Comparing User QoE via Physiological and Interaction Measurements of Immersive AR and VR Speech and Language Therapy Applications

Conor Keighrey
Department of Electronics & Informatics
Athlone Institute of Technology
Athlone, Co. Westmeath, Ireland
c.keighrey@research.ait.ie

Ronan Flynn
Department of Electronics & Informatics
Athlone Institute of Technology
Athlone, Co. Westmeath, Ireland
rflynn@ait.ie

Siobhan Murray
Health Service Executive
Primary Care Centre, Flancare Business Park
Ballyminion, Longford, Co. Longford
Siobhan.murray1@hse.ie

Sean Brennan
Department of Electronics & Informatics
Athlone Institute of Technology
Athlone, Co. Westmeath, Ireland
s.brennan@research.ait.ie

Niall Murray
Department of Electronics & Informatics
Athlone Institute of Technology
Athlone, Co. Westmeath, Ireland
nmurray@research.ait.ie

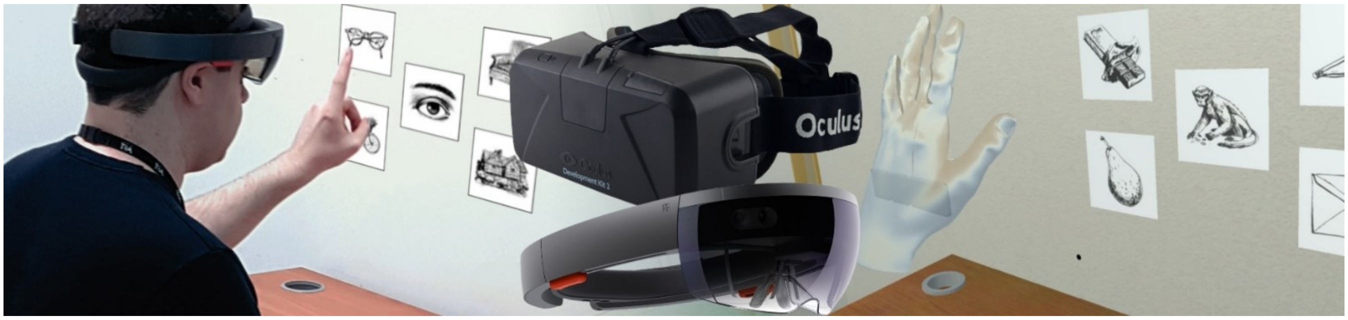

Figure 1: (Left) Third person view of Immersive Augmenetd Reality Speech and Language Assesment (Right) First Person view of Immersive Virtual Reality Speech and Language Assesment.

ABSTRACT

Virtual reality (VR) and augmented reality (AR) applications are gaining significant attention in industry and academia as potential avenues to support truly immersive and interactive multimedia experiences. Understanding the user perceived quality of immersive multimedia experiences is critical to the success of these technologies. However, this is a multidimensional and multifactorial problem. The user quality of experience (QoE) is influenced by human, context and system factors. Attempts to understand QoE via multimedia quality assessment has typically involved users reporting their experiences via post-test questionnaires. More recently, efforts have been made to automatically collect objective metrics that can quantitatively reflect user QoE in terms of physiological measurement methods. In this context, this paper presents a novel comparison of objective quality measures of immersive AR and VR applications through physiological: (electrodermal activity (EDA) and heart rate (HR)); and interaction (response times (RT), incorrect responses, and miss-click) metrics. The analysis shows consistency in terms of physiological ratings and miss-click metrics between the AR and VR groups. Interestingly, the AR group reported lower response times and less incorrect responses compared to the VR group. The difference between the AR and VR groups was statistically significant for the incorrect response metric and in 45.5% of the cases tested for response times metric, they were statistically significant with 95% confidence levels.

CCS CONCEPTS

Human-centered computing → Human computer interaction (HCI) • *Human-centered computing → Mixed / augmented reality* • *Human-centered computing → Virtual reality*

KEYWORDS

Multimedia; Virtual Reality; Augmented Reality; Perception; Physiological; Speech Language Therapy; Semantic Memory;

ThematicWorkshops'17, October 23–27, 2017, Mountain View, CA, USA

ACM 978-1-4503-5416-5/17/10 $15.00
https://doi.org/10.1145/3126686.3126747

1 INTRODUCTION

Immersive multimedia applications aim to support enhanced user immersion and interaction above and beyond what is possible with traditional media. Typically, the term immersive multimedia reflects technologies such as virtual reality (VR) and augmented reality (AR). Recently this has been extended to include sensory (multi) experience [1], as true immersion is naturally multisensory.

VR is the creation of fully simulated environments that replicate real or imaginary environments [2]. AR is concerned with the overlaying of contextual information upon a real-world object [3]. According to many of the marketing and media promises, VR and AR applications will support greater degrees of emotional attachment, a sense of reality, naturalistic interactions and user engagement like never experienced before. Both technologies claim to support naturalistic interaction with virtual objects, hence blurring the lines between the real and virtual worlds. Critical to the success of both technologies, is the fact that on a per application basis, it is crucial to understand the perceptual user and the quality of experience (QoE).

