Numerous studies have assessed the effect of specific aspects of interferers and their spatial distribution on speech intelligibility. It is known that the spectro-temporal properties of interfering sources influence speech intelligibility. For example, when a few interfering speakers (fluctuating interferers) are employed, frequent temporal gaps in the interferers will occur which allow listening into these gaps to hear the target speaker better and to improve intelligibility (e.g., [20, 21]). The spatial separation between interfering speech and attended speech improves intelligibility (e.g., [22–26]). This spatial benefit in speech intelligibility, however, is reduced in reverberant environments and may strongly depend on the orientation and position with respect to reflecting surfaces nearby [27, 28]. Depending on the specific real-life acoustic environment, these effects can be assumed to occur in specific combinations and to affect speech intelligibility in a particular manner. Therefore, the development of complex acoustic environments representative of a large variety of real-life everyday environments can be highly relevant for obtaining ecologically valid estimates of speech intelligibility. In addition, the assessment of the effectiveness of algorithms for hearing devices to be expected in real-life may depend on having a realistic audio–visual environment that provides a well-defined *Context of participation* which can elicit natural behavior of the participant, such as head movements. Several studies have shown that “knowing where to attend to”, i.e., a predictable as opposed to an unpredictable stimulus location, can improve speech intelligibility [25, 29, 30]. Visual cues

can guide the spatial attention of the listener [31, 32] and can affect the self-motion of listeners, which can in turn influence speech intelligibility, e.g., as a consequence of altered head orientation [33]. Moreover, lip reading, or speech reading, can specifically contribute to speech intelligibility in noisy situations [34–36].

One option to adopt these aspects naturally are field tests in real daily-life situations. However, in contrast to laboratory experiments, control over the precise acoustic condition and stimulus properties is typically very limited. This might affect evaluation of the results as well as the development of auditory models which can support interpretation of the results and development of hearing supportive algorithms. An alternative are spatial and dummy-head recordings of realistic scenes (e.g., [33, 37, 38]), allowing for the reproduction of existing acoustic scenes in the laboratory, however, with limited flexibility regarding the controlled modification of the scenes. Dummy-head recordings will also limit interactive behaviour in the scene, such as natural head movements that can improve intelligibility [33].

One more recent option is the use of virtual (acoustic) environments (VE) to produce realistic audio–visual scenarios for which all properties can be measured and controlled (e.g., [39–41]). Virtual reality techniques which synthesize the acoustic scene by (room) acoustics simulation and render the visual scene using computer graphics allow to systematically manipulate and interact with the scene. Several systems for (room) acoustics simulation and auralization exist that can all be combined with a visual component (e.g., [42–46]). When a high enough degree of realism is reached, virtual audio–visual environments thus offer the opportunity to precisely control and reproduce certain stimulus properties to, e.g., probe hypotheses about auditory processing, and to test the effectiveness of hearing-aid algorithms, while at the same time reaching a high degree of ecological validity, ideally exactly as in the corresponding real-life scenario.

For classical psychoacoustic experiments, certain methodologies have been widely used across labs (e.g., transformed up–down, [47]; matrix sentence test, [48–50]), which, combined with the acoustic calibration of broadly similar loudspeaker or headphone setups, lead to comparable results. Such established methods and measures, forming “de-facto” standards and enabling comparison of results across different research sites, do not yet exist for virtual audio–visual environments with applications in hearing research. Especially, the high complexity of VR systems and particularly the differences in setups, and acoustic and visual rendering techniques used across labs will likely lead to the use of different solutions for investigating the same problem across labs. As a consequence, reproducibility and comparability across labs requires a special effort.

For this reason, the current contribution presents a framework for defining and documenting complex audio–visual environments and embedded realistic communication “scenes” for hearing research, with the aim to stipulate increased reproducibility and comparability of research across labs. In this context, the environment refers to a specific audio–visual surrounding typically encountered in

real life, such as a living room. The term “scene” refers to a specific (communication) situation in a given environment, e.g., a conversation between two people seated on sofa chairs in the living room. Within an environment, a multitude of scenes can be defined. The proposed framework defines the required formats in which information about the environment and the scene needs to be provided to (i) enable recreation in different labs and to (ii) be extendable with further environments and scenes for new experiments. For visual and acoustic rendering of the environment, a geometric model provides detailed information for the visual representation, and coarser information for the acoustic representation. Simple (albedo) textures define visible surfaces, while acoustic surfaces are characterized by their absorptive and scattering properties. These definitions should be independent of the systems they are rendered with in order to provide greatest flexibility for use in different laboratories and over time.

Within the proposed framework, three example audio-visual environments (an underground station, a pub, and a living room) are specified according to the suggested format and supplemented with acoustic “ground truth” measurements obtained in the corresponding real-life environments. These three environments represent relevant daily-life situations in which speech recognition may be challenging.

Within these three environments, two scenes each define representative source-receiver positions and orientations. One scene is motivated by an audiological standard-test distribution of sources as far as reasonable in the context of the environment, while the other scene varies additional parameters, e.g., distances or interferer positions. Independent variables are the angular position and distance of sources. In addition, representative interfering source positions and/or signals are provided fitting the context of the environment. To verify acoustic and visual rendering methods, three types of acoustic measurements (omni-directional, Ambisonics, and dummy-head recordings) of selected source-receiver combinations from the existing real-life counterparts of the environments are provided. Static images are provided for visual verification.

The current contributions with documentations are hosted in a dedicated channel on the open Zenodo platform (https://zenodo.org/communities/audiovisual_scenes/), and new contributions from the community are cordially invited. With the provided information, the current contributions should enable researchers to reproduce the same virtual environments using their preferred visual and acoustical rendering methods. Based on future evaluation of the suggested and additionally contributed environments across different research laboratories, this contribution can serve as a starting point towards establishing standardized test cases for ecologically valid hearing research in complex scenes.

2 Audio-visual environments and scenes

Three audio-visual environments (underground station, pub, and living room; left to right in Fig. 1), modelled after

real-life enclosed spaces are provided. They cover a large variety of everyday communication situations with different degrees of complexity and reverberation time. For each environment, a visual model is provided (upper row of Fig. 1) as well as at least one simplified geometry model suited for real-time room acoustics simulation (middle row of Fig. 1). In each of the three environments, two audio-visual scenes are defined. In each scene, specific combinations of acoustic sources and receivers are used, resembling distances and spatial configurations typical for communication in the respective environments. The lower row shows the floor plan for each of the environments including the viewing direction (in red) used for the images in the upper row. In the first scene, several source positions are arranged in a circular manner around the listener, adapted to the geometry and natural communication distances in the environment. Source and listener are located at 1.6 m height. This scene represents a spatial configuration commonly used in audiology research. In the second scene, for a specific angular position, the focus is on different distances to the source appropriate for the environment. For each scene, acoustic measurements in the real-life spaces were performed, so that the scenes can be acoustically recreated using recorded impulse responses from the source position to the receiver or modelled using room acoustics simulation with the acoustic measurements serving as reference. In addition, we also provide room acoustic parameters of the environments.

A more detailed description as well a selection of characteristic measurements of these three audio-visual environments will be given in the remainder of this section. In addition, in Section 3, a comparison between room acoustic parameters of the three environments is provided. For the measurements, the reverberation time was computed according to ISO 3382-1 (T30 see para 6, EDT see A.2.2) using the decay between 5 dB and 35 dB below the stationary level extrapolating the time for a 60-dB (constant exponential) decay. The early decay time (EDT) was derived from the decay curve between 0 dB and -10 dB. For these calculations of T30 and EDT, the ITA toolbox [51] was used which implements the ISO 3382-1 standard. Additional descriptions as well as all measurements and models of the environments are provided as part of the freely available dataset structured as described in Section 4, under Hladek and Seeber [52], Grimm et al. [53], and Schütze et al. [54].

