Towards Realistic Immersive Audiovisual Simulations for Hearing Research

Capture, virtual scenes and reproduction

Gerard Llorach[†1,2] **Giso Grimm**[2] **Maartje M.E. Hendrikse**[2] **Volker Hohmann**[1,2]
[1]Hörzentrum Oldenburg GmbH
Oldenburg, Germany
{g.llorach, v.hohmann}@hoerzentrum-oldenburg.de
[2]Medizinische Physik, Cluster of Excellence 'Hearing4all'
Universität Oldenburg, Germany
{g.grimm , maartje.hendrikse}@uni-oldenburg.de

ABSTRACT

Most current hearing research laboratories and hearing aid evaluation setups are not sufficient to simulate real-life situations and to evaluate future generations of hearing aids that might include gaze information and brain signals. Thus, new methodologies and technologies might need to be implemented in hearing laboratories and clinics in order to generate audiovisual realistic testing environments. The aim of this work is to provide a comprehensive review of the current available approaches and future directions to create audiovisual realistic immersive simulations for hearing research. Additionally, we present the technologies and use cases of our laboratory, as well as the pros and cons of such technologies: From creating 3D virtual simulations with computer graphics and virtual acoustic simulations, to 360º videos and Ambisonic recordings.

CCS Concepts: • Information systems ~ Multimedia content creation • Computing methodologies ~ Virtual reality • Human-centered computing ~ Interactive systems and tools • Hardware ~ Emerging interfaces

Keywords: Virtual Reality; Audiovisual Capture; Audiovisual Reproduction; Hearing Research

ACM Reference format:

Gerard Llorach, Maartje M.E. Hendrikse, Giso Grimm and Volker Hohmann. 2018. Towards realistic immersive audiovisual simulations for hearing research: Capture, virtual scenes and reproduction. In *Proceedings of 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia (AVSU'18), October 2018, Seoul, Republic of Korea. ACM, New York, NY, USA, 9 pages.* https://doi.org/10.1145/3264869.3264874

0 Introduction

There are several reasons for hearing research laboratories that work on the development and/or evaluation of hearing devices and their algorithms to move towards realistic audiovisual simulations. To begin with, it has been shown that the established procedures in the laboratories do not reflect real-life situations [1][2], one of the possible reasons being that head and gaze behaviors are different when vision cues are added [3], thus affecting speech intelligibility [4] and probably the performance of the hearing aid [5].

Another important factor is that recent advances in hearing technology have shown that the horizontal direction of the eyes can be measured inside the ear [6]. This information can be used to improve hearing aid algorithms [7][8]: gaze plays an important role in attention, as humans tend to look at the speaker a little more than two thirds of the time when listening [9]. Even more, brain activity recorded with electrodes around the ear can be used to find out which sound source is attended [10], probably with more ecologically valid results when visual cues are available. Having said that, it can be hypothesized that in the following generations of hearing aids gaze information and brain activity may be used. In order to be able to develop and evaluate these new algorithms, immersive surrounding audiovisual stimulations are needed.

Nevertheless, most current facilities for carrying out experiments with hearing impaired subjects do not provide any immersive visual stimuli, i.e., head-mounted displays (HMD), display screens or CAVE systems. Furthermore, acoustic conditions are usually limited to relatively easy stationary conditions with a few loudspeakers spatially arranged around the listener, which does not enable the reproduction of the complex dynamic acoustic scenes commonly experienced in daily life [11][12]. Some hearing research laboratories are starting to consider adding visual cues in their simulations [13][14][15] to solve some of the aforementioned issues.

The aim of this work is to provide a description of the existing approaches one could follow to create surrounding immersive realistic audiovisual simulations for hearing research. In the methodologies sections we describe the different methodologies to create or capture stimuli (*1. Methods I*), to design virtual scenes (*2. Methods II*) and to reproduce and display audiovisual simulations (*3. Methods III*). In section *4. Used technology and implementation* we describe the technologies that we used in our simulations in order to provide examples and specific solutions. In section *5. Conclusions* we use specific implementations of audiovisual environments to exemplify how to choose the right methodologies and finally, in section *6. Future work* we remark

AVSU'18, October 26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5977-1/18/10 $15.00
https://doi.org/10.1145/3264869.3264874

the importance of ecological validity and describe some of our ongoing work.

1 Methods I – Creation/Capture

In the following section we will describe how to create or capture the audiovisual stimuli. Some of the methodologies are more centered on doing an exact capture of a real scene to achieve a copy of a real situation. Other methodologies are more focused on creating malleable realistic simulations which can be adapted to different experiments. We divided the capture and creation between visual and acoustic stimuli, as these can be combined regardless of the method.

1.1 Creating visual stimuli

There are two main approaches to create visual stimuli, either video recordings or computer graphics. The advantages, disadvantages and properties of the methods are described in this section and summarized in Table 1.

