IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 30, 2021, accepted August 16, 2021, date of publication August 27, 2021, date of current version September 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3108446

The Role of Physiological Responses in a VR-Based Sound Localization Task

ADRIELLE N. MORAES[1], RONAN FLYNN[1], (Member, IEEE), ANDREW HINES[2], (Senior Member, IEEE), AND NIALL MURRAY[1], (Member, IEEE)

[1]Department of Computer and Software Engineering, Athlone Institute of Technology, Athlone, N37 HD68 Ireland
[2]School of Computer Science, University College Dublin, Dublin 4, D04 V1W8 Ireland

Corresponding author: Adrielle N. Moraes (a.nmoraes@research.ait.ie)

This work was supported in part by the Science Foundation Ireland through the ADAPT Centre and the European Regional Development Fund under Grant 12/RC/2106.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Athlone Institute of Technology Research Ethics Committee under Application No. 20190602, and performed in line with the Declaration of Helsinki.

ABSTRACT Virtual reality (VR) has recently emerged as a platform that can be employed in the context of e-health applications. Even though the majority of VR-based applications focus on visual stimuli as the main content, audio also plays a very important role. If someone has an issue with auditory processing, the comprehension of auditory information is compromised. This condition reflects negatively on one's quality of life, given the impact of audio perception on one's ability to communicate effectively or to differentiate and ignore ''noise''. This work aims to design a VR application that can be used to: (a) optimize multimodal VR experiences based on user trials (b) extract user data continuously throughout the experiment (c) evaluate a user's auditory processing ability. To accomplish this goal, participants are required to localise a sound source in space in the presence of multiple listening conditions with simple and complex sound stimuli configurations. Data collected from users consists of physiological and objective metrics. The results of this study highlight the relationship between user behaviour (head movement, fixation points) and performance in the sound localisation task. This information can be used to design future applications with the purpose of training one's auditory localisation ability. In addition, the evaluation compared the impact of using two interaction methods to perform this task: using a pointer or eye gaze to indicate the location of the target source. The findings from this study show statistically significant differences in terms of physiological response when subjects are exposed to different interaction methods, with greater immersion and performance for the pointer group.

INDEX TERMS Quality of experience, spatial audio, virtual reality.

I. INTRODUCTION

Virtual reality (VR) is a tool to present computer-generated content to its users. It aims to be interactive, providing a sense of being immersed in a virtual environment [1]. Recent advances in immersive media experience [2] technologies have reduced costs in terms of hardware and software [3]. Consequently, the use of VR applications has increased over the last few years.

VR experiences are multimodal by their nature, and are dependent on a virtual environment, which contains all of the content and objects to be displayed. The perception of these environments objects is given via multiple sensory pathways (e. g. visual, auditory, sensory), as they share the same properties as real objects such as shape, colour, texture and temperature [4].

Another factor to be noted is the interaction by a user with the environment and the objects therein. Such interactions are usually manipulated via controllers and interfaces. Therefore, the evaluation of the available interaction methods becomes important when immersed in a VR environment, enhancing the sense of presence [5], [6].

Although VR applications are often associated with the entertainment industry, there are a number of examples of applications of VR systems in education, manufacturing and health domains. For instance, VR-based education is a

The associate editor coordinating the review of this manuscript and approving it for publication was Chin-Feng Lai.

popular use case for VR applications [7]. In these scenarios, physiological responses can be measured to infer emotional response related to cognitive load and moments of insight [8], [9]. Examples of where these e-health applications are used include the diagnosis, assessment, or treatment of multiple disorders (panic, generalised anxiety, autism spectrum) [10], [11]. The increased interest in VR for the development of e-health applications is related to its flexibility and adaptability in generating a virtual world [12]. Using VR technology, it is possible to simulate many different scenarios in a safe and repeatable manner that includes interaction and high quality audio-visual content.

When implementing these VR experiences, spatialised audio plays a significant role [13]. However, people with central auditory processing disorders (CAPDs) are unable to process sounds properly [14]. This condition affects sound perception, impacting negatively on one's quality of life. Many daily functions require the ability to process complex audio stimuli. The most relevant function is related to oral communication, a scenario when one must focus one's attention on a person and filter the received speech, ignoring any other noise or distractors. Also, spatialised audio signals provide auditory feedback, which is fundamental for interacting with the surrounding environment.

As mentioned, VR is a powerful tool when it comes to simulate immersive audio-visual in the context of e-health applications. However, the evaluation of these technologies is crucial to increase the quality of experience (QoE) of the user [15]. QoE is related to the level of user satisfaction with an application or service [16]. Furthermore, it can be expressed with explicit and implicit metrics. Implicit objective metrics include collecting physiological data (i.e. heart rate, electrodermal activity, temperature) whilst explicit subjective metrics are usually obtained via questionnaire responses [17]. For this reason, the QoE framework can be applied to understand how users interact with the environment and learn how to perform the required task, optimizing the assessment protocol.

Although VR shows great promise in terms of presentation of the stimuli and enabling user interaction with the system, the majority of the applications designed to assess spatial auditory abilities are based only on performance metrics. This approach leaves a considerable gap in terms of amount of data that can be collected from users when performing this task. Therefore, this paper presents a VR application designed to assess the spatial auditory abilities of a listener. The experimental design is inspired by the LISN test [18] and contains 3 different listening conditions. Each condition presents a different challenge for the listener in terms of the complexity of the audio stimuli. The novelty of the work presented is an extensive analysis of the collected data from the listener. Data includes psychophysiological responses which can be used to analyse other relevant metrics to the experience besides performance, such as cognitive load and user behaviour. To achieve this goal, the QoE framework is applied to continuously assess user state throughout the experiment. Furthermore, this work compares and evaluates two different interaction methods used by a listener to correctly locate the sound source. The first method is based on using a pointer and the second one is based on eye gaze. This analysis gives an insight into which method can improve user performance and provide a more natural interaction in a VR environment.

II. AUDITORY LOCALISATION ABILITY

Our ability to perceive sounds in space is fundamental when interacting with objects and other people, both in the real or virtual world [19]. When listening to any sound, the human brain tries to answer two questions:

- *Identification:* What is this sound?
- *Localisation:* Where is this sound located?

Even though this may appear as a simple task, it requires high-level processing by the human brain [20]. Since it is an ability vital for survival, the human species has developed dedicated neural circuits to identify and localise sounds as fast and accurate as possible [20]. However, some challenges arise when localising a sound source. The most common is related to the nature of the sound wave. The sound that reaches the ear is a sum of all audio sources that surrounds an individual. This combination of multiple sound sources in different locations leads to different acoustic patterns at the two ears. Consequently, each sound source can be described by its spectral content and its directional information [20].