2.1 Underground station

This environment represents the platform of the underground (U-Bahn) station Theresienstraße in Munich, Germany (see left column of Fig. 1). The detailed floor plan is provided in Figure 2, showing the strongly elongated and large environment with overall dimensions of 120.00 m × 15.70 m and a ceiling height of 4.16 m from the platform, extending to 11.54 m around the escalators. The environment involves the lower platform and a part of the upper floor around the escalators. Only the lower platform was used to place sound sources and receivers. The volume of the platform space is 8555 m³, which increases to 11,084 m³ when the area around stairs and escalators is

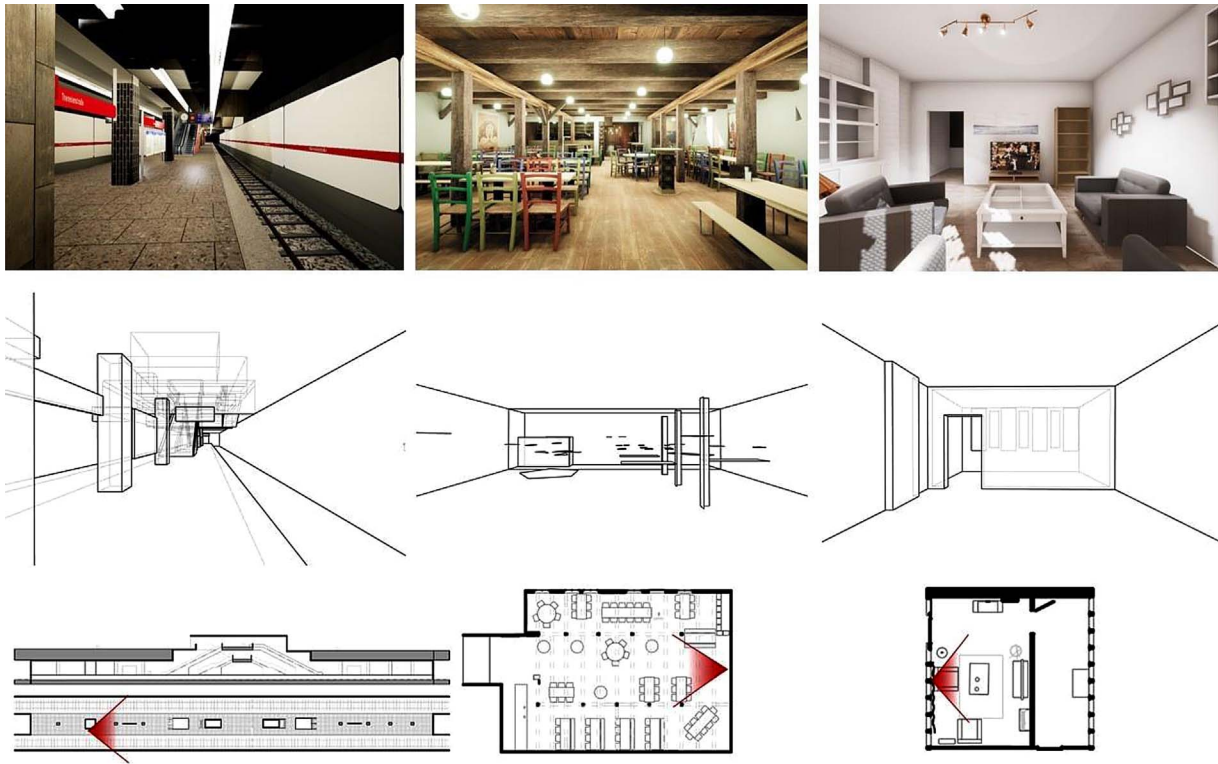


Figure 1. The three example environments underground station, pub, and living room (left to right) modelled after an underground station in the center of Munich, the OLS Brauhaus in Oldenburg, and the Living room Laboratory created at the University of Oldenburg for research purposes. Upper row: Visual rendering in Unreal engine. Middle row: Wire frame representation of the simplified geometry for room acoustics simulation. Lower row, the view frustum of the renderings is indicated in the floor plans which are depicted in more detail in [Figures 2–4](#).

included. The floor, column and track surfaces are composed of hard, acoustically reflective materials (stone tiles, concrete, crushed rock track ballast), while side walls and ceiling are covered with paneling and acoustical treatment. The reverberation time T_{30} ranges from 2.44 s at 250 Hz to 0.65 s at 8 kHz and the EDT changes between 1.46 s and 0.46 s in the same frequency range.

The environment is typically noisy, with fans providing air circulation and cooling the video projection system while escalators rumble – in the absence of the noise from incoming trains and announcements. Selected background sounds were recorded individually to recreate the environment’s noise with a room acoustic simulation. The escalators and the elevator were recorded with a directional condenser microphone placed near the sources (1 m – 1.8 m). Additional recordings with a multi-microphone array at the receiver position preserve spatial properties of the sound and can be used to recreate the acoustic background at the receiver position, e.g., with Ambisonics rendering.

Single-channel room impulse responses were measured from all source positions (1–17; see [Fig. 2](#)) to receiver positions R2–R5 and from sources positions (1, 3, 11) to receiver position R1 and can be used to verify the acoustic simulation. Multi-channel room impulse responses from various source positions (1, 3, 11, 16) to the listener position R1 can be used to present spatialized sources or interferers to the listener without the use of room simulation techniques.

Scene 1: Nearby communication – equidistant sources with semi-distal background noise sources

The first scene resembles communication with one or more nearby persons standing on the platform (see lower panel of [Fig. 2](#); receiver R1, source positions 1–12). The listener is in the center and the talker could stand in 1 of 12 possible positions spaced in equidistant 30° steps around the listener. The positions are in 1.6 m distance from the listener. This arrangement represents frequently used configurations in audiology research. Semi-distal noise sources are distributed at four angles (30° , 150° , 210° , 330°) at a distance of 2.53 m from the receiver, which could be used to create interference from other people or other noise sources on the platform.

Scene 2: Approaching person – radially spaced sources

The second scene represents a situation where a person is approaching or receding from the listener (see lower panel of [Fig. 2](#), receiver R1, source positions 13, 1, 14, 15, 16, 17). The sources are radially distributed along one line at distances from 1 m to 10 m such that the level change of the direct sound is 4 dB. Scene 1 and scene 2 are arranged around the same listener position R1 and share source position 1 on the circle in front of the listener and the interferer positions 18–21.