The term quality of experience is defined in [4] as: "the degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the person's evaluation of the fulfillment of his or her expectations and needs with respect to the utility and/or enjoyment in the light of the person's context, personality and current state". It is a measurement-centered reflection of a users' perception of an application, system, or service. The influencing factors of QoE are typically categorized as being within human, system, and context categories.

To capture and understand user QoE, the literature reports quality assessment approaches based on both implicit and explicit evaluations [5]. Explicit evaluations require the user to report perceived quality using predefined scales (e.g. mean opinion score), or open-ended questions. This type of approach has been dominant in efforts to capture user QoE of many types of immersive experiences [1] [6] [7] [8]. However, the literature also highlights numerous issues associated with explicit evaluations such as: the time required; bias in subjective responses; and inaccuracies in responses due to external factors [9] [10]. Implicit evaluations aim to analyze the relationship between captured physiological measures and user QoE. Implicit evaluations have gained traction due to their real-time continuous nature and context-based unbiased feedback.

Psychophysiology-based QoE assessment, as discussed in [11], is also proposed as a method to gain a deeper understanding of QoE. It considers the captured physiological data along with the psychological bases of perceptual and cognitive processes [11]. Furthermore, related to the context of the quality assessment approach proposed here, the use of interaction measures is another approach to capturing QoE [5]. Some of the influencing factors to be considered here are: learning, effort required, response times, interaction, errors and satisfaction. These all fall within the human, system, and context domains and provide valuable objective data on user QoE from an interaction perspective.

In this work, we developed immersive AR and VR Speech and Language Therapy (SLT) applications. SLT involves the diagnosis and mitigation of speech and language difficulties. The AR and VR application developed, focuses on assessing semantic memory, it aims to evaluate an individual's language processing abilities on a cognitive level. It was created to mimic the paper based assessment as outlined in the Comprehensive Aphasia Test (CAT) [12]. A healthy population was divided into two independent groups; one group experienced an AR SLT application and another group experienced a VR SLT application. This paper presents, compares, and analyses physiological (Electrodermal Activity (EDA) and Heart Rate (HR)) and interaction (response time (RT), and interaction errors) measures of the two groups. To the best of the authors' knowledge, this is the first work that performs a physiological and interaction measurement comparison of AR and VR applications.

The paper is structured as follows: Section 2 provides an overview of related work from VR, AR and user perception perspectives; Section 3 presents the experimental method employed for this study, as well as an overview of the AR and VR SLT applications, the measurement scales used and the participants; in Section 4, the results of the physiological and interaction measurements of the two groups are presented, contrasted, and the findings are discussed. Finally, Section 5 concludes this research.

2 RELATED WORKS

Generally speaking, the methods used to perform quality assessments of immersive AR and VR applications have involved borrowing aspects from methods designed for traditional media components i.e. explicitly (through variations of the ITU-T standards [13] [14]) and implicitly via eye measurements, EDA, EEG, ECG, heart rate as highlighted in [11].

In [10], a hybrid approach combining implicit and explicit quality assessment was taken in the evaluation of immersive 2D and 3D multimedia content. The participants in the study completed a subjective rating analysis and were monitored in terms of brain activity and peripheral physiological responses. EEG data was correlated with QoE, but there was little correlation of respiration and ECG data with QoE.

User QoE levels in immersive VR and non-VR environments were compared implicitly and explicitly in [15]. A sample size of thirty-three participants, divided randomly into two groups, answered questionnaires (post experience) and provided physiological metrics of EDA and HR (during the experience). Analysis showed that HR and EDA levels were elevated in the immersive VR environments compared with the non-VR environment. Similarly, [8] investigated the correlation between physiological measures (EDA and heart rate) and subjective data as users experienced a virtual environment in a video game. The subjects were exposed to three first-person shooters for a twenty-minute time period and asked to complete an in-game experience questionnaire (iGEQ) every five minutes. The results reported a statistically significant correlation between heart rate and the subjective data gathered from the iGEQ across seven dimensions of gameplay.

In [5], an interactive AR application which emulated tasks carried out in the field of neuronavigation was assessed (implicit and explicit metrics). The implicit metrics considered were: task completion time, error rates and accuracy; whilst the explicit data was the response times to questionnaires. The authors conducted a pilot study that displayed some promising results. Objective metrics with respect to the time it took a participant to complete a neuronavigation task fell from minutes during the first user interaction to seconds by the final interaction. The paper also contributed a survey of existing assessment approaches for AR applications.

A theoretical evaluation framework for user experience in AR applications was presented in [16]. The user experience was considered to be a function of four distinct categories: *Input* - focusing on visual, auditory, tactile, and kinesthetic data; *Output* - focusing on the output of visual and audio content, and haptic feedback; *Context Awareness* - sense of immersion, health; *Safety and Integrity Privacy and Security*. It also discussed [17] which defined user experience as "a person's perceptions and responses that result from the use or anticipated use of a product, system or service".