The auditory system is then responsible for classifying each sound source in the environment, and it completes this task based on a top-down approach [21]. Rather than identifying each sound source in the environment, the human brain sets up a priority list, targeting only the most relevant sounds. A similar approach is used when performing a visual searching task. For the sound-based task, an attentional filter is applied to identify target sounds in the middle of competing sounds by answering the following:

- What are the most relevant sound sources?
- What can be classified as a target or distracter sound?
- What is considered to be background noise?

Advances in technology have made it possible to provide acoustic information to answer these questions and to reproduce 3D audio efficiently through headphones and speakers [22]. This allows for the development of immersive virtual reality applications to accurately simulate real-world scenarios [23]. Within this context, it is possible to perform sound localisation and obtain good results, applying mathematical functions to present the audio stimuli, including height and depth information.

A. RENDERING SPATIAL AUDIO IN VR

In order to perform sound localisation, the brain uses the properties of the sound and the anatomy of the human body. Therefore, three main cues are used to model this process [24]. Firstly, the difference in the intensity of the sound, the interaural level difference (ILD), reaching each ear, which is dominant for frequencies below 1.5KHz. The second cue

is the time difference between the arrival of the stimuli from one ear to another, the interaural time difference (ITD), which is dominant for frequencies above 1.5kHz. Both the ILD and the ITD are dominant to recognise the azimuth angle of the sound source [23]. Finally, to discriminate the elevation of a sound, the human brain uses the spectral cues encoded in the shapes of our pinnae, the shape of the head, and the acoustical characteristics of the environment [25].

While performing spatial listening, some people may experience front-back confusion, or up-down confusion, when they hear a stimulus symmetric with respect to the axis of the ears to its actual position in space [26]. In those cases, both the ITD and the ILD assume the same values. This phenomenon is called the *cone of confusion*, where humans are not able to distinguish the location of perceived sound sources [25]. However, listeners with normal hearing can distinguish two different sound sources with a minimum distance of 5 degrees in azimuth or horizontal planes between each other [4]. This process is a result of a phenomenon called binaural listening [27]. Binaural audio techniques led to the development of 3D auditory environments that can be simulated through headphones, giving a realistic perception of the auditory space [5]. This auditory environment can render a series of auditory events, giving insight into a range of listening conditions, from simple to complex audio stimuli.

B. CHALLENGES WHEN LISTENING TO SPATIAL AUDIO

Auditory processing disorders affect how the sound is perceived and how it is processed by the brain [28]. As mentioned in Section I, CAPD is a term used to describe dysfunctions in how sound information is processed at a central nervous system level [29]. People who suffer from this disorder have normal peripheral hearing, but they cannot process complex sounds. Additionally, their speech reception threshold (SRT) is higher, meaning that perceived audio has a lower signal-to-noise ratio (SNR) that gives rise to the following [14]:

- difficulties hearing in noisy or reverberant environments.
- difficulties following long conversations.
- difficulties learning other languages or vocabulary.
- difficulties in directing and sustaining attention.
- auditory memory deficits.
- spelling and reading difficulties.

CAPD may be a result of a post-traumatic brain injury [30], stroke [9], ASD or due to other congenital reasons [31]. Tests designed to evaluate this condition assess the brain's ability to process audio with multiple concurrent auditory events [32].

In order to evaluate auditory processing ability, the benchmark is the **L**isten **I**n **S**patialized **N**oise (LISN) test, developed by [33] to be a 3D auditory environment simulated under headphones. This test consists of an adaptive speech test with a target stimulus combined with noise sentences. The main objective is to focus on the target sound, ignoring the background noise.

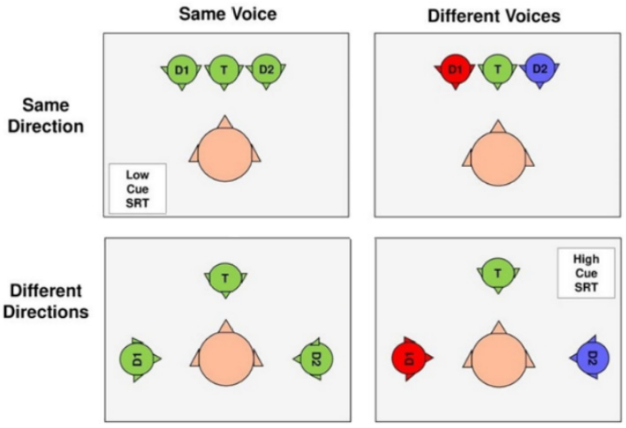

FIGURE 1. LISN test conditions. The listener has to identify the target sound (T) while ignoring other distractors (D1 and D2). The speech reception threshold (SRT) is directly related to the level of noise in speech (Adapted from [34]).

The LISN test contains four conditions as illustrated in Fig. 1. The bottom-right condition embraces the total advantage, meaning that the listener receives cues from the different positions of the distractions and target, as well as the voice difference of each source. The top-right shows the talker advantage condition, with all sound sources localized in the same position, each with a different voice stimulus. The bottom-left illustrates the spatial advantage, meaning that each stimulus is localized in a different position in space, but each with the same voice stimulus. The top-left represents the low cue speech reception threshold, given by the similarity of voice stimulus and location of the target and the distractions.

III. RELATED WORK

Within the health domain, the ability to process audio may be affected by several conditions like autism spectrum disorder (ASD) or attention deficit hyperactivity disorder (ADHD). However, a review of the literature suggests that investigations of the effect of spatialised audio in e-health applications is very limited. How the auditory difficulties experienced by subjects diagnosed with ASD affected their interaction with a VR therapy application was the focus of [7]. For this purpose, 29 participants were exposed to a virtual scenario consisting of an enchanted forest at night. The participants were free to move around a pre-defined tracked space while wearing a head mounted device (HMD). The first phase of the experiment tested for spatial audio attention using 8 different non-speech sound sources with no background noise. Participants were instructed to identify the location of each sound source by walking to the place they perceived to be the location of the sound. The second phase analysed sound source localisation in the presence of competing background noise. In this case, 8 different speech utterances were used as sound sources. Participants performed the same localisation strategy as in the first phase. Although the experimental design included an immersive virtual scenario and competing sounds, only the performance data was used to measure the

behavioural response to spatial audio. Such issues highlight the importance of applying a QoE framework to assess user behaviour during the test. Therefore, other metrics intrinsic to a localisation task can be analysed, like head rotation or gaze data.

