AVT-ECoClass-VR: An open-source audiovisual 360° video and immersive CGI multi-talker dataset to evaluate cognitive performance

Stephan Fremerey[1*], Carolin Breuer[2], Larissa Leist[3], Maria Klatte[3], Janina Fels[2] and Alexander Raake[1]
[1] *Audiovisual Technology Group, Technische Universität Ilmenau, Germany*
[2] *Institute for Hearing Technology and Acoustics, RWTH Aachen University, Germany*
[3] *Center for Cognitive Science, RPTU Kaiserslautern-Landau, Germany*
[*] *Corresponding author: stephan.fremerey@tu-ilmenau.de*

***Abstract*—The paper is part of a project to assess how complex visual and acoustic scenes affect cognitive performance in classroom scenarios, across age groups from children to adults. Here, the potential of audiovisual virtual environments for systematic user studies is explored. As of now, most studies have examined rather simple acoustic and visual representations, which do not reflect the reality of school children. An adapted version of the audiovisual scene analysis paradigm introduced by Ahrens et al. is presented, focusing on the localization and identification of talkers within a scene. The dataset includes two audiovisual scenarios (360° video and computer-generated imagery) and two implementations for dataset playback. The paper details the recording and post-processing of the content. The 360° video part of the dataset features 200 video and single-channel audio recordings of 20 speakers reading ten stories, and 20 videos of speakers in silence, resulting in a total of 220 video and 200 audio recordings. The dataset also includes one 360° background image of a real primary school classroom scene, targeting young school children for subsequent subjective tests. All stories were recorded in the German language with native German speakers. The second part of the dataset comprises 20 different 3D models of the speakers and a computer-generated classroom scene, along with an immersive audiovisual virtual environment implementation that can be interacted with using an HTC Vive controller. Both implementations also include a Unity plugin to connect and interact with the Virtual Acoustics auralization software. As a proof of concept, the dataset includes example output data collected from ongoing perception tests. There, subjects have the task of identifying which talker in the scene is reading out which story, using the story-to-speaker mapping input system developed within this paper.**

***Index Terms*—360, 360° , degree, immersive, CGI, computer-generated imagery, 8K, dataset, scene analysis, cognitive performance, classroom, classrooms, binaural, audio, quality of experience, qoe**

I. Introduction

The project "Evaluating cognitive performance in classroom scenarios using audiovisual virtual reality (ECoClass-VR)" aims to investigate the effectiveness of audiovisual virtual environments in assessing cognitive performance in classroom-like settings, realizing complex visual and acoustic scenes in a controlled setting. Until now, most existing paradigms have focused on relatively simple acoustic and visual representations, which do not reflect the reality of school children. To enhance the validity of cognitive performance research in such environments, the realism of such paradigms shall be progressively increased in terms of cognitive task and audiovisual scene complexity. For this purpose, the suitability of the so-called "audiovisual scene analysis test paradigm" by Ahrens et al. [1] is investigated. In our paper, we introduce a modified version and implementation of this paradigm. To increase the realism of the experimental procedure and to compare two different visualization approaches, new 360° videos and computer-generated imagery (CGI) models of classrooms were created. The audiovisual recordings were done in the blue box of the virtual studio at TU Ilmenau, utilizing chroma-keying techniques in post-production. Recognizing the lack of high-quality audiovisual immersive content available for public use to evaluate cognitive performance, we decided to record the AVT-ECoClass-VR dataset. We also aimed to increase the complexity of acoustic scenes to better represent a typical classroom setup with chairs arranged in a circle. The published dataset includes two different types of content: 360° video and CGI with single-channel speech recordings, which can be arranged in the virtual scene at the chairs and hence different spatial positions. The dataset features the following contents:

- 220 different 360° video recordings at 8K resolution (20 silent and 200 representing the video components of the single-channel audio recordings, see next item)
- 200 different single-channel audio recordings (audio counterpart of the 200 video recordings, see above)
- One 360° image of a real classroom scene
- CGI-based scene of a classroom
- Rigged 3D models and 3D scans of 20 different persons
- Example output data from 5 subjects per IVE

Additionally, it offers two setups for an audiovisual immersive virtual environment (IVE) that simulates a virtual classroom scenario, allowing users to interact with both the 360° video and CGI content through an accompanying input system for participants' responses. It also includes a Python script that uses FFmpeg to create video stimuli for each participant, placing the recorded speakers in random chair positions in the 360° video sequence. To advance research in analyzing

cognitive performance using IVEs, we have made the dataset publicly available with this paper[1].