Considering existing works on quality assessment approaches to determine user QoE of immersive multimedia applications, the novelty of the work presented in this paper lies in the implicit comparison of users' QoE in AR and VR by way of physiological and interaction measurements. In the next section, the experimental method by the authors to achieve this task is presented.

3 Experimental Method

In section 3, we provide an in-depth overview of the quality assessment protocol, the virtual SLT applications, the AR and VR hardware employed in the evaluations, the physiological measurement devices, and finally participant screening.

3.1 Quality Assessment Protocol

Four key phases were defined for the quality assessment protocol: information and screening, resting, training, and testing. Completion of all four phases required approximately 30-35 minutes.

3.1.1 Information & Screening Phase

Each participant was greeted upon arrival and guided to the waiting area. The experiment was described in full and an information sheet was provided. Participants completed a consent form and then took part in a screening process. The screening process assessed participants for visual defects with respect to visual acuity and color perception. Color perception with respect to red-green color deficiencies was screened using the Ishihara test [18]. Visual acuity deficiencies were screened using the Snellen test [19]. More details on how the Ishihara and Snellen tests were executed are available in previous work [20] for the interested reader. Typically, the information and screening phase lasted 12 minutes.

3.1.2 Resting Phase

Due to the variability of participant EDA and heart rate measurements, each participant was required to participate in a resting phase. The resting phase focused on gathering baseline metrics for EDA and HR for comparative analysis. During the resting phase, participants were introduced to the physiological sensors. Heart rate data was captured using a Fitbit Charge HR [21]. A biosensor called the PIP [22] was employed to capture EDA. To ensure measurement consistency across participants, the Fitbit was configured to gather data from a participant's non-dominant hand. Additionally, it was requested that users hold the PIP in their non-dominant hand allowing them to focus on interaction using their dominant hand. Over a five-minute period, baseline metrics were recorded from each user. On average, the resting phase was completed in 8 minutes.

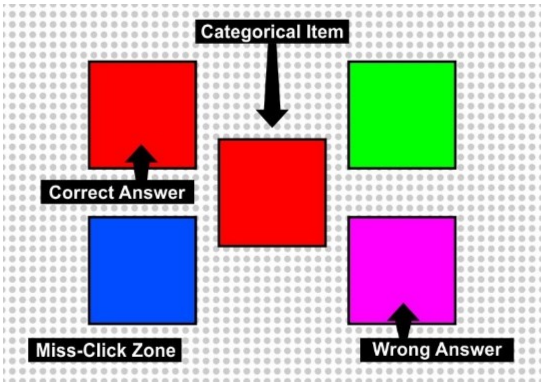

Figure 2: Example of virtual training excercise

3.1.3 Training Phase

The training phase focused on three aspects: introduction of participants to the head mounted display (HMD); gesture based interaction; and an overview of the SLT activity. A series of training videos were created for each HMD in a first person perspective and were viewed on a computer monitor. Three short videos focused on assisting the participants to understand user interaction in terms of the virtual curser, hand gestures, and the virtual speech and language assessment. Subsequent to the training videos, the HMD was fitted to the participant and they took part in the training activity. The virtual training exercise consisted of eleven slides in an identical layout to the main test. Participants progressed throughout the training exercise by simply matching colors in the presented stimuli (as illustrated in Fig. 2). Completion of this phase took approximately 10 minutes. EDA and HR were recorded throughout the training phase for comparative purposes.

3.1.4 Testing Phase

The testing phase was composed of eleven slides (Fig. 3) as per the Comprehensive Aphasia Test (CAT) [12], which is discussed in section 3.2. In accordance with the CAT assessment procedure, a successful choice resulted in audio feedback. This feedback aimed to replicate the positive reinforcement techniques that are used by speech language therapists in practice. After the test, participants were asked to complete the subjective questionnaire [23]. 14 questions were rated using the absolute category rating (ACR) system as outlined in ITU-T P913 [24]. Questions were inspired from [15] and aimed to evaluate QoE from four key perspectives: user interaction, immersion, discomfort, and enjoyment. Results of this aspect of the study can be found at [20]. The virtual SLT assessment as well as the time to complete the questionnaire was approximately 5 minutes.