A game to integrate therapy techniques into game mechanics for children diagnosed with ASD was developed in [35]. Individuals with ASD usually experience auditory hypersensitivity, which raises levels of anxiety and other fear responses. Therefore, the focus of [35] was to reduce those feelings of fear and anxiety, exposing subjects to averse binaural-based spatial audio. Results from this study indicated a decreased level of reported anxiety levels after 4 weeks. Also, participants reported a positive response in terms of engagement with the system.

Both [7] and [35] demonstrated positive results for the assessment of auditory disorders in VR. However, other factors that influence user behaviour were not deeply explored. The perceived QoE of multimedia applications is influenced by human, system, and context factors [36]. Therefore, understanding these factors is crucial to the success of a VR application [15]. For this reason, it is important to evaluate participants continuously while they are still performing the experiment's task. Implicit metrics, to help assess the QoE, can be gathered efficiently. This approach allows fill in the gaps resulted from the lack of participant data when using only reported information. Such implicit metrics include physiological responses such as eye gaze [37], electrodermal activity (EDA) [38], heart rate (HR) [39], and electroencephalography (EEG) [40]. The use of implicit metrics complements explicit metrics, which are usually gathered through post-experience questionnaires, to build a complete assessment of the QoE.

IV. METHODS

A. SYSTEM

The virtual environment was designed using the Unity game engine (version 2018.2.15f1) [41]. The headset to present the VR environment was the HTC VIVE with integrated Tobii eye-tracking [42]. This headset has a resolution of $1,080 \times 1.200$ pixels per eye, a field of view of 110°, and a refresh rate of 90Hz. To facilitate the localisation task, a wireless adaptor was attached to the headset. This configuration provides listeners with a natural interaction with the system, with 6 degrees of freedom (DoF).

Before the experiment begins, the headset initial position was calibrated with the same reference point for all participants. Therefore, all visual stimuli appeared at eye-level and with the same depth. Since this study investigated the effects of gaze data in a sound localisation task, the headset lenses were adjusted for each participant, ensuring that both visual content and data collection were performed correctly.

The Steam Audio plugin [43] for Unity 3D was used to render audio. It is a tool developed by Valve that has a software integration with Unity 3D. It is a head-related transfer function (HRTF) based binaural rendering tool that provides high quality 3D audio. In addition, it is also physics-based audio, with rotational and positional tracking to create immersive VR experiences. In order to standardise the listening experience for all users, a generic HRTF was used. The audio was reproduced by Beyerdynamic DT 990 PRO Studio Headphones [44], with diffuse-field equalisation.

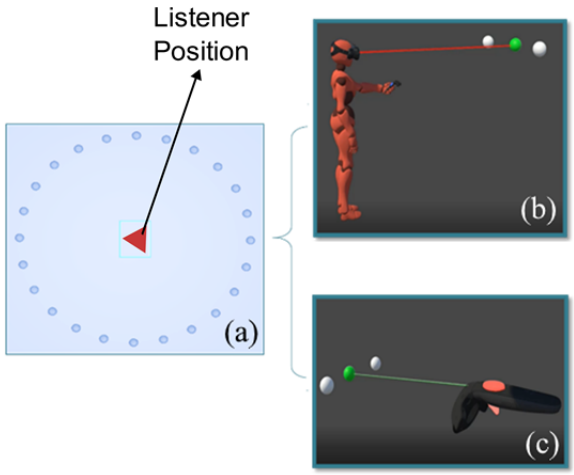

FIGURE 2. Top view of the virtual environment (a). Each sphere represents a possible sound source. Images on the right represent interaction methods for this experiment. The top one (b) is pointing with gaze and the bottom one (c) is pointing with the controller.

The virtual environment, as shown in Fig. 2(a) consisted of an open field with 24 white spheres surrounding the listener. All spheres are equally spaced from each other by a distance of 15 degrees. Additionally, the participant is positioned at the centre of the circle of spheres, with a distance (depth) of 10m between them. This configuration allows evaluating the auditory localisation within the human ear resolution.

In this work, two interaction methods were used to evaluate user interaction with the system. Each participant experienced only one of the pointing methods, which were:

Gaze Pointing (GP): this interaction is through selection with eye gaze and a controller button-click to confirm the selection and is displayed in Fig. 2(b).

Pointer Pointing (PP): this interaction is though selection using a laser pointer triggered by the listener and a button-click to confirm the selection. This interaction method is displayed in Fig. 2(c).

B. EXPERIMENTAL METHODOLOGY

The experimental method applied in this research is inspired by other QoE studies [2], [31] containing a screening and a tutorial phase before the localisation task begins. In the first phase participants were given a consent form and an information sheet with a description of the study and its objectives. Once the consent form was signed, users were ready to begin the screening phase, which consisted of two steps:

Checking for Hearing Impairments: since this study is related to auditory abilities, participants were required to pass a hearing test. For this experiment, it was an online tool

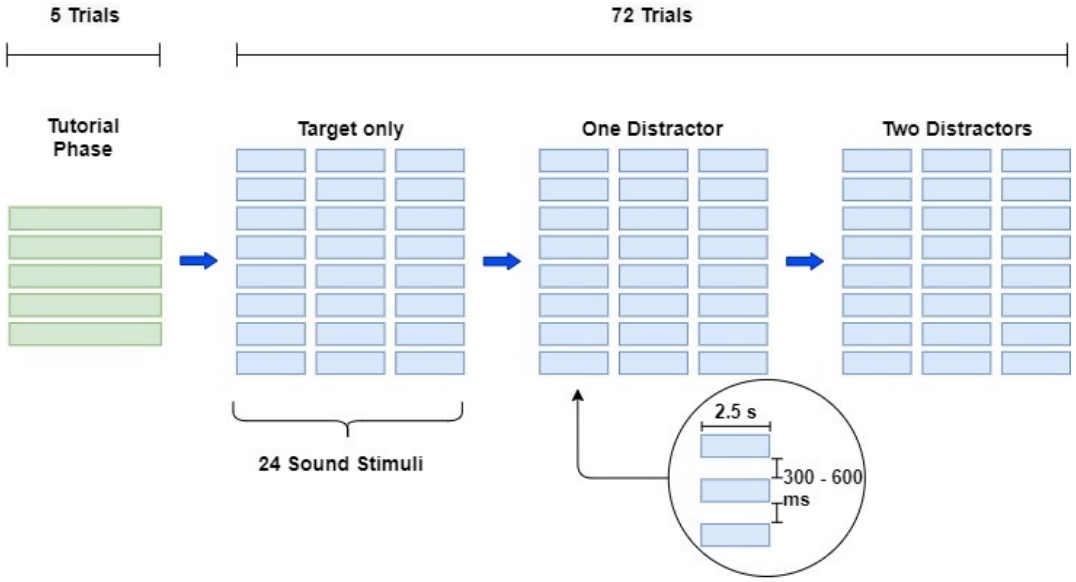

FIGURE 3. Experiment protocol.

developed by WIDEX [46] and participants took an average of 3 minutes to complete this test. It aimed to determine how well a participant could hear different frequency levels, asking users to adjust the volume of the presented pitch/tone until an audible level was reached.