II. State of the Art

Most laboratory studies on auditory cognition research use simple visual and acoustic reproductions that lack real-life complexity, missing crucial elements like realistic visualization, as well as a binaural acoustic representation. There is a growing interest in studying auditory cognition through more immersive methods, such as virtual audiovisual environments that incorporate these elements. Virtual Reality (VR), in particular, has become increasingly popular in cognitive psychology studies over the past two decades (cf. [2], [3]), offering new ways to explore how the characteristics of such virtual audiovisual environments affect auditory perception.

A. Scene Analysis

Scene analysis involves studying the ability of humans to focus on a single sound source or visual object within a complex scene and selectively process the information that relates to that sound source or object. Auditory scene analysis research has shown that the auditory system can segregate and group signal components that belong to the attended sound source, based on auditory grouping principles such as common frequency and amplitude modulation, as well as on spatial cues [4]. The relationship between an increasing number of sources and noise levels, and the performance of the participants in localization and speech comprehension tasks has been investigated by Freyman et al. [5] and Best et al. [6]. A summary of research relating to the spatial distribution of interfering sources and their effects on speech perception can be found in [7]. The availability of spatial cues has been shown to improve auditory processing of complex scenes, specifically for speech intelligibility [8]. In complex acoustic environments, it will be interesting to know how the listener's capacity to process complex acoustic cues is influenced by using interactive audiovisual virtual environments. In this paper, two different close-to-real-life audiovisual representations of a virtual classroom scene including 20 different people are created to better investigate the cognitive performance of humans in interactive audiovisual virtual environments.

B. Audiovisual Scene Analysis

The audiovisual scene analysis paradigm was initially developed by Ahrens et al. [1] and Lund et al. [9] and it has already been used in various studies. In their study, the authors presented a paradigm where the audiovisual scenes were varied in complexity. 21 talkers were represented as schematic avatar silhouettes arranged in a semi-circle around the listener between -90° and 90° in 30° steps, which is shown in Fig. 1. During each trial, two to ten talkers read stories simultaneously, but from different positions. In a perception test, subjects were asked to assign the respective stories to the corresponding visual locations of the talkers (cf. Fig. 1), while the talker and location were drawn randomly on each trial.

Fig. 1: View of the virtual visual scene by Ahrens et al. [1].

In the tests, the participants were able to correctly match the stories of up to six simultaneous talkers. When more speakers were presented, the performance of the participants decreased, while distance perception was not found to vary with the amount of talkers. Another study by Ahrens et al. [10] had a similar task as in the previously described study by Ahrens et al. and Lund et al., but the reverberation of the room was changed. The visual details of the room either matched or diverged from the acoustic characteristics of the room. It was found that incongruent visual information of the room did not impact the capability of subjects to assign the respective stories to their corresponding visual locations. When there were only a few speakers, people could quickly and accurately pick out a speaker, even in tough sound conditions. However, when four or more speakers were involved, reverberation began to slow down response times. This was especially true when the reverberation time was high. The presence of five or more simultaneous speakers significantly impacted the ability to detect the correct talker, indicating that reverberation increased the task difficulty. A further study by Ahrens et al. [11] explored the impact of head- and eye-steered beamformers in VR on the ability to analyze audiovisual scenes with multiple talkers and varying levels of reverberation. Preliminary results indicated that beamforming technology, compared to an omnidirectional setting, significantly reduced the time required to locate the target talker, especially in environments with a higher number of concurrent speakers.