1.1.1 Video recordings. With video recordings, one can achieve very realistic visual stimuli. Nevertheless, recorded videos cannot be modified easily once they are recorded, i.e. changing the position of a person, and the participants cannot interact with the environment, i.e. moving an object of the scene. If the desired scene or simulation is not too complex in terms of production (number of actors, vehicles, public scenarios with anonymity issues ...) video recordings can be quite effective and fast. Nowadays it is possible to record 360º videos with relatively affordable cameras, which can produce surrounding and immersive videos without noticeable artifacts (Figure 1). Additionally, technologies such as light-field cameras and some 360º 3D cameras can capture depth and, instead of one single point of view, permit a small area around the recording spot where the participant can move.

Figure 1. 360º video capture of a street scene. Acoustics recordings include binaural recordings, spot microphone recordings combined with GPS tracks and Ambisonic recordings.

1.1.2 Computer graphics. One of the advantages of computer graphics is that once the scenario is created, it can be used for different purposes and simulations. These simulations are malleable and easy to modify and reuse, but if very high realistic quality is desired, the effort to create them escalates quite quickly, even more when virtual animated characters are to be in the simulations. In order to create and design an environment, real scenarios can be captured and reconstructed using different techniques and technologies such as Lidar scans, depth cameras and reconstructions from still images [16]. Nevertheless, the post-production implied with these methods can be very effortful and complex to achieve scenarios without artifacts and with the desired final quality. Creating and designing environments with 3D editing software is a more progressive and safe approach, as the quality and complexity of the scene can be increased over time.

In order to generate visual stimuli with computer graphics, there are two different rendering approaches: real-time graphics or offline renderings. With the first approach, each frame of the simulation is rendered in real-time, which allows the scene to change during its reproduction. Users can move inside a virtual environment and can interact with the environment. With the second approach, offline renderings, each frame is rendered offline to create a final video or render. Offline renderings can achieve realistic results but, as each simulation needs to be previously rendered before being displayed, the production pipeline might be slower i.e. if we want to change the color of an object we would have to render again all the simulations that use that object. Offline renderings can achieve the quality and fidelity of a video recording (Figure 2), nevertheless the effort to create this type of offline renderings is very high in terms of production and time, even more when virtual characters, living beings or complex scenes are to be simulated.

Figure 2. Different visual qualities for Tarkin, a character from the Star Wars saga. From left to right: video recording from Star Wars: Episode IV - A New Hope, realistic offline rendering from Rogue One: A Star Wars Story and cartoon-like offline rendering from Star Wars Rebels – "Call to Action".

1.2 Creating acoustic stimuli

In order to create acoustic stimuli, there are several methods and technologies one can use, which sometimes are closely related to the reproduction system and scene. In this work we only present recording strategies with microphones and not synthesis methods such as text-to speech. Table 1 summarizes some of the properties of the recording methods presented.

1.2.1 Anechoic/clean recordings. These recordings are done in the best anechoic acoustic possible condition possible, to avoid any room effects and to obtain a clean signal. Anechoic/clean

recordings lack from any acoustic properties of the environment (reverberation, reflections, etc.), meaning that these acoustic properties can be added later in the simulation. If a sound would be recorded in a reverberant room, it could only be used in a simulation where the sound should have similar properties to that reverberant room. Otherwise, if that sound would be recorded in an anechoic condition, one could use it in any environment and simulation by convolving it with impulse responses from different environments for example.

When recording speech, sometimes it is not enough to just record the speaker in a silent/anechoic environment. If the recorded speech is to be spoken by a virtual human in a noisy environment, the speech should have different properties as humans change the way they speak in noisy environments. These speech transformations, the so-called Lombard speech [17], can be induced during the recording: the subjects are recorded while wearing headphones and hearing noise, their own voice and the voice of other participants of the conversation. The levels of the noise and the other speech signals are adjusted so the participants of the conversation can understand each other but have to speak with Lombard speech.

1.2.2 Ambisonic recordings and microphone arrays. Ambisonic microphones and microphone arrays can record an acoustic signal with some spatial information. They are usually composed of several microphones facing different directions in the same position. Although they provide better spatial resolution at higher orders (more microphones), the cost, complexity and artifacts escalate quite quickly when raising the order. The signals from the individual microphones can be later processed to obtain standard formats such as the B-format. Although Ambisonic microphones provide some spatial information, they only record the acoustic signals in one specific position of the scene.

Table 1. Properties of different audiovisual capture and creation methods. The last column (Virtual scene complexity) and the last row (Spatialized anechoic recordings) are explained in the following section (Methods II – Virtual scenes).

		Fidelity	Effort	Reusability to create new scenes	Virtual scene complexity
Visual stimuli	**Video recordings**	High	High/Low depending on the acted scene	Low	Low
	Real-time computer graphics	Low	Medium/High	High	High
	High quality offline renderings	High	High	High	High
Acoustic stimuli	**Ambisonic recordings**	High	High/Low depending on the post-processing of the mic signals	High for background diffuse sounds. Low for specific scene recordings	High if used for spatial decomposition. Low otherwise
	Spatialized anechoic recordings	High/ Low	High	High	High

2 Methods II – Virtual scenes

Virtual scenes create a spatial representation of virtual objects and sources. This virtual scene is necessary to render the virtual objects with their corresponding properties depending on the position of the virtual camera/listener. For video recordings and spatial audio, such as Ambisonic recordings, virtual scenes are not always required. Nevertheless, for computer graphics and object-based audio, i.e. anechoic recordings and virtual sources, a spatial representation needs to be present in order to create 3D effects such as distance attenuation and reflections.