Checking for Visual Impairments: participants were also screened for visual impairments using a Snellen chart placed 6 meters away from the participant. During this test, subjects covered one eye and read all letters row by row, repeating the procedure for the other eye. A score of 20/20 was required form all participants to pass. The final screening process was an Ishihara test for colour blindness. A set of coloured dotted plates were presented to participants, showing a number or a path. Subjects with normal vision are able to see and report the correct information displayed on the plate.

The screening phase was followed by the baseline phase. During this step, participants were asked to wear the E4 Empatica wristband, which started the acquisition of physiological data. Then, participants were left alone for 5 minutes to collect baseline data, which was used to perform comparisons between user states throughout the experiment. The data acquired with this device was continuously collected throughout the experiment.

After the baseline data was acquired, participants were instructed on how to interact with the environment. Two tutorial videos were prepared for this purpose, one for each interaction method. Also, there was a demonstration of the sound stimuli, reproducing the target stimulus for 2.5 seconds followed by a reproduction of the distractor stimulus for 2.5 seconds. At this stage, the participants were ready to start the listening task.

The room was equipped with a swivel chair, providing a 6 DoF, and a wireless adaptor for the head-mounted display and headphones. The first scene was the same one that participants interact with until the end of the testing phase. However, a tutorial scene was first presented to guarantee that users would be able to operate the system since it may be the first time experiencing a VR environment for some of the participants.

The sound stimuli was a 2.5s-white-noise presented at 55dB sound pressure level (SPL). The competing sound stimulus consisted of a low-frequency, 1500 Hz pure tone sine wave presented at the same SPL as the target stimulus. This frequency value represents the cut off frequency which best describes the interaural level difference and interaural time different cues.

The test consisted of three main phases, inspired by the Listen and Learn in Spatialised Noise test (LiSN) [34]. Each phase represented a different advantage for the listener in terms of spatial location and target discrimination:

1. Target-only
2. Target and Distracter presented at the same location.
3. Target and Two distracters presented $+/-90$ degrees apart from the target.

Fig. 3 contains a diagram illustrating the experiment protocol. It consists of four blocks: tutorial phase, target-only phase, one distractor phase, and two distractors phase.

During the tutorial phase, the participants were asked to select different spheres in the environment based on their colour. A random sphere was highlighted in green and users had to identify its location in the virtual environment and make a selection based on their interaction method (Gaze Pointing or Pointer Pointing). This step phase was repeated a minimum of 5 times and there were no time limit to complete this part of the experiment.

The other 3 phases are related to the sound localisation task. Each phase had a total of 24 sound stimulus presented one at a time. Participants were instructed to make a selection as fast as they could, but there was no time limit to select the location of the sound source after the stimulus was presented. However, after a selection was made, there was a small pause of a random value between 300 and 600 ms.

In total, participants attempted to select a sphere 72 times (24 for each phase). To identify front-back confusion and localisation blur type of errors, the position of the sound stimulus follows the pattern described below:

1. The first stimulus was always presented at a random position amongst the 24 possible spawn points.
2. The next stimulus had 3 possible future locations, randomly applied, and based on the location of the previous target location:
   - At the opposite side of the previous target location (180° apart)
   - To the right of the target location (15° apart clockwise)
   - To the left of the target location (15° apart counterclockwise)

It is important to note that each sphere reproduced a sound only once per phase. For this reason, the algorithm selected a random sphere if none of the possibilities described in step 2 was available.

C. PARTICIPANTS

A total of twenty subjects took part in the experiment (10 for each interaction group). The average age was 29 years, with 8 female and 12 male subjects. Six of the participants had never experienced VR before. All of the participants were healthy subjects and none of them reported any hearing issue.

D. DATA ANALYSIS

This section provides an overview of the data collected during the experiment.:

Gaze and head pose data was collected directly from the built-in eye-tracking and G-sensors in the HMD. Eye tracking data is sampled at a rate of 120Hz containing information on pupil size, gaze origin, gaze direction, and head pose. For the gaze data, a second-order low-pass filter was used to remove any spikes on the signal. All blink artefacts were removed considering a window of 0.1 s before and 0.25s after a blink is detected.

The detection of saccades and fixations is determined by a dispersion-based algorithm [47]. This approach takes as parameters the size of the window and a threshold radius considering the participant's field of view and the distance between virtual objects and the viewer.

Physiological data was acquired using the E4 Empatica wristband [48]. This set of data includes EDA sampled at 4Hz, heart rate sampled at 1Hz, temperature sampled at 4Hz, blood volume pulse (BVP) sampled at 64Hz, and acceleration sampled at 32Hz. Participants' physiological data and pupillary response were normalised in order to compare participants' metrics. This normalisation procedure takes as parameters a participant's maximum and minimum value during the experiment. As a result, normalised data are presented on a scale with values ranging from 0% to 100%.

Performance data was based on a participant's (i) number of selection attempts, (ii) number of correct selections, (iii) informed location of the target, (iv) the real location of the target and (v) time to complete each trial. Each participant's performance during the experiment was classified regarding the angular distance (α) between the indicated position of the sound source and the correct location of the sound source:

- Correct Response: $-10° < \alpha < 10°$
- Localisation Blur: $-45° < \alpha < 45°$
- Front-back confusion: $\alpha < -135°$ and $\alpha > 145°$
- Neither: otherwise

Questionnaire Responses: At the end of the experiment, listeners were asked to fulfill a questionnaire regarding their experience with the application. Participants completed the NASA-TLX questionnaire [49] to assess user workload during the task. It is important to highlight that instructions on how to complete the questionnaire were given to all participants. NASA-TLX questionnaire results were obtained using the weighting scores from the paired-choice and the individual scores for each workload factor.

V. RESULTS AND DISCUSSION

The findings of this research are presented in this section. Since each group contained 10 participants, all statistical analysis was done with a non-parametric Mann-Whitney test, performed with a 95% confidence level.