As stated in the work by Owens et al. [12], the visual and audio components of a video signal should be modelled jointly through a combined multisensory representation. Further, according to Bowman et al. [13] and Kothgassner et al. [14], IVEs should offer a high scenario fidelity, which is especially important in our case when realistic or close-to-realistic experiences are targeted. Because this was not yet fully achieved in the studies by Ahrens et al., in this paper, we introduce a modified version of the audiovisual scene analysis paradigm, focusing on creating two close-to-real-life audiovisual virtual classroom scenes.

III. Dataset

In the following, the adaptation of the audiovisual scene analysis paradigm by Ahrens et al. [1] and the generation process of the dataset AVT-ECoClass-VR will be presented. AVT-ECoClass-VR consists of two main parts, a 360° video

[1]https://github.com/Telecommunication-Telemedia-Assessment/AVT-ECoClass-VR

and a CGI-based visual representation, both with single-channel audio recordings that can be spatially rendered and presented using binaural, stereophonic loudspeaker or sound-field synthesis techniques [15].

A. Audiovisual Scene Analysis Paradigm

The first step in the content creation of this work involved translating the stories from the original paradigm by Ahrens et al. from Danish to German, which was carried out using DeepL and subsequent cross-checking by a native speaker. This translation was necessary because one of the target groups includes young school children, for whom the stories are intended. Additionally, we expanded the stories to ensure an average reading duration of 120 s and changed the name of one story from "Jimi Hendrix" to "Gitarre" (German for *guitar*). This renaming was done to more accurately match the story's content with its title, making it more suitable for future use with school children. However, based on findings from Ahrens et al. [1] that the perception of distance does not change with the number of speakers, we decided for a different arrangement. We positioned the speakers in a circle of 20 chairs, each occupied by one of the talkers, to aim for a more typical classroom-like scenario, whereby we have refrained from arranging the speakers behind each other. We have decided for 20 persons and not 21 as done by Ahrens et al. [1] to maintain an azimuth angle of 18° per speaker, aiming for a setup that allows for a clear distinction between each talker's position. Modularity is a key in the dataset, allowing for personalized scene versions in perception tests for each participant. This is essential for both 360° video and CGI IVEs, leading to different requirements for each environment. It involves varied speaker arrangements and the stories narrated by them, leading to a complex creation of AVT-ECoClass-VR to avoid repetitive speaker setups for test subjects.

B. 360° IVE

The following sections detail the recording and post-processing of the 360° IVE. Additionally, the processes for implementing the 360° video-based IVE is explained.

1) Recording: The recording process involved recruiting speakers from the student body of TU Ilmenau. The ethical committee of TU Ilmenau granted approval on March 14, 2023. Before beginning the recording, the speakers gave written consent to the publication and use of the audiovisual recordings created of them. To facilitate a more effective randomization of conditions for later perception tests, each story was narrated by 20 speakers, comprising an equal number of females and males. The audiovisual stimuli were recorded in the virtual studio part (at that time a blue box) of the media lab of TU Ilmenau, which is designed for producing content at a TV studio level. To keep the acoustic influence of the environment minimal, audio recordings of the speakers were captured using the close-talking headset microphone Shure MX153, offering a cardioid characteristic and a frequency range between 20 and 20000 Hz. They were obtained in Waveform Audio File Format (WAV) in pulse code modulation (PCM), using a sample rate of 48 kHz and one channel with a sampling depth of 24 bits. This bow microphone was chosen as a compromise, considering it would later be visible in the video recordings of the speakers. It was also selected to capture high-quality and relatively dry audio recordings, essential for feeding the audio files into the spatial audio auralization software later on. To avoid the speakers having to hold papers, which could lead to noise from paper rustling or visual reflections that might reduce the quality of chroma-keying, the text of the stories was displayed on an Autoscript TFT 12 EB-N teleprompter. The recording manager carefully operated the teleprompter, simultaneously acting as quality control instance for the text being read by the speaker. 360° videos were captured with an Insta360 Pro 2 camera at a high resolution of 7680×3840 pixels and a framerate of 29.97 fps in front of the virtual studio's blue screen, which would later allow chroma-keying of the recorded contents. The requirement for a blue screen for subsequent chroma-keying and proper scene illumination was the primary reason why the recordings were not conducted in a mainly acoustically optimized environment. While recording, speakers were seated on a primary school chair to enable an even more realistic classroom-like scenario, matching the aforementioned classroom background image. At the start of each recording session, the talkers needed to clap their hands to synchronize audio and video during the later stages of post-processing. The camera was set at a height of 120 cm, aligning with the DIN CEN ISO/TR 7250-2 standard [16], which states the average eye height when seated is between 120 and 125 cm measured from the floor's surface. Additionally, an 360° image was recorded at a school using the same 360° camera setup. Speakers were required to narrate the stories in a single take, with an average duration of 120 s, without the option to re-read sections, as doing so would compromise the consistency of the audiovisual experience. If a talker made a significant reading error or stuttered, the recording manager interrupted the current recording to maintain the quality across all audio recordings as uniformly as possible. In such cases, the speaker was asked to restart the narration of the specific story from the beginning.