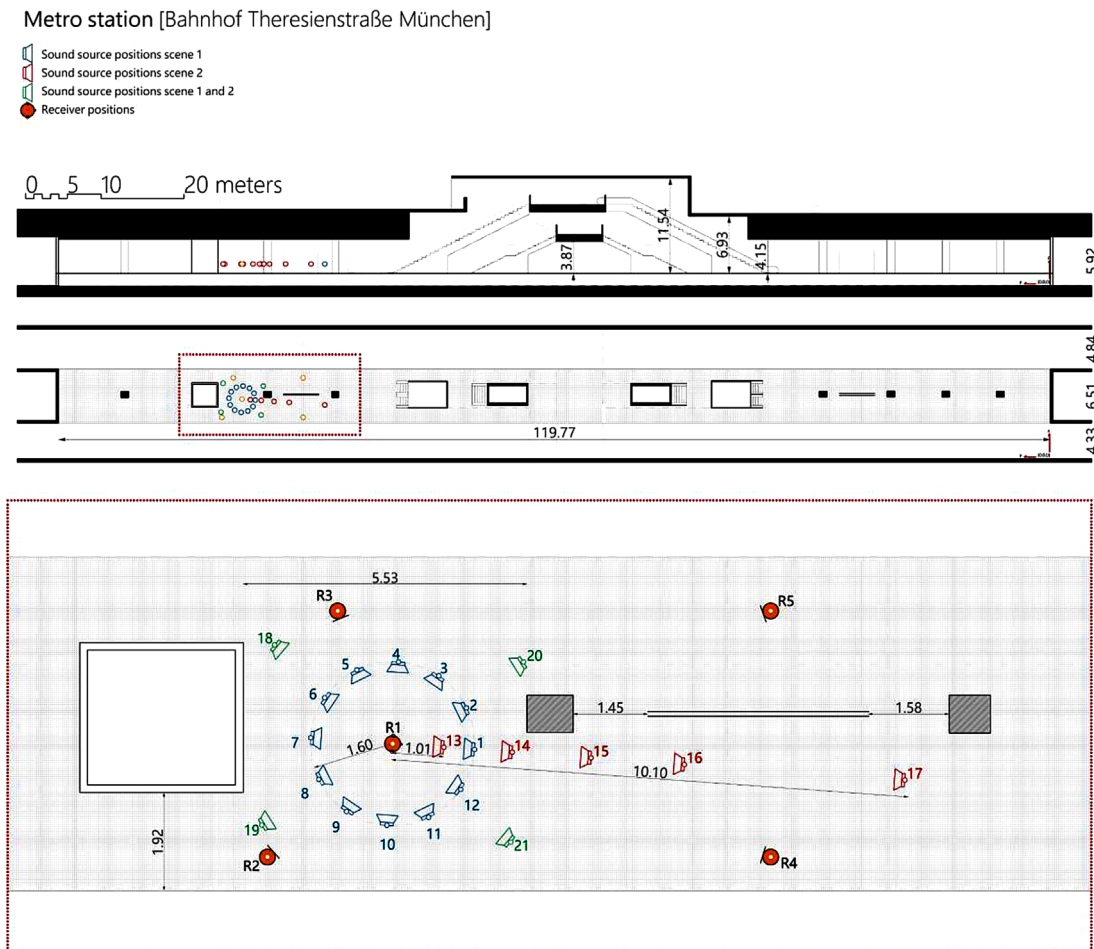


Figure 2. Upper panels: Cross section (top) and floorplan (middle) of the underground station environment with dimensions in meters. Lower panel: Magnified view of the area with receivers (denoted with R) (R1–5, yellow head and microphone symbols) and sources (loudspeaker symbols) indicating their orientation in the horizontal plane; all were located at 1.6 m height. The blue loudspeakers are part of scene 1 and the red loudspeakers of scene 2. Green loudspeakers can be used in both scenes for, e.g., interfering sources.

A more detailed description of the underground scene acoustics along with speech intelligibility data obtained with the binaural recordings for different source positions is given in Hládek and Seeber [55]. There, the scene rendering with the rtSOFE system [43] is compared to the “ground truth” measurements in the real space.

2.2 Pub

This environment is modelled after the Pub OLs Brauhaus (Rosenstraße) in the city of Oldenburg. The floor plan is shown in Figure 3 indicating overall dimensions of about 15 m × 10 m. The volume of the whole environment is about 442 m³. The walls are made of plaster, the floor of oiled wood and the ceiling of raw wood supported by rough irregular wooden beams. The pub is equipped with wooden tables and chairs, and a bar desk in one corner. The resulting reverberation time, T30, is 0.66 s.

The Pub resembles an environment in which people participate in social conversations and in which they might

experience challenges understanding one another because of babble noise and music in the background.

Impulse responses from many different source positions at neighboring tables were recorded, so that a babble-noise background can be generated. Moreover, impulse responses from a playback system for music in the pub were recorded, so that music can be added to the acoustic model. For both scenes, the receiver position R1 in Figure 3 serves as a listener sitting at the table. Three additional receiver positions were mainly intended for the room characterization and can be used as alternative listener positions. No background sounds or near-field sounds were recorded to avoid privacy issues. The impulse responses recorded from the other N, S, and P positions can be used to generate background babble noise, where the speech material presented from the N positions is probably understandable because of the close proximity to the listener position. The impulse responses recorded from the PA1 and PA2 positions can be used to add background music to the acoustic environment or other nearfield sounds representative for a pub environment.

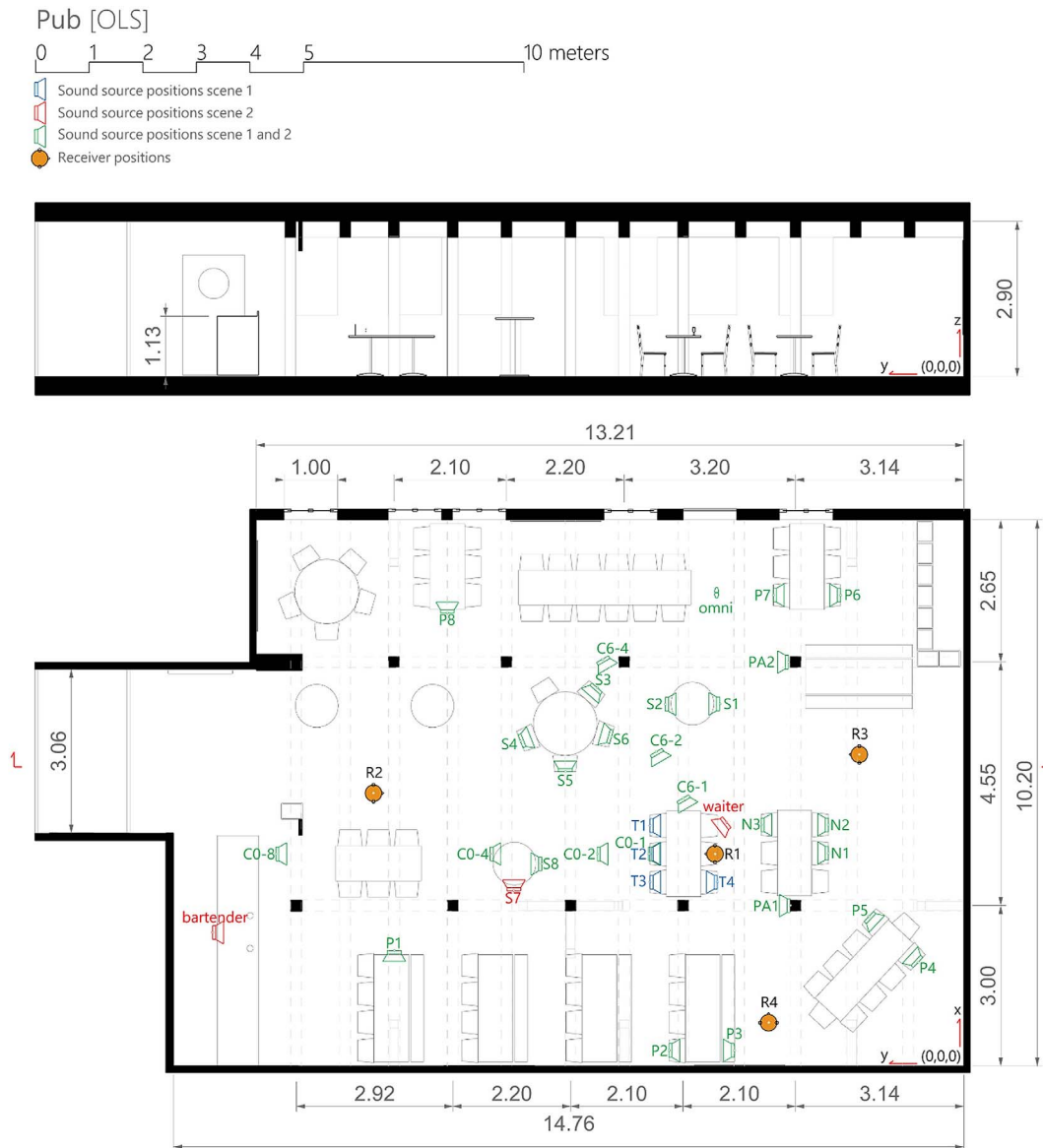


Figure 3. Cross section (top) and floorplan (bottom) of the pub environment in the same style as in Figure 2. Position and orientation in the horizontal plane of receivers (denoted with R1 to R4) is indicated by head symbols. Position and orientation of sources is indicated loudspeaker symbols. Sources indicated in blue belong to scene 1, sources indicated in red belong to scene 2. Here, T represents typical target sources at the table. Sources in green belong to both scenes. The letter N is used for sources at the neighbour table, which would typically be mostly intelligible, the letters S and P are used for sources further away. C refers to sources standing at various distances from R1. Furthermore, there are a bartender and waiter source, and an omni-directional source for measurements of room acoustic parameters.

Scene 1: Communication at a table

The first scene represents a person sitting at a table in the pub (R1) and source positions T1–4 can be used to represent a conversation with other people at the same table.

Scene 2: Communication with the waiter/waitress

The second scene resembles the situation in which a waiter/waitress approaches the table to take the order. In addition, two alternative receiver positions are provided,

resembling a person standing at overall three different positions in the pub.