2.1 Visual virtual scene

When using video recordings sometimes it is not necessary to have any specific virtual scene, especially when they are played through desktop displays or flat screens. Nevertheless, when surrounding videos and videos with depth are used, some kind of spatial representation is needed. For example, there are some established approaches for representing 360º videos in virtual scenes, such as using a sphere with a video texture surrounding the virtual camera. When using 3D computer graphics, a spatial virtual representation is inherently present: the geometry of the scene. Video recordings can also be displayed together with computer graphics i.e. 2D videos can be displayed by adding video textures to flat planes in the 3D scene.

2.2 Acoustic virtual scene

In order to create realistic environments, the effects of the acoustic environment need to be taken into account (reverberation, reflections) as well as the effects caused by the moving objects such as vehicles (Doppler effect, distance attenuation, directivity). Ambisonic recordings already provide some spatial information about the scene and the effects of the environment from the recorded spot. Anechoic and clean microphone recordings do not provide any spatial information, as they are meant to capture the sound of an object without any

reference to the external world, thus there is the need to spatialize the microphone recordings. In the same way as computer graphics, there are real-time and offline approaches to spatialize sounds. Real-time approaches commonly simulate effects such as the Doppler effect, distance attenuation and directivity as well as other room effects such as pre-computed reverb effects, delays and first order reflections. The acoustics with real-time simulations are always approximations to the real acoustics, as all the reflections of the sounds and room acoustics cannot be computed in real-time. With offline approaches, acoustic simulations can be more realistic as a higher order of reflections can be computed. Nevertheless, this approach fails when scenes are dynamic and they can change their properties in real-time i.e. if a reflecting surface changes (a moving truck for example) all the room acoustics will change and will need to be computed again.

Another approach is to extract the spatial information from Ambisonic recordings and microphone arrays. For example, it is possible to use Ambisonic microphone recordings to obtain a direction-of-arrival and a diffuseness parameter on the frequency domain [18] and to improve the spatial resolution of Ambisonic recordings [19].

3 Methods III – Reproduction

In this section we will speak about the methods to reproduce audiovisual stimuli in the context of hearing research. We will focus on single-user setups for small laboratories and clinics, although some of the setups could be multi-user and scaled to bigger dimensions. Table 2 and Table 3 summarize the properties of different approaches for visual and acoustic reproduction respectively.

3.1 Visual reproduction

The most simplistic approach to reproduce visual stimuli is to use desktop displays. The number and size of displays can be increased progressively to surround the user, but up to a point it can be more effective to use image projectors as one projector can cover a larger area. The most common immersive setup using multiple projectors is the cave automatic virtual environment (CAVE) [20]. This setup usually consists on a squared/rectangular room where each wall has a projected image. The projectors are outside this room, thus more space around it is required. The user's viewpoint is tracked and the stereoscopic projectors and shutter glasses permit depth perception.

A more simplistic approach is to surround the user with an acoustically transparent cylindrical screen, without visual cues on the floor and the ceiling (Figure 3), similar to [15]. In comparison to a CAVE, the field of view is smaller but there are no edges in the corners as the screen is circular, not squared. One of the big disadvantages of CAVEs is that the room shape does not help the acoustics: a squared/rectangular room with flat surfaces creates lots of reflections. Additionally, some CAVEs use hard surfaces for the walls, which increase the acoustic reflectivity. Although there have been approaches to do spatialized audio with loudspeakers in CAVEs [21], the complexity of the acoustic system and installation increases. When using a cylindrical screen, the loudspeakers can be placed behind the screen with standard setups for 3D audio reproduction.

These systems with surrounding images and projectors require lots of calibration and expertise to set up. In comparison, head-mounted displays (HMD) have become consumer available and are relatively affordable and simplistic to set up. Simulations with HMDs can be easily installed in different spaces in less than an hour. The disadvantages are that the user has to wear something bulky in the head and that the device can cause some acoustic distortions and reflections when using loudspeakers for sound reproduction. Regarding the preference of the user, CAVEs and HMDs have similar preference ratings [22].

Table 2. Properties of different surrounding visual display systems

	CAVE	Surrounding screen	HMD
Cost	High	High	Low
Implementation effort	High	High	Medium-Low
Portability	Low	Low	High
Immersion	High	Medium	High
Room acoustics quality	Low	Medium-High	Medium-High

Figure 3. Visual reproduction with an acoustically nearly transparent cylindrical screen. The cafeteria scenario is used in [53].

3.2 Acoustic reproduction

Acoustic signals can be reproduced either through loudspeakers or headphones. How to choose between them can depend on many factors such as the room acoustics and size, the visual reproduction system, if the participant is wearing a hearing aid/cochlear implant, the stimuli/recordings used and the 3D audio software. For example, binaural reproductions are usually played via headphones with head-tracking. Binaural reproduction synthesizes how the acoustic signal is received on each ear using head-related transfer functions (HRTFs). Hearing aid simulations with normal hearing are easier to test with headphones because arrays of loudspeakers and acoustically treated rooms are not

required. Nevertheless, there are some existing issues, as each person has different HRTFs and the spatialization does not work for everybody [23].