2) Post-processing: The initial content stitching was performed using the Insta360 Pro Stitcher software v3.0.0 Beta for Windows. Videos were rendered using the visually lossless ProRes 422 HQ codec to minimise potential encoding artefacts during the stitching process. All speakers were aligned in the same position to facilitate a properly randomized arrangement in the virtual classroom scenario later on. Despite the Insta360 Pro 2 camera already delivering high-quality 360° content, the recorded content was somewhat noisy, which is typical for the Insta360 Pro 2 camera, leading to the need for further enhancement of the visual quality. This was achieved using Topaz Video Enhance AI v2.6.4. In informal pre-tests it was shown to be effective at noise removal, thereby improving the overall video quality essential for the subsequent chroma-keying process. The parameters used for the model included

medium video quality, progressive video type, noise as the video artefact type, and the "Artemis Medium Quality" setting. The output size was maintained without upscaling, and the output video codec remained ProRes 422 HQ to not have any further influence by transcoding the video sequences. For further video editing, including synchronization of the video with the audio, chroma-keying, and rendering video sequences, DaVinci Resolve 17 was used. The video sequences were rendered out with the GoPro CineForm HD (CFHD) codec. This codec was chosen for its visually lossless quality and support for transparency, which is a crucial aspect for the next step, where the 20 single speakers are combined with the 360° image to the classroom scene. The source resolution was kept at 7680×3840 pixels with a frame rate of 29.97 fps.

The audio post-processing was conducted as follows. An FFT filter from Adobe Audition 1.5 with an FFT size of 1892 was utilized for filtering, specifically targeting frequencies above 13-14 kHz. This resulted in a flattened audio profile to mitigate ambient noises from e.g. lamps and cameras. The filtering varied with the position of the bow microphone. Additionally, the denoiser noise reduction plugin from Adobe Audition was minimally applied. The latter was only necessary for 2-3 speakers, further audio peaks have been smoothed out using Adobe Audition. The final step involved normalization with FFmpeg 6.0, utilizing the parameters loudnorm with an integrated loudness target of -23, a loudness range target of 7 and a maximum true peak of -2. This normalization process adheres to the EBU R128 standard [17].

The rendering of the 360° image together with the 20 single chroma-keyed video streams of the talkers was performed using FFmpeg 6.0. A Python script included in the dataset was used to generate the necessary FFmpeg commands to encode individual 360° videos per subject for the perception tests and individual JavaScript Object Notation (JSON) files, representing the speaker-to-story mappings on a per-subject basis. This information will be crucial for the correct arrangement of sound sources in the 360° IVE scene later on. The videos were encoded using the High Efficiency Video Coding (HEVC) implementation of FFmpeg 6.0 (libx265) with a Constant Rate Factor (CRF) of 1 and chroma subsampling of 4:2:0, resulting in a visually almost lossless encoding while supporting hardware-accelerated decoding for smooth playback. Further, attention was given to ensuring gender balance among the speakers. For instance, in scenes with ten active speakers, an equal division was maintained, featuring 5 female and 5 male speakers. Since the reading speed and length of the stories vary, the audio was played twice to guarantee that talkers who required less than 2 minutes per story would continue speaking for the entire 2-minute duration of the test. Hence, the output length for each video was limited to 120 s. The paper also features an example still image of a generated 360° video scene in Fig.2. It features the speakers arranged in a circle of chairs, as well as the created classroom-type scenery as a background using chroma-keying.