2.3 Living room

The third environment is a living room with a connected room (kitchen; via a regular door made of tubular planks) that is part of the lab infrastructure of the University of Oldenburg. The floor plan is provided in Figure 4. The dimensions are 4.97 m × 3.78 m × 2.71 m (width × length × height), the volume of the living room

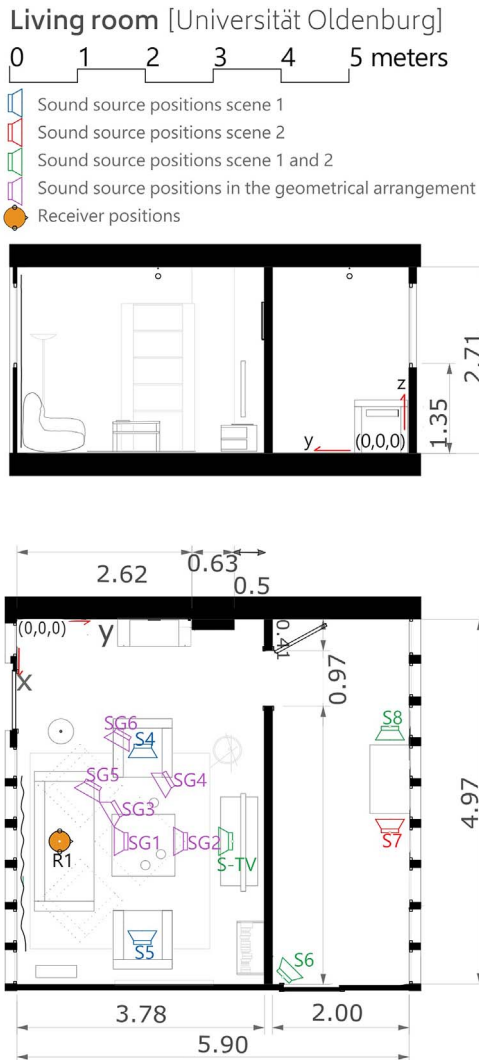


Figure 4. Cross section (top) and floorplan (bottom) of the living room environment in the same style as in Figure 2. Positions and orientations of receiver (letter R) and sources (letters S) are indicated by head, and loudspeaker symbols including their orientation in the horizontal plane. Source positions indicated in blue are attributed to scene 1, red to scene 2, green to both scenes. S-TV refers to the position of the television set.

is 50.91 m^3 , while the coupled room with the dimensions $4.97 \text{ m} \times 2.00 \text{ m} \times 2.71 \text{ m}$ has a volume of 26.94 m^3 , resulting in an overall volume of 77.85 m^3 . The walls are comprised of different materials: drywall material covered with wood chip wall paper can be found for three walls and the ceiling. The upper wall in the floorplan consists of bricks. The floor covering is laminate, which is partially covered by a 6 m^2 carpet made of Polypropylene and a layer of vinyl rubber on the backside. Window fronts are located on one side of the living room as well as for the opposite side of the kitchen. In the living room, a seating arrangement consisting of a textile couch and two textile armchairs can be found, arranged around a glass coffee table. In front of the brick wall of the living room is a cabinet, filled with glasses and decoration. Opposite to the couch a TV bench

with a TV is located. To the right of the TV, there is a bookcase filled with books. In the coupled room, a table and two chairs are placed next to the wall with the window front.

The reverberation time of the living room environment with open door between the living room and the coupled room ranges from $T_{30} = 0.55 \text{ s}$ at 250 Hz to 0.46 s at 8 kHz .

The Living room resembles an environment that people encounter frequently in their private homes and in which speech intelligibility may not necessarily be impaired – unless the television is turned on or the source speaker turns away from the listener. More challenging, however, are possible communication scenarios involving the connected room.

Background sounds were not recorded for this environment. Any audio from a TV show can be used for the TV source, while for the connected room typical kitchen sounds such as a dishwasher or a fridge may be chosen if stationary background noise is desired.

Scene 1: Television set and communication

The listener is seated on the chair/sofas in a conversation-like situation. The listener is listening to a second speaker, while at the same time various other sound sources are active (e.g., a television set).

Scene 2: Communication across rooms

Here, the source distance is increased and as a further aspect the listener is addressed from the neighbouring room with an obstructed direct sound path. Recent studies suggest that such an acoustically-coupled room setting can provide extra challenges for speech intelligibility [56, 57].

3 Room acoustical comparison of the environments

The provided scenes in the three environments represent different communication conditions typical for the respective environments. They follow the same principle of varying angular positions of targets and maskers in the horizontal plane (scene 1) or varying the distance to the target (scene 2), each embedded in the acoustical conditions of the environment and the respective background noise. For a comparative overview of the acoustic conditions in the three environments, Table 1 lists the broadband reverberation time in seconds (RT, estimated as T_{30}), early decay time, and the broadband direct-to-reverberant ratio (DRR) for a specific source receiver condition of the two scenes side by side. For a general comparison of the different acoustics in the environments, these measures are provided for a fixed source-receiver distance of about 1 m , in addition to the average values across all available source-receiver positions. As can be expected, the reverberation time is largest for the Underground station (with an average of 1.7 s) and smallest for the Living room (average 0.56 s). Because of the connected room, the reverberation time in the living room is not that much different to that of the Pub (average

Table 1. Various acoustic parameters for a comparable set of source-receiver combinations defined and measured within the three environments. In addition, room averages are calculated from a number of omnidirectional microphones and omnidirectional sources (except for the Pub, where all loudspeakers were used). Shown are the distance between source and receiver (occluded path length for the source in the adjacent kitchen of the Living room), the reverberation time, the EDT, and the DRR, in columns 3 to 6, respectively.

Descriptor		Dist. S-R [m]	RT (T30) [s]	EDT [s]	DRR [dB]
Under-ground	Room average		1.68	0.74	
	R1-S1 (Sc. 1), talker on “circle”	1.60	1.11	0.31	2.7
	R1-S13 (Sc. 2), talker very close	1.01	1.23	0.01	5.6
Pub	R1-S15 (Sc. 2), talker somewhat distant	4.02	1.25	0.35	-3.2
	Room average		0.66	0.68	
	R1-T2 (Sc. 1), talker on the same table	0.97	0.67	0.17	2.8
	R1-wtr (Sc. 2), „waiter” talking to listener	0.90	0.92	0.46	-0.4
Living room	R1-C0-4, talker at medium distance	4.00	0.65	0.56	-2.9
	Room average (door open)		0.56	0.46	
	R1-SG1, loudspeaker in 1 m distance	1.01	0.49	0.13	2.3
	R1-S-TV (Sc. 1), television running	2.51	0.49	0.23	-5.6
	R1-S7 (Sc. 2), talker in adjacent kitchen, occluded sound path	5.69	0.60	0.25	

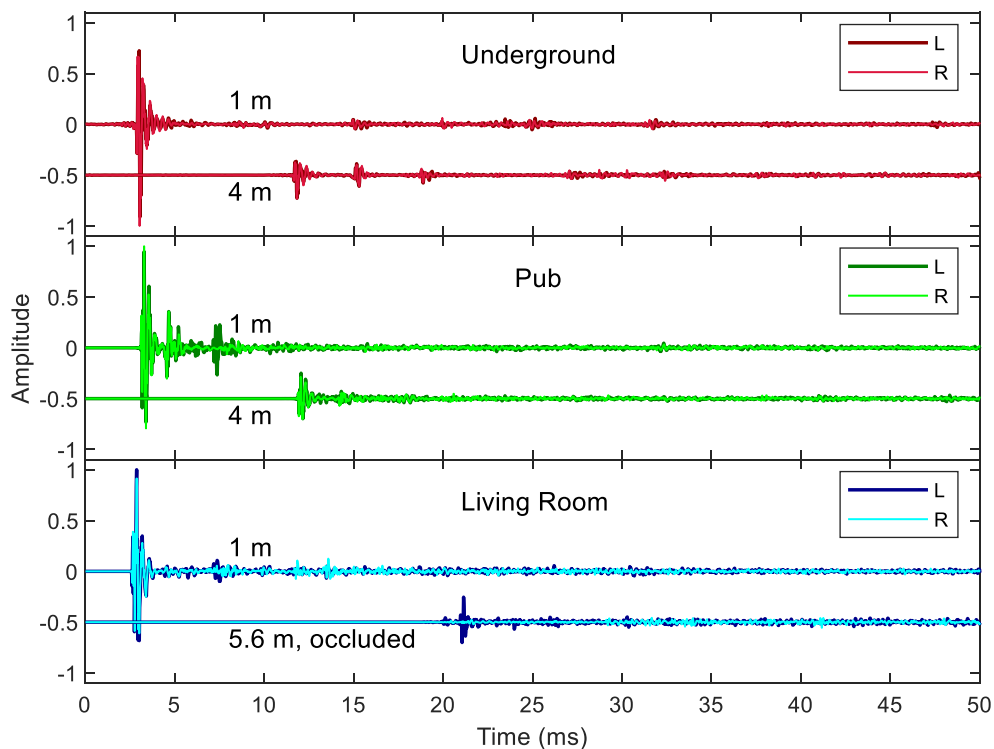


Figure 5. Impulse responses recorded in the three environments at 1 m distance (upper track) and at 4 m (5.6 m for the occluded path in the Living room) distance (lower track, offset by -0.5) as shown in Table 1. All impulse responses were normalized to the maximum of the 1-m condition. The distance-related amplitude reduction can be observed in the lower tracks. Note that for better readability the lower track in the Living room lab (lower panel) was scaled up by a factor of 2.

