Towards an AI-Driven Talking Avatar in Virtual Reality for Investigative Interviews of Children

Syed Zohaib Hassan
syed@simula.no
SimulaMet

Pegah Salehi
pegah@simula.no
SimulaMet

Ragnhild Klingenberg Røed
rar@oslomet.no
OsloMet

Pål Halvorsen
paalh@simula.no
SimulaMet

Gunn Astrid Baugerud
gunnba@oslomet.no
OsloMet

Miriam Sinkerud Johnson
mirsin@oslomet.no
OsloMet

Pierre Lison
plison@nr.no
Norwegian Computing Center

Michael Riegler
michael@simula.no
SimulaMet

Michael E. Lamb
mel37@cam.ac.uk
University of Cambridge

Carsten Griwodz
griff@ifi.uio.no
University of Oslo

Saeed Shafiee Sabet
saeed@simula.no
SimulaMet

ABSTRACT

Artificial intelligence (AI) and gaming systems have advanced to the stage where the current models and technologies can be used to address real-world problems. The development of such systems comes with different challenges, e.g., most of them related to system performance, complexity and user testing. Using a virtual reality (VR) environment, we have designed and developed a game-like system aiming to mimic an abused child that can help to assist police and child protection service (CPS) personnel in interview training of maltreated children. Current research in this area points to the poor quality of conducted interviews, and emphasises the need for better training methods. Information obtained in these interviews is the core piece of evidence in the prosecution process. We utilised advanced dialogue models, talking visual avatars, and VR to build a virtual child avatar that can interact with users. We discuss our proposed architecture and the performance of the developed child avatar prototype, and we present the results from the user study conducted with CPS personnel. The user study investigates the users' perceived quality of experience (QoE) and their learning effects. Our study confirms that such a gaming system can increase the knowledge and skills of the users. We also benchmark and discuss the system performance aspects of the child avatar. Our results show that the proposed prototype works well in practice and is well received by the interview experts.

GameSys '22, June 14, 2022, Athlone, Ireland

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/3534085.3534340

CCS CONCEPTS

• **Human-centered computing** → *User studies*; **Virtual reality**; • **Computing methodologies** → *Natural language processing*; **Animation**.

KEYWORDS

Avatar, Virtual Reality (VR), AI, Dialogue Model, Child Protection Services (CPS), Generative Adversarial Networks (GANs), Quality of Experience (QoE)

ACM Reference Format:
Syed Zohaib Hassan, Pegah Salehi, Ragnhild Klingenberg Røed, Pål Halvorsen, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Pierre Lison, Michael Riegler, Michael E. Lamb, Carsten Griwodz, and Saeed Shafiee Sabet. 2022. Towards an AI-Driven Talking Avatar in Virtual Reality for Investigative Interviews of Children . In *2nd edition of the Game Systems Workshop (GameSys '22), June 14, 2022, Athlone, Ireland.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3534085.3534340

1 INTRODUCTION

Sexual and physical abuse of children is unfortunately a widespread problem in our society, with estimated prevalence rates suggesting that such abuse affects the lives of 300 million children around the world [41]. Although abuse seldom leads to the death of the child, it often causes long-term psychological and physical health problems [3, 5, 35, 38]. Recognizing and preventing child abuse is thus of vital importance for the child protection services (CPS) and police personnel. In cases of child abuse, children are not only the victims of the alleged crime, but often also the sole witnesses. Thus, in order to obtain reliable and detailed accounts of the alleged crimes and to help these children, investigative agencies must correctly perform high-quality interviews [18].

Research over the past three decades on children's cognitive development, combined with knowledge derived from many thousands of interviews conducted by police officers, social workers, medical professionals, and clinicians, has contributed to the development of 'best practice investigative interview guidelines. These

recommendations and guidance advise police and social service agencies around the world on how to conduct better investigative interviews with young victims [17, 22, 33]. Today, there is an international consensus among researchers and professionals that best practice is to ask open-ended question (e.g., "what happened after that?" where the child gets no clues that could influence the answer it gives). On the other hand, avoid asking suggestive or leading questions (e.g., "the man touched your private part, didn't he?" where clues are given and can lead the child to tell a certain story). Following the best practice guidelines is the only way to elicit reliable and detailed accounts of the incidents which are admissible in court [16]. It also decreases the risk of false reports, which can result in an innocent persons being charged and convicted [17]. Hence, the interview skills of CPS and police personnel greatly affect the outcomes of courts trials regarding abused children.

However, studies conducted in countries all over the world show that the quality of the interviews conducted with maltreated children is consistently poor, and interviewers often fail to follow the recommended best practices [1, 4, 13, 37]. Several studies have shown that only 2-8% of the questions asked in the average investigative interviews are open-ended, whereas suggestive or leading questions comprise from 16-60% of the questions asked [17].

Studies have shown that computer-based interactive learning can further improve investigative interviewing skills [26], this is not enough. However, the development of dynamic avatars has opened a new dimension for the transfer of knowledge. Currently, systems that employ such avatars lack generality and realism, using a multiple-choice questions format in which interviewers choose the best option offered. Although this does help interviewers select the right questions to ask, it does not help interviewers to verbalize and remember questions or question types [40]. Another issue with current avatars is that they do not seem visually realistic and do not show any emotions [23, 43].

