A Quality of Experience and Visual Attention Evaluation for 360° videos with non-spatial and spatial audio

Amit, Mr., Hirway

Department of Computer and Software Engineering, Technological University of the Shannon – Midlands Midwest, Athlone, Ireland, a.hirway@research.ait.ie

Yuansong, Dr., Qiao

Software Research Institute, Technological University of the Shannon – Midlands Midwest, Athlone, Ireland, ysqiao@research.ait.ie

Niall, Dr., Murray

Department of Computer and Software Engineering, Technological University of the Shannon – Midlands Midwest, Athlone, Ireland, nmurray@research.ait.ie

This research presents the results of an empirical study that aimed to investigate the influence of various types of audio (spatial and non-spatial) on the quality of experience (QoE) of and visual attention in 360° videos. The study compared the eye gaze, head pose, pupil dilations, and heart rate of 73 users who watched ten 360° videos with different sound configurations. The configurations evaluated were: no sound; non-spatial (stereo) audio; and two spatial sound conditions (first and third-order ambisonics). The videos covered various categories and presented both indoor and outdoor scenarios. An analysis of the collected data was performed using various data visualization and statistical techniques. The results of the study show that users have different viewing patterns and physiological responses when watching 360° videos with different sound conditions. Specifically, users paid more attention to areas of the videos with spatial audio and had higher pupil dilation when watching videos with third-order ambisonics. The study also found that users' gaze fixations were more evenly distributed when watching videos with spatial audio. These findings have important implications for the development of effective techniques for optimizing processing, encoding, distributing, and rendering content in VR and 360° videos with spatialized audio. The study also suggests that the use of spatial audio improves visual attention and enhances over all user experience and engagement in immersive experiences.

CCS CONCEPTS • Information Systems -> Multimedia Streaming

Additional Keywords and Phrases: 360° videos, Ambisonics, Visual attention, Quality of Experience

1. INTRODUCTION

In recent years, Virtual Reality (VR) applications have become increasingly popular, with 360° videos being a particularly appreciated form of content. These videos are often viewed through Head Mounted Display (HMD) technology, which allows users to experience a more immersive and interactive viewing experience. However, streaming these videos to HMDs is very challenging due to limitations in bandwidth, storage, and computation. Additionally, there are other factors that need to be considered such as low motion-to-photon delay expectations, complex view adaptation, complex rules and metadata for viewing 360° content, and limited understanding of 360° video Quality of Experience

(QoE) [1]. To address these challenges and optimize 360° video streaming applications, significant research has been undertaken to understand visual attention (VA) [2]. Visual attention refers to the selective process by which our brains focus on specific elements of information from our environment for further processing and analysis. This biological mechanism allows us to understand our surroundings efficiently and quickly by extracting relevant information from the available data. Researchers have developed various techniques for visual attention modelling, which are based on either bottom-up or top-down architectural designs [3]. These techniques consider saliency feature vectors, object-based attention, and salient maps to identify and select important elements for further analysis. Thus, visual attention plays a critical role in determining which parts of the video content are most important to users and where their focus lies. By understanding visual attention, it is possible to optimize video content and delivery to enhance the user experience [4].

Audio plays a critical role in creating an immersive experience for users. It is known to contribute significantly to the feeling of presence in a virtual environment [43]. Despite this, there has been limited research considering the impact of audio (and in particular spatial audio) on visual attention in immersive media [5]. Currently, there are numerous public datasets available that capture user behaviors, including head and eye tracking data, while they watch 360° videos, the majoring of which are without spatial audio [6]. Previous research on "Audio-Visual Attention" has mainly focused on analyzing attention in conventional non-spatial sound videos, rather than in more immersive multimedia [7]. Recently, spatial audio has been gaining more attention in industry and academia due to its advanced features. Its influence on VR experiences should not be overlooked [8]. By adding spatial audio to the environment, users' head movements, the direction of their focus, and the content they remember after each session may all be influenced [9]. As a result, there is a need to further investigate the role of spatial audio in visual attention in immersive media to enhance the overall VR quality of experience [10].

Ambisonics is a technique used for capturing and reproducing 360° audio. It has gained popularity in recent years due to its ability to provide a highly immersive audio experience in VR [11]. The technique involves recording sound from all directions and encoding it into a multi-channel format, which can then be played back through a surround sound system or headphones. The most common format for Ambisonics is first-order B-format, which uses four channels - W, X, Y, and Z - to capture the full sound sphere [12]. However, higher-order Ambisonics formats are now available, providing even greater spatial resolution and immersion. Second-order Ambisonics uses up to nine channels, while third-order Ambisonics uses up to 16 channels. The highest resolution available is sixth-order Ambisonics, which uses up to 49 channels. These higher-order formats enable more precise localization of sound sources and can enhance the sense of presence and realism in VR experiences [13].

With the widespread adoption of VR and 360° videos, researchers have become increasingly interested in understanding the Quality of Experience (QoE) in multimedia applications. QoE refers to the overall subjective experience of a user while interacting with a multimedia application, and it is influenced by several factors such as content quality [14], network conditions [15], device performance [16], and user expectations [17]. In the case of VR and 360° videos, QoE plays a crucial role in determining the level of immersion and presence experienced by the user [18]. It is important to understand the QoE because it can provide insights into the user's preferences [19], behavior [20], and perception of the multimedia content [21], which can be used to design more effective and engaging applications that provide a more enjoyable user experience.

This paper extends the authors' existing research on evaluating the influence of spatial audio on visual attention and Quality of Experience (QOE) in 360° videos. In a previous study [22], the authors explored how non-spatial (stereo) and spatial (third-order ambisonics) audio can impact visual attention in 360° videos. The study found that when spatial sound was used in videos, participants tended to pay attention to a wider range of visual elements compared to non-spatial sound.

This was observed for both indoor and outdoor categories. Furthermore, participants had a larger maximum pupil size when third-order sound was used. Building on this work, a new empirical study using the same videos and introducing two additional sound conditions (no sound and first order ambisonics sound) was undertaken. In terms of analysis, head pose and eye gaze fixations, pupil diameter, and heart rate are captured and analyzed. 53 additional participants were recruited (bringing the total to 73). The video categories include 360° videos captured in indoor and outdoor locations with four different sound conditions: no-sound, first-order, higher-order, and stereo sound. The study concludes with a discussion of the analysis, which provides insights into the impact of spatial audio on visual attention and QoE in 360° videos.

The rest of the work is organized as follows: section 2 presents a brief description of the prevailing visual attention models, section 3 elaborates the experimental setup, section 4 specifies the research methodology, and section 5 gives a detailed description of the dataset. Section 6 presents the data visualization and analysis of the objective (eye gaze, head pose, pupil diameter) and physiological data (heart rate) captured in the four different sound conditions and two environments. Section 7 presents the statistical analysis of the subjective questionnaire. Finally, the paper is concluded in section 8.

2. LITERATURE REVIEW

Visual attention in 360-degree videos has been the focus of several studies over the past decade. Lo et al. [23] investigated the visual attention of viewers in 360-degree videos and produced a dataset with sensor data and content data. The dataset contained ten videos from YouTube classified into three categories (Computer Generated, fast-paced; Natural Image, fast-paced, and Natural Image, slow-paced), and OpenTrack, an open-source head tracking tool, was used to capture viewer orientations from the HMD sensors. David et al. [24] presented a public dataset of head and eye movements derived from a free-view trial with participants wearing a VR headset with an embedded eye tracker. The dataset contained 360-degree videos played without audio, and it also had saliency maps and scan paths of users in addition to the videos.

The influence of sound on visual attention in 360-degree videos was explored by Min et al. in [25]. They conducted eye-tracking tests with 60 videos in various test conditions: videos with audio-video (AV) and without audio (V) soundtracks. They concluded that the effect of sound relied on the consistency between visual and audio signals. Interestingly, they found that audio had little or no influence on visual attention in cases where the sound sources were precisely the salient objects in the video. However, in cases where sound sources were different from salient objects, they were likely to attract attention of the participants.

Marighetto et al. in [26] investigated the influence of audio on visual attention in non-360-degree videos. Their eye-tracking dataset contained the eye positions collected during four eye-tracking studies. They recorded observers watching video in different audio conditions (with or without sound) and visual groups (moving objects, landscapes, and faces). They reported that there was always less dispersion for audiovisual content than visual content. The presence or absence of sound did influence the spatial distribution of the eye gaze, prominently in the face's category. Wu et al. in [27] presented a head tracking dataset with user behavior patterns in VR applications. The dataset consisted of 48 users watching 18 spherical videos from five categories. They recorded user's head movements, directions of focus, and content remembered by users after each session. However, the videos were accompanied by non-spatial sound.

Finally, Almquist et al. [28] investigated the viewing behaviors of subjects while they experienced various 360-degree videos from different categories through an HMD. Data related to head orientation and rotation speed was collected through the HMD sensors as the subjects watched the videos. They reported that the viewing angle distribution was highly dependent on the content of the video, with viewers spending significant time looking at the front of the video in the Static

Focus and Rides categories. The Exploration and Moving Focus categories had a nearly linear distribution, and yaw-rotation was reported to be the most common rotation compared to pitch and roll rotations.

The novelty of the work presented in this paper lies in its approach to evaluating visual attention and QoE in 360 videos. Specifically, the paper uses head pose and eye gaze data to evaluate visual attention. This approach goes beyond simply analyzing head movements or eye-tracking data to understand which parts of the video viewers are paying attention to. By utilizing eye gaze data along with head pose data, a deeper understanding of how viewers are engaging with the content in the 360° videos can be obtained. In addition, the paper also evaluates QoE through pupil dilations and subjective questionnaires. Pupil dilation is a physiological response that has been linked to cognitive and emotional processing. By measuring pupil dilation, it is possible to gain insights into viewers' cognitive and emotional responses to the 360° videos. This approach provides a more holistic understanding of the user experience beyond traditional subjective questionnaires alone. Furthermore, the paper investigates the effect of different types of audio on visual attention and QoE in 360° videos. This is an important aspect to consider, as audio can greatly impact the viewer's experience of the video.