0.66 s). At the fixed distance of 1 m, the DRR drops from 5.6 dB in the underground to 2.8 dB and 2.3 dB in the Pub and Living room, respectively.

Figure 5 shows impulse responses recorded in the three environments at 1 m distance (top traces) and at 4 m distance (5.6 m for the occluded path in the Living room) as the lower traces with an offset of -0.5 , for the respective configurations provided in Table 1. For the Underground station (upper panel), distinct reflections with relatively

large temporal separation are obvious. For the larger distance (4 m), the direct sound and the first reflection (likely from the platform floor) are much closer spaced (about 3 ms) and more similar in amplitude, as can be expected. For the Pub (middle panel), a prominent early reflection from the table between source and receiver is visible for the short distance of 1 m. In the Living room (lower panel), multiple scattered and early reflections are visible reflecting the overall smaller volume with furniture. For the source in

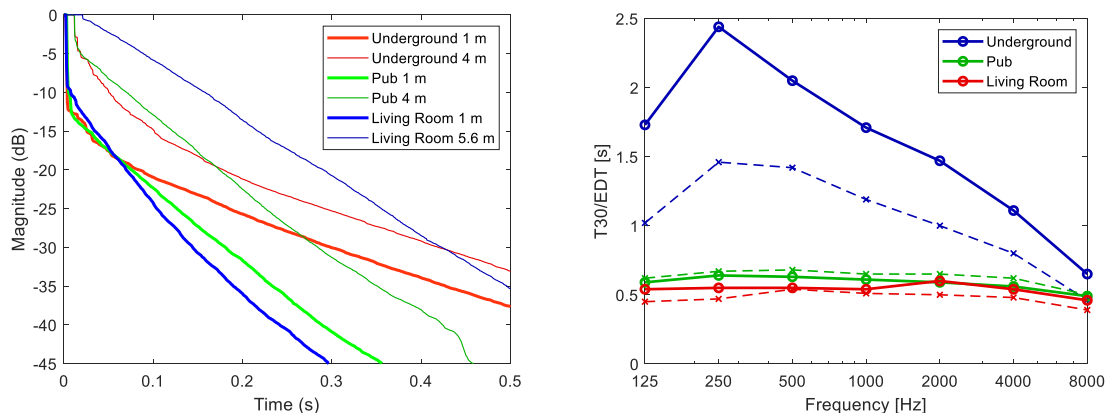


Figure 6. EDCs (left panel) for the three environments and distances (1.00 m, thick; 4.00 or 5.60 m, thin) as in Figure 5 and Table 1. In addition, reverberation times (T30, solid lines) and early decay times (dashed lines) per octave band averaged over all loudspeaker and receiver positions for the three environments are shown in the right panel. These estimates are comparable to the broadband estimates provided in Table 1.

the neighboring kitchen room (lower trace), the diffracted direct sound is weaker than the first reflection which is directly followed by dense reverberation from the coupled room.

For further analysis, Figure 6, left panel, shows the EDCs calculated for the same conditions as depicted in Figure 5 for 1 m (thick traces) and for the larger distances (thin traces). While the Underground (red) shows a dual-slope decay, likely related to local reverberation on the platform and a slower decay of the tube and coupled (escalator) volumes, the Pub (green) shows dominantly a single-sloped decay. For the Living room (blue), a dual-slope decay is observed for the short distance (thick trace) related to the coupled room, whereas the decay process of the coupled room dominates the condition with occluded direct sound (thin line). The left panel of Figure 6 shows the reverberation times (solid) and EDTs (dashed) for octave bands for the three environments. Comparable to the average broadband RTs and EDTs in Table 1, these measures were derived using all available loudspeaker and receiver positions. The Living room and Pub show fairly constant reverberation times with a slight decrease at 4 and 8 kHz, while the underground has a markedly increased reverberation time at low frequencies. In a similar fashion, a comparable frequency-dependent trend as for the reverberation times is observed for the EDTs.

Taken together, the comparison of RTs, EDTs, BRIRs, and EDTs demonstrates the large variety of acoustic conditions covered by the three environments. Moreover, the current comparative analysis only covers a small excerpt of all conditions available for each of the environments in the accompanying dataset.

4 Environment description and data files

In order to create a sustainable platform for the exchange of audio-visual environments for hearing research, a recommended standard set of data is defined

here to be provided for each environment (c.f. [58]). The environment and the data are described in a human readable *environment description document* (EDD), which extends the short descriptions provided above. The EDD contains all guidelines necessary to recreate the audio-visual scenes and the information provided in all files. The first part of the EDD contains all general information about the nature of the environment, the intended purpose of the scene(s), as well as specific data such as a floorplan, volume, and T30 times. The EDD also includes details about the recordings of background sounds and measurements conducted to obtain, e.g., impulse responses and directional characteristics of sources. The second part of the EDD describes the directory structure and file names of all provided *technical data files* (TDFs) for visual and acoustical rendering and the verification of rendering quality. Data files from specific measurements, like impulse responses, are included that allow to optimize and verify the acoustic simulation. Additionally, depending on the environment, original noise sources were recorded that can be rendered at different positions within the environment, or they were recorded for direct spatial reproduction at the respective locations of receivers in multi-channel spatial or binaural audio formats. The three environments with their six scenes as represented by their EDDs and TDFs have been published on Zenodo (https://zenodo.org/communities/audiovisual_scenes/).

Although for each environment two specific scenes are provided that entail sets of defined source-receiver combinations, within each audio-visual environment, other combinations of the available source-receiver positions or additional positions created with virtual acoustics can be used to address specific research questions. The defined scene positions nevertheless serve as “anchor points” for reproducible research since the simulations can be compared against recorded binaural room impulse responses in the corresponding real environment and against further baseline measures taken in our labs, e.g., of speech intelligibility.

4.1 Environment description document

The Environment Description Document provides all information about the environment as structured and easily assessable human readable information. The EDD contains two sections with the following information:

ENVIRONMENT DESCRIPTION

- 1 OVERVIEW
 - 1.1 Location
 - 1.2 Scenes
- 2 FLOORPLAN
- 3 VOLUME
- 4 SOURCE-RECEIVER POSITIONS AND ORIENTATIONS
 - 4.1 Scene 1 – equidistant circular sources with semi-distal background noise sources
 - 4.2 Scene 2 – radially spaced sources
- 5 SCENE SOUNDS (optional)
 - 5.1 Single channel recordings
 - 5.2 Multi-channel recordings
- 6 ACOUSTICAL SPACE DESCRIPTION
 - 6.1 Reverberation time
 - 6.2 Direct-to-reverberant energy ratio
- 7 MEASUREMENT DESCRIPTION
 - 7.1 Measurement conditions
 - 7.2 Sound sources
 - 7.3 Measurement microphone type
 - 7.4 Sound source signal
 - 7.5 Measurement equipment
 - 7.6 Averaging procedure
- 8 ACOUSTIC and VISUAL MODELS
- 9 ATTRIBUTION
- 10 REFERENCES

TECHNICAL DATA FILE (TDF) DESCRIPTION

- 1 MAIN DIRECTORY STRUCTURE
- 2 ACOUSTIC MODEL
 - 2.1 Description
 - 2.2 Files
 - 2.3 Application note

Detailed information about the requirements of each of the information items can be found in the supplementary information provided on (https://zenodo.org/communities/audiovisual_scenes/).

4.2 Technical data files

The audio-visual environments are defined in at least two separate models, one detailed model for the visual rendering, and (at least) one coarser model that allows real-time acoustic simulation and rendering. To ensure maximum compatibility, the widely used Wavefront object (.obj) format was chosen for 3D geometry which also has the advantage to be human readable and is thus easily editable with a text editor. For convenience, Blender (.blend) files as well as Unreal engine projects are optionally provided (for Unreal Engine 4.25). For the visual model, textures are provided in separate files referenced in

the .obj files or in accompanying material .mtl files. For simplicity, only simple (colormap or albedo) textures are provided. For the acoustic model, absorption (and if available scattering) properties are provided in a separate .txt file linked to the material names referenced for the surfaces in the acoustic .obj model.

