# Generation of audiovisual immersive virtual environments to evaluate cognitive performance in classroom type scenarios

Stephan Fremerey[1], Carolin Reimers[2], Larissa Leist[3], Jan Spilski[3], Maria Klatte[3], Janina Fels[2] and Alexander Raake[1]

[1] *Audiovisual Technology Group, Technische Universität Ilmenau, 98693 Ilmenau, Germany*

[2] *Institute for Hearing Technology and Acoustics, RWTH Aachen University, 52062 Aachen, Germany*

[3] *Cognitive and Developmental Psychology, Technische Universität Kaiserslautern, 67663 Kaiserslautern, Germany*

*Corresponding author: stephan.fremerey@tu-ilmenau.de*

## Abstract

In the project ECoClass-VR, which is part of the AUDICTIVE priority programme, we investigate the suitability of audiovisual Immersive Virtual Environments (IVEs) for a "real-world" assessment of cognitive performance of adults and children in classroom-type environments under different visuospatial and acoustic conditions. Existing knowledge in this area comes predominantly from auditory experimental paradigms with typically simple acoustic replications. So far, only limited attention has been paid to visual processing, without considering relevant audiovisual aspects. In the project ECoClass-VR, it is planned to successively increase the realism of three existing psychological test paradigms that are first translated into audiovisual tasks and subsequently increased in terms of their realism with regard to cognitive tasks and IVEs.

To do so, we use two different types of scene visualization: immersive 360° video, hence captured 360° video scenes and Computer Generated Imagery (CGI) based scenes generated by a software such as e.g. Unity. Within this paper, we present first results and explain how we captured and generated the corresponding IVEs. We describe the challenges as well as the technical solutions to achieve close-to-photorealism for two virtual representations of a classroom-type scene, and to enable the flexibility required for conducting the targeted cognitive performance tests.

## Introduction

The focus of the project called "Evaluating cognitive performance in classroom scenarios using audiovisual virtual reality (ECoClass-VR)" is to investigate the suitability of audiovisual IVEs to assess possible impacts on cognitive performance of adults and children in classroom-type settings.
To achieve this, the realism of the experimental procedures used in the project will be increased in terms of the cognitive tasks used and the settings of the audiovisual representations. Most of the current experimental paradigms on auditory cognition lack in appropriate audiovisual considerations. Especially the visual representations of such paradigms often are relatively poor.

In Figure 1, we illustrate the approach of the ECoClass-VR project, where we want to increase the perceived realism step by step.

**Figure 1:** Approach of the ECoClass-VR project

In this paper, we introduce two different ways of creating scene visualization: 360-degree video and CGI. We present first insights on how we captured the corresponding content and how we generated the corresponding IVEs. After giving a short overview on the related work, the creation process of both types of scene visualization is presented. Subsequently, we present the results of the generation of audiovisual contents, conclude the paper and give an outlook on the next steps.

## Related work

Previous studies such as e.g. [4, 5] used auditory experimental paradigms with typically simple acoustic representations. In other studies, the auditory representation was adapted to be more realistic, e.g. using binaural / spatial sound reproductions [3, 7, 6, 8], however the visual part only played a subordinate role.
Within ECoClass-VR, three auditory experimental paradigms are considered: The Auditory Selective Attention (ASA) paradigm by Koch et al. [4], the Listening Comprehension (LC) paradigm by Klatte et al. [3] and the Scene Analysis (SA) paradigm by Ahrens et al. [1].
Within this paper, we focus on the modification of the SA paradigm initially proposed by Ahrens et al. in [1], while we call our modified paradigm Audio-Visual Scene Analysis (AV-SA). The SA paradigm was specifically designed for auditory cognition-type research in IVEs. In a multi-talker scene with 21 schematic avatar silhouettes up to ten of which are active talkers, participants are asked to match presented tales to little icons for all active talkers. Performance indicators are the number of correct associations and the task completion time.

## Creation of immersive virtual environments

As the overall and future goal of ECoClass-VR is to evaluate the cognitive performance of children in classroom scenarios, the adaptation of the AV-SA task needs to carefully consider the applicability for children. Further, in the present version of the SA paradigm, the tales are prepared in Danish language. In ECoClass-VR, we will prepare and record German tales in the context of classroom-type scenarios, with respect to arrangement and balancing of words indicating the belonging item.
As follows we will explain how we will generate the respective IVEs for the immersive 360° captured video and for the CGI based scenes. For both virtual representations, we aim for close-to-photorealism. Further, our goal is to enable flexibility required to conduct the targeted cognitive performance tests, hence with small modifications the scenes can be used for all three experimental paradigms. Furthermore, we plan to use Unity [1] to generate the corresponding IVEs for both scenes.