3. EXPERIMENTAL SETUP

This section provides information on the experimental setup, including details on the laboratory design in section 3.1, the presentation system used for immersive media in section 3.2 and 3.3, and the techniques employed to capture user head pose, gaze, and physiological data in section 3.4.

3.1 Laboratory Design

In Fig. 1, a test subject can be seen participating in the experiment within our laboratory. The design of the laboratory follows the guidelines outlined in ISO 8589:2007[29], which provides recommendations for designing test rooms specifically for sensory analysis. Table 1 details the specifics of the hardware and software components used for the experiment.

Figure 1: Participant experiencing the 360° environment in our lab

Table 1 Experiment setup components

Component	Vendor/Specifications	Used For
PC	Intel Core™ i5 – 4590 CPU @ 3.30GHz, 10.0 GB RAM with a 16GB nVidia GTX 970 Graphics Card running Windows 10	Running the hardware software for the immersive environment
HMD	HTC Vive with Tobii Pro VR Integration [30]	Watching 360° videos
Headphones	Beyerdynamic DT 990 Pro [31]	Listening to non-spatial/ spatial audio
360° Player	GoPro VR Player [32]	Playing 360° videos to the HMD, obtaining head orientation as yaw, pitch and roll
360° videos	[33]	Audio-visual presentation to participants
E4 wristband	[34]	Collecting physiological data

3.2 Presentation System

3.2.1 Selection of 360° Videos

For the experiment, ten 360° videos with first and third-order Ambisonics sound from a pool of recordings available at [33] were chosen. The selection criteria included factors such as video length, content, resolution, and Ambisonics sound order. The videos were categorized as indoor and outdoor, with five videos in each category. The videos were further subcategorized as Opera, Instrument, Riding, and Exploration. To create a non-biased presentation for the participants, the videos were randomly stitched together to form two 300-second (60-sec * 5) segments (one for indoor and one for outdoor) using the ffmpeg [35] tool. The selected videos were processed to set their duration to 60 seconds, stitched together, and the Ambisonics sound was converted to stereo for non-spatial audio experience and to no sound. The videos did not contain any narrative or subtitles.

The five videos which belonged to the Indoor category, featured an Opera performance with actors on an elevated platform and an orchestra performing below the platform (as per Fig. 2a-e). The camera was stationary and positioned between the platform and the orchestra, with each video starting with the participant facing the stage where the actors were performing. The other five videos belonged to the Outdoor category and were exploratory in nature (as per Fig. 2f-j), lacking any specific object of visual interest for participants to focus on immediately. The videos contained sound-emitting objects, some of which were stationary, such as a clock tower or a person playing a musical instrument while seated, while others were in motion, such as people talking while walking or ducks quacking while wading in water.

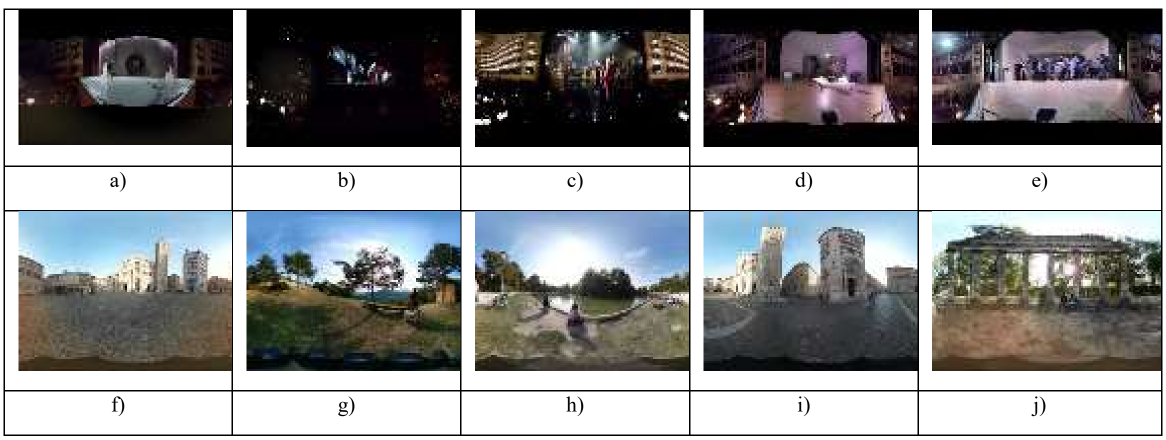
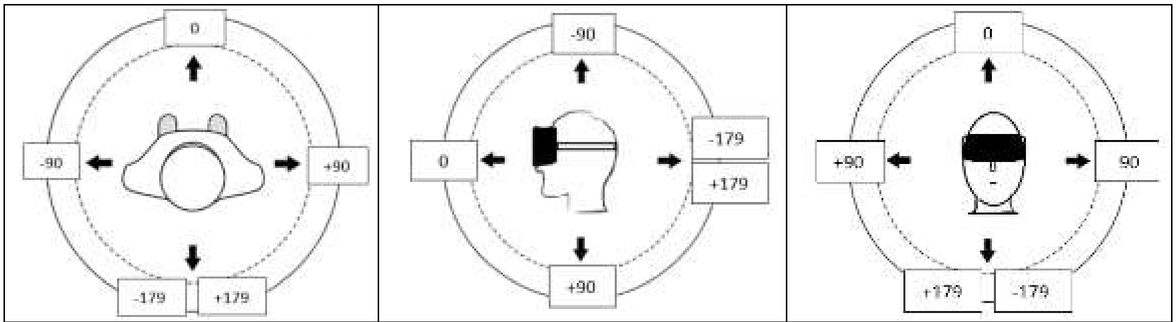


Figure 2: Representative frames for videos in the Indoor (a-e) and Outdoor (f-j) categories

3.2.2 Video Presentation

The GoPro VR player [32], a free 360° video player, was used to present the videos on the HTC Vive with an integrated Tobii pro eye tracker. The VR player transmitted 360° video playback information, such as camera orientation, video URL, playback status, and playback position, to a port on the system on which it was running. This allowed viewers to watch the 360° videos at any orientation, which were recorded as yaw, pitch, and roll (refer to Fig. 3 a, b and c).



a) Yaw	b) Pitch	c) Roll

Figure 3: Yaw, Pitch and Roll angles in degrees adapted from [28].

3.2.3 Audio Presentation

Although the HTC Vive came with an audio strap providing integrated earphones, the Beyerdynamic DT 990 Pro headphones [31] were used to present the various forms of audio. These headphones have larger open-back ear cups, which allows for more natural and clear hearing as suggested by the manufacturer.

3.3 Head pose, Gaze, Pupil Diameter and Heart-rate Data Acquisition

Python scripts were developed to record pose, gaze and pupil diameter data for each frame of the 360° video sequence. The pose data included the yaw, pitch and roll angles measured in radians. The gaze data included the X and Y coordinates of the point where the participant was looking on the screen, while the pupil diameter data recorded the size of the participant's pupils. The data was saved in a CSV file for each participant, which was later used for data analysis. The integration of the Tobii Pro eye tracker with the HTC Vive headset allowed for simultaneous recording of pose and gaze data, which helped in analyzing the correlation between head movement and visual attention.

To collect heart rate data, we used the E4 wristband [34]. This medical grade wearable device offers real-time data acquisition and software for in-depth analysis and visualization. We followed the official installation guide to set up the hardware, and the E4 Connect application was used to capture physiological data during the experiment.

4. RESEARCH METHODOLOGY

This research utilizes an experimental method of evaluation that draws upon various sources, including [36], [37], and [38], as well as ITU-T recommendation P.913 [39].

4.1 Participants

For this study, a convenience sampling approach was employed to recruit a total of 73 participants with an average age of 29 years, consisting of 45 men and 28 women. Among the participants, 26 participants had prior experience with VR, 26 did not have prior experience. VR experience was not captured for the remaining 19 participants and these were the initial participants when we had started the study.

4.2 Assessment Protocol

The assessment protocol for this study was divided into five key phases: an information phase (10-min), a screening phase(10-min), a training phase(5-min), a testing phase(10-min) and 5-10 minutes to answer a subjective questionnaire.

During the information phase, participants were provided with details about the experiment and given the opportunity to ask questions before signing a consent form. The screening phase involved evaluating the participant's visual and auditory acuity, as well as screening for color perception using the Snellen [40] and Ishihara [41] tests and an auditory test [42]. In the training phase, participants watched a 60-second 360° video to become familiar with the VR environment and underwent a calibration process. Further, in the testing phase, participants watched two, 300-second 360° video segments, one recorded indoors and the other outdoors, with either no sound, stereo or first-order or third-order spatial sound. Finally, the assessment ended with every participant answering a subjective questionnaire (5-10 min). The entire assessment lasted approximately 40-50 minutes on average

4.3 Questionnaire and Rating Scale

A questionnaire consisting of twenty questions was developed to evaluate the participants' perception of presence, immersion, and spatiality of sound after watching the stimuli. Inputs from [44] and [45] were considered in the development of the questionnaire. The absolute category rating (ACR) system, as outlined in [39], was used to rate each question. The rating system used a five-point Likert scale, where the participants indicated whether they agreed or disagreed with each statement.