For the acoustic model, on the one hand, the level of detail should be sufficiently accurate to allow a faithful simulation of the room acoustics, this means that single surfaces that create strong individual reflections should be specified. On the other hand, the level of detail should be sufficiently low to allow for a reasonable computational effort in simulating the room acoustics in a real-time system. The minimum requirement for a faithful rendering is not precisely known. Based on our experience with current real-time room acoustical rendering software, the following is proposed:

- Use a total of about 25 surfaces to define the boundaries of the acoustic environment simulated.
- Simulate additional surfaces close to sources and receivers when they create a respective solid angle to the source or receiver that is more than $36^\circ \times 36^\circ$ (this entails about 1% reflective surface of a full 4π spatial angle).
- Use expert insight about the particular simulated environment to finalize the level of detail required for the simulation.

For sources, directionality data can be provided in the SOFA-format ([59], AES69-2015; [60]). This is of specific relevance for the room acoustic simulation to be able to compare measured impulse responses to simulated impulse responses. In addition to source directivity, receiver directivity can be included in SOFA-format. Specifically, for the measurements made with an artificial head, the availability of a full set of head-related transfer functions (HRTFs) will allow comparing the measured binaural room impulse responses (BRIRs) with the simulated BRIRs using the acoustical simulation and rendering method the user chooses.

Linked to the specified source-receiver positions, a set of measurements is provided that were made in the real location to allow for optimization and verification of the acoustic rendering. The measurements support three purposes:

The first purpose is that BRIRs are measured for at least one specific source-receiver combination. Once the specific scene is simulated and rendered, it can be inspected whether the auralization matches the measured BRIRs. In this case, optimally, headphone rendering should be used, while that auralization is performed with an HRTF-set of the same dummy head which was used for the BRIR measurements.

The second purpose is that room impulse responses allow determining frequency dependent T30 times. Based on T30 times, the acoustic simulation of an environment can be optimized, for example, by adjusting absorption coefficients of the environment's surfaces. For measuring

T30 times, we recommend generally following ISO – 3382-2:2008, possibly with a reduction in the number of measurements such as using at least one source and two receiver positions.

The third purpose is that, ideally, recordings of typical background sounds of the environment are provided. These environment or scene specific sounds would specifically entail background sounds that would be regarded as interfering sources in a communication scenario. The background sounds can be recorded in two manners: One option is that recordings are made near a background source (or a number of background sources). In this case, the recordings can be rendered as, effectively, anechoic recordings that are placed in the scene at their respective places, and room acoustics is added as part of the simulation. The second option would be to capture the spatial sound field with a multi-channel microphone for reproduction via, e.g., higher-order Ambisonics, which allows the rendering of the captured spatial sound field either via headphones or via a loudspeaker setup.

Besides the above recordings, calibration files are provided as well. These are recorded signals of a calibrator placed on the measurement microphones in order to relate the recorded signals to a specified sound pressure level.

4.3 Contributions to the environment data set

The Zenodo channel on which the AV environments have been made publicly available is open for contributions from the community via the Zenodo channel https://zenodo.org/communities/audiovisual_scenes/. Potentially, some contributions will be made containing environments of particular research interest, which, however, do not exist in real life. In this case the corresponding measurement data will not be obtainable and can be discarded as part of the environment description document and as part of the environment data files. A template of the Environment Description Document with further submission information is provided in the Zenodo channel.

5 Discussion

A framework was presented to define audio–visual environments applicable for hearing research in complex acoustical environments. The framework contains visual and acoustical models that can be rendered with room acoustic simulation methods and visual rendering engines. In addition, each environment is supplemented with a range of measurements that allow to optimize and verify the acoustic rendering. Within each of the presented three environments two “scenes” are defined which represent specific source–receiver combinations that are typical for such an environment. Furthermore, sound recordings of background sources are included that are typical for such an environment. The presented framework, which can be retrieved via an online repository (https://zenodo.org/communities/audiovisual_scenes/), is open for future contributions from the general scientific community.

The implementation of the current environments in acoustic and visual virtual-reality rendering engines requires special attention to achieve reproducible results in auditory–visual research studies. For this reason, the Environment Description Document available for each environment provides detailed guidelines regarding the process of rendering the environment with the acoustic and visual models, namely: origin reference position, Z orientation, normal face orientation, material IDs, surface textures, receiver location and direction, source location and direction. The document further details the content of the Technical Data Files which contain the information for rendering, e.g., the acoustical models and impulse responses measured at scene positions.

Related efforts regarding virtual acoustical environment emerge in the literature. Brinkmann et al. [61] conducted a round robin to evaluate the state-of-the-art in room acoustical simulation and auralization. In this study, the focus was on evaluation of various existing simulation methods in terms of technical accuracy, and in terms of perceived plausibility and authenticity. Llorca-Bofi et al. [62] investigated the use of 3D photogrammetry to support acoustic measurements and to derive geometries for simulation. Their work relies on the photographic data – from concert halls, auditoriums and theatres-, to extract geometric information by triangulation algorithms, as well as acoustic material definition of surfaces objects via deep learning methods.

The present work will support future research into different rendering methods and their suitability to assess hearing abilities in more complex real-life environments. Further extension will also be needed when users can interactively move within the provided environments. The acoustic model can be extended to have a variable level of detail to be able to incorporate the variable effect close-by objects have on the perceivable sound field. For example, the underground scene’s acoustic models are provided in three versions differing in detail. Related to this, Llorca-Bofi and Vorländer [63–65], published a multi-detailed 3D architectural framework for sound perception research in Virtual Reality.

5.1 Future relevance for hearing research

The presented framework is envisioned to strengthen research on speech intelligibility and more general hearing research. The availability of ground-truth data for each complex acoustic environment will allow verifying, e.g., speech intelligibility in the acoustic simulations, permitting to make stronger assertions based on the findings of experiments performed in such virtual environments. In general, the virtual environments will allow to obtain subjective data in contexts more similar to real-life. Currently, surveys are exploring situations and environments in which persons with reduced speech recognition ability experience most challenges (e.g., hearing-impaired listeners, [66]). With the proposed framework, these environments and situations can be used in “lab-based” experiments and will be available across laboratories to allow to deepen understanding on

speech intelligibility in such complex environments within the scientific community.

The framework provided here will allow active participation of subjects in the environment. More specifically, the role of head movements in response to an active conversation can be investigated. Factors revolving about auditory attention, supported by visual information can be taken into account within a complex environment that is relevant in daily life. The interaction of head movements with hearing-aid processing can be studied within an audio-visual environment that should elicit much of the typical head-movement behavior that would also be observed in daily life [32].

Having a virtual acoustic rendering of complex acoustic environments, will allow to specifically manipulate auditory environments to get a better understanding about their relevance. Factors such as conservation of spatial Interaural Time Delay (ITD), Interaural Level Difference (ILD), and Inter Aural Cross Correlation (IACC) cues in binaural hearing aids can be investigated in relevant daily-life settings (e.g., [67, 68]). Whereas ITD and ILD cues are relevant for the perceived location of sound sources, IACC influences the perceived width of a sound source. In addition, it is possible within such a virtual environment to create the “perfect” hearing aid, that amplifies a single source, even in an interactive setting.

Complex auditory-visual environments that simulate every day settings will allow to better probe cognitive factors involved in processing speech information by (hearing impaired) listeners in such settings. It stands to reason that the complexity of every-day acoustic environments will be of relevance for the way cognitive resources are used by hearing-impaired listeners.

Finally, also for more basic hearing-related questions, such as the precedence effect, perception of moving sound sources, and distance perception, complex acoustic environments provide acoustic signals that will help gaining a better understanding about the perceptual mechanisms underlying these perceptual phenomena, specifically within everyday complex acoustic environments.

Conflict of interest

The authors declare no conflict of interest.

Data availability statement

The audio-visual environments and all relevant data accompanying these environments are available on https://zenodo.org/communities/audiovisual_scenes/.

Acknowledgments

We thank the reviewers for the detailed comments made to a previous version of this manuscript. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 352015383 – SFB 1330, projects A4, B1, C2, C4, and C5.