3) Implementation of 360° IVE: The implementation of the 360° video-based IVE was carried out using Unity

Fig. 2: Equirectangular still image of one 360° video.

2019.4.17f1. A key feature of this implementation is the player, which utilizes the Unity video player component to play back the 360° videos. These final videos, rendered for each subject using FFmpeg, are provided as input to the Unity IVE. Our implementation of the 360° video scenario of the audiovisual scene analysis paradigm is detailed as follows: The Unity environment processes a playlist file, which initially contains 4 training video sequences followed by 9 test video sequences. These sequences feature 20 persons in total and between two and ten simultaneous active speakers in a randomized setting, hence people were sitting at different positions in the circle of chairs. The audio files were always combined with the correct video files that match the lip-sync of the recording. For spatial audio playback, the single-channel audio files (speaker and story combination) were used as input for VAUnity [18], a plugin designed to integrate Virtual Acoustics (VA) [19] into Unity. VA is a real-time auralization framework developed by the Institute for Hearing Technology and Acoustics (IHTA) at RWTH Aachen University.

For the auralization, a binaural free field renderer was used which accounts for the spatial position of the sound sources and the receiver. A generic head-related transfer function (HRTF) of the IHTA artificial head [20] with an angular resolution of 5°×5° is provided in the dataset. The receiver position and orientation were updated according to the tracking data of the HMD to account for head movements. This ensures that the sound sources are perceived as coming from the same positions as the visualized speakers.

To correctly assign audio sources to the speakers, JSON files containing speaker-to-story mappings generated from the Python script, as previously described in section III-B2, are loaded for each subject. This requires the generation of the appropriate audio signals, which involves using audio WAV files as input and connecting the sound sources with the correct audio files. Sound sources are then assigned to the corresponding visual positions of the speakers. All these preparations are carried out before the runtime of the Unity scene and VA, and are automatically done by a script located in the "Editor" folder within Unity.

The virtual environment introduces a novel story-to-speaker mapping input system and graphical user interface, operated with an HTC Vive controller, as depicted in Fig. 3. To match a story symbol to the corresponding speaker, subjects first select the symbol using the controller's touch trackpad and then aim

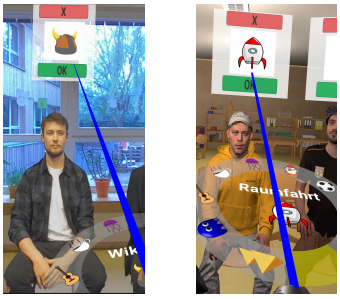

Fig. 3: Story-to-speaker mapping system in the 360° IVE (left) and the CGI IVE (right).

the controller and visible blue interaction ray at the "OK" button positioned above the talker associated with that story and click the controller's trigger button. This results in the selected story's symbol appearing within the designated white box located above each corresponding speaker. If necessary, selections can be undone by clicking the red "X" button in the same way. Two conditions lead to a scene exit: The first possibility is that participants of later perception tests are instructed to press the controller's menu button once they think they have correctly assigned a story to each active speaker. Upon the participant pressing the menu button, a prompt appears, requesting the user to remove the HMD. The second possibility to exit the scene is a timeout after 2 minutes, following the task of assigning all stories to their respective speakers. Then, all selected stories are saved, and a prompt appears, again requesting the user to remove the HMD. Speaker-to-story mappings are stored in a JSON file. Additionally, viewing-related data, such as time, the positions and Euler angles of the HMD and controller can be simultaneously recorded.