5. DATASET DESCRIPTION

Every participant watched ten videos, five each in indoor and outdoor categories. Every video had a total duration of 60 seconds (including a brief transition splash screen). With a total of ten videos, the overall duration of the videos watched by the participant was 600 seconds; 300-sec in in indoor and 300-sec in the outdoor category. The videos in the indoor category were numbered as 1,3,5,6,7 and those in the outdoor category were numbered as 1,2,4,5,6. The order in which the videos were watched was randomized within the indoor and outdoor category. The indoor sequence was played first followed by the outdoor sequence. Each participant only experienced one sound condition across all the video's they viewed. i.e. videos with no sound (ns) or stereo (st) or first order ambisonics (fo) or high order ambisonics (ho) i.e. third order. The sound condition selection per participant was also randomized.

In terms of the dataset structure, it contains eight folders: foin, fout, hoin, hout, nsin, nsout, stin, and stout. These correspond to the four sound conditions in indoor and outdoor environments. Here, foin and fout indicate the data acquired by making the participant watching videos with first-order sound in indoor and outdoor environments respectively. Similarly, the case for the other 3 sound conditions. Additionally, each folder contains three subfolders, such as: 1) Gaze data (_gazedata); 2) Heart rate data (_HR); 3) Pose data (_posedata) with the respective .csv files. The data regarding gaze and pupil diameter for both the left and right eyes can be found in the _gazedata folder. The data pertaining to the yaw and pitch can be found in the _posedata folder. Whereas, the heart rate data is available in the _HRdata folder. The folders include multiple files in .csv format, which comprise the data captured during the time the participant watched the videos. For example, consider a file "gazedata-20211215-153657_16735.csv" in foin/_gazedata subfolder. Here, the first term "gazedata" indicates the type of data contained in the file, "20211215_153657" indicates that the gaze and pupil diameter information was acquired from the participant on 15th December 2021 at the time 15:36:57. Further, the last term "16735" specifies that the participant was made to watch videos in the sequence 1,6,3,7, and 5. Files that have pose-related information have the prefix 'pose' and those that store heart-rate data have the prefix 'HR'. The gaze and pose data were captured at a sampling rate of 120 samples/second, and hence for a total playback time of 300 seconds, approximately 36000 samples can be found in each file. The heart rate data was captured at a sampling rate of 1 sample/second, and hence for a total playback time of 300 seconds, 300 samples can be found in each file. The dataset is available at <link to be added>.

5.1 Gaze Data

A total of 146 files are present in the dataset which contains information regarding the gaze data under eight different sub-folders. The total files present in each folder are: 18 files each in foin and fout, 18 files each in hoin, and hout, 19 files each in stin and stout, and 18 files each in nsin and nsout. Thus, the information regarding the gaze data is recorded for a total of 43,800 seconds of video playback time including both indoor and outdoor environments, and every file in the _gazedata folder comprises approximately 36000 samples.

5.2 Heart Rate Data

In total, 73 files with heart rate related data are contained in the dataset. Each file has information regarding heart rate of the participant in the baseline condition, heart rate when the participant watched videos in the Indoor category and heart rate when the participant watched videos in the Outdoor category along with the corresponding timestamp for a particular sound condition. The fo, ho, and ns folders comprise 18 files each, whereas the st folder includes a total of 19 files with heart rate data.

5.3 PoseData

The pose data gives information regarding the pose of the participants while watching videos. Pitch indicates the movement of the head along the horizontal axis, while the movement in the vertical axis is given by yaw. Head movements can be used to identify the area where the visual attention of the user is concentrated, as the position of the head changes with the variation in the displayed visuals (and perhaps the sound and its nature viz. non-spatial/spatial). The pose data recorded are stored in the folder named _posedata, which has eight sub-folders. In this dataset, a total of 146 files are present with information containing the pose data, with 18 files contained in foin, 18 files in fout, 18 files each in hoin, and hout, 19 files in both stin and stout, and 18 files in nsin, and nsout, correspondingly, similar to the gaze data. Similar to the _gazedata, the files in the _posedata folder also contain around 36000 samples, recorded from when the participants were made to watch the sequence of 5 videos in indoor or outdoor conditions.

6. RESULTS AND DISCUSSION

In this section, the results and analysis from a data visualization perspective on the implicit metrics and from a statistical analysis perspective on the self-reported explicit metrics.

6.1 Data Visualization

Here, the data visualization is executed by considering the values of the different parameters such as gaze, pose, pupil diameter and heart rate captured at various instances. Further, the values obtained for the various parameters are analyzed in both indoor and outdoor environments with the four different sound conditions, such as no sound, stereo, first-order, and higher-order.

It is important to capture and analyze gaze and pupil diameter for the left and right eyes separately when studying visual attention because not only the two eyes work together to form a single visual image, but they also have separate roles in visual perception, such as in binocular vision and stereopsis [46]. This means that the left and right eyes may have different contributions to visual attention and processing, which can impact the interpretation of eye-tracking data. Studies have shown that analyzing gaze and pupil diameter separately for each eye can provide more accurate and detailed information about visual attention and cognitive processing in 360° video environments [47] [48] [49]. Also, research suggests that the left and right eyes may have different fixation durations and patterns, which can affect the perception and understanding of the visual content [47]. Similarly, analyzing pupil diameter separately for each eye can provide insights into the cognitive and emotional processing of visual information in immersive environments [48] [49].

6.1.1 Indoor category Analysis

Fig. 4a-d below show participant's head pose. eye gaze, pupil diameter and heart rate as they watched a video in the Indoor category in various sequences from videos 1,3,5,6 and 7 in the no sound, stereo, first order and high order sound conditions. Table 2 summarizes the overall analysis for head pose, eye gaze, pupil diameter and heart rate based on Fig.4a-d. Section 6.1.1.1 gives a detailed analysis of the head pose yaw and pitch movement and angles for each individual video in the sequence considering the content of the video. Section 6.1.1.2 has a discussion on the pupil diameter for each sound condition followed by a discussion on heart rate in section 6.1.1.3.

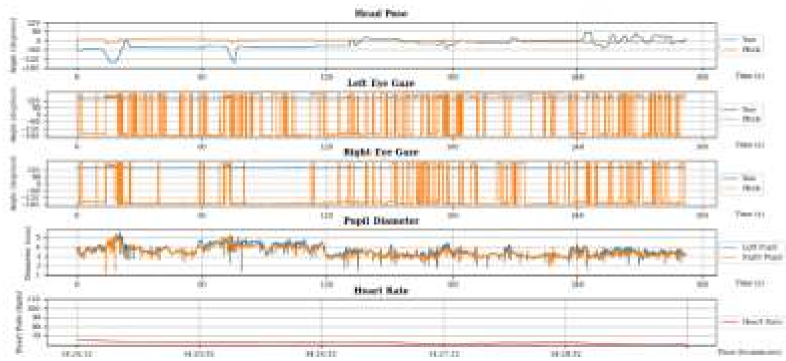


Figure 4a: No Sound, Indoor – Videos 1,3,5,6,7 (each video is 60-sec duration; total duration 300-sec)

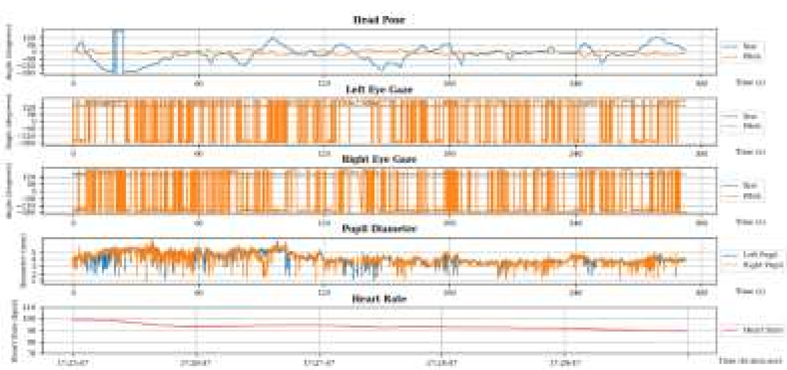


Figure 4b: Stereo sound, Indoor – Videos 1,3,5,6,7 (each video is 60-sec duration; total duration 300-sec)

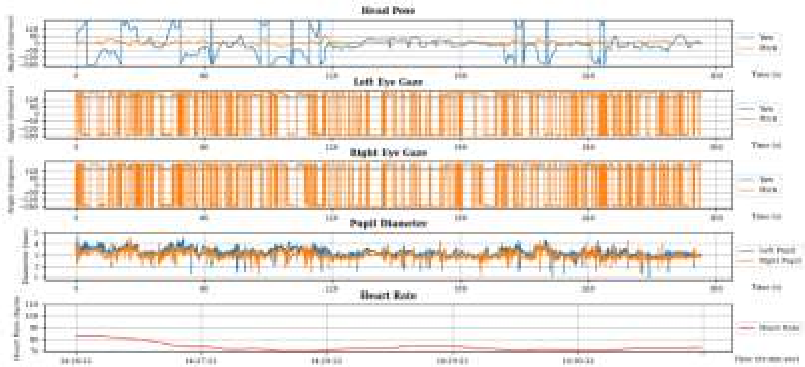


Figure 4c: First-order sound, Indoor – Videos 1,3,5,6,7 (each video is 60-sec duration; total duration 300-sec)