Table 1: Virtual Semantic Memory Assessment Slide Content

	Categorical Item	Target	Distractor	Distractor	Distractor
Slide P (Practice)	Monkey	Banana	Pear	Chocolate	Envelope
Slide 1	Glasses	Eye	Ear	Mouth	Elephant
Slide 2	Hand	Mitten	Sock	Jersey	Lighthouse
Slide 3	Matches	Candle	Light Bulb	Radio	Star
Slide 4	Pillow	Bed	Chair	Stool	Flag
Slide 5	Eskimo	Igloo	Hut	House	Sunshade
Slide 6	Watch	Arm	Leg	Neck	Tortoise
Slide 7	Nun	Church	School	Factory	Skate
Slide 8	Camping Tent	Fire	Torch	Rocket	Picture
Slide 9	Mask	Clown	Ballerina	Priest	Sheep
Slide 10	Flower	Watering Can	Bucket	Shower	Anchor

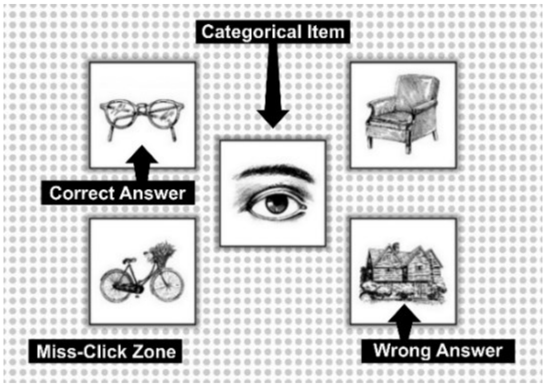

Figure 3: Example of virtual semantic memory exercise

Figure 4: (a) PIP Biosensor, (b) Fitbit Charge HR

3.2 Immersive AR & VR SLT Assessments

3.2.1 Virtual SLT Assessment

The virtual SLT assessment was developed using the Unity [25] game engine. The focus of this assessment is on the analysis of receptive language, more specifically in the context of semantic memory. The assessment content and administration procedures used in this work have been inspired by those outlined in the CAT [12]. The CAT is often used in clinical settings to assess individuals who have experienced traumatic brain injury, resulting in symptoms of aphasia. Ten multiple-choice slides and one additional practice slide (Slide P) make up the contents of the test. Each slide contains five images as shown in Fig. 3. The central image, which can be described as the categorical item, forms a relationship with one of the four outer images. The remaining three images serve as distractors. Hence, this test focuses on evaluating cognition in terms of speech and language. Administration of this type of test requires participants to use gestures to identify the correct relationship between the images as opposed to a verbal response.

3.2.2 Immersive AR & VR systems

The immersive AR and VR SLT applications were evaluated using the Microsoft HoloLens [26] and the Oculus Rift (OR) Development Kit 2 (DK2) [27] HMDs. Currently, there is no VR HMD on the market that allows for natural hand gesture interaction similar to the Microsoft HoloLens. However, hand tracking can be accomplished using third-party equipment such as the Leap Motion [28]. The Leap Motion facilitates naturalistic interaction and tracks hand movement using an IR camera system. The gestures of the HoloLens were emulated using the Leap Motion in the VR environment, thus providing a like-for-like experience between the AR and VR test groups.

3.3 Measurement Scales

3.3.1 Electrodermal Activity

Electrodermal activity is the measure of physiological changes in skin conductivity. This measurement can be divided into two distinct categories: tonic change and phasic change [29]. Tonic change corresponds to a steady or slow change in skin conductance, be it positive or negative. Typically, this is referred to as skin conductance level (SCL). Signals from the autonomic nervous system in terms of physiological arousal are reflected through phasic change. This change is also known as the skin conductivity response (SCR). Phasic events correspond to short-term peaks in the skin conductance that are accompanied by varied rates of decline.

SCR signals can be triggered through the presence of environmental stimuli such as sound, smell, or sight. In this study, skin conductivity changes were monitored using the PIP biosensor (see Fig. 4 (a)) [22] . This handheld non-intrusive device provides a stream of data over Bluetooth. EDA metrics were captured 8 times per second. The wireless Bluetooth device is held between the thumb and index finger by a user.

3.3.2 Heart Rate

In addition to EDA, HR was monitored throughout the experiment as a measure of emotional arousal. Reactions to HR and EDA are both triggered within the autonomic nervous system, which is a division of the peripheral nervous system [11]. Previous studies have indicated that EDA and HR are correlated with data captured through subjective post-test questionnaires [8]. The Fitbit Charge HR (see Fig. 4 (b)) [21] was used to capture HR data. The non-invasive wireless activity tracker utilizes an optical sensor located under the device to monitor blood volume changes. Internal Fitbit algorithms are used to convert this data stream into beats per minute (BPM). The sensitivity of these algorithms is effected by which hand (dominant or non-dominant) the participant wore the device on [30]. To ensure an accurate and consistent measure was gathered throughout the experiment, all participants were requested to wear the device on their non-dominant hand as per the devices default setting. The device monitors the wearer's heart rate on a per second basis.

3.3.3 Response Times

User interaction in terms of response time (RT) was captured throughout the training and test phases of the experiment. In previous studies [5], RT was used in a similar manner to monitor user interaction. Furthermore, this objective metric provides opportunity for more precise analysis from an SLT assessment perspective. Although user interaction time is often monitored on paper based SLT assessments, it is purely from the subjective view of the therapist. The inclusion of an accurate response time tracks user interaction in a precise way, thus allowing for performance monitoring. In this experiment, the RT was calculated by considering the presentation time of the visual stimuli (i.e. a slide) with the duration of time it took for a user to identify the correct response.