C. CGI IVE

The following sections detail the recording and post-processing of the CGI IVE content. Additionally, the processes for implementing the CGI-based IVE is explained.

1) Recording: After the 360° video recordings of the speakers, the same persons were scanned using an Artec Leo 3D scanner. During the scanning process, speakers were asked to stand in a T-pose and minimize movement as much as possible. The handheld 3D scanner is highly sensitive to movements, though minor movements are typically not problematic and can be handled during later post-processing. Some speakers were required to cover their heads because the scanner struggled with scanning certain types of hair. This was visible from the scanner's preview screen, displaying a pre-rendered version of the model and highlighting any areas that were not scanned or scanned with low quality. Using headwear made it possible to complete the 3D scans. The scanning process began with the body parts most likely to move, starting with the face and the back of the head, and then proceeded in a spiral-shaped pattern downwards: first scanning the shoulders, then the arms and fingers, followed by the torso and back, and finally the legs and shoes. Among all body parts, the fingers were the most challenging part of the body to scan due to their potential for movement. Based on the visual quality of the pre-rendered view of the 3D scan, the 3D scanning process was repeated from the beginning as often as necessary until the scan of each person reached an acceptable quality level. The assessment of quality was conducted on-site by the recording manager.

2) Post-processing: Initially, post-processing was conducted in Artec Studio 16, a proprietary software included with the 3D scanner. Upon importing the raw scan data from the Artec Leo 3D scanner, the registration step was executed using "Global registration" with the "Features" parameter set to "Geometry", keeping further parameters at their default settings. The global registration algorithm merges surfaces from individual frames into a unified coordinate system, using data on the relative positions between pairs of surfaces. The next step was "Fusion", where we opted for "Smooth fusion" with default parameters, as this method is well-suited for scanning humans as it compensates for slight movements. For some models, "Sharp fusion" with default parameters yielded better results, so this method was used instead. Subsequently, unnecessary parts of the 3D scan, such as scanning errors and the floor, were removed using the appropriate selection tool. At times, the scanning conditions might have not allowed for a complete capture of all body parts, resulting in the fused 3D model having holes. These were addressed in the following step using the "Automatic Hole Filling" functionality. The mesh produced after fusion and hole filling was too complex for later use in the Unity IVE, due to a high polygon count that increased the model's complexity and memory usage. Hence, "Fast mesh simplification," targeting 600,000 polygons was applied. In the final step in Artec Studio, the texture was applied to the fused and simplified model and exported in the Wavefront OBJ data format including textures. The subsequent step involved importing the avatar into Blender 3.6 for some arrangement adjustments concerning the position of the avatars. This step is essential for the subsequent phase of rigging, which is carried out using Adobe Mixamo. Distinctive body parts like the chin, wrists, elbows, groin and knees are marked, then Mixamo provides the rigged avatar as output. This does not include rigged facial bones or blend shapes. Further, an avatar in a seated position for later inclusion in the classroom scenario was exported from Mixamo. As part of the dataset published in this paper, rigged 3D models of all 20 speakers are made available in FilmBox (FBX) data format, while non-rigged 3D scans are published in the Wavefront OBJ format.

3) Implementation of CGI IVE: The classroom was designed using SketchUp Make 2017, as this version was the only one available for free at the time. Additionally, it was selected to facilitate the future integration of the classroom scene into the Room Acoustics for Virtual ENvironments (RAVEN) room acoustic simulation environment software [21] developed by the IHTA at RWTH Aachen University. This integration ultimately enables real-time visualization of parameters and auralization based on the materials defined in the SketchUp scene. The SketchUp scene is also made available in AVT-

ECoClass-VR in both the SketchUp file format (SKP) and the COLLADA Digital Asset Exchange (DAE) format, which can be imported into Unity.