Loudspeaker setups are especially important for testing listeners with hearing aids and cochlear implants as these populations cannot be tested with headphones. There are several methods for rendering 3D audio with loudspeakers. Wave field synthesis involves using large arrays of loudspeakers to create artificial wave fronts [24]. The main advantage of wave field synthesis is that there is no sweet spot and the participant can move inside the room without head-tracking. Most wave field synthesis work only on one plane (no elevation, only azimuth) and they require a large number of loudspeakers to avoid artifacts. Moreover, the acoustic frequency limit is typically low (in the oder of 1.5-2 kHz), which is perceptually fine, but is not sufficient for hearing devices which depend on a correct sound field even at higher frequencies. Vector base amplitude panning (VBAP [25]) is another method for 3D audio rendering. It requires multiple loudspeakers, but these don't need to have any specific configuration, as this algorithm will find the best positioned loudspeakers to reproduce a virtual source. The main issue with VBAP is that the coloration of a virtual source can change depending on the position of the virtual source and the loudspeakers layout: if a virtual source is in the same place as a loudspeaker, only one loudspeaker will be used, but if it is not, a maximum of three loudspeakers will be used to reproduce the virtual source, creating spectral distortions. With Ambisonic reproduction, this issue is solved, as the diffuseness is constant regardless of the position of the virtual source. Nevertheless, the 3D audio algorithm suffers from spatial resolution at lower orders [26] but works great for diffuse sources and environmental sounds. Another approach that uses Ambisonic recordings is Directional Audio Coding (DirAC) [27], although its use is not as extended as the other 3D audio methods. Table 4 shows existing literature comparing different 3D audio rendering methods.

Table 3. Properties of different 3D audio rendering techniques.

		Hearing aids	Spatial resolution	Sweet spot area
Loudspeaker arrays	**VBAP**	Yes	High	Medium-High
	High Order Ambisonics	Yes	High	Medium
	Low Order Ambisonics	Yes	Low	Low
	Wave Field Synthesis	Yes/No	High	High
Headphones with head-tracking	**Binaural reproduction with HRTFs**	No	Variable between individuals	-

Table 4. Studies comparing different 3D audio rendering techniques.

	High Order Ambisonics	Low Order Ambisonics	Wave Field Synthesis
VBAP	Simon et al. 2017 [55], Grimm et al. 2015 [28],	Pulkki and Hirvonen 2005 [29]	-
High Order Ambisonics	-	Bertet et al. 2013 [30]	Daniel and Nicol 2003 [31], Spors and Ahrens 2008 [32]
DirAC	Politis et al. 2017 [33]	-	-

4 Used technology and implementation

In the following section we will explain the technologies and methodologies that we chose for our hearing research laboratory.

4.1 Creation/Capture

4.1.1 Visual stimuli. In our work, we created most of the simulations with computer graphics. The 3D environments that we created are common listening scenarios (cafeteria, living room, train station, lecture hall, street) and they were used in different experiment setups. We created the 3D environments using Blender [34] and spatial references such as site plans, satellite images and still images from the real spaces. We used two different tools to create virtual characters: Makehuman [35] and Autodesk Character Generator [36]. The characters made with Makehuman had the automated behaviors of blinking, breathing, lip-syncing [37] and turning the head towards the current virtual character speaking. The virtual characters created with the Autodesk Character Generator were used to fill the scenes as crowd and had animations such as walking, cycling and waiting as well as automated behaviors such as blinking, breathing and randomly looking around. As all the characters made with the same tool share the same base properties, body animations and behaviors can be easily applied to different characters e.g. we used the same walking animation at different speeds for different crowd characters. We are currently considering using Adobe Fuse [38] and Mixamo [39] for the next generation of virtual characters because of the quality, simplicity and the database of animations.

For a different experiment [3], we recorded videos with two cameras next to each other (Canon EOS 700D). With the two reflex camera setup we could achieve a panoramic video with a horizontal field of view of 100º approximately and a resolution of 3840x1080px. We used these cameras to record a conversation between four persons with a uniform background [40].

In an ongoing experiment we used a low-cost 360º camera, the Xiaomi Mi Sphere Camera, to replicate an exact real scene into the laboratory (Figure 1). The 360º camera that we are using has two wide-angle fish eye cameras that can compose a final video of

3.5K. The images are stitched and processed with the dedicated software of the camera manufacturer without noticeable artifacts.

4.1.2 Audio stimuli. Most of the sounds were recorded individually in the best acoustic conditions possible to obtain a clean signal. Speech and conversations were recorded in an acoustically treated room with individual microphones and with headphones in the case of Lombard speech conversations. Additionally we created new speech stimuli by applying some distortion and filters to anechoic recordings, as in some of our simulations, a speech signal would not be spoken by a virtual human in the virtual environment but through a loudspeaker i.e. the television newscast in a living room or the announcements in a train station. We applied these distortions and filters to simulate the distortion of the low-end loudspeaker.