Response time varied throughout the test for each participant, therefore the amount of HR and EDA data that was relevant for a specific slide also varied. Hence, the RT metric was used to extract only the HR and EDA data whilst the user interacted with a stimulus.

3.3.4 Incorrect Responses & Miss-Clicks

As per the CAT guidelines, on administration of the semantic memory assessment, errors in selection were noted. Incorrect response events were triggered when a participant did not form the correct categorical link within the virtual SLT test. Additionally, mistakes were monitored throughout the training phase as a method of monitoring user interaction levels. Similar to the tracking of incorrect responses and RT, information on how often a user missed the target was captured. A miss-click event was triggered when a user made an interaction gesture whilst not taking note of where they were looking e.g. if the cursor was pointing towards the dotted area in Fig. 3.

3.4 Participants

Forty-six participants were recruited for this study with an average age of twenty-seven years. A convenience sampling method resulted in participants from a wide variety of backgrounds: students, post-graduate researchers, academic staff, and members of the public. Due to incomplete data, the results of four participants were removed. Additionally, despite having a complete set of results, two participants were omitted due to the identification of visual defects as outlined in screening process (as outlined in Section 3.1.1). Twenty participants experienced the immersive VR environment, out of which six had used virtual reality technologies in the past. Similarly, the AR group consisted of twenty participants. Seven of these had experienced AR before in the form of mobile augmented reality. No participants had any prior experience in subjective quality evaluations of multimedia content.

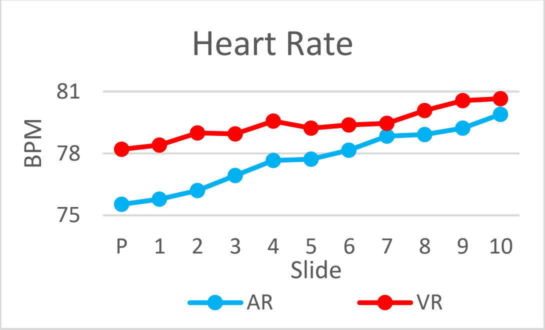

Figure 5: Average user heart rate throughout the virtual speech language therapy assessment

4 RESULTS & DISCUSSION

In this section, the results of the physiological and interaction metrics are presented. Statistical analysis was carried out using a multivariate analysis of variance (MANOVA) at a 95% confidence level using IBM SPSS [31]. In terms of test design, it was hypothesized that the VR group would have a higher level of physiological arousal. It was also assumed that the VR group would have more naturalistic interactions with the virtual objects, since all the interaction was within the virtual world. Unlike for the AR group, where the slides were overlaid upon a real-world environment.

4.1 Physiological Measurements

4.1.1 Heart Rate

Table 2 provides a comparative breakdown of the average user HR recorded over the resting (baseline), training and testing (Slides P-10) periods. During the test phase, the HR data for both groups is provided on a per slide basis. From Table 2, there were no statistically significant differences between the AR and VR groups. Notably, there is commonality in the trajectory (as seen in Fig. 5) of HR readings for both the AR and VR groups as participants progressed from slide P to slide 10 of the immersive SLT assessment. In terms of standard deviation, each group remains relatively consistent throughout the slides. However, the SD within the AR group was higher than the VR group, with an average SD of 13.69 BPM for the AR group and an average SD of 9.64 BPM for participants in the VR group.

4.1.2 EDA

Table 3 provides a comparative breakdown of EDA data recorded throughout the test. As was the case for the HR results, the breakdown of the results is provided on a per slide basis averaged

Table 2: Comparison of the User Heart Rate Levels in AR and VR Groups

	Augmented Reality		Virtual Reality		Between Subjects	
	Heart Rate (BPM)	SD (BPM)	Heart Rate (BPM)	SD (BPM)	F	Sig.
Baseline	77.948	13.224	78.868	9.982	.575	.443
Slide P (Practice)	75.534	13.770	78.203	9.894	.674	.417
Slide 1	75.780	13.577	78.407	9.864	.641	.429
Slide 2	76.210	13.413	78.995	9.776	.604	.442
Slide 3	76.928	13.406	78.943	9.640	.332	.568
Slide 4	77.659	13.451	79.566	9.614	.249	.620
Slide 5	77.719	13.472	79.228	9.438	.155	.696
Slide 6	78.158	13.711	79.377	9.320	.142	.708
Slide 7	78.843	14.260	79.467	9.035	.060	.808
Slide 8	78.911	14.205	80.083	9.437	.120	.732
Slide 9	79.225	13.998	80.559	10.076	.106	.746
Slide 10	79.898	13.857	80.658	9.798	.037	.849