### 360° video

In Figure 2, our approach to capture the immersive 360° captured video based scenes is shown.

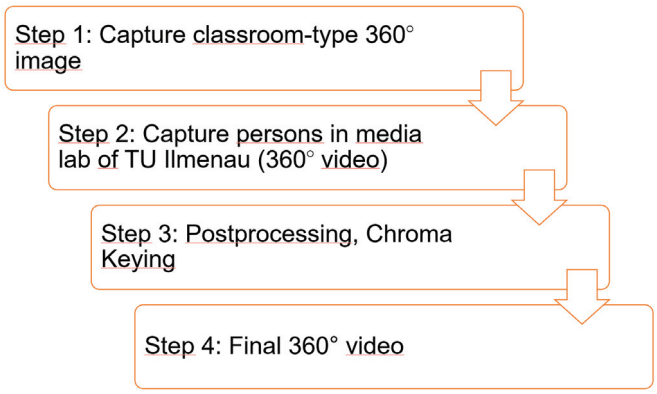


**Figure 2:** Approach to generate the immersive 360° captured video based scenes

The first step is to capture a classroom-type 360° image. This was done in the "Franz von Assisi" school in Ilmenau, Germany in a primary school classroom. We captured the 360° image shown in Figure 3 with an Insta360 Pro 2 camera[2] at a resolution of 7680x3840 pixels using the uncompressed .dng image format.

The second step is to individually capture all 21 persons, 10 of them being active speakers, sitting on a primary school chair in our media lab with an appropriate lighting setting as shown in Figure 4.

To do so, we will use the same 360° camera we used to capture the 360° image of the classroom. Each speaker will be captured individually in the blue box of our media lab with the Insta360 Pro 2 360° camera. After stitching, resulting videos will have a framerate of 60 fps and are encoded with the ProRes 422 HQ video codec at a resolution of 7680x3840 pixels.

**Figure 3:** 360° captured image of a primary school classroom for the AV-SA paradigm

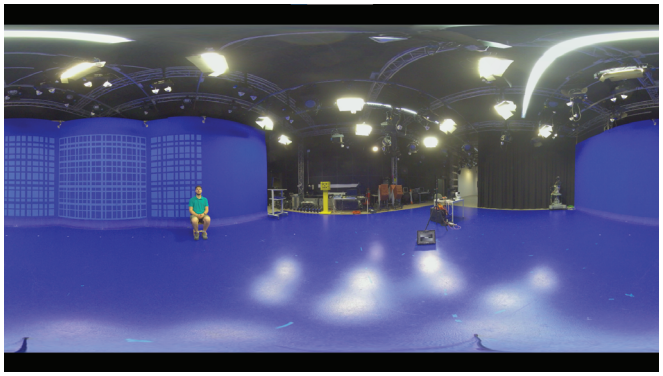

**Figure 4:** 360° captured image in the media lab of TU Ilmenau

When doing the final 360° recordings, we will equip each person with a bow microphone, enabling a dry audio recording needed for later simulation of the room acoustics. Corresponding to the study presented by Ahrens et al. [1], each of the German speakers will read out a story of up to 120 seconds duration, while during the later subjective test only 10 persons of them will be active speakers at once. Each of the 21 persons will be then inserted into the 360° captured image of the classroom. To do so, the chroma keying technology will be used as it was e.g. done in a study done by Sermon et al. [9]. Further, the active speakers will be randomized.

## CGI

The second approach we will use in the ECoClass-VR project is to generate the corresponding IVE by using CGI technology.
The first step is to generate the corresponding 3D model of the real classroom, aiming for close-to-photorealism. To do so, we used SketchUp Make 2017 [3] because in later subjective tests, we will use RAVEN, an audio plugin for SketchUp developed by the Institute for Hearing Technology and Acoustics (IHTA), RWTH Aachen University. The mentioned software is used for room acoustic simulations [4].

[1] https://unity3d.com
[2] https://www.insta360.com/de/product/insta360-pro2
[3] https://help.sketchup.com/de/downloading-older-versions
[4] https://www.akustik.rwth-aachen.de/go/id/dwoc/file/94244

In Figure 5, the current state of the CGI-based scene is shown.