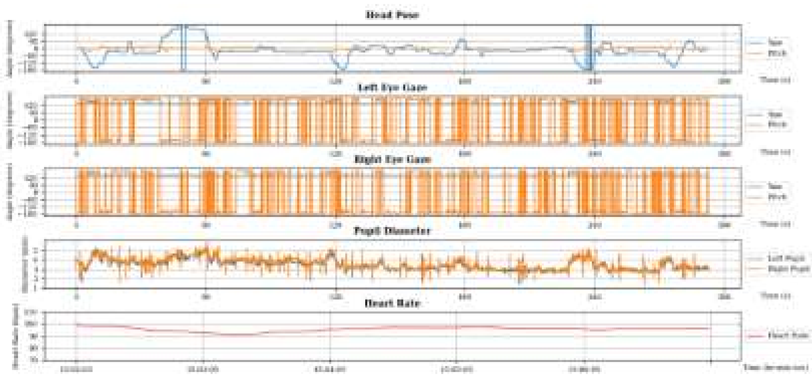


Figure 4d: High Order sound, Indoor – Videos 1,3,5,6,7 (each video is 60-sec duration; total duration 300-sec)

Table 2: Overall Analysis – Indoor category

No Sound	**Stereo**	**First order**	**High order**
Pupil diameter is between 2-5 mm.	Pupil diameter between 3-5 mm.	Pupil diameter between 1-5 mm.	Pupil diameter between 1-6 mm.
Heart rate less than 70 bpm	Heart rate is between 70-110 bpm	Heart rate is between 70-80 bpm	Heart rate is between 90-100 bpm
Most fluctuations in head pose yaw for videos 5 and 7 where many actors are present on the stage and there is motion	More fluctuation in head pose yaw compared to pitch throughout the video and compared to no sound condition. Most fluctuations in head pose yaw for Video 5 and	More fluctuation in head pose yaw compared to pitch throughout the video and compared to no sound and stereo conditions. Head pose yaw found to be varying for	More fluctuation in head pose yaw compared to pitch throughout the video. Head pose yaw found to be varying for all videos except 3 and 5. The variation is gradual and

	Video 7 where many actors are present on the stage and there is motion	Videos 1,3 and 6 which is not the case in the previous two sound conditions	not abrupt unlike first-order sound condition
Sudden changes in head pose yaw cause the gaze to fluctuate across all sound conditions. Rapid head movements, such as sudden turns or rotations, can cause the visual scene in a 360° video to shift or become blurry, resulting in a delay in the brain's ability to adjust to the new visual information. This may cause the user's gaze to fluctuate or become unstable [50] [50]			

6.1.1.1 Head Pose

a) Video 1:

The video (Fig.2a) shows two actors on stage singing at the same location throughout the video. There is an orchestra performing below the stage. The camera and mic position are fixed between stage and orchestra. There was very limited movement of the head pose for both yaw and pitch directions when the video was accompanied with no sound (Fig.4a). Also, the yaw angle did not vary much. The yaw angle remained close to the equator. In the stereo sound condition (Fig.4b), the movement in the yaw direction increased while the pitch remained stable. A larger variation was observed for the yaw angle compared to the no sound condition. In the first order condition (Fig.4c), a significant movement was the head pose was found in terms of the yaw direction and also the yaw angle was found to vary largely compared to the previous two condition. Both the pitch direction and angle increased but the increase was lesser compared to the yaw. For the high-order sound condition (Fig.4d), both the movement of the head pose in the yaw direction and the variation in the yaw angle was found similar to that of the first order sound condition; however, the variation in the yaw angle was gradual as opposed to abrupt in the first order sound condition. The variation in terms of the pitch direction and angle was similar to that of the first order sound condition.

When the video was watched with no sound, there was very limited movement of the head pose for both the yaw and pitch directions. The yaw angle also did not vary much, remaining close to the equator. This suggests that the participant was relatively stationary and focused on the visual aspects of the performance. When stereo sound was added to the video, there was an increase in the movement of the head pose in the yaw direction. This indicates that the participant was turning their head to track the sound, possibly to locate the source of the sound. The pitch direction remained stable, suggesting that the participant was primarily focused on the visual aspects of the performance.

In the first order sound condition, there was a significant increase in the movement of the head pose, particularly in the yaw direction. The yaw angle also varied largely compared to the no sound and stereo sound conditions. This suggests that the participant was actively tracking the sound and moving their head to locate the source of the sound. Both the pitch direction and angle also increased, although to a lesser extent than the yaw direction. In the high-order sound condition, the movement of the head pose in the yaw direction and the variation in the yaw angle were found to be similar to that of the first order sound condition. However, the variation in the yaw angle was gradual as opposed to abrupt in the first order sound condition. The variation in terms of the pitch direction and angle was similar to that of the first order sound condition.

Overall, the participant's head movements suggest that they were actively engaged in the video, particularly in response to the addition of sound.

b) Video 3:

This video (Fig.2b) had a single actor on stage singing at the same location throughout the video. The background was dark with only the singer visible clearly. There was an orchestra performing below the stage. Both the camera and mic were positioned between the stage and the orchestra. A behavior similar to Video 1 was observed for the no sound, stereo and first order sound conditions (Fig.4a-c). As regards the high order sound condition (Fig.4d), surprisingly, both the movement and angle for the yaw were limited. The yaw angle remained close to the equator. There was very little variation in pitch as well. Thus, it is possible that the similarity between Video 1 and Video 3 contributed to the participant's similar head movements across the different sound conditions.

c) Video 5:

The video (Fig.2c) had multiple actors on stage in different locations with more than one actor singing at different times. Some actors moved slowly across the stage. An orchestra performed below the stage. The camera and mic position were between the stage and orchestra. In the no sound condition (Fig.4a), there was more movement in terms of the yaw direction compared to the previous two videos. The pitch remained stable. For the stereo condition (Fig.4b), the movement in the yaw direction increased while the pitch remained stable. A larger variation was observed for the yaw angle compared to the no sound condition. Both the yaw movement and angle were relatively limited compared to the Videos 1 and 3 for the first order sound condition (Fig.4c). In the high order sound condition (Fig.4d), both the movement and angle for the yaw were limited. There was very little variation in pitch as well.

In the no sound condition, the participant exhibited more movement in the yaw direction compared to the previous two videos. However, the pitch remained stable. This suggests that the participant was more engaged with the visual aspects of the performance, possibly due to the presence of multiple actors and movement across the stage. In the stereo condition, the movement in the yaw direction increased while the pitch remained stable. The larger variation in the yaw angle compared to the no sound condition suggests that the participant was more engaged with the audio aspects of the performance. For the first order sound condition, both the yaw movement and angle were relatively limited compared to Videos 1 and 3. This suggests that the participant was less engaged with the audio compared to the previous videos, possibly due to the complexity or unfamiliarity of the sound. In the high order sound condition, both the movement and angle for the yaw were limited, and there was very little variation in pitch. This suggests that the participant was less engaged with the audio and that the complexity of the sound may have been overwhelming or distracting.

Overall, the participant's head movements in response to the different sound conditions suggest that they were more engaged with the visual aspects of the performance than the audio, particularly in the no sound and stereo conditions. However, the presence of multiple actors and movement across the stage may have increased their engagement with the visual aspects. The complexity and unfamiliarity of the audio in the first order and high order sound conditions may have limited the participant's engagement with the performance.

d) Video 6:

Video 6 (Fig.2d) showed a single actor on stage singing at the same location throughout the video. The singer was very close to camera compared to other videos in this category and clearly visible. An orchestra performed below the stage. The camera and mic were positioned between the stage and orchestra. A behavior similar to Video 1 and 3 was observed for the no sound, stereo and first order sound conditions (Fig.4a-c). In the high order sound condition (Fig.4d), both the movement and angle for the yaw were limited and a very little variation in pitch as well. The participant's head movements in response to the different sound conditions were similar to those observed in Videos 1 and 3 for the no sound, stereo, and first order sound conditions. This suggests that the participant was similarly engaged with the performance's audio and visual aspects. In the high order sound condition, both the movement and angle for the yaw were limited, and there was very little variation in pitch. This suggests that the complexity of the audio in the high order sound condition may have limited the participant's engagement with the performance, causing them to focus more on the visual aspects. Additionally, the close proximity of the singer to the camera may have also limited the participant's head movements due to the singer's proximity to the camera and their limited movements across the stage.

Overall, the participant's head movements suggest that they were engaged with the audio and visual aspects of the performance, with the high order sound condition limiting their engagement with the audio. However, the close proximity of the singer to the camera may have also limited the participant's head movements.

e) Video 7:

The video (Fig.2e) had multiple actors on stage in different locations with only one actor singing from the same location throughout. There was a lot of motion across the stage with an orchestra performing below the stage. A camera and mic were positioned between the stage and orchestra. The largest variation in head pose yaw movement and angle in the no sound condition (Fig.4a) was observed for this video. For the stereo condition (Fig.4b), a larger variation was observed for the yaw angle compared to the no sound condition. The pitch remained stable. Both the yaw movement and angle were relatively limited compared to the Videos 1 and 3 for the first order sound condition (Fig.4c). This observation is similar to that for Video 5 in the first order sound condition. In the high order sound condition (Fig.4d), the yaw movement continued to vary gradually and pitch did not show much variation.

The participant's head movements in response to the different sound conditions indicate that there was the largest variation in head pose yaw movement and angle in the no sound condition for this video. This suggests that the participant was more engaged with the visual aspects of the performance in the absence of audio. For the stereo condition, a larger variation was observed for the yaw angle compared to the no sound condition, while the pitch remained stable. This suggests that the participant's engagement with the audio increased with the addition of stereo sound.

For the first order sound condition, both the yaw movement and angle were relatively limited compared to Videos 1 and 3, and the pitch remained stable. This suggests that the complexity of the audio in this condition may have limited the participant's engagement with the performance's audio and visual aspects. This observation is similar to that for Video 5 in the first order sound condition. In the high order sound condition, the yaw movement continued to vary gradually, while the pitch did not show much variation. This suggests that the participant's engagement with the audio was still limited in this condition, but the gradual variation in yaw movement suggests that the audio complexity may have been more engaging than in the first order sound condition.