A. PERFORMANCE

Table 1 shows the number of localisation errors according to the listener's performance for each testing phase. As expected, the usage of a VR application to present stimuli had an impact on the distribution of the localisation type of errors. One of the advantages of using a VR headset is the increased number of degrees of freedom. This type of application allows users to explore the virtual environment, which is not possible when the assessment is employed via traditional headphones-based experiences. However, the majority of the localisation errors were classified as localisation blur. As shown in Table 1, the mean localisation error angle is around 30 degrees for the PP group and 36 degrees for the GP group. However, it is interesting to note that the standard deviation for each group in each phase corresponds to missing the target stimulus by one sphere (15 degrees).

Results from Table 2 are interesting to evaluate each phase separately. During phase 1, there was no difference in localisation errors distribution for both groups. This result is expected as all participants were getting used to the proposed interaction method. However, during phase 2, it was found a significant difference between the number of correct responses for each group ($U = 24.5$, $p = 0.049$).

TABLE 1. Mean localisation error angle for each group.

Group	Phase 1	Phase 2	Phase 3
PP	36.3 ± 11.5	26.9 ± 10.3	31.1 ± 13.5
GP	34.1 ± 11.1	35.4 ± 11.5	41.9 ± 14.1

TABLE 2. Distribution of localisation error types for each group.

	Group	Phase 1	Phase 2	Phase 3
Correct	PP	8.50 ± 3.89	9.67 ± 2.05*	8.67 ± 3.91*
	GP	5.90 ± 4.58	5.70 ± 2.58*	5.40 ± 4.40*
Blur	PP	10.88 ± 2.90	12.00 ± 1.15	10.22 ± 3.73
	GP	11.90 ± 3.60	11.50 ± 1.86	10.80 ± 3.65
Front-back confusion	PP	1.50 ± 0.84	3.00 ± 0.71	2.00 ± 1.41
	GP	2.67 ± 2.25	2.83 ± 0.97	3.38 ± 2.00
Neither	PP	4.00 ± 4.28	3.50 ± 1.50	4.75 ± 4.50
	GP	5.11 ± 3.06	6.25 ± 1.06	5.56 ± 2.83

* $p < 0.05$

During phase 3, the PP continues to score better than the GP group ($U = 24.5$, $p = 0.05$). This result indicates that the interaction method with the environment affects the number of correct responses. The same can be observed when looking at Table 1. During phase 2 and 3, the mean localisation error angle decreases for the PP group but increases for the GP group. This result suggests that even when the task becomes more challenging, the PP group was able to perform better and keep a higher score than the GP group.

Analysing each group separately, we can observe that the average number of correct answers from the GP group does not improve during the test while the number of front-back confusion localisation errors increase when compared with phase 1. Those results indicate that this interaction method may not be optimal for localising sound sources out of users' field of view, as this interaction is extremely dependent on gaze.

B. IMPLICIT METRICS

Fig. 4 illustrates the pupillary response for both interaction groups. It contains data for each sound stimulus, using the average pupil diameter of the interval when the sound starts and when users make a selection. The baseline value was obtained with a time window of the last 5s values of the tutorial phase instead of counting for the entire testing period. The length of this time window represents an approximation of the time participants take to make a selection.

According to [17], the pupillary response is positively correlated with the level of cognitive load, visual attention, and memory. Therefore, increasing values suggest a higher cognitive load level for the task, indicating that the task is more challenging for the user [50].

Before the experiment begins (baseline), users had a small value of the pupillary diameter compared with the rest of the

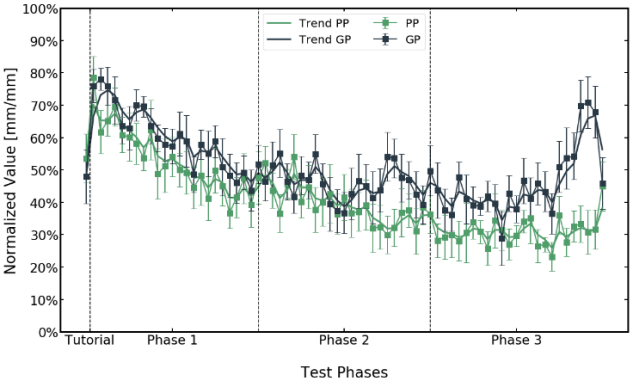

FIGURE 4. Pupillary response for both interaction groups. To evaluate changes, all values were normalised using the participant's highest pupil diameter size. The first data point represents the pupil size before the experiment begins (tutorial).

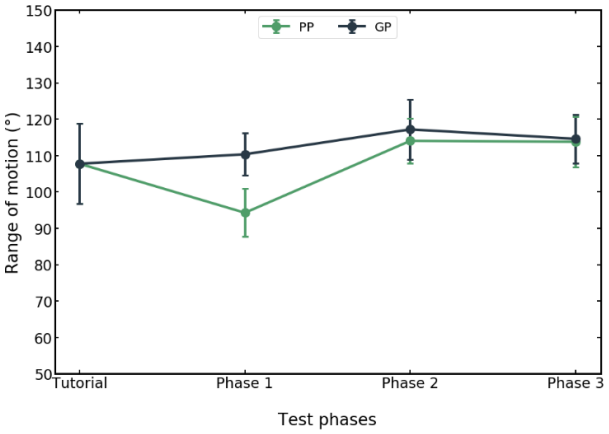

FIGURE 5. Range of motion for head movements for each test phase. Values on this graph are related to the head's yaw rotation.

experiment. After the experiment begins, there is an increase in this value explained by the novelty of the task, and users were still trying to understand how to operate the system. This statement is also supported by the decreasing value during the middle of the task, which indicates that users were getting familiar with the system.

For the GP group, this value increased during phase 3, explained by the difficulty of this phase, which is the most challenging one. For this phase, there is a statistical difference between groups ($U = 10$, $p = 0.007$). This result can also be validated by the performance from both groups, as the GP group had an inferior score when compared to the PP group.

Fig. 5 illustrates the range of motion for head movements for each test phase. It is important to note that performing head movements improves listeners ability to localise sounds on the horizontal plane [51].