For sounds of objects of the environment i.e. fire crackling, beeps of a machine, pens writing, etc. we used recordings available in sound databases such as freesound.org [41] or we recorded them ourselves with a handheld recorder such as the Zoom H6. In the case of moving objects, such as vehicles i.e. cars, trucks, trains, etc. we used spot microphones attached to those vehicles with magnets to record the sound of the engine and the sound of the wheels individually (Figure 1). If the vehicles were moving, we recorded GPS coordinates to estimate the velocity and relate it to the recorded sound. If these vehicles were kept at the same velocity, the recorded sounds could be looped and used in other simulations.

We also used a first-order Ambisonic microphone, a tetrahedral microphone, to record diffuse sounds such as the background noise of a cafeteria. We also used it together with the 360º camera, to record the sound received at the position of the camera with some spatial information (Figure 1). These recordings are used to time align video recordings and other spot microphone recordings or to do acoustic scene analysis and spatial upsampling.

4.2 Virtual scenes

4.2.1 Visual virtual scene. In our work we used real-time game engines for computer graphics simulations. We decided not to use specific offline rendering tools as in most of our simulations the viewpoint of the user is tracked and can change. Nevertheless we used some of the techniques of offline renderings to improve the performance and quality of real-time graphics such as light baking (pre-rendering all the lights and shadows to objects that are static). We used two different real-time game engines: Unity [54] and Blender. Although Unity offers better tools and performance as a game engine, the production pipeline is faster with Blender, as the 3D editing and the game engine are one single application. To display 360º videos in the cylindrical screen, we used common video players such as VLC Lan and MPlayer. We used the zoom and crop options of the aforementioned video players to adjust the videos to the cylindrical screen. We plan to use Unity to display 360º videos with the HMD or 360º video media players.

4.2.2 Virtual acoustic scene. Virtual acoustic environment engines try to simulate the room acoustics of a virtual space. One of the approaches is to compute offline the impulse response for each virtual source for the position of the virtual listener. Engines such as EASE [42] and ODEON [43] are very powerful in this area, but are mainly specialized in architectural acoustics and loudspeaker layout design. In our case we opted to use TASCAR [44], an engine specialized for hearing research. Aside from providing relevant 3D audio effects such as the Doppler effect and distance attenuation, TASCAR has the advantage of being interactive and real-time. The software uses the first order image source model [45] to create reflections for each virtual source and first order Ambisonics (FOA) for diffuse sources and late reverberation. Although the image source model creates an approximation of the room acoustics, with this approach virtual sources and receivers can move as well as acoustic reflectors. The engine has also the possibility to render impulse responses and to integrate plugins from a simulator of hearing aids, the openMHA [46]. Other available open-source acoustic engines are the 3D Tune-In Toolkit [47], with hearing aid simulations and integration in Unity and the SoundScape Renderer [48].

Another approach that we are implementing is to use first-order Ambisonic recordings to create virtual sources and improve the spatial resolution through spatial upsampling. Although this method would only work in specific acoustic conditions [18], it would reduce the effort, cost and complexity in some scene capture and recordings: just with a 360º camera and a first-order Ambisonic microphone one could capture both surrounding video and 3D audio in one spot, as done by [50].

4.3 Reproduction

4.3.1 Visual reproduction. To reproduce surrounding visual stimuli we use either a cylindrical screen or a HMD (HTC Vive). We render the projector images with the Blender Game Engine and we use Unity for the rendering of the HMD, as there are many tools and packages for HMDs in this software. The acoustically transparent cylindrical screen surrounding the user receives the images of three projectors, achieving a 300º horizontal field of view (Figure 3). The user's viewpoint is tracked with infrared cameras and markers on the head of the user. Inside the same laboratory the HTC Vive with the HTC Base Stations are installed.

4.3.2 Acoustic reproduction. In order to reproduce surrounding and 3D audio, we use a horizontal array of 16 loudspeakers and another additional 12 loudspeakers distributed above and on the floor. The horizontal array of loudspeakers is placed behind the cylindrical screen in a way that the loudspeakers cannot be seen. With TASCAR one can choose the reproduction and rendering method desired: nearest speaker panning, VBAP, Ambisonics and binaural rendering. In our experiments we used a 7^{th} order Ambisonic panning with max-rE decoding [49] for virtual sources and for diffuse sources, such as background noise, we used FOA reproduction. We also tried to reproduce FOA recordings in combination of 360º videos, but the spatial resolution of these Ambisonic recordings was too poor. We are currently exploring techniques for reproducing spatially upsampled FOA recordings.

Table 5. Description of different methodologies used in two different scenarios. The cafeteria scenario (Figure 3) is created with computer graphics, thus different situations can be created i.e. two simultaneous conversations in a table, only two speakers, etc. The street scene (Figure 1) is an acted scene were cars/road vehicles pass through a street with different velocities.