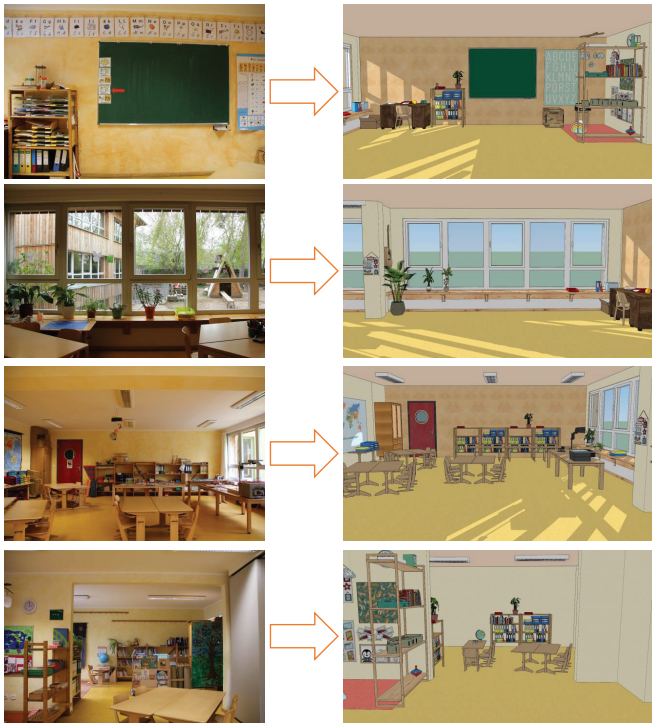

**Figure 5:** CGI generated classroom scene in SketchUp Make 2017

While creating the CGI-based scene, we payed attention that the CGI model reflects the real classroom as good as possible. A few of the elements were taken from the SketchUp 3D Warehouse store, however we designed a few elements on our own, examples are e.g. the school chair, the overhead projector or the doors of the classroom.

We investigated photogrammetry and 3D scanning as an alternative to CGI based scene generation. Further, we want to use 3D scanning to obtain individual 3D models of all 21 persons from the AV-SA scenario. At first, we have tried out Meshroom [5], which is a free 3D reconstruction software. To retrieve a 3D model, e.g. of a chair, at first one needs to capture multiple images of an object using a standard camera, while it is also possible to use a recent smartphone to do so.
We used the Pixel 4 XL smartphone to capture images from multiple angles. After post-processing the images in Meshroom we obtained high quality scans of textured objects as e.g. a rock. Further, we found the whole 3D reconstruction process of Meshroom intuitive and straightforward to use.
However, we did not manage to create a 3D model of a complete room with Meshroom, as the resulting 3D model was not satisfying. Further, we were not able to create a 3D model of a human by using Meshroom, this is why we have went for a different approach which will be described in the next chapter.

[5] https://github.com/alicevision/meshroom

## Results

To create an artificial immersive 360° video scene with chroma keying has a few advantages. At first, we don't need to capture the full 360° video scene in the school room itself, which would be a big challenge. Further, it is not really feasible to capture a 360° video with 21 persons, while 10 of them being active speakers, in a way that each audio recording does not contain any audio from another speaker and no room acoustics. Additionally, with our modular approach we can randomize the active talkers. The chosen approach also allows to apply it for other paradigms from the ECoClass-VR project like AV-ASA and AV-LC. Last but not least, with the described 360° video approach we will achieve the highest possible amount of close-to-photorealism.
Besides its advantages, the chosen approach also has its disadvantages, e.g. the colour temperature of the lamps in the studio is warmer compared to the relatively cold light temperature in the classroom. Hence, in post-processing we also need to do some color grading and adaptation. Further, at least for the AV-SA paradigm, the positioning of 21 persons in a circle of chairs is a challenging task, but possible. Compared to the CGI-based approach, the immersive 360° video approach offers less flexibility in terms of changing the scene after recording.