For the tutorial phase, the values for both groups remain the same. However, there is a difference in the distribution of the points for the first phase of the test ($U = 146.0$, $p = 0.002$). During this phase, the PP group had a smaller range of motion when compared to the tutorial phase. As expected, participants should increase their range of

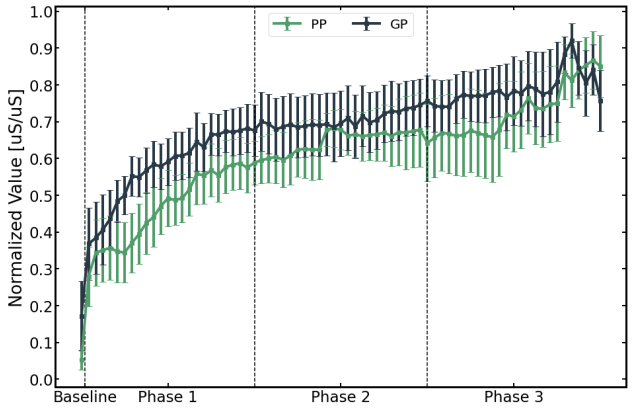

FIGURE 6. Normalised EDA for both groups.

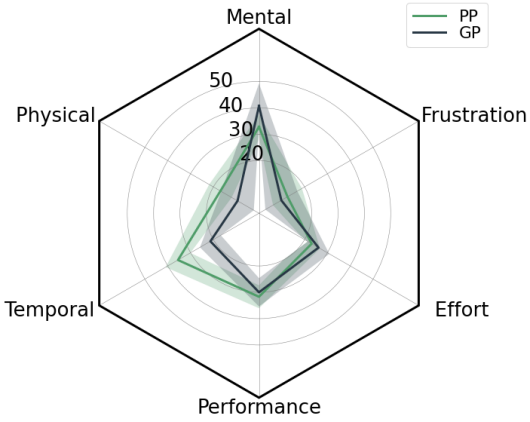

FIGURE 7. NASA-TLX questionnaire responses for both groups.

motion for head movements across test phases. The first reason is to allow users to explore the VR environment. The second is related to the nature of the task. Since users must locate the correct source of the sound stimulus, it was expected that they would rotate their heads to search for the source and indicate a location. However, for the GP group, the range of motion did not increase over the test phases, which can explain the reason why this group's performance did not improve.

Another implicit metric collected during the experiment is the electrodermal activity (EDA). There is no statistical difference between the collected EDA values for both groups during the test. However, it is interesting to note that values for this metric increase over time. EDA is commonly associated with the level of cognitive activity [17]. Localising a sound source requires a high level of processing from the brain, which explains the EDA curve of the first phase when EDA values increased drastically when compared to the baseline values.

C. EXPLICIT METRICS

Explicit metrics were collected in the format of post-experiment questionnaires. Since the sound localisation task is very demanding in terms of cognitive load [20], the NASA-TLX questionnaire was employed to understand the distribution of influence factors on the level of cognitive load.

Fig. 7 summarises the responses to the NASA-TLX questionnaire. Data were weighted considering the contribution of each factor to the total workload of the test. Given that spatialized audio is a complex stimulus, a greater value was expected for both groups when analysing the mental influence factor. Consequently, this factor had the highest value in comparison with other workload factors. The higher value for the mental demand was also observed for the Pupillary response and EDA metrics (Fig. 4 and Fig. 6 respectively), where even though there was no statistical difference between the groups, the GP group always had a higher value in comparison with the PP group.

It is also interesting to note that the performance factor received a higher score than the frustration factor. This result suggests that users had no difficulties in interacting with the virtual environment, and the events were compatible with their expectations.

VI. CONCLUSION

This work explored the analysis of explicit and implicit metrics using a QoE framework for a sound localisation task in VR. It is inspired by the state-of-the-art assessment protocol for spatial auditory abilities, where subjects are exposed to simple and complex audio stimuli. The key objective of this research was to present how metrics apart from traditional experiment designs like performance and reported information can be used to measure cognitive load and give insight into user behaviour.

Findings from this work show that physiological metrics are good indicators of cognitive load and immersion for a sound localisation task. During the experiment, EDA values and pupillary response increased over time, suggesting that the task was challenging and requires a higher mental demand. Those results are validated by the responses obtained with the NASA-TLX questionnaire, with a higher mental demand for users in comparison with other workload factors.

Another interesting finding is related to the relationship between the head movement range of motion and the performance data. When subjects had a higher range of motion, they performed better the localisation task. This was observed for the testing phase with one distractor only, when subjects from the PP group increased the number of correct responses even when the task was more challenging over time. This result highlights the importance of the user interaction with the environment, and it can be used to investigate multiple listening conditions. This is particularly relevant for applications designed to improve one's spatial localisation auditory abilities in which listeners have to rotate their heads to search for the source of the stimuli. In addition, the present study investigates a comparison between two interaction methods. According to experimental results, the PP group performed better in terms of the number of correct selections, which reinforces that the interaction with the virtual environment

has an impact when users are performing a sound localisation task.

In this research, a limited sample size was used and both groups had 10 participants each. As result, only non-parametric statistical tests were applied. Whilst this is a valid approach, future work will extend the sample size, with the objective of validating the results obtained and covering parametric analysis. Additionally, subjects only take part on the experiment once, which affects the analysis on the learning effect. Therefore, future work will also include a broader comparison of the test phases, adding a revalidation procedure to reinforce the obtained results.