Table 3: Comparison of the User EDA Levels in AR and VR Groups

	Augmented Reality		Virtual Reality		Between Subjects	
	EDA (mho)	SD (mho)	EDA (mho)	SD (mho)	F	Sig.
Baseline	3.140	1.292	2.895	1.225	.851	.362
Slide P (Practice)	3.456	0.990	3.386	1.225	.232	.633
Slide 1	3.608	1.064	3.369	1.110	.804	.376
Slide 2	3.685	1.108	3.355	1.101	1.18	.285
Slide 3	3.640	1.151	3.373	1.133	.829	.369
Slide 4	3.663	1.153	3.367	1.133	.974	.330
Slide 5	3.592	1.178	3.350	1.145	.772	.385
Slide 6	3.573	1.200	3.377	1.119	.520	.475
Slide 7	3.578	1.207	3.480	1.160	.274	.604
Slide 8	3.577	1.247	3.473	1.193	.275	.603
Slide 9	3.552	1.233	3.505	1.165	.109	.743
Slide 10	3.615	1.184	3.488	1.211	.309	.582

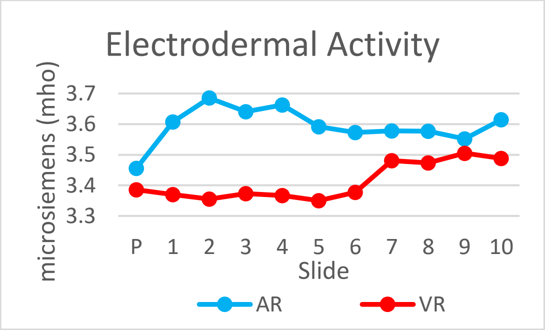

Figure 6: Average levels of electrodermal activity throughout the virtual speech language therapy assessment

over the resting, training, and testing periods. Similar to the HR data, Table 3 shows no statistically significant differences between the AR and VR groups. However, there was a greater degree of consistency in terms of standard deviation across both groups compared with the HR data. This indicates similar levels of emotional arousal between the groups, which was not expected during the experimental design. Fig. 6 reveals a large increase in EDA as AR users began the activity. This increase stabilized to a consistent level over the course of the assessment. Interestingly, the VR group remained stable throughout the initial exposure to the exercise, but large increases occurred for slides 7 through 10. This increase coincided with the levels of stress which may be associated with a reduction in performance (similar findings are reported later with respect to RT in section 4.2.1).

4.2 Interaction Measurements

4.2.1 Response Times

Response time is an important indicator in measuring user interaction with the immersive SLT applications. Along with error rates, it gives objective performance-related data on the two groups. In Table 4, a comparative analysis is provided for user RT for both the AR and VR groups. The RT column indicates the average time each user took to identify the correct answer to the presented stimuli.

Table 4: Comparison of User Response Time in AR and VR Groups

	Augmented Reality		Virtual Reality		Between Subjects	
	RT (seconds)	SD (seconds)	RT (seconds)	SD (seconds)	F	Sig.
Slide P (Practice)	4.3866	2.2190	8.8081	9.4245	3.250	.080
Slide 1	7.1770	1.4103	6.0663	6.2263	.069	.795
Slide 2	4.4943	1.6066	4.6561	2.5255	.488	.489
Slide 3	4.7581	1.8111	7.0791	4.1774	5.626	.023
Slide 4	4.1037	1.3663	6.0513	7.4761	3.415	.073
Slide 5	4.1453	1.4955	5.1346	3.1170	1.255	.270
Slide 6	4.6940	2.5771	6.5548	5.4282	6.609	.014
Slide 7	4.9904	1.9737	8.9596	7.5801	5.975	.020
Slide 8	6.8113	3.7966	11.4593	11.4789	5.576	.024
Slide 9	7.4249	5.6753	15.8381	15.8663	4.481	.041
Slide 10	5.3192	2.7804	6.2156	2.9452	.776	.384

Table 5: Comparison of the User Missed Targets and Incorrect Responses in AR and VR Groups

	Augmented Reality		Virtual Reality		Between Subjects	
	Other	SD	Other	SD	F	Sig.
Miss-Clicks	.20	.523	.65	1.137	1.458	.235
Incorrect Response	.35	.933	1.15	1.461	6.746	.014

From Table 4 there is a statistically significant difference at a 95% confidence level, between the AR and VR groups for five of the eleven slides. These differences were with slide 3 (p=0.023); slide 6 (p=0.014); slide 7 (p=0.020); slide 8 (p=0.024); and slide 9 (p=0.041). Hence, from an interaction perspective, the AR group outperformed the VR group significantly on these specific questions.

Notably, for slides 8 and 9 there were much larger RTs for the VR group compared to the AR group. The VR group spent on average 4.2 seconds longer responding to slide 8, and 8.3 seconds longer responding to slide 9. The standard deviation for slide 8 and 9 was also much higher in comparison to previous questions to the respected groups. In the AR group the SD was 3.79 seconds for slide 8, whilst a little higher, this falls in line with the previous slides. However, in the VR group there is a much higher SD of 11.47 seconds for slide 8. Slide 9 also saw a large increase in SD, with an average of 15.8 seconds RT for VR and 7.42 seconds for the AR group. Slide 9 also reported the highest SD for AR with a deviation of 5.67 seconds between users; interestingly the VR group reported almost triple that with an SD of 15.86 seconds for slide 9.