Overall, the participant's head movements suggest that they were more engaged with the visual aspects of the performance in the absence of audio, and that their engagement with the audio increased with the addition of stereo sound. However, the complexity of the audio may have limited their engagement in the first order sound condition, and their engagement with the audio was still limited in the high order sound condition.

6.1.1.2 Pupil Diameter

Table 2 provides a summary of the changes in pupil dilation observed across all four sound conditions in the indoor category, as depicted in Figures 4a-d. In the no sound condition, the pupil diameter was found to be between 2-5 mm (Fig.4a). This range indicates a moderate level of arousal and attention, with the smaller diameter values indicating lower arousal and the larger diameter values indicating higher arousal. The relatively narrow range of pupil diameter in this condition suggests that the absence of sound may have resulted in a relatively stable physiological response. In the stereo condition (Fig.4b), the pupil diameter ranged from 3-5 mm, which is slightly narrower compared to the no sound condition. This suggests that the addition of stereo sound may have led to a slight increase in arousal and attention, but not enough to result in a significant change in the range of pupil diameter. It is worth noting that this condition may still have a moderate impact on the viewer's physiological response, as even small changes in pupil diameter can indicate changes in attention and arousal.

In the first order condition (Fig.4c), the range of pupil diameter was found to be between 1-5 mm, with a wider range compared to the previous two conditions. The wider range of pupil diameter in this condition may be attributed to the fact that first-order sound has a basic left-right stereo effect, which may have resulted in a more immersive and engaging audio experience, potentially leading to greater fluctuations in arousal and attention.

In the high order condition (Fig.4d), the range of pupil diameter was found to be between 1-6 mm, which is the widest range observed among the four sound conditions. The wider range of pupil diameter in this condition may be attributed to the more advanced spatial audio techniques used to create a more immersive and realistic audio experience by the high order. The wider range of pupil diameter in this condition suggests that it may have the greatest impact on the viewer's physiological response among the four sound conditions, potentially resulting in greater fluctuations in arousal and attention.

Overall, the pupil diameter analysis suggests that the presence and type of sound can have an impact on the viewer's physiological response, with wider ranges of pupil diameter potentially indicating greater fluctuations in arousal and attention.

6.1.1.3 Heart Rate

In accordance with Figures 4a-d, Table 2 presents a summary of the variations in heart rate that were observed across all four sound conditions in the indoor category. In the no sound condition (Fig.4a), the heart rate was found to be less than 70 bpm. This indicates a low level of arousal and attention, suggesting that the absence of sound may have resulted in a relatively stable physiological response.

In the stereo condition (Fig.4b), the heart rate was found to be between 70-110 bpm. This range indicates a moderate level of arousal and attention, with the higher heart rate values suggesting a greater level of engagement and immersion. The increase in heart rate compared to the no sound condition suggests that the addition of stereo sound may have led to a more engaging and immersive audio experience, resulting in increased physiological response.

In the first order condition (Fig.4c), the heart rate was found to be between 70-80 bpm, which is similar to the stereo condition. However, the heart rate was found to be more stable and consistent in this condition compared to the stereo condition, suggesting that the type of sound used may have resulted in a less immersive and engaging audio experience.

In the high order condition (Fig.4d), the heart rate was found to be between 90-100 bpm, which is slightly higher compared to the first order and stereo conditions. The higher heart rate values suggest a greater level of engagement and immersion, potentially resulting from the more advanced spatial audio techniques used to create a more immersive and realistic audio experience.

6.1.2 Outdoor category analysis

Fig. 5a-d below show participant's head pose. eye gaze, pupil diameter and heart rate as they watched a video in the outdoor category in various sequences from videos 1,2,4,5 and 6 in the no sound, stereo, first order and high order sound conditions. Table 3 summarizes the overall analysis for head pose, eye gaze, pupil diameter and heart rate based on Fig.5a-d. Section 6.1.2.1 gives a detailed analysis of the head pose yaw and pitch movement and angles for each individual video in the sequence considering the content of the video. Section 6.1.2.2 has a discussion on the pupil diameter for each sound condition followed by a discussion on heart rate in section 6.1.2.3

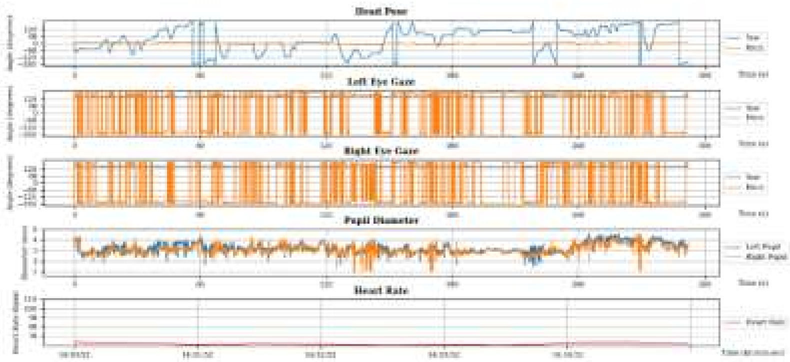


Figure 5a: No Sound, Outdoor – Videos 1,6,5,4,2 (each video is 60-sec duration; total duration 300-sec

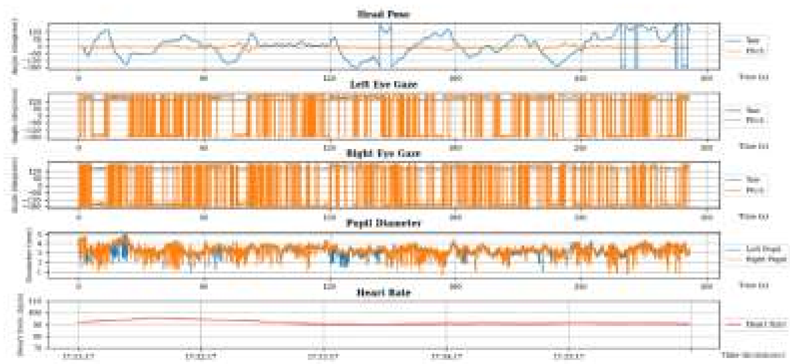


Figure 5b: Stereo sound, Outdoor – Videos 1,6,5,4,2 (each video is 60-sec duration; total duration 300-sec)

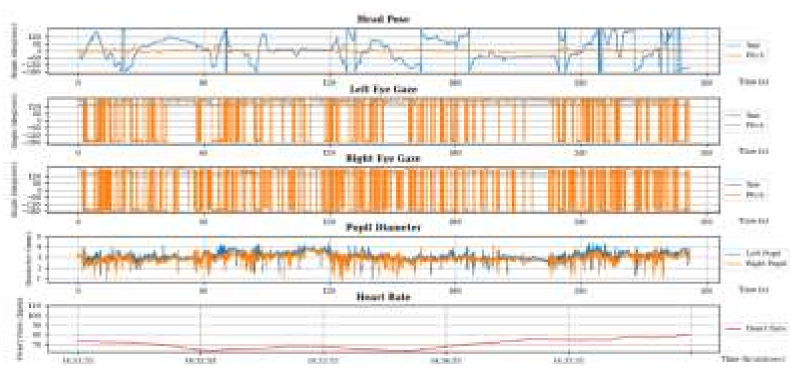


Figure 5c: First order sound, Outdoor – Videos 1,6,5,4,2 (each video is 60-sec duration; total duration 300-sec

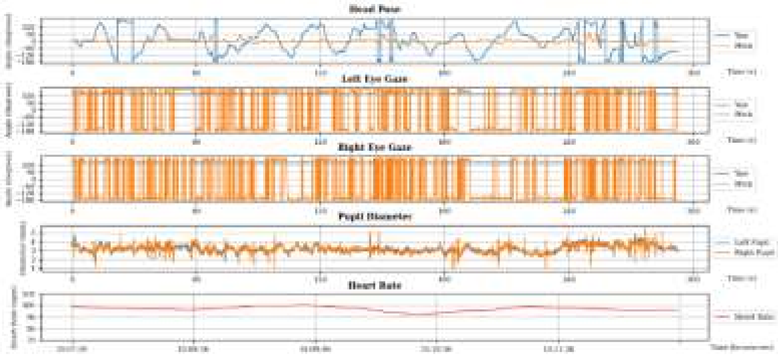


Figure 5d: High order sound, Outdoor – Videos 6,5,4,1,2 (each video is 60-sec duration; total duration 300-sec)

Table 3: Overall Analysis – Outdoor category

No Sound	**Stereo**	**First order**	**High order**
Pupil diameter is between 1-5 mm.	Pupil diameter between 3-5 mm.	Pupil diameter between 1-5 mm.	Pupil diameter between 2-5 mm.
Heart rate less than 70 bpm	Heart rate is between 90-100 bpm	Heart rate is between 60-80 bpm	Heart rate is between 90-100 bpm
Head pose yaw found to be varying for all videos with abrupt changes	Head pose yaw found to be gradual compared to no sound	More fluctuation in head pose yaw compared to the previous two sound conditions	Pitch is also seen to vary along with head pose yaw throughout the video, unlike previous sound conditions.
Sudden changes in head pose yaw cause the gaze to fluctuate. If there is a sudden change in head pose, such as a rapid turn or rotation, the visual scene may shift or blur before the brain has a chance to adjust, which can lead to gaze fluctuations [50].			

6.1.2.1 Head Pose

a) Video 1:

The video (Fig.2f) shows a market square with many people moving across. The main source of sound is a clock tower which is in an elevated fixed location. There are faint sounds from people moving across. Both the camera and mic position are static.