REFERENCES

[1] S. E. Kober and C. Neuper, "Using auditory event-related EEG potentials to assess presence in virtual reality," *Int. J. Hum.-Comput. Stud.*, vol. 70, no. 9, pp. 577–587, 2012, doi: 10.1016/j.ijhcs.2012.03.004.
[2] A. Perkis and C. Timmerer, "QUALINET white paper on definitions of immersive media experience (IMEx)," in *Proc. 14th QUALINET Meeting Eur. Netw. Qual. Exper. Multimedia Syst. Services*, May 2020.
[3] C. Keighrey, R. Flynn, S. Murray, S. Brennan, and N. Murray, "Comparing user QoE via physiological and interaction measurements of immersive AR and VR speech and language therapy applications," in *Proc. Thematic Workshops ACM Multimedia Thematic Workshops*, 2017, pp. 485–492, doi: 10.1145/3126686.3126747.
[4] M. Mihelj, D. Novak, and S. Beguš, *Virtual Reality Technology and Applications*, vol. 68. Dordrecht, The Netherlands: Springer, 2014.
[5] C. Zhang, A. S. Hoel, A. Perkis, and S. Zadtootaghaj, "How long is long enough to induce immersion?" in *Proc. 10th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, May 2018, pp. 1–6, doi: 10.1109/QoMEX.2018.8463397.
[6] J. Kreimeier, S. Hammer, D. Friedmann, P. Karg, C. Bühner, L. Bankel, and T. Götzelmann, "Evaluation of different types of haptic feedback influencing the task-based presence and performance in virtual reality," in *Proc. 12th ACM Int. Conf. Pervas. Technol. Rel. Assistive Environ.*, Jun. 2019, pp. 289–298, doi: 10.1145/3316782.3321536.
[7] J. Pottle, "Virtual reality and the transformation of medical education," *Futur. Healthc. J.*, vol. 6, no. 3, p. 181, Oct. 2019, doi: 10.7861/FHJ.2019-0036.
[8] J. Collins, H. Regenbrecht, T. Langlotz, Y. Said Can, C. Ersoy, and R. Butson, "Measuring cognitive load and insight: A methodology exemplified in a virtual reality learning context," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Oct. 2019, pp. 351–362, doi: 10.1109/ISMAR.2019.00033.
[9] I. Horvath, "Evolution of teaching roles and tasks in VR/AR-based education," in *Proc. 9th IEEE Int. Conf. Cognit. Infocommunications (CogInfoCom)*, Aug. 2018, pp. 355–360, doi: 10.1109/CogInfoCom.2018.8639907.
[10] V. Russell, R. Barry, and D. Murphy, "HAVE experience: An investigation into VR empathy for panic disorder," in *Proc. IEEE Games, Entertainment, Media Conf. (GEM)*, Aug. 2018, pp. 167–172.
[11] D. Johnston, H. Egermann, and G. Kearney, "Measuring the behavioral response to spatial audio within a multi-modal virtual reality environment in children with autism spectrum disorder," *Appl. Sci.*, vol. 9, no. 15, p. 3152, Aug. 2019, doi: 10.3390/app9153152.
[12] M. M. E. Hendrikse, G. Llorach, V. Hohmann, and G. Grimm, "Movement and gaze behavior in virtual audiovisual listening environments resembling everyday life," *Trends Hearing*, vol. 23, Jan. 2019, Art. no. 233121651987236, doi: 10.1177/2331216519872362.
[13] E. R. Hoeg, L. J. Gerry, L. Thomsen, N. C. Nilsson, and S. Serafin, "Binaural sound reduces reaction time in a virtual reality search task," in *Proc. IEEE 3rd VR Workshop Sonic Interact. Virtual Environ. (SIVE)*, Mar. 2017, pp. 1–4, doi: 10.1109/SIVE.2017.7901610.
[14] M. D. Barker and S. C. Purdy, "An initial investigation into the validity of a computer-based auditory processing assessment (Feather Squadron)," *Int. J. Audiol.*, vol. 55, no. 3, pp. 173–183, Mar. 2016, doi: 10.3109/14992027.2015.1074734.
[15] C. Keighrey, R. Flynn, S. Murray, and N. Murray, "A physiology-based QoE comparison of interactive augmented reality, virtual reality and tablet-based applications," *IEEE Trans. Multimedia*, vol. 23, pp. 333–341, 2021, doi: 10.1109/tmm.2020.2982046.
[16] J. Pokhrel, N. Kushik, B. Wehbi, N. Yevtushenko, and A. R. Cavalli, "Multimedia quality of experience," in *Emerging Research on Networked Multimedia Communication Systems*. Hershey, PA, USA: IGI Global, 2016, pp. 250–284.
[17] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J. N. Antons, K. Y. Chan, N. Ramzan, and K. Brunnström, "Psychophysiology-based QoE assessment: A survey," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 6–21, Feb. 2017, doi: 10.1109/JSTSP.2016.2609843.
[18] S. Cameron, H. Dillon, and P. Newall, "Development and evaluation of the listening in spatialized noise test," *Ear Hearing*, vol. 27, no. 1, pp. 30–42, Feb. 2006, doi: 10.1097/01.aud.0000194510.57677.03.
[19] L. Pisha, J. Warchall, T. Zubatiy, S. Hamilton, C.-H. Lee, G. Chockalingam, P. P. Mercier, R. Gupta, B. D. Rao, and H. Garudadri, "A wearable, extensible, open-source platform for hearing healthcare research," *IEEE Access*, vol. 7, pp. 162083–162101, 2019, doi: 10.1109/ACCESS.2019.2951145.
[20] J. Van Opstal, *The Auditory System and Human Sound-Localization Behavior*. Amsterdam, The Netherlands: Elsevier, 2016.
[21] U. Pomper and M. Chait, "The impact of visual gaze direction on auditory object tracking," *Sci. Rep.*, vol. 7, no. 1, pp. 1–16, Dec. 2017, doi: 10.1038/s41598-017-04475-1.
[22] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D audio—The new standard for coding of immersive spatial audio," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 5, pp. 770–779, Aug. 2015, doi: 10.1109/JSTSP.2015.2411578.
[23] A. Lecharpentier, P. Vielh, P. Perez-Moreno, D. Planchard, J. C. Soria, and F. Farace, "Detection of circulating tumour cells with a hybrid (epithelial/mesenchymal) phenotype in patients with metastatic non-small cell lung cancer," *Brit. J. Cancer*, vol. 105, no. 9, pp. 1338–1341, Oct. 2011, doi: 10.1038/bjc.2011.405.
[24] J. C. Middlebrooks, *Sound Localization*, vol. 129, 1st ed. Amsterdam, The Netherlands: Elsevier, 2015.
[25] O. Balan, F. Moldoveanu, A. Morar, and V. Asavei, "Experiments on training the sound localization abilities: A systematic review," in *Proc. 10th Int. Sci. Conf. eLearn. Softw. Educ.*, 2014, p. 34.
[26] Y.-H. Wu and A. Roginska. (2019). *Analysis and Training of Human Sound Localization Behavior With VR Application*. [Online]. Available: http://www.aes.org/e-lib
[27] T. Walton, "The overall listening experience of binaural audio," in *Proc. 4th Int. Conf. Spat. Audio*, Sep. 2017, pp. 2–9.
[28] S. Cameron, H. Glyde, and H. Dillon, "Efficacy of the LiSN & learn auditory training software: Randomized blinded controlled study," *Audiology Res.*, vol. 2, no. 1, pp. 86–93, Sep. 2012, doi: 10.4081/audiores.2012.e15.
[29] D. Tomlin and G. Rance, "Maturation of the central auditory nervous system in children with auditory processing disorder," *Semin. Hear.*, vol. 37, no. 212, pp. 74–83, Feb. 2016.
[30] C. F. B. Murphy, G. Stavrinos, K. Chong, T. Sirimanna, and D.-E. Bamiou, "Auditory processing after early left hemisphere injury: A case report," *Frontiers Neurol.*, vol. 8, pp. 1–6, May 2017, doi: 10.3389/fneur.2017.00226.
[31] *Diagnosis, Treatment and Management of Children and Adults With Central Auditory Processing Disorder*, American Academy of Audiology, Reston, VA, USA, 2010.
[32] S. Ratib, D. R. Moore, M. A. Ferguson, A. M. Edmondson-Jones, and A. Riley, "Nature of auditory processing disorder in children," *Pediatrics*, vol. 126, no. 2, pp. 382–390, 2010, doi: 10.1542/peds.2009-2826.
[33] S. Cameron and H. Dillon, "Development of the listening in spatialized noise-sentences test (LISN-S)," *Ear Hearing*, vol. 28, no. 2, pp. 196–211, Apr. 2007, doi: 10.1097/AUD.0b013e318031267f.
[34] D. K. Brown, S. Cameron, J. S. Martin, C. Watson, and H. Dillon, "The north American listening in spatialized noise—Sentences test (NA LiSN-S): Normative data and test-retest reliability studies for adolescents and young adults," *J. Amer. Acad. Audiol.*, vol. 21, no. 10, pp. 629–641, Nov. 2010, doi: 10.3766/jaaa.21.10.3.
[35] D. Johnston, H. Egermann, and G. Kearney, "SoundFields: A virtual reality game designed to address auditory hypersensitivity in individuals with autism spectrum disorder," *Appl. Sci.*, vol. 10, no. 9, p. 2996, Apr. 2020, doi: 10.3390/app10092996.
[36] K. Brunnström *et al.*, "Qualinet white paper on definitions of quality of experience," 2013.