From an observational standpoint, these delays were noticeable throughout the testing. However, the degree of delay did not appear as extreme as displayed in the objective data. Interestingly, as part of an open-ended discussion after testing, participants described difficulty viewing detailed slide content towards the end of the test. This could be equated to the finer details within the final set of slides. More often than not, finer details can be hindered by the screen door effect [32] often experienced by users of VR headsets. While the delay in response was not as evident in the AR users group it is clear that slides 8 and 9 also caused them some difficulty. Interestingly, a correlation can be observed between the RT and EDA for the VR group with respect to slides 7-10. The delay in associative language recall with respect to user interaction can be viewed as an increase in cognitive load as users spend more time thinking about the correct response. As a result of this, it would appear that the average user EDA increased in synchronous to this delay in response. This observational correlation builds on previous work of Kilpatrick et al. [29] which describes a tonic change in skin conductance as a function of increased cognitive activity.

4.2.2 Miss-clicks & Wrong Answers

Performance metrics in terms of miss-clicks and incorrect responses provided by participants are presented in Table 5. A statistically significant result (p=0.014) was found in terms of incorrect responses between the AR and VR groups. This is directly reflected in the VR scores, which had on average 1.15 wrong answers provided throughout the test with a standard deviation of 1.461. On the other hand, participants in the AR group had only 0.35 wrong answers on average with a standard deviation of 0.933. Again, this was unexpected based on the hypothesis that users would be more engaged in the VR environment compared to the AR.

While not as significant, a difference is also noted in terms of miss-clicks. In the AR group there was an average of 0.2 miss-clicks with a SD of 0.523 compared to that of the VR group who experienced on average 0.65 miss-clicks with a larger SD of 1.137. From an interaction perspective, with respect to miss-clicks, the larger SD offers insight into a larger window for error for VR participants. This demonstrates that there is an interactive difference between AR and VR. However, there is no statistical significance, and the results do not provide a definitive answer.

5 CONCLUSIONS

User perception of quality of immersive multimedia experiences is influenced by a combined relationship of human, system and context factors. As a step towards understanding user perceived quality in immersive AR and VR experiences, this paper presented the results of an implicit user QoE comparison of immersive augmented and virtual reality applications. Physiological and interactive measures in the form of heart rate, electrodermal activity, response times, incorrect responses, and miss-clicks were captured during participant AR and VR experiences.

The results reveal that from a physiological point of view, AR and VR users experienced a similar reaction in terms of HR elevation throughout the virtual SLT assessment. However, analysis of the EDA data expose an unexpected result. VR users experience a rise in EDA which coincides with increased cognitive load as reflected through increased response time. The AR group revealed unexpected levels of physiological arousal at the start of the activity. This rise and fall in EDA could be associated with users becoming more accustomed their environment however further analysis is required.

The interaction metrics explored response times and interaction errors. They indicated that VR users experienced more difficulty in terms of a delay in response times and interaction errors (incorrect responses) whilst experiencing the virtual SLT assessment. Future work will involve further analysis of the physiological measures, specifically with respect to the SD of HR and EDA readings. This work will also be extended to the development and evaluation of SLT diagnostics and interventions based on AR technologies.

ACKNOWLEDGMENTS

The work presented in this paper has been supported by the Irish Research Council (grant number: GOIPG/2016/1493).

REFERENCES

[1] C. Timmerer, M. Waltl, R. Benjamin and N. Murray, "Sensory Experience: Quality of Experience beyond Audio-Visual," in *Quality of Experience: Advanced Concepts, Applications and Methods*, Springer, 2014, pp. 351-365.

[2] J. Steuer, "Defining Virtual Reality: Dimensions Determining Telepresence," *Journal of Communication,* vol. 42, no. 4, pp. 73-93, 1992.

[3] R. Azuma, Y. Baillot, S. Feiner , S. Julier, B. MacIntyre and R. Behringer, "Recent Advances in Augmented Reality," *IEEE Computer Graphics and Applications,* vol. 21, no. 6, pp. 34-47, 2001.

[4] . S. Möller and A. Raake , "Quality and Quality of Experiance," in *Quality of Experience: Advanced Concepts, Applications and Methods*, Springer, 2014, p. 19.

[5] J. Puig, A. Perkis, F. Lindseth and T. Ebrahimi, "Towards an Efficient Methodology for Evaluation of Quality of Experience in Augmented Reality," in *Quality of Multimedia Experience (QoMEX)*, 2012.

[6] J. Cha, M. Eid, A. Barghout, A. M. Rahm and A. El Saddik, "HugMe: Synchronous Haptic Teleconferencing," in *ACM international conference on Multimedia*, 2009.