A gradual movement in the head pose yaw with angles changing gradually as well for the no sound condition (Fig.5a). The yaw movement is more gradual and the change is angles is also subtle in the stereo condition (Fig.5b). Pitch is also found to vary compared to the no sound conditions. The movement is close to the equator. A big change in yaw and pitch movement and angles is observed in the first order condition (Fig.5c). The same observation holds true for the high order sound condition (Fig.5d) but the changes are gradual in the high order sound condition as compared to abrupt in the first order.

In the no sound condition, the viewer experiences a gradual movement in the head pose yaw, with angles changing gradually as well. This yaw movement suggests that the viewer is scanning the scene in a slow and deliberate manner, taking in the details of the environment. The pitch, which refers to the vertical movement of the head, is also found to vary compared to the no sound condition, indicating that the viewer is actively exploring the scene. When the stereo sound is introduced, the viewer experiences a similar gradual movement in the head pose yaw, but with more subtle changes in angles. This could be due to the influence of the stereo sound on the viewer's attention, which may be more focused on the source of the sound. The variation in pitch is also present in the stereo condition, suggesting that the viewer is still actively exploring the scene.

In the first-order condition, there is a sudden and significant change in the viewer's head movement, with a big change in yaw and pitch movement and angles. This change could be due to the sudden loud noise (sound from the clock tower), which grabs the viewer's attention and shifts their focus. This sudden change in head movement highlights the role of sound in capturing attention and guiding the viewer's focus in a dynamic outdoor environment. Similarly, in the high-order sound condition, there is also a big change in yaw and pitch movement and angles, but the changes are more gradual compared to the first-order condition. This gradual change suggests that the viewer is still reacting to a significant sound event, but the reaction is not as sudden or intense as in the first-order condition.

b) Video 2:

The video (Fig.2g) shows the top of a hill with a man driving a motorbike across the hill and three dogs barking and following him around. The sound sources (the rider and the dogs) are moving across a large area. Both the camera and mic position are fixed.

The very nature of video 2 with lots of movement is seen to result in large head pose changes in terms of the yaw movement and angles in the no sound condition (Fig. 5a). This behavior is found to continue in the stereo condition. The difference being there is more variation in the yaw angles along with more movement since the video is accompanied with sound (Fig.5b). A similar pattern is observed for the first (Fig.5c) and high order (Fig.5d) conditions, but there are variations in head pose pitch angles as well in the high order sound condition

In the no sound condition, the viewer experiences large head pose changes in terms of yaw movement and angles. This suggests that the viewer is actively tracking the movement of the rider and the dogs, trying to keep them in view. The large head pose changes indicate that the viewer is making rapid movements to follow the action, reflecting the dynamic nature of the scene. When the stereo sound is introduced, the viewer continues to experience large head pose changes, with more variation in the yaw angles and more movement overall. This indicates that the sound is helping to guide the viewer's attention and focus, with the viewer's head movements tracking the sound sources as they move across the scene.

In the first-order condition, a similar pattern is observed, with the viewer making large head pose changes to track the movement of the rider and dogs. However, the viewer's movements are more sudden and abrupt in response to the realism of the sound events. This highlights the role of sound in capturing attention and guiding the viewer's focus in a dynamic outdoor environment. In the high-order sound condition, there are variations in both yaw and pitch angles of the head pose, suggesting that the viewer is reacting to changes in the frequency and amplitude of the sound. This could be due to changes in the tone and loudness of the dogs' barking or the engine noise of the motorbike.

c) Video 4:

Fig.2h shows people and birds near a water body. The sound sources not very widely spread. The camera and mic positions are fixed.

In the no sound condition (Fig.5a), the yaw movement and angle are stationary for some time and then fluctuates briefly before being stationary again. The yaw varies more gradually in terms of movement and angle in the stereo condition (Fig.5b). Again, one can see a lot of variation in terms of yaw movement and angles in the first order sound condition

(Fig.5c). The largest variation in terms of pose movement and angles can be found in the high order sound condition (Fig.5d).

In the no sound condition, the viewer's head pose exhibits relatively little movement, remaining stationary for some time before experiencing brief fluctuations and returning to a stationary state. This suggests that the viewer is relatively passive and not particularly engaged with the content. In the stereo condition, where the sound is presented in a more immersive manner, the viewer's head pose exhibits more gradual and subtle variations in yaw movement and angle, indicating a greater level of engagement and immersion in the scene. This effect is even more pronounced in the first-order sound condition, where the sound is presented in a more directional and realistic manner.

The largest variations in head pose movement and angles are observed in the high-order sound condition, where the sound is enhanced with additional layers of complexity and spatial cues. This suggests that the viewer is not only more engaged with the content, but also actively exploring and responding to the added layers of information presented in the sound.

d) Video 5:

Fig.2i shows a market square with many people moving across. The main source of sound is a man playing an instrument at a fixed location but not immediately visible. There are faint sounds from people moving across. Both the camera and mic are in a fixed position.

The behavior of the head pose yaw movement and angles is similar to that of Video 1(Fig.2f) across all four sound conditions (Fig.5a-d). In Video 1, the yaw movement and angle of the head pose is relatively stable and stationary in all four sound conditions. This could be because the main source of sound, the clock tower, is stationary and does not move across the listener's field of view. Additionally, the faint sounds of people moving across do not appear to have a significant effect on the viewer's head pose. Similarly, in Video 5, the behavior of the head pose yaw movement and angles is also relatively stable and stationary across all four sound conditions. This could be due to the fact that the main source of sound, a man playing an instrument at a fixed location, is also not moving across the viewer's field of view. However, it is worth noting that the presence of many people moving across the market square could potentially cause more variability in head pose movements, but this does not appear to be the case.

Overall, the results suggest that when the main source of sound is stationary and not moving across the viewer's field of view, there is less variability in head pose movements. In contrast, when the sound sources are more dynamic and move across the viewer's field of view, there is more variability in head pose movements.

e) Video 6:

Refer to Fig. 2j which shows two performers, one playing a musical instrument and one clapping, in the same location throughout, sitting below a monument. The sound sources are concentrated. There are a few spectators watching the performance. The camera and mic are static and these are located between the performers and the spectators.

In the no sound condition, the yaw angle changes a couple of times to return back to the equator for the rest of the time (Fig.5a). It stays near the equator for most of the time after a slight variation in the angle in the stereo condition (Fig.5b). The yaw angle is seen to vary more in the first order condition with a very slight variation in pitch (Fig.5c). The largest variation in the yaw angle is found in the high order sound condition (Fig, 5d). From this, it can be deduced that the user's head pose yaw and pitch angles are affected by the different sound conditions in Video 6. The presence of sound sources with higher spatial complexity, such as in the first-order and high-order sound conditions, leads to more variation in the user's head pose yaw angle compared to the no sound and stereo sound conditions. However, the pitch angle remains relatively stable across all sound conditions. This suggests that the user may be focusing more on the spatial location of the sound sources and less on their elevation.

6.1.2.2 Pupil Diameter

With reference to Fig.5a-d, Table 3 summarizes the pupil dilation across all four sound conditions in the outdoor category. The pupil diameter in the outdoor condition ranges from 1-5 mm for the no sound condition (Fig.5a). This is a relatively wide range and could indicate that the participants are experiencing some level of stress or discomfort. It's also possible that the outdoor environment itself is contributing to the variation in pupil diameter. Without any sound, there may be other environmental factors that are influencing the participants' arousal levels. With stereo sound, the pupil diameter is between 3-5 mm (Fig.5b). This range is narrower than the no sound condition and indicates that the stereo sound may be having a calming effect on the participants. This could be due to the fact that stereo sound is more natural and immersive compared to no sound. In the first order sound condition, the pupil diameter ranges from 1-5 mm (Fig.5c). This is similar to the no sound condition and could indicate that the first order sound is not having a significant effect on the participants' arousal levels. It's also possible that other environmental factors in the outdoor setting are outweighing the effects of the first order sound. With high order sound, the pupil diameter ranges from 2-5 mm (Fig.5f). This range is slightly narrower than the no sound condition and may indicate that the high order sound is having a slight calming effect on the participants. However, it's worth noting that the range is still relatively wide, so it's possible that other factors in the outdoor environment are contributing to the variation in pupil diameter.

Overall, it's difficult to attribute the variation in pupil diameter solely to the sound conditions in the outdoor category. The outdoor environment itself can have a significant impact on arousal levels, and other factors such as lighting, and physical activity can also influence pupil diameter. However, the narrower ranges of pupil diameter in the stereo and high order sound conditions suggest that these sound conditions are having a slight calming effect on the participants.

6.1.2.3 Heart Rate

Table 3 provides summarizes the variation of heart rate across the four sound conditions in the outdoor category, as illustrated in Figures 5a-d. In the outdoor category with no sound, the heart rate is less than 70 bpm (Fig.5a). This indicates a relatively calm state with minimal physical or emotional stress. The lack of auditory stimuli may have contributed to the lower heart rate. When stereo sound is present in the outdoor category, the heart rate is between 90-100 bpm (Fig.5b). This suggests increased arousal or excitement, possibly due to the auditory stimuli. It is important to note that this range is still within a normal heart rate range for adults. In the outdoor category with first-order sound, the heart rate is between 60-80 bpm (Fig.5c). This is a lower heart rate compared to the stereo sound condition, which could indicate a state of relaxation or even boredom. It is also possible that the sound quality in this condition was not as engaging as in the stereo condition. In the outdoor category with high-order sound, the heart rate is between 90-100 bpm (Fig.5d). This is similar to the heart rate observed in the stereo condition, which suggests that high-order sound also contributes to increased arousal or excitement in the outdoor environment.

Overall, it seems that the presence and quality of sound can significantly impact heart rate in the outdoor environment. Stereo and high-order sound conditions appear to increase arousal and excitement, while the first-order sound condition appears to have a less stimulating effect on heart rate. The lack of sound, on the other hand, seems to result in a calmer state with a lower heart rate.