References

1. S.T. Goverts, H.S. Colburn: Binaural recordings in natural acoustic environments: estimates of speech-likeness and interaural parameters. *Trends in Hearing* 24 (2020) 1–19. <https://doi.org/10.1177/2331216520972858>.
2. R. Plomp: Auditory handicap of hearing impairment and the limited benefit of hearing aids. *The Journal of the Acoustical Society of America* 63 (1978) 533–549. <https://doi.org/10.1121/1.381753>.
3. A. Warzybok, T. Brand, K.C. Wagener, B. Kollmeier: How much does language proficiency by non-native listeners influence speech audiometric tests in noise? *International Journal of Audiology* 54, sup2 (2015) 88–99. <https://doi.org/10.3109/14992027.2015.1063715>.
4. S.D. Ewert: Defining the proper stimulus and its ecology – “mammals”. *The senses: a comprehensive reference*, Elsevier, 2020, pp. 187–206. <https://doi.org/10.1016/B978-0-12-809324-5.24238-7>.
5. H.J.M. Steeneken, T. Houtgast: A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America* 67 (1980) 318–326. <https://doi.org/10.1121/1.384464>.
6. ANSI: S3.5 (R2007), American National Standard Methods for the Calculation of the Speech Intelligibility Index. Acoustical Society of America, New York, 1997.
7. K.S. Rhebergen, N.J. Versfeld, W.A. Dreschler: Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *The Journal of the Acoustical Society of America* 120 (2006) 3988–3997. <https://doi.org/10.1121/1.2358008>.
8. R. Beutelmann, T. Brand, B. Kollmeier: Prediction of binaural speech intelligibility with frequency-dependent interaural phase differences. *The Journal of the Acoustical Society of America* 126 (2009) 1359–1368. <https://doi.org/10.1121/1.3177266>.
9. S. Jørgensen, S.D. Ewert, T. Dau: A multi-resolution envelope-power based model for speech intelligibility. *The Journal of the Acoustical Society of America* 134 (2013) 436–446. <https://doi.org/10.1121/1.4807563>.
10. T. Biberger, S.D. Ewert: Envelope and intensity based prediction of psychoacoustic masking and speech intelligibility. *The Journal of the Acoustical Society of America* 140 (2016) 1023–1038. <https://doi.org/10.1121/1.4960574>.
11. M. Cord, D. Baskent, S. Kalluri, B. Moore: Disparity between clinical assessment and real-world performance of hearing aids. *Hearing Review* 14 (2007) 22.
12. J. Jerger: Ecologically valid measures of hearing aid performance. *Starkey Audiology Series* 1 (2009) 1–4.
13. S. Kerber, B.U. Seeber: Towards quantifying cochlear implant localization performance in complex acoustic environments. *Cochlear Implants Intl.* 12 (2011) S27–29. <https://doi.org/10.1179/146701011X13074645127351>.
14. K.M. Miles, G. Keidser, K. Freeston, T. Beechey, V. Best, J.M. Buchholz: Development of the everyday conversational sentences in noise test. *The Journal of the Acoustical Society of America* 147 (2020) 1562–1576. <https://doi.org/10.1121/10.0000780>.
15. M.T. Cord, R.K. Surr, B.E. Walden, O. Dyrland: Relationship between laboratory measures of directional advantage and everyday success with directional microphone hearing aids. *Journal of the American Academy of Audiology* 15 (2004) 353–364. <https://doi.org/10.3766/jaaa.15.5.3>.
16. R.A. Bentler: Effectiveness of directional microphones and noise reduction schemes in hearing aids: a systematic review of the evidence. *Journal of the American Academy of Audiology* 16 (2005) 473–484. <https://doi.org/10.3766/jaaa.16.7.7>.

17. G. Keidser, G. Naylor, D.S. Brungart, A. Caduff, J. Campos, S. Carlile, M.G. Carpenter, et al. The Quest for Ecological Validity in Hearing Science: What it is, why it matters, and how to advance it. *Ear & Hearing* 41 (2020) 5S–19S. <https://doi.org/10.1097/AUD.0000000000000944>.
18. B.U. Seeber, S. Clapp: Auditory room learning and adaptation to sound reflections. In: J. Blauert, J. Braasch (Eds.), *The Technology of Binaural Understanding*, Springer, 2020, pp. 203–222. https://doi.org/10.1007/978-3-030-00386-9_8.
19. A. Weisser, J.M. Buchholz, G. Keidser: Complex acoustic environments: review, framework, and subjective model. *Trends in Hearing* 23 (2019) 2331216519881346. <https://doi.org/10.1177/2331216519881346>.
20. D.S. Brungart, N. Iyer: Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. *The Journal of the Acoustical Society of America* 132 (2012) 2545–2556. <https://doi.org/10.1121/1.4747005>.
21. A. Lingner, B. Grothe, L. Wiegrebe, S.D. Ewert, Binaural glimpses at the cocktail party? *Journal of the Association for Research in Otolaryngology* 17 (2016) 461–473. <https://doi.org/10.1007/s10162-016-0575-7>.
22. A.W. Bronkhorst, R. Plomp: The effect of head-induced interaural time and level differences on speech intelligibility in noise. *The Journal of the Acoustical Society of America* 83 (1988) 1508–1516. <https://doi.org/10.1121/1.395906>.
23. J. Peissig, B. Kollmeier: Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. *The Journal of the Acoustical Society of America* 101 (1997) 1660–1670. <https://doi.org/10.1121/1.418150>.
24. V. Best, C.R. Mason, G. Kidd, N. Iyer, D.S. Brungart: Better-ear glimpsing in hearing-impaired listeners. *The Journal of the Acoustical Society of America* 137 (2015) EL213–EL219. <https://doi.org/10.1121/1.4907737>.
25. E. Schoenmaker, T. Brand, S. van de Par: The multiple contributions of interaural differences to improved speech intelligibility in multitalker scenarios. *The Journal of the Acoustical Society of America* 139 (2016) 2589–2603. <https://doi.org/10.1121/1.4948568>.
26. S.D. Ewert, W. Schubotz, T. Brand, B. Kollmeier: Binaural masking release in symmetric listening conditions with spectro-temporally modulated maskers. *The Journal of the Acoustical Society of America* 142, 1 (2017) 12–28. <https://doi.org/10.1121/1.4990019>.
27. R. Beutelmann, T. Brand: Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America* 120 (2006) 331–342. <https://doi.org/10.1121/1.2202888>.
28. T. Biberger, S.D. Ewert: The effect of room acoustical parameters on speech reception thresholds and spatial release from masking. *The Journal of the Acoustical Society of America* 146 (2019) 2188–2200. <https://doi.org/10.1121/1.5126694>.
29. G. Kidd, T.L. Arbogast, C.R. Mason, F.J. Gallun: The advantage of knowing where to listen. *The Journal of the Acoustical Society of America* 118, 6 (2005) 3804–3815. <https://doi.org/10.1121/1.2109187>.
30. R. Teraoka, S. Sakamoto, Z. Cui, Y. Suzuki: Effects of auditory spatial attention on word intelligibility performance, in 2017 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP17), Guam, USA, 2017, 485–488.
31. V. Best, E.J. Ozmerai, B.G. Shinn-Cunningham: Visually-guided attention enhances target identification in a complex auditory scene. *Journal of the Association for Research in Otolaryngology* 8 (2007) 294–304. <https://doi.org/10.1007/s10162-007-0073-z>.
32. M.M.E. Hendrikse, G. Grimm, V. Hohmann: Evaluation of the influence of head movement on hearing aid algorithm performance using acoustic simulations. *Trends in Hearing* 24 (2020) 1–20. <https://doi.org/10.1177/2331216520916682>.
33. J.A. Grange, J.F. Culling: The benefit of head orientation to speech intelligibility in noise. *The Journal of the Acoustical Society of America* 139, 2 (2016) 703–712. <https://doi.org/10.1121/1.4941655>.
34. W.H. Sumby, I. Pollack: Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America* 26 (1954) 212–215. <https://doi.org/10.1121/1.1907309>.
35. A. MacLeod, A.Q. Summerfield: Quantifying the benefits of vision to speech perception in noise. *British Journal of Audiology* 21, 4 (1987) 131–141. <https://doi.org/10.3109/03005368709077786>.
36. J.-L. Schwartz, F. Berthommier, C. Savariaux: Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition* 93, 2 (2004) B69–B78. <https://doi.org/10.1016/j.cognition.2004.01.006>.
37. H. Kayser, S.D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, B. Kollmeier: Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. *EURASIP Journal on Advances in Signal Processing* 2009, 1 (2009) 298605. <https://doi.org/10.1155/2009/298605>.
38. J.F. Culling: Speech intelligibility in virtual restaurants. *The Journal of the Acoustical Society of America* 140, 4 (2016) 2418–2426. <https://doi.org/10.1121/1.4964401>.
39. S. Kerber, B.U. Seeber: Localization in reverberation with cochlear implants: predicting performance from basic psychophysical measures. *Journal of the Association for Research in Otolaryngology* 14, 3 (2013) 379–392. <https://doi.org/10.1007/s10162-013-0378-z>.
40. F. Pausch, L. Aspöck, M. Vorländer, J. Fels: An extended binaural real-time auralization system with an interface to research hearing aids for experiments on subjects with hearing loss. *Trends in Hearing* 22 (2018). <https://doi.org/10.1177/2331216518800871>.
41. M. Blau, A. Budnik, M. Fallahi, H. Steffens, S.D. Ewert, S. van de Par: Toward realistic binaural auralizations – perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario. *Acta Acustica* 5 (2021) 8–12. <https://doi.org/10.1051/aacus/2020034>.
42. E. Hafter, B. Seeber: The simulated open field environment for auditory localization research, in: Proc. ICA 2004, 18th Int. Congress on Acoustics, Kyoto, Japan, 4–9 April, Int. Commission on Acoustics, Vol. 5, 2004, pp. 3751–3754.
43. B.U. Seeber, S. Kerber, E.R. Hafter: A system to simulate and reproduce audio-visual environments for spatial hearing research. *Hearing Research* 260 (2010) 1–10. <https://doi.org/10.1016/j.heares.2009.11.004>.
44. D. Schröder, M. Vorländer: RAVEN: A real-time framework for the auralization of interactive virtual environments, in: Presented at the Forum Acusticum, Aalborg, Denmark, 2011, pp. 1541–1546.
45. T. Wendt, S. van de Par, S.D. Ewert: A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation. *Journal of the Audio Engineering Society* 62, 11 (2014) 748–766. <http://www.aes.org/e-lib/browse.cfm?elib=17550>.
46. G. Grimm, B. Kollmeier, V. Hohmann: Spatial acoustic scenarios in multichannel loudspeaker systems for hearing aid evaluation. *Journal of the American Academy of Audiology* 277 (2016) 557–566. <https://doi.org/10.3766/jaaa.15095>.