The implementation of the CGI IVE was carried out using Unity 2019.4.17f1 as well. After importing the classroom scene in DAE file format, the rigged and seated speakers in the FBX data format were imported. Utilizing the specific JSON file for speaker-to-story mapping, the speakers were again automatically visually arranged in a circle at the chair positions using Unity and acoustically using VAUnity for auralization. The principle of automatically assigning audio sources and sound sources follows the same methodology as already described in section III-B3. The CGI environment also employs the same input system for story-to-speaker mapping as already outlined in section III-B3, which is illustrated in Fig. 3. Similar to the 360° immersive virtual environment (IVE), it is also possible to simultaneously record viewing-related data, including time, the positions and Euler angles of the HMD and controller.

In Fig. 4, a viewport of the final CGI scene is shown. It features the rigged and seated 3D models of 20 different

Fig. 4: Viewport of the CGI scene.

speakers, each positioned on a chair, as well as the created classroom-type CGI scenery. While there is the possibility to navigate within the CGI scene using the touch trackpad of a second controller, this feature is disabled by default to maintain consistency with the static position of the 360° scenario. Further, we chose not to consider lip-syncing at this stage. Informal pre-tests indicated that it is hard to find a publicly available lip-sync solution that provides accurate and acceptable results for this case. In turn, for cognitive performance tests, the absence of lip-sync is a relevant technical feature distinguishing the CGI from the 360° case.

IV. Evaluation

The AVT-ECoClass-VR dataset is initially evaluated in terms of time needed to create it and its suitability to investigate cognitive performance. Creating the dataset required a significant amount of time. Each recording session, including the 3D scan, took about 3 hours, accumulating to a total of 60 hours for all sessions. The process of stitching and rendering demanded approximately 2 hours per recorded video, resulting in 440 hours. Video enhancement with Topaz Video Enhance AI took around 5 hours per video, leading to a substantial 1100 hours of rendering time. Additionally, 2 hours were allocated for post-processing tasks such as chroma-keying, cutting, and rendering video sequences, leading to another 440 hours of time spent. The FFmpeg command used to generate the final individual video sequences for each subject required at least 270 GB of RAM and about 3 hours processing per command. For 36 subjects and 13 videos per subject, the processing time was approximately 1400 hours. Furthermore, post-processing and rigging of the 3D models took about 3 hours per 3D scan, leading to 60 hours spent.

A. Initial Evaluation Results

A brief evaluation of the story-to-speaker mapping input system, the primary interactive component in the IVE scenes, is presented as follows. First subjective tests indicate that both implementations of the adapted version of the audiovisual scene analysis paradigm can successfully be used to assess cognitive performance in classroom-type settings, especially the novel story-to-speaker mapping input system. The story-to-speaker mappings are correctly saved in JSON file format. Viewing-related data recorded simultaneously, such as time, positions, and Euler angles of the HMD and controller, are available for further analysis in Comma-separated values (CSV) file format. Although tests with participants younger than 18 years (a key target group of the ECoClass-VR project) were rather informal yet, they have shown that this age group is capable of interacting with the IVE and the input system effectively.

V. Conclusion and Outlook

This paper introduces the AVT-ECoClass-VR dataset, created to assess cognitive performance in classroom-type audiovisual multi-talker settings. To achieve this, two distinct immersive virtual environments (IVEs) were developed: one based on 360° video and the other on CGI. The dataset includes a variety of content components, such as different video and audio recordings, rigged 3D models and 3D scans of all 20 speakers, and the final Unity scenes, along with example output data from 5 subjects for each IVE. The findings from this paper indicate that AVT-ECoClass-VR is suitable to evaluate cognitive performance in classroom-like settings. In future research, subjective tests will be conducted using both IVEs, for instance, to compare the cognitive performance of subjects in each virtual environment and explore how factors related to the audiovisual IVE such as lip-sync, 360° video resolution or reverberation affect performance. Additionally, future tests will involve young school children as participants.

Acknowledgment

This work is part of two projects funded by the German Research Foundation (DFG): ECoClass-VR, number 444697733 and ILMETA, number 438822823. Further, the authors would like to thank the speakers, the test participants, the “Freie Reformschule Franz von Assisi Ilmenau” for allowing the recordings to take place, and Ana Garcia Romero for her contributions as a student assistant.