7. ANALYSIS OF VARIANCE (SUBJECTIVE QUESTIONNAIRE)

Table 4 represents the statistical analysis of the subjective questionnaire aimed at capturing the user QoE of the 73 participants who watched the ten 360° videos under four different sound conditions: no sound, stereo, first order ambisonics, and high order ambisonics across the Indoor and Outdoor categories. The questionnaire consisted of 20 questions. An ANOVA with 95% confidence level was executed on the subjective data between the different sound condition groups. significant results are those with a p-value less than 0.05, which indicates that there is a low probability of obtaining the observed difference in means by chance alone.

The table shows the average responses for each question and sound condition, as well as the p-value, degrees of freedom (df), and F-value for each question. From the table, we can see that questions 1, 2, 4, 5, 9, 11 and 17 have statistically significant differences in means across the four sound conditions.

Table 4 Analysis of Variance (ANNOVA) - Subjective Questionnaire

No.	**Question**	**st**	**fo**	**ho**	**ns**	**p-value**	**df**	**F**
1	**Retaining Attention**	4.4	4.6	4.4	3.7	**0.01**	3,68	4.11
2	**Conscious Awareness of real world**	2.7	2.8	2.5	3.7	**0.02**	3,68	3.23
3	Separation from real world	3.9	4	3.4	3.1	0.09	3,68	2.23
4	**Experiencing vs. watching**	4	4.2	3.9	2.8	**0.02**	3,68	5.27
5	**Enjoying the experience**	4.7	4.5	4.8	3.8	**0.007**	3,68	4.3
6	Motivation to continue watching	4.4	4.5	4.4	3.6	0.08	3,68	2.27
7	Proximity to objects	3.8	3.5	3.5	3.5	0.79	3,68	0.33
8	Involvement in experience	4	4.1	4.2	3.4	0.25	3,68	1.3
9	**Engagement of senses**	4.1	4	4.3	3.1	**0.0009**	3,68	6.13
10	Awareness of real world	3.1	2.3	2.7	2.9	0.41	3,68	0.96
11	**Naturalness of interaction**	3.7	3.8	4	2.8	**0.01**	3,68	3.55
12	Presence in virtual space	4.4	4.3	4.4	3.9	0.45	3,68	0.87
13	Sound identification	4.7	4.8	4.8	1.5	<0.001	3,68	200.1
14	Sound location	4.4	4.5	4.3	1.4	<0.001	3,68	82.3
15	Sound Stressfulness	1	1.2	1.2	1.2	0.73	3,68	0.42
16	**Sound Realism**	4	4	4.4	1.5	**<0.001**	3,68	65.7
17	Sound Loudness	2.3	1.8	2.1	1.1	<0.001	3,68	6.36
18	Sound Clarity	3.8	3.9	4.3	1.3	<0.001	3,68	96.5
19	Exploration within environment	4.4	3.7	3.9	4	0.15	3,68	1.8
20	Adjustment to experience	3.9	4.2	4.3	4	0.63	3,68	0.57

For question 1, "Retaining Attention," the mean response was highest for the first-order ambisonics condition (4.6), followed closely by stereo (4.4), and then high-order ambisonics (4.4). The no-sound condition had the lowest mean response (3.7). The ANOVA showed a significant difference between the means (p-value = 0.01), indicating that the sound conditions influenced the users' ability to retain their attention during the experience. This is an important consideration for creators of immersive media who aim to create compelling and engaging experiences for their audience.

For question 2, "Conscious Awareness of real world," the mean response was highest for the no-sound condition (3.7), followed by stereo (2.7) and then first-order ambisonics (2.8). High-order ambisonics had the lowest mean response (2.5). The ANOVA showed a significant difference between the means (p-value = 0.02), indicating that the sound conditions influenced the users' conscious awareness of the real world during the experience. The absence of sound may increase users' conscious awareness of the real world, as indicated by the highest mean response for the no-sound condition. This finding is consistent with the idea that sound can create a sense of immersion and may distract users from their external environment.

For question 4, "Experiencing vs. watching," the mean response was highest for the first-order ambisonics condition (4.2), followed by stereo (4.0), high-order ambisonics (3.9), and then no sound (2.8). The ANOVA showed a significant difference between the means (p-value = 0.02), indicating that the sound conditions influenced the users' perception of experiencing the content rather than just watching it. The use of sound can significantly enhance the user's perception of experiencing the content rather than just watching it. Specifically, the participants reported higher mean responses for the sound conditions (first-order ambisonics, stereo, and high-order ambisonics) compared to the no-sound condition, indicating that the presence of sound can enhance the user's immersion and sense of presence within the virtual environment.

For question 5, "Enjoying the experience," the mean response was highest for the high-order ambisonics condition (4.8), followed by stereo (4.5), first-order ambisonics (4.7), and then no sound (3.8). The ANOVA showed a significant difference between the means (p-value = 0.007), indicating that the sound conditions influenced the users' enjoyment of the experience. The finding that the high-order ambisonics condition received the highest mean response suggests that the use of spatially accurate sound can enhance the overall immersive experience for users.

For question 9, "Engagement of senses," the mean response was highest for the high-order ambisonics condition (4.3), followed by stereo (4.1), first-order ambisonics (4.0), and then no sound (3.1). The ANOVA showed a significant difference between the means (p-value = 0.0009), indicating that the sound conditions influenced the users' engagement of their senses during the experience. The use of high-order ambisonics can provide a more immersive and engaging experience by capturing the three-dimensional sound field and allowing for more accurate localization of sounds. This can enhance the user's sense of presence and immersion in the virtual environment.

For question 11, "Naturalness of interaction," the mean response was highest for the high-order ambisonics condition (4.0), followed by first-order ambisonics (3.8), stereo (3.7), and then no sound (2.8). The ANOVA test showed a significant difference among the means of the four sound conditions (p-value = 0.01). The finding that the high-order ambisonics condition had the highest mean response for "Naturalness of interaction" indicates that adding sound to a virtual experience can enhance the users' perception of the naturalness of their interaction with the content. This is an important consideration for developers of virtual reality and other immersive technologies, as creating a sense of naturalness in the user's interaction with the content is crucial for achieving a more immersive and engaging experience.

In conclusion, the questionnaire investigated the effect of different sound conditions (no sound, stereo, first-order ambisonics, and high-order ambisonics) on users' presence and immersive experience while watching the 360-degree videos in different sound conditions. The results showed that sound conditions had a significant effect on users' ability to retain attention, conscious awareness of the real world, perception of experiencing the content, enjoyment of the experience, engagement of senses, and naturalness of interaction. Specifically, high-order ambisonics condition consistently showed the highest mean response for these measures, followed by first-order ambisonics and stereo, while no sound condition consistently had the lowest mean response. These findings suggest that high-quality spatial audio can significantly enhance users' immersive experiences in virtual reality.

8. CONCLUSION AND FUTURE WORK

Visual attention models are crucial in understanding human behavior in various environments, but the lack of proper datasets makes it a challenging process. This research work presents a detailed analysis of a dataset collected for studying visual attention in 360° videos. In this study, participants were asked to watch various 360° videos shot in indoor and outdoor environments, with different sound conditions, and multi-modal data comprising of gaze, pupil diameter, heart rate, pitch, and yaw were recorded. The dataset was analyzed using data visualization. In conclusion, the study found that viewers exhibited different viewing patterns and physiological responses when watching 360° videos with different sound conditions. Viewers had higher heart rate and pupil dilation when watching videos with third-order ambisonics sound. Additionally, pose and gaze fixations were more evenly distributed when watching videos with spatial audio. Furthermore, the high-order sound condition was rated as the most realistic and clearest sound condition, indicating its superiority over stereo and first-order sound. Also, the complexity and richness of the sound had an impact on the participants' heart rate, with the high-order sound condition resulting in the highest heart rate values, followed by the stereo and first-order sound conditions. These findings have significant implications for the development of techniques to optimize processing, encoding, distributing, and rendering content in VR and 360° videos, ultimately improving the overall user experience. Incorporating spatial audio with high-order ambisonics can enhance visual attention and user engagement in immersive experiences, and should be considered as a valuable tool in VR and 360° videos. The available dataset used for this research is expected to be a valuable resource for researchers in the field of VR and 360° video.

In our future work, we plan to conduct joint correlation analysis to investigate the interrelationships and dependencies among participants' head pose, pupil diameter, eye gaze, and heart rate data in immersive experiences. By analyzing the joint correlations, one can determine if there are any patterns or trends in the data that might not be immediately obvious from examining individual correlations alone. In the context of this research, it can provide valuable insights into viewers' behavior and physiological responses in immersive experiences, which can be used to optimize the processing, encoding, distribution, and rendering of VR and 360° videos. By identifying the factors that contribute to viewers' engagement and attention, this analysis can inform the development of more effective and engaging immersive experiences.

ACKNOWLEDGMENTS

This research is supported by Science Foundation Ireland and the ADAPT Centre under Grant Number 12/RC/2106.

REFERENCES

[1] M. I. M. Shojib, M. S. Kaiser, S. M. R. Islam, and A. Al Mahmud, "360-Degree Video Streaming: State-of-the-Art and Future Directions," IEEE Access, vol. 9, pp. 30685-30708, 2021.

[2] Y. Huang and R. Wang, "Visual Attention Modeling for Virtual Reality Videos," IEEE Transactions on Multimedia, vol. 22, no. 7, pp. 1711-1725, 2020.

[3] J. Li and M. D. Levine, "Visual Attention in Videos," in Proceedings of the 24th ACM International Conference on Multimedia, 2016, pp. 1356-1365.