47. H. Levitt: Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America* 49 (1971) 467–477. <https://doi.org/10.1121/1.1912375>.
48. B. Hagerman: Sentences for testing speech intelligibility in noise. *Scandinavian Audiology* 11 (1982) 79–87. <https://doi.org/10.3109/01050398209076203>.
49. K.C. Wagener, T. Brand, B. Kollmeier: Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil III: Evaluation des Oldenburger Satztests. *Zeitschrift für Audiologie* 38 (1999) 86–95.
50. B. Kollmeier, A. Warzybok, S. Hochmuth, M.A. Zokoll, V. Uslar, T. Brand, K.C. Wagener: The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International Journal of Audiology* 54, sup2 (2015) 3–16. <https://doi.org/10.3109/14992027.2015.1020971>.
51. P. Dietrich, M. Guski, M. Pollow, M. Müller-Trapet, B. Masiero, R. Scharrer, M. Vorländer: ITA-Toolbox – An Open Source MATLAB Toolbox for Acousticians, in: 38. Jahrestagung der Deutschen Gesellschaft für Audiologie, Darmstadt. 2012. <https://git.rwth-aachen.de/ita/toolbox>.
52. L. Hladek, B.U. Seeber: Underground station environment (1.1) [Data set]. Zenodo, 2022. <https://doi.org/10.5281/zenodo.6025631>.
53. G. Grimm, M. Hendrikse, V. Hohmann: Pub environment [Data set]. Zenodo, 2021. <https://doi.org/10.5281/zenodo.5886987>.
54. J. Schütze, C. Kirsch, K.C. Wagener, B. Kollmeier, S.D. Ewert: Living room environment (1.1) [Data set], Zenodo, 2021. <https://doi.org/10.5281/zenodo.5747753>.
55. H. Hládek, B.U. Seeber: Communication conditions in virtual acoustic scenes in an underground station. *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, Bologna, Italy, 2021, pp. 1–8. <https://doi.org/10.1109/I3DA48870.2021.9610843>.
56. M. Schulte, M. Vormann, M. Meis, K. Wagener, B. Kollmeier: Vergleich der Höranstrengung im Alltag und im Labor, in: 16. Jahrestagung der Deutschen Gesellschaft für Audiologie (Rostock), 2013.
57. A. Haessler, S. van de Par: Crispness, speech intelligibility, and coloration of reverberant recordings played back in another reverberant room (Room-In-Room). *The Journal of the Acoustical Society of America* 145, 2 (2019) 931–944. <https://doi.org/10.1121/1.5090103>.
58. D. Leckschat, C. Epe, M. Kob, B. Seeber, S. Spors, S. Weinzierl, F. Zotter, DEGA-Memorandum Nr. VA 1201 zur Durchführung und Dokumentation von Audio-Produktionen für wissenschaftliche Anwendungen in der Akustik, DEGA VA 1201, 2020. <https://doi.org/10.5281/ZENODO.3597238>.
59. P. Majdak, Y. Iwaya, T. Carpentier, R. Nicol, M. Parmentier, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorf, H. Ziegelwanger, M. Noisternig: Spatially oriented format for acoustics: a data exchange format representing head-related transfer functions, in: *Audio Eng. Society Convention*, Paper 8880, 2013. <https://www.aes.org/e-lib/browse.cfm?elib=16781>.
60. P. Majdak, F. Zotter, F. Brinkmann, J. De Muynke, M. Mihocic, M. Noisternig: Spatially oriented format for acoustics 2.1: introduction and recent advances. *Journal of the Audio Engineering Society* 70 (2022) 565–584. <https://doi.org/10.17743/jaes.2022.0026>.
61. F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, S. Weinzierl: A round robin in room acoustical simulation and auralization. *The Journal of the Acoustical Society of America* 45 (2019) 2746. <https://doi.org/10.1121/1.5096178>.
62. FN Llorca-Bofi, I. Witew, E. Redondo, M. Vorländer: 3D modelling photogrammetry to support acoustic measurements and derive geometries for simulation, in: Presented at the Auditorium Acoustics, Hamburg, Germany, 2018. <https://doi.org/10.5281/zenodo.2583195>.
63. J. Llorca-Bofi, M. Vorländer: Multi-detailed 3D architectural framework for sound perception research in Virtual Reality. *Frontiers in Built Environment* 7 (2021) 687237. <https://doi.org/10.3389/fbuil.2021.687237>.
64. J. Llorca-Bofi, M. Vorländer: IHTAclassroom. Multi-detailed 3D architecture model for sound perception research in Virtual Reality [Data set]. Zenodo, 2021. <https://doi.org/10.5281/zenodo.4629716>.
65. J. Llorca-Bofi, M. Vorländer: IHTApark. Multi-detailed 3D architectural model for sound perception research in Virtual Reality [Data set]. Zenodo, 2021. <http://doi.org/10.5281/zenodo.4629760>.
66. P. Gablenz, U. Kowalk, J. Bitzer, M. Meis, I. Holube: Individual hearing aid benefit in real life evaluated using ecological momentary. *Trends in Hearing* 25 (2021) 1–18. <https://doi.org/10.1177/2331216521990288>.
67. C. Kirsch, J. Poppitz, T. Wendt, S. van de Par, S.D. Ewert: Spatial resolution of late reverberation in virtual acoustic environments, *Trends in Hearing* 25 (2021) 233121652110549. <https://doi.org/10.1177/23312165211054924>.
68. C. Kirsch, J. Poppitz, T. Wendt, S. van de Par, S.D. Ewert: Computationally efficient spatial rendering of late reverberation in virtual acoustic environments, in: 2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA). 2021, pp. 1–8. <https://doi.org/10.1109/I3DA48870.2021.9610896>.

Cite this article as: van de Par S. Ewert SD. Hladek L. Kirsch C. Schütze J, et al. 2022. Auditory-visual scenes for hearing research. *Acta Acustica*, 6, 55.