		Creation/Capture	Virtual scene	Reproduction
Cafeteria scenario (Figure 3)	**Visuals**	3D modelling (Blender)	Blender	Cylindrical screen with user's viewpoint tracking
			Unity	HMD
	Acoustics	Anechoic/clean recordings for virtual sources	TASCAR	7th Order Ambisonic panning with max-rE decoding
		FOA recordings (tetrahedral microphone) for diffuse sources (background noise)	TASCAR	FOA reproduction
Street scenario (Figure 1)	**Visuals**	360º video recordings	None (video player)	Cylindrical screen
			Unity /video player	HMD
	Acoustics	Spot microphones and GPS tracks for moving sources (vehicles)	TASCAR	7th Order Ambisonic panning with max-rE decoding
		FOA recordings for synchronization	-	-
		FOA recordings for spatial upsampling	TASCAR – *In development*	*Spatial upsampling techniques- In development*

5 Conclusions

In this work we present different methodologies and specific solutions to produce audiovisual realistic simulations. We aim at providing a guide for laboratories that want to start implementing audiovisual simulations. Visual and acoustic approaches are presented separately, as they can be combined in different ways. For example, in the street scenario (Table 5) we considered different methods: video recordings or computer graphics; combined with object-based audio and virtual sources or spatially upsampled FOA recordings; and reproduced either through a curved screen or a HMD.

Which approach and method to follow depends much on the specific topic to research, the experiment setup and the expertise and facilities available. Commonly one would aim at the best quality/effort ratio among the available methods. In the aforementioned street scenario (Table 5), our choice is video recordings and spatially upsampled FOA recordings in order to avoid the effort to create any virtual scene. We are comparing the perceived loudness and acoustic annoyance of land vehicles between the real-life acted scene and the laboratory simulation, thus we don't need to create stimuli that is malleable as we want an exact copy of the real-life experiment. As the experiment might be reproduced in a second facility without a cylindrical screen, the visual reproduction will be both done with a cylindrical screen and a HMD. In the case of the cafeteria scenario (Table 5), the cafeteria 3D model is used in several experiments that study head and gaze behaviors and that do simulations of hearing aids with beamformers and gaze attention models. In this case, using computer graphics and object-based audio makes much more sense, as we can modify the scene depending on the research question.

In [50] had to recreate several realistic scenes with multiple sources involved such a street, a market and a floorball. Instead of virtualizing all these sources of each scene with computer graphics and virtual acoustics, they avoided any virtual scene by recording the scenes with a 360º camera and an Ambisonic microphone for DirAC, similar to our street scenario. Although the recorded scenes were not controlled to the detail, they answered their research questions.

Another good example is [13]. The scenes, which included characters speaking, had to be reproduced with different conditions (different tasks, acoustic stimuli and visual information) to study listening effort, speech intelligibly and dual task paradigms. In this case, computer graphics was a better choice, as the effort to record videos for each trial would be much higher than changing a parameter in the simulation software. In their experimental setup it was not needed to use any virtual acoustic scene, as they used a loudspeaker for each virtual character, placed at the corresponding physical position of the virtual character's image.

With the overview provided in this work, we hope that the reader will be able to compare and select the right techniques to fit his/her situation.

6 Future work

When doing realistic simulations, the experiments are more difficult to control, as the complexity of the stimuli is higher. For example, if we were studying loudness perception, we would have to take into account the color of the objects in the scene (red trains are perceived louder) and the culture of the participants, as explained in [51]. In addition, the ecological validity of the stimuli and experiment setup should always be taken into account if the realistic simulations are meant to reflect everyday environments.

Thus, there is still a lot of work to do in order to validate these new audiovisual technologies and environments in the field of hearing research. Currently we are conducting several experiments to be able to do ecologically valid experiments with realistic audiovisual environments: a) video recordings and different behaviors of virtual characters were compared for gaze

and head behaviors, showing that virtual characters only require any kind of lip-syncing animation to induce the same gaze and head behaviors as video stimuli [3]; b) we compared a surrounding cylindrical screen and a HMD for gaze and head behaviors and for acceptance of the technology; c) we are preparing an experiment to validate the speech intelligibility contribution of the lip-sync strategy [37] used in our computer graphics simulations, with a similar setup as [52]; d) an experiment analyzing gaze and head behaviors in realistic environments has been conducted [53] and results are being analyzed; e) and an experiment with different audiovisual complexities to estimate attention with EEG is being prepared.

ACKNOWLEDGMENTS

This work has received funding from EU's H2020 research and innovation programme under the MSCA GA 675324 (ENRICH) and from the DFG research unit FOR1732 "Hearing Acoustics". We thank Anita Wagner for the useful suggestions and comments.

REFERENCES

[1] Bentler, R.A., 2005. 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(7), pp.473-484.

[2] Cord, M.T., Surr, R.K., Walden, B.E. and Dyrlund, O., 2004. Relationship between laboratory measures of directional advantage and everyday success with directional microphone hearing aids. Journal of the American Academy of Audiology, 15(5), pp.353-364.

[3] Hendrikse, M.M., Llorach, G., Grimm, G. and Hohmann, V., 2018. Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters. Speech Communication.

[4] Bronkhorst, A.W., 2000. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acta Acustica united with Acustica, 86(1), pp.117-128.

[5] Tessendorf, B., et al., 2011. Recognition of hearing needs from body and eye movements to improve hearing instruments. Pervasive Comput. 6696 LNCS 314–331. Thiemann, J., Escher, A., van de Par, S., 2015. Multiple model high-spatial resolution HRTF measurements. In: Proceedings of the German Annual Conference on Acoustics (DAGA). Nürnberg, Germany,. pp. 797–798.