[7] M. Obrist, C. Velasco, C. Vi, N. Ranasinghe, A. Israr, A. Cheok, C. Spence and P. Gopalakrishnakone, "Sensing the future of HCI: touch, taste, and smell user interfaces," Sussex Research Online, 2016.

[8] A. Drachen, L. E. Nacke, G. Yannakakis and A. L. Pedersen, "Correlation between Heart Rate, Electrodermal Activity and Player Experience," in *SIGGRAPH Symposium on Video Games*, 2010.

[9] T. Hoßfeld, R. Schatz and S. Egger, "SOS: The Mos Is Not Enough!," in *Quality of Multimedia Experience (QoMEX)*, 2011.

[10] E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek and T. Ebrahimi, "Modeling Immersive Media Experiences by Sensing Impact on Subjects," *Multimedia Tools and Applications,* vol. 75, p. 12409–12429, 2016.

[11] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J.-N. Antons, K. Y. Chat, N. Ramzan and K. Brunnström, "Psychophysiology-Based QoE Assessment: A Survey," *IEEE Journal of Selected Topics in Signal Processing,* 2017.

[12] K. Swinburn, G. Porter and D. Howard, Comprehensive Aphasia Test, Psychology Press, 2004.

[13] "ITU-T BT.500 : Methodology for the subjective assessment of the quality of television pictures," 2017. [Online]. Available: https://www.itu.int/rec/R-REC-BT.500. [Accessed 07 04 2017].

[14] "ITU-T P.910 : Subjective video quality assessment methods for multimedia applications," [Online]. Available: https://www.itu.int/rec/T-REC-P.910-199909-S/en. [Accessed 01 02 2017].

[15] D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer and N. Murray, "An evaluation of Heart Rate and ElectroDermal Activity as an objective QoE evaluation method for immersive virtual reality environments," in *Quality of Multimedia Experience (QoMEX)*, 2016.

[16] P. D. Ritsos, D. P. Ritsos and A. S. Gougoulis, "Standards for Augmented Reality: a User Experience perspective," in *International AR Standards Meeting*, 2011.

[17] "ISO 9241-210:2010: Ergonomics of human-system interaction -- Part 210: Human-centred design for interactive systems," International Organization for Standardization, [Online]. Available: https://www.iso.org/standard/52075.html. [Accessed 04 07 2017].

[18] Committee on Vision, Assembly of Behavioral and Social Sciences, National Research Council, "Procedures for Testing Color Vision: Report of," NATIONAL ACADEMY PRESS, 1981.

[19] C. H. Graham, N. R. Bartlett, J. L. Brown, C. G. Mueller, Y. Hsia and L. A. Riggs, Vision and Visual Perception, John Wiley & Sons Inc., 1965.

[20] C. Keighrey, R. Flynn, S. Murray and N. Murray, "A QoE Evaluation of Immersive Augmented and Virtual Reality Speech & Language Assessment Applications," in *QoMEX 2017 – 9th International Conference on Quality of Multimedia*, Erfurt, Germany, 2017.

[21] "Fitbit Charge HR," Fitbit, [Online]. Available: https://www.fitbit.com/chargehr. [Accessed 10 02 2017].

[22] "PIP Biosensor," PIP, [Online]. Available: https://thepip.com/en-eu/. [Accessed 01 02 2017].

[23] "ACM MM: Subjective Questionnaire," [Online]. Available: http://bit.ly/ACM-Multimedia-2017. [Accessed 11 04 2017].

[24] ITU-T, "ITU-T 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: https://www.itu.int/rec/T-REC-P.913/en. [Accessed 29 04 2017].

[25] "Unity - Game Engine," Unity3D, [Online]. Available: https://unity3d.com/. [Accessed 07 04 2017].

[26] "Microsoft HoloLens," Microsoft, [Online]. Available: https://www.microsoft.com/microsoft-hololens/en-us. [Accessed 02 02 2017].

[27] "Oculus Rift Development Kit 2," Oculus, [Online]. Available: https://www.oculus.com/en-us/dk2/ . [Accessed 02 02 2016].

[28] "Leap Motion," Leap Motion, [Online]. Available: https://www.leapmotion.com/. [Accessed 07 04 2017].

[29] D. G. Kilpatrick, "Differential responsiveness of two electrodermal indices to psychological stress and performance of a complex cognitive task," *Psychophysiology,* vol. 9, no. 2, pp. 218-226, 1972.

[30] Fitbit, "Help article: How accurate are Fitbit trackers?," Fitbit, 25 05 2017. [Online]. Available: https://help.fitbit.com/articles/en_US/Help_article/1136#wrist. [Accessed 25 05 2017].

[31] "IBM SPSS - IBM Analytics," IBM, [Online]. Available: https://www.ibm.com/analytics/us/en/technology/spss/. [Accessed 07 04 2017].

[32] I. Goradia, J. Doshi and L. Kurup, "A Review Paper on Oculus Rift & Project Morpheus," *International Journal of Current Engineering and Technology,* vol. 4, no. 5, pp. 3196-3200, 2014.