[37] L. Zhang, J. Wade, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Cognitive state measurement from eye gaze analysis in an intelligent virtual reality driving system for autism intervention," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 532–538, doi: 10.1109/ACII.2015.7344621.
[38] A. Drachen and A. L. Pedersen, "Correlation between heart rate, electrodermal activity and player experience in first-person shooter games," *ACM*, vol. 475, no. 3, pp. 121–123, 2010, doi: 10.1016/j.neulet.2010.03.050.
[39] S. Laborde, E. Mosley, and J. F. Thayer, "Heart rate variability and cardiac vagal tone in psychophysiological research—Recommendations for experiment planning, data analysis, and data reporting," *Frontiers Psychol.*, vol. 8, pp. 1–18, Feb. 2017, doi: 10.3389/fpsyg.2017.00213.
[40] J. Fernández-Vargas, T. V. Tarvainen, K. Kita, and W. Yu, "Effects of using virtual reality and virtual avatar on hand motion reconstruction accuracy and brain activity," *IEEE Access*, vol. 5, pp. 23736–23750, 2017, doi: 10.1109/ACCESS.2017.2766174.
[41] Unity Technologies. *Unity 3D*. Accessed: Jan. 27, 2020. [Online]. Available: https://unity.com/
[42] HTC Corporation. *HTC Vive*. Accessed: Jan. 27, 2020. [Online]. Available: https://www.vive.com/us/product/vive-virtual-reality-system/
[43] Valve. *Steam Audio*. Accessed: Nov. 25, 2019. [Online]. Available: https://valvesoftware.github.io/steam-audio/
[44] BeyerDynamic. *DT 990 PRO*. Accessed: Jan. 24, 2020. [Online]. Available: https://europe.beyerdynamic.com/dt-990-pro.html
[45] E. Hynes, R. Flynn, B. Lee, and N. Murray, "A quality of experience evaluation comparing augmented reality and paper based instruction for complex task assistance," in *Proc. IEEE 21st Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2019, pp. 1–6.
[46] Widex. (2020). *Online Hearing Test*. Accessed: Jan. 24, 2020. [Online]. Available: https://global.widex.com/en/online-hearing-test#3
[47] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proc. Symp. Eye tracking Res. Appl. (ETRA)*, 2000, pp. 71–78.
[48] Empatica. *E4 Sensors*. Accessed: Jan. 27, 2020. [Online]. Available: https://www.empatica.com/en-gb/research/e4/
[49] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Adv. Psychol.*, vol. 52, pp. 139–183, Apr. 1988.
[50] G. Porter, T. Troscianko, and I. D. Gilchrist, "Effort during visual search and counting: Insights from pupillometry," *Quart. J. Exp. Psychol.*, vol. 60, no. 2, pp. 211–229, 2007, doi: 10.1080/17470210600673818.
[51] H. Wallach, "The role of head movements and vestibular and visual cues in sound localization," *J. Express Psychol.*, vol. 27, no. 4, pp. 339–368, 1940, doi: 10.1037/h0054629.

ADRIELLE N. MORAES received the B.Sc. degree in biomedical engineering from the Federal University of Uberlandia, Brazil, in 2018. She is currently pursuing the Ph.D. degree with the ADAPT Centre, Athlone Institute of technology. She investigates what are the key factors for developing applications to assess spatial auditory abilities. Her research interest includes understanding user quality of experience of multimedia applications.

RONAN FLYNN (Member, IEEE) received the B.E. degree in electronic engineering from University College Dublin, the M.Eng. degree in computer systems from the University of Limerick, and the Ph.D. degree from the National University of Ireland, Galway. He is currently a Lecturer and a Researcher with the Faculty of Engineering and Informatics, Athlone Institute of Technology. He has industrial experience in telecommunication product design and development for international markets, having previously worked with companies in Germany and Ireland. His research interests include speech recognition, speech enhancement, and multi-modal affective computing.

ANDREW HINES (Senior Member, IEEE) is currently the Director of Research, Innovation, and Impact for the School of Computer Science, University College Dublin, Ireland. He is also a Funded Investigator in both the SFI ADAPT, CONNECT, and INSIGHT research centers. His QxLab research group has research interests in multimedia quality of experience and applying machine learning for applications in speech, audio, and video signal processing. He is a Committee Member of the Audio Engineering Society (Ireland).

NIALL MURRAY (Member, IEEE) is currently a Lecturer with the Faculty of Engineering and Informatics, Athlone Institute of Technology (AIT), Ireland. He is also the Founder, in 2014, and the Principal Investigator (PI) in the truly Immersive and Interactive Multimedia Experiences (tIIMEx) Research Group in AIT. He is a Science Foundation Ireland (SFI) Funded Investigator (FI) with the Confirm Centre for Smart manufacturing and an FI with the SFI Adapt Centre for AI enabled Digital Content. He is an Associate PI with the Enterprise Ireland Funded Technology Gateway COMAND. His current research interests include immersive and multisensory multimedia communication and applications, multimedia signal processing, quality of experience, and wearable sensor systems (further information available at: www.niallmurray.info).