[6] Hládek, Ľ., Porr, B. and Brimijoin, W.O., 2018. Real-time estimation of horizontal gaze angle by saccade integration using in-ear electrooculography. PloS one, 13(1), p.e0190420.

[7] Kidd G Jr., Favrot S, Desloge JG, Streeter TM, Mason CR. Design and preliminary testing of a visually guided hearing aid. J Acoust Soc Am. Boston, MA, United States.: Acoustical Society of America; 2013; 133: EL202–EL207.

[8] Hart J, et al. The Attentive Hearing Aid: Eye Selection of Auditory Sources for Hearing Impaired Users. Lecture Notes in Computer Science . 2009. pp. 19–35.

[9] Vertegaal, R., Slagter, R., Van der Veer, G. and Nijholt, A., 2001, March. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 301-308). ACM.

[10] Bleichner, M.G., Mirkovic, B. and Debener, S., 2016. Identifying auditory attention with ear-EEG: cEEGrid versus high-density cap-EEG comparison. Journal of neural engineering, 13(6), p.066004.

[11] Grimm, G., Luberadzka, J. and Hohmann, V., 2018. Virtual acoustic environments for comprehensive evaluation of model-based hearing devices. International journal of audiology, 57(sup3), pp.S112-S117.

[12] Grimm, G., Kollmeier, B., & Hohmann, V. (2016). Spatial acoustic scenarios in multichannel loudspeaker systems for hearing aid evaluation. Journal of the American Academy of Audiology, 27(7), 557-566.

[13] Devesse, A., Dudek, A., van Wieringen, A. and Wouters, J., 2017. The AVATAR-approach: how real-life listening affects speech intelligibility and listening effort.

[14] Seeber, B.U. and Clapp, S.W., 2017. Interactive simulation and free-field auralization of acoustic space with the rtSOFE. The Journal of the Acoustical Society of America, 141(5), pp.3974-3974.

[15] Bolaños, J.G. and Pulkki, V., 2012, April. Immersive audiovisual environment with 3D audio playback. In Audio Engineering Society Convention 132. Audio Engineering Society.

[16] Kim, H. and Hilton, A., 2013. 3d scene reconstruction from multiple spherical stereo pairs. International journal of computer vision, 104(1), pp.94-116.

[17] Lombard, E. (1911). "The sign of the elevation of the voice," Ann. Diseases Ear, Larynx, Nose, Pharynx 37, 101–119, available at http://paul.sobriquet.net/wp-content/uploads/2007/02/lombard-1911-p-h-mason-2006.pdf

[18] Pulkki, V., Politis, A., Laitinen, M.V., Vilkamo, J. and Ahonen, J., 2017. First-order directional audio coding (DirAC). Parametric Time-Frequency Domain Spatial Audio, pp.89-138.

[19] Politis, A., Tervo, S., Lokki, T. and Pulkki, V., 2018, May. Parametric Multidirectional Decomposition of Microphone Recordings for Broadband High-Order Ambisonic Encoding. In Audio Engineering Society Convention 144. Audio Engineering Society.

[20] Cruz-Neira, C., Sandin, D.J., DeFanti, T.A., Kenyon, R.V. and Hart, J.C., 1992. The CAVE: audio visual experience automatic virtual environment. Communications of the ACM, 35(6), pp.64-72.

[21] Kohnen, M., Stienen, J., Aspöck, L. and Vorländer, M., 2016, July. Performance Evaluation of a Dynamic Crosstalk-Cancellation System with Compensation of Early Reflections. In Audio Engineering Society Conference: 2016 AES International Conference on Sound Field Control. Audio Engineering Society.

[22] Philpot, A., Glancy, M., Passmore, P.J., Wood, A. and Fields, B., 2017, June. User experience of panoramic video in CAVE-like and head mounted display Viewing Conditions. In Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video (pp. 65-75). ACM.

[23] Rumsey, F., 2011. Whose head is it anyway? Optimizing binaural audio. Journal of the Audio Engineering Society, 59(9), pp.672-675.

[24] Berkhout, A.J., de Vries, D. and Vogel, P., 1993. Acoustic control by wave field synthesis. The Journal of the Acoustical Society of America, 93(5), pp.2764-2778.

[25] Pulkki, V., 1997. Virtual sound source positioning using vector base amplitude panning. Journal of the audio engineering society, 45(6), pp.456-466.

[26] Bertet, S., Daniel, J., Parizet, E. and Warusfel, O., 2013. Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources. Acta Acustica united with Acustica, 99(4), pp.642-657.

[27] Pulkki, V., 2007. Spatial sound reproduction with directional audio coding. Journal of the Audio Engineering Society, 55(6), pp.503-516.

[28] Grimm, G., Ewert, S. and Hohmann, V., 2015. Evaluation of spatial audio reproduction schemes for application in hearing aid research. Acta Acustica United with Acustica, 101(4), pp.842-854.

[29] Pulkki, V. and Hirvonen, T., 2005. Localization of virtual sources in multichannel audio reproduction. IEEE Transactions on Speech and Audio Processing, 13(1), pp.105-119.