References

[1] A. Ahrens et al. *Audio-visual scene analysis in reverberant multi-talker environments*. Universitätsbibliothek der RWTH Aachen, 2019.

[2] N. Foreman. "Virtual reality in psychology". In: *Themes in Science and Technology Education* 2.1-2 (2010), pp. 225–252.

[3] S. Schnall et al. "The Immersive Virtual Environment of the digital fulldome: Considerations of relevant psychological processes". In: *International Journal of Human-Computer Studies* 70.8 (2012), pp. 561–575.

[4] A. S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[5] R. L. Freyman et al. "Effect of number of masking talkers and auditory priming on informational masking in speech recognition". In: *The Journal of the Acoustical Society of America* 115.5 (2004), pp. 2246–2256.

[6] V. Best et al. "The influence of spatial separation on divided listening". In: *The Journal of the Acoustical Society of America* 120.3 (2006), pp. 1506–1516.

[7] A. W. Bronkhorst. "The cocktail-party problem revisited: early processing and selection of multi-talker speech". In: *Attention, Perception, & Psychophysics* 77.5 (2015), pp. 1465–1487.

[8] A. W. Bronkhorst. "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions". In: *Acta Acustica united with Acustica* 86.1 (2000), pp. 117–128.

[9] K. D. Lund et al. "A method for evaluating audio-visual scene analysis in multi-talker environments". In: *International Symposium on Auditory and Audiological Research: Auditory Learning in Biological and Artificial Systems*. The Danavox Jubilee Foundation. 2020, pp. 357–364.

[10] A. Ahrens et al. "Auditory spatial analysis in reverberant multi-talker environments with congruent and incongruent audio-visual room information". In: *The Journal of the Acoustical Society of America* 152.3 (2022), pp. 1586–1594.

[11] A. Ahrens et al. "Audio-visual scene analysis in conditions with head-and eye-steered beamformers in virtual reality". In: *The Journal of the Acoustical Society of America* 153.3_supplement (2023), A48–A48.

[12] A. Owens et al. "Audio-visual scene analysis with self-supervised multisensory features". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 631–648.

[13] D. A. Bowman et al. "Virtual reality: how much immersion is enough?" In: *Computer* 40.7 (2007), pp. 36–43.

[14] O. D. Kothgassner et al. "A virtual training tool for giving talks". In: *Entertainment Computing-ICEC 2012: 11th International Conference, ICEC 2012, Bremen, Germany, September 26-29, 2012. Proceedings 11*. Springer. 2012, pp. 53–66.

[15] S. Spors et al. "Spatial sound with loudspeakers and its perception: A review of the current state". In: *Proceedings of the IEEE* 101.9 (2013), pp. 1920–1938.

[16] D. ISO. *TR 7250-2; DIN SPEC 91279: 2011-07: Wesentliche Maße des menschlichen Körpers für die technische Gestaltung–Teil 2: Anthropometrische Datenbanken einzelner Bevölkerungen von ISO-Mitgliedsländern (ISO/TR 7250-2: 2010); Deutsche Fassung CEN ISO*. Tech. rep. TR 7250-2: 2011, 2011.

[17] R. EBU. "Loudness normalisation and permitted maximum level of audio signals". In: *EBU Recommendation, Geneva* (2023).

[18] Institute for Hearing Technology and Acoustics, RWTH Aachen University. *Virtual Acoustics Unity Package*. Available online: (accessed 27 July 2023). 2023. eprint: https://git.rwth-aachen.de/ita/vaunity_package.

[19] Institute for Hearing Technology and Acoustics, RWTH Aachen University. *Virtual Acoustics - A real-time auralization framework for scientific research*. Available online: (accessed 27 July 2023). 2023. eprint: http://virtualacoustics.de/VA/.

[20] A. Schmitz. "Ein neues digitales Kunstkopfmeßsystem". In: *Acta acustica united with acustica* 81.4 (1995), pp. 416–420.

[21] D. Schröder. *Physically based real-time auralization of interactive virtual environments*. Vol. 11. Logos Verlag Berlin GmbH, 2011.