With respect to the CGI-based approach, we currently face some difficulties in importing the 3D model from SketchUp into Unity. Apart from that we found SketchUp to be a suitable and intuitive solution to create a corresponding 3D model of the classroom. Especially because we need a SketchUp model as input for the subjective tests, where we will use the RAVEN plugin for SketchUp developed by the (IHTA) at RWTH Aachen University to do the appropriate rendering of the room acoustics.
As already described in the previous chapter, we didn't find any appropriate solution to scan a whole room and create a corresponding CGI model from it.
We plan to generate CGI models from the persons we will capture in the media lab. To do so, we will use the Artec Leo 3D scanner [6] in combination with the Artec Studio software for post-processing and generation of the respective 3D models. Hence, each person will at first be captured in front of the blue box of our media lab and afterwards the same person will be scanned using the mentioned 3D scanner. The 3D models will be post-processed using Artec Studio, exported as .obj file and imported into Unity. The animation and rigging process will be very likely done using Blender [7]. For lip sync we are currently investigating various solutions as e.g. VOCA (Voice Operated Character Animation) [8] published in a study by Cudeiro et al. [2] or Blender Rhubarb Lip Sync developed by Daniel S. Wolf [9].

[6] https://www.artec3d.com/de/portable-3d-scanners/artec-leo-v2
[7] https://www.blender.org
[8] https://github.com/TimoBolkart/voca
[9] https://github.com/scaredyfish/blender-rhubarb-lipsync

## Conclusion and outlook

To conclude, the adaptation of the SA paradigm initially proposed by Ahrens et al. [1] using two visual representations, immersive 360° video and CGI, is possible. Further, we found out that despite of the challenges regarding lighting and chroma resolution of the Insta 360 Pro 2 camera, chroma keying also works for 360° video.
Additionally, SketchUp Make 2017 is suitable for CGI-modelling, however we found that the import into Unity can be a challenging task. To create a corresponding 3D model from the persons which we will record in our media lab, we will use the Artec Leo 3D scanner.

## Acknowledgments

This work was funded by the German Research Foundation under the project ID 444697733 with the title "Evaluating cognitive performance in classroom scenarios using audiovisual virtual reality (ECoClass-VR)".
Further, this work has partially been supported by the CYTEMEX project funded by the Free State of Thuringia, Germany (FKZ: 2018-FGI-0019).

Funded by

DFG Deutsche Forschungsgemeinschaft
German Research Foundation

EFRE EUROPE FOR THURINGIA
EUROPEAN REGIONAL DEVELOPMENT FUND

EUROPEAN UNION

We also want to thank Ana Garcia Romero for her work as a student assistant in the course of this project.
Last but not least we want to thank the "AG Wissenschaftliches Rechnen", especially Henning Schwanbeck, and the technical employees from the Institute of Media Technology of the Technische Universität Ilmenau for their support.

## References

[1] A. Ahrens, K. D. Lund, and T. Dau. *Audio-visual scene analysis in reverberant multi-talker environments.* Universitätsbibliothek der RWTH Aachen, 2019.

[2] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black. "Capture, learning, and synthesis of 3D speaking styles". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10101–10111.

[3] M. Klatte, T. Lachmann, M. Meis, et al. "Effects of noise and reverberation on speech perception and listening comprehension of children and adults in a classroom-like setting". In: *Noise and Health* 12.49 (2010), p. 270.

[4] I. Koch, V. Lawo, J. Fels, and M. Vorländer. "Switching in the cocktail party: Exploring intentional control of auditory selective attention." In: *Journal of Experimental Psychology: Human Perception and Performance* 37.4 (2011), p. 1140.

[5] V. Lawo, J. Fels, J. Oberem, and I. Koch. "Intentional attention switching in dichotic listening: Exploring the efficiency of nonspatial and spatial selection". In: *Quarterly Journal of Experimental Psychology* 67.10 (2014), pp. 2010–2024.

[6] J. Oberem, I. Koch, and J. Fels. "Intentional switching in auditory selective attention: Exploring age-related effects in a spatial setup requiring speech perception". In: *Acta psychologica* 177 (2017), pp. 36–43.

[7] J. Oberem, V. Lawo, I. Koch, and J. Fels. "Intentional switching in auditory selective attention: Exploring different binaural reproduction methods in an anechoic chamber". In: *Acta acustica united with acustica* 100.6 (2014), pp. 1139–1148.

[8] J. Oberem, J. Seibold, I. Koch, and J. Fels. "Intentional switching in auditory selective attention: Exploring attention shifts with different reverberation times". In: *Hearing research* 359 (2018), pp. 32–39.

[9] P. Sermon, C. Gould, and J. Ambrose. "Out of Sight, Out of Mind". In: (2019).