[30] Bertet, S., Daniel, J., Parizet, E. and Warusfel, O., 2013. Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources. Acta Acustica united with Acustica, 99(4), pp.642-657. [4] Ian Editor (Ed.). 2018. *The title of book two* (2nd. ed.). University of XXX Press, City, Chapter 100. DOI: http://dx.doi.org/10.1000/0-000-00000-0.

[31] Daniel, J., Moreau, S. and Nicol, R., 2003, March. Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging. In Audio Engineering Society Convention 114. Audio Engineering Society.

[32] Spors, S. and Ahrens, J., 2008, October. A comparison of wave field synthesis and higher-order ambisonics with respect to physical properties and spatial sampling. In Audio Engineering Society Convention 125. Audio Engineering Society.

[33] Politis, A., McCormack, L. and Pulkki, V., 2017, October. Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing. In Applications of Signal Processing to Audio and Acoustics (WASPAA), 2017 IEEE Workshop on (pp. 379-383). IEEE.

[34] Roosendaal T., 1995. Blender. Available at: https://www.blender.org/.

[35] MakeHuman_Team, 2016. MakeHuman. Available at: http://www.makehuman.org/.

[36] Autodesk, 2014. Autodesk Character Generator. Available at: https://charactergenerator.autodesk.com/.

[37] Llorach, G., Evans, A., Blat, J., Grimm, G. and Hohmann, V., 2016, September. Web-based live speech-driven lip-sync. In Games and Virtual Worlds for Serious Applications (VS-Games), 2016 8th International Conference on (pp. 1-4). IEEE.

[38] Adobe, 2018. Adobe Fuse CC (beta). Available at: https://www.adobe.com/products/fuse.html/.

[39] Adobe, 2018. Mixamo. Available at: https://www.mixamo.com/.

[40] Hendrikse, M.M.E, Grimm, G., Llorach, G. and Hohmann V., 2018. Audiovisual recordings of acted casual conversations between four speakers in German. Zenodo. Available at: http://doi.org/10.5281/zenodo.1203198/.

[41] Akkermans, V., Font Corbera, F., Funollet, J., De Jong, B., Roma Trepat, G., Togias, S. and Serra, X., 2011. Freesound 2: An improved platform for sharing audio clips. In Klapuri A, Leider C, editors. ISMIR 2011: Proceedings of the 12th International Society for Music Information Retrieval Conference; 2011 October 24-28; Miami, Florida (USA). Miami: University of Miami; 2011.. International Society for Music Information Retrieval (ISMIR).

[42] AFMG Technologies GmbH, 2011. EASE. Available at: http://ease.afmg.eu/.
[43] Naylor, G.M., 1993. ODEON—Another hybrid room acoustical model. Applied Acoustics, 38(2-4), pp.131-143.
[44] Grimm, G., Luberadzka, J., Herzke, T. and Hohmann, V., 2015. Toolbox for acoustic scene creation and rendering (TASCAR)-Render methods and research applications. In Proceedings of the Linux Audio Conference, Mainz.
[45] Allen, J.B. and Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. The Journal of the Acoustical Society of America, 65(4), pp.943-950.
[46] Herzke, T., Kayser, H., Loshaj, F., Grimm, G. and Hohmann, V., 2017. Open signal processing software platform for hearing aid research (openMHA). In Proceedings of the Linux Audio Conference (pp. 35-42).
[47] Levtov, Y., Picinali, L., D'Cruz, M. and Simeone, L., 2016, May. 3D Tune-In: The Use of 3D Sound and Gamification to Aid Better Adoption of Hearing Aid Technologies. In Audio Engineering Society Convention 140. Audio Engineering Society.
[48] Levtov, Y., Picinali, L., D'Cruz, M. and Simeone, L., 2016, May. 3D Tune-In: The Use of 3D Sound and Gamification to Aid Better Adoption of Hearing Aid Technologies. In Audio Engineering Society Convention 140. Audio Engineering Society.
[49] Daniel, J., Rault, J.B. and Polack, J.D., 1998, September. Ambisonics encoding of other audio formats for multiple listening conditions. In Audio Engineering Society Convention 105. Audio Engineering Society.
[50] Rummukainen, O., 2016. Reproducing reality: Perception and quality in immersive audiovisual environments. PhD Thesis
[51] Fastl, H. and Florentine, M., 2011. Loudness in daily environments. In Loudness (pp. 199-221). Springer, New York, NY.
[52] Schreitmüller, S., Frenken, M., Bentz, L., Ortmann, M., Walger, M. and Meister, H., 2018. Validating a Method to Assess Lipreading, Audiovisual Gain, and Integration During Speech Reception With Cochlear-Implanted and Normal-Hearing Subjects Using a Talking Head. Ear and hearing, 39(3), pp.503-516.
[53] Hendrikse, M.M.E., Llorach, G., Grimm, G. and Hohmann, V., 2018. Realistic virtual audiovisual environments for evaluating hearing aids with measures related to movement behavior. The Journal of the Acoustical Society of America, 143(3), pp.1745-1745.
[54] Unity Technologies, 2018. UNITY. Available at: https://unity3d.com/.
[55] Simon, L.S., Wuethrich, H. and Dillier, N., 2017. Comparison of Higher-Order Ambisonics, Vector-and Distance-Based Amplitude Panning using a hearing device beamformer.