[4] W. Xiong et al., "Visual Attention Driven Adaptive 360-Degree Video Streaming System for Virtual Reality," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 2, pp. 336-349, 2020.

[5] J. Kim, J. Han, and S. Lee, "Effects of 3D Audio on Presence, Emotional Responses, and Behavioral Intentions in Virtual Reality," Multimodal Technologies Interact., vol. 4, no. 4, p. 80, 2020.

[6] X. Wang et al., "A Benchmark Dataset and Evaluation for Non-360° Spatial Audio Object Localisation," IEEE Access, vol. 7, pp. 32983-32996, 2019.

[7] A. Rasheed et al., "Visual Attention and User Engagement in a 360 Video Environment," in Proceedings of the 2018 International Conference on Cyberworlds, 2018, pp. 9-16.

[8] A. Xylakis, S. Mellado, and M. S. Nixon, "Immersive Audio Rendering Techniques for Virtual Reality: A Survey," ACM Trans. Multimedia Comput. Commun. Appl., vol. 16, no. 1s, pp. 1-23, 2020.

[9] L. Lee et al., "The Effects of Spatial Audio on Visual Attention and Memory in Virtual Reality," in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1-13.

[10] M. Garau et al., "Spatial Audio for Virtual Reality: An Overview," in Proceedings of the 14th Sound and Music Computing Conference, 2017, pp. 1-8.

[11] G. P. Innes and J. H. W. Shellard, "Ambisonic reproduction of 3D soundfields," Journal of the Audio Engineering Society, vol. 53, no. 11, pp. 1022-1046, Nov. 2005.

[12] M. A. Gerzon, "Ambisonics in Multichannel Broadcasting and Video," in Audio Engineering Society Convention 102, Munich, Germany, 1997.

[13] A. Politis, S. Siltanen, and V. Pulkki, "Higher-Order Ambisonics," in Immersive Sound, Berlin, Germany: Springer, 2018, pp. 51-102.

[14] A. Raake, "Quality of Experience: What it is and Why it Matters," in Understanding and Improving Quality of Experience in Multimedia Communications, Springer, 2014, pp. 1-15.

[15] H. Xue, et al., "Subjective Quality Assessment of Video: A Tutorial Review," IEEE Signal Processing Magazine, vol. 28, no. 6, pp. 29-48, 2011.

[16] K. Singh and K. Kapoor, "A Study on Quality of Experience and Quality of Service of YouTube Videos on Mobile Devices," Wireless Personal Communications, vol. 102, no. 3, pp. 2253-2273, 2018.

[17] S. Z. Li, et al., "User Expectation-Aware Mobile Video Streaming," IEEE Journal on Selected Areas in Communications, vol. 34, no. 4, pp. 982-992, 2016.

[18] R. Wang and W. Li, "Objective and Subjective Quality Assessment of 360-Degree Video Streaming: A Survey," IEEE Communications Surveys & Tutorials, vol. 23, no. 1, pp. 416-445, 2021.

[19] A. T. M. Shahriar and N. Shiratori, "Assessing Users' Quality of Experience with Smartphone Applications: An Empirical Study," in Proceedings of the 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, Liverpool, UK, 2015, pp. 52-59.

[20] N. Ahmed, et al., "Quality of Experience-Aware Dynamic Resource Allocation for Cloud Gaming," IEEE Transactions on Cloud Computing, vol. 7, no. 1, pp. 239-252, 2019.

[21] Y. Wang, et al., "Quality of Experience (QoE) for Video Streaming: A Review," in Proceedings of the 2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Chengdu, China, 2016, pp. 248-255.

[22] Amit Hirway, Yuansong Qiao, and Niall Murray. 2022. Spatial audio in 360° videos: does it influence visual attention? In Proceedings of the 13th ACM Multimedia Systems Conference (MMSys '22). Association for Computing Machinery, New York, NY, USA, 39–51.

[23] Lo, Wen-Chih & Fan, Ching-Ling & Lee, Jean & Huang, Chun-Ying & Chen, Kuan-Ta & Hsu, Cheng-Hsin. (2017). 360° Video Viewing Dataset in Head-Mounted Virtual Reality. 211-216.

[24] David, Erwan & Gutiérrez, Jesús & Coutrot, Antoine & Perreira Da Silva, Matthieu & Le Callet, Patrick. (2018). A dataset of head and eye movements for 360° videos. 432-437.

[25] Min, Xiongkuo & Zhai, Guangtao & Gao, Zhongpai & Hu, Chunjia & Yang, Xiaokang. (2014). Sound influences visual attention discriminately in videos. 2014 6th International Workshop on Quality of Multimedia Experience, QoMEX 2014. 153-158.

[26] Marighetto, Pierre & Coutrot, Antoine & Riche, Nicolas & Guyader, Nathalie & Mancas, Matei & Gosselin, Bernard & Laganiere, Robert. (2017). Audio-Visual Attention: Eye-Tracking Dataset and Analysis Toolbox.

[27] Wu, Chenglei & Tan, Zhihao & Wang, Zhi & Yang, Shiqiang. (2017). A Dataset for Exploring User Behaviors in VR Spherical Video Streaming. 193-198.

[28] Almquist, Mathias and Viktor Almquist, "Analysis of 360° Video Viewing Behaviours", (2018).

[29] ISO 8589:2007 Sensory analysis — General guidance for the design of test rooms," International Standards Organization, [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:8589:ed-2:v1:en.

[30] Tobii Pro. 2018. Tobii Pro VR Integration – based on HTC Vive Development Kit Description. [ONLINE] Available at: https://www.tobiipro.com/siteassets/tobii-pro/product-descriptions/ tobii-pro-vr-integration-product-description.pdf/?v=1.7. [Accessed 27 March 2023].

[31] Beyerdynamic. 2020. Beyerdynamic DT990 Pro. [ONLINE] Available at: https://europe.beyerdynamic.com/dt-990-pro.html. [Accessed 27 March 2023].

[32] GoPro. 2020. GoPro VR Player for Desktop FAQ. [ONLINE] Available at: https://gopro.com/help/articles/block/gopro-vr-player-for-desktop-faq. [Accessed 27 March 2023].

[33] Angelo Farina. 2020. Index of /Public. [ONLINE] Available at: http://www.angelofarina.it/Public/. [Accessed 7 March 2021].

[34] Empatica. "E4 wristband support page." Available: https://support.empatica.com/hc/en-us/categories/200023126-E4-wristband. [Accessed 27 March 2023]

[35] Ffmpeg.org. 2021. FFmpeg. [online] Available at: <https://ffmpeg.org/> [Accessed 27 March 2023].

[36] Keighrey, Conor & Flynn, Ronan & Murray, Siobhan & Murray, Niall. (2017). A QoE Evaluation of Immersive Augmented and Virtual Reality Speech & Language Assessment Applications. 10.1109/QoMEX.2017.7965656.

[37] Hynes, Eoghan & Flynn, Ronan & Lee, Brian & Murray, Niall. (2019). A Quality of Experience Evaluation Comparing Augmented Reality and Paper Based Instruction for Complex Task Assistance. 1-6.

[38] Egan, Darragh & Brennan, Sean & Barrett, John & Qiao, Yuansong & Timmerer, Christian & Murray, Niall. (2016). An evaluation of Heart Rate and ElectroDermal Activity as an objective QoE evaluation method for immersive virtual reality environments. 1-6.

[39] International Telecommunications Union. 2016. P.913: Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment. [ONLINE] Available at: https://www.itu.int/rec/T-REC-P.913/en. [Accessed 27 March 2023].

[40] Provisu.ch. 2021. [online] Available at: <https://www.provisu.ch/images/PDF/Snellenchart_en.pdf> [Accessed 27 March 2023].

[41] Colblindor. 2021. [online] Available at: <https://www.color-blindness.com/ishiharas-test-for-colour-eficiency-38-plates-edition/> [Accessed 27 March 2023].

[42] D. Pigeon, "Online Hearing Test & Audiogram Printout", Hearingtest.online, 2020. [Online]. Available: https://hearingtest.online/. [Accessed: 27 March 2023].

[43] Poeschl-Guenther, Sandra & Wall, Konstantin & Döring, Nicola. (2013). Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence. Proceedings - IEEE Virtual Reality. 129-130.

[44] J. M. Rigby, S. J. J. Gould, D. P. Brumby, and A. L. Cox, Development of a questionnaire to measure immersion in video media: The Film IEQ, TVX 2019 - Proc. 2019 ACM Int. Conf. Interact. Exp. TV Online Video, pp. 35–46, 2019.

[45] U. C. Lab, “Sheet_PRESENCE QUESTIONNAIRE(PQ),” 2004.

[46] Schreiber, K.M., & Hillis, J.M. (2017). Effects of eye dominance on binocular rivalry with continuous flash suppression. Vision Research, 130, 76-86.

[47] Chen, X., Zhou, X., Xu, P., Xu, K., & Liu, Y. (2020). Gaze behavior and cognitive processing in 360-degree virtual reality videos: An eye-tracking study. Journal of Visual Communication and Image Representation, 73, 102778.

[48] Bruder, G., Steinicke, F., Ritter, H., & Hinrichs, K. (2013). Pupil dilation indicates cognitive processing load during 360° video playback in head-mounted displays. In Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology (pp. 43-50).

[49] Sugano, Y., Matsushita, Y., & Okabe, T. (2016). Self-calibrating eye tracking for free-viewpoint head-mounted display using egocentric visual attention analysis. ACM Transactions on Graphics, 35(6), 192.

[50] Larsson, P., Västfjäll, D., & Kleiner, M. (2017). Gaze behavior in stereoscopic and 360-degree video. Journal of Eye Movement Research, 10(1), 1-12