This paper proposes an interview-training system using realistic avatars created using a game engine and virtual reality (VR). This system offers effective training of professionals in CPS and law enforcement, enabling them to consistently conduct high-quality investigative interviews of children who are alleged victims of sexual or physical violence. Users of this system will talk to an artificial abused child, having a story to tell, with the goal to elicit information which can be used in the court as evidence. Success of the interaction is judged based on the type of questions asked and closure of the story. In particular, we focus on the system challenges and quality aspects of such an avatar. Moreover, we describe the architecture of the proposed child avatar system, experiment with the system performance for both functionality and resource usage, and investigate the system's potential to enhance the competence of CPS and law enforcement professionals effectively. Our system is evaluated objectively by assessing its performance aspects and subjectively by conducting a user study targeting the various learning and quality of experience (QoE) aspects. A test based on the ITU-T Recommendation P.809 [11] was designed and conducted: an interactive test using experts in the field, to investigate the user experience and potential of the system for training CPS personnel. It shows that the proposed prototype is working well in practice and is well received by the experts.

In summary, the main contributions of our work are:(1) Development of the dialogue model to mimic an abused child; (2) Development of the prototype of an artificial intelligence (AI)-based child avatar in VR using the Unity game engine; (3)Evaluating the prototype in terms of system performance, user experience and learning effects.

2 RELATED WORKS

A large amount of research has been conducted to enhance the training of the CPS professionals. Powell et al [26] introduce a program to improve child investigative interviewer performance made up of computer-based training activities like watching videos, reading material and quizzes. These activities are followed by 10 minute mock assessment interviews where a trained actor play the role of a child. Empowering Interviewer Training (EIT) [25] is another program that simulates investigative interview. In EIT, child memories and responses are predefined, and rule-based algorithms are employed to select responses. An operator chooses a prerecorded video clip with different emotions which is shown to the users, based on the selected response. Moreover, Linnæus university and AvBIT Labs have developed an online interview setting, using prerecorded videos of a child avatars and audio responses. An avatar controller using the Wirecast software controls emotion and responses that should be shown to the user via the Skype interface [6, 12]. Although these systems have shown improvement in transferring investigative interviewing skills, these system are too rigid in the response generation, lack generality and have a human input during the interactions which make them expensive and harder to operate. In this paper, we propose an AI-driven system in which the talking avatar can dynamically handle the questions and can provide a higher realism for the interviewers. In addition, by removing human input in the system, it would be cost effective by saving the cost of training or hiring operator/actors. Lack of the need for human operators will lead to 24/7 availability for frequent training sessions. Such a system is realizable to conceive by employing different system components from the field of AI such as dialogue models, generation of digital visual avatars, and presenting that in a VR environment. In the rest of this section, we discuss the state of the art in each of these system components.

2.1 Conversations

Dialogue models have attracted significant attention in recent years for replacing systems that communicate with humans. They have been around for years now and have mainly been employed in the customer service, commercial, and health sectors. We are exploring the use of dialogue models in educating and enhancing learning. Most research has employed them as assistants that answer students' questions and help them accordingly [9, 20, 36].

The introduction of transformers [39], a deep learning model which employs an encoder-decoder framework with a self-attention mechanism, accelerated the development of language generation models like the GPT series [2, 27, 28] and the Bidirectional Encoder Representations from Transformers (BERT) model [7]. Before the introduction of transformers, natural language processing (NLP) models mostly relied on recurrent neural networks like gated recurrent units (GRUs) and Long short-term memory (LSTMs) with

an additional attention mechanism. Dialogue models developed using these models and large open-domain conversational corpora have improved the quality of responses generated. However, non-coherent, off-topic, and uninformative response generation, especially for domain-specific cases, remains a challenge for these models.

2.2 Vitual Reality

Virtual Reality is an immersive media technology that users experience the virtual environment created entirely in a computer simulation. Qualinet Whitepaper [24] classifies the immersive applications based on the level of interaction and number of senses involved in the application. Although more senses do not necessarily lead to higher immersion [8], information can be conveyed faster using more senses, which accelerates the creation of immersion. The levels of interaction while consuming media can be anywhere between being passive and interactive. A proper interactive system can make media consumption more immersive [8]. However, it does not mean that interaction always increases the level of immersion [8], i.e., giving freedom to the users sometimes can make the story even chaotic [29], and as a result, less immersive. VR has been used in many applications such as gaming, omnidirectional video [42], remote control and industrial applications, and health. In this paper, we will show how it can be merged with a dialogue model to create an interactive storytelling system that is beneficial for training CPS professionals.

3 THE AI-BASED CHILD AVATAR

This section discusses details about the developed child avatar, with different components integrated. Figure 1 shows the high-level overview of the architecture, where language module is a dialogue model developed on Rasa[1] is placed in the back-end, the font-end visual component is running in an Oculus Quest device developed in the Unity game engine, and the auditory component is based on the IBM Watson cloud service.

3.1 Visual

The visual part of the child avatar system was developed using the Unity game engine, with results shown to users using Oculus Quest 2. The child avatar was developed using the open-source project Unity Multipurpose Avatar (UMA)[2] in which character customization to combine meshes and textures are possible. Afterward, the voices were matched to the avatar using the Unity game engine asset Salsa Suit[3] to generate eye, head, and lip movements synced to the voices. In addition to facial movements, animations were added to the avatar's hand and neck to create gestures similar to a child talking by employing some simple animations which were repeating.

3.2 Language

A dialogue model was developed and deployed at the back-end using Rasa, an open-source framework to develop solutions to automate text or voice-based conversations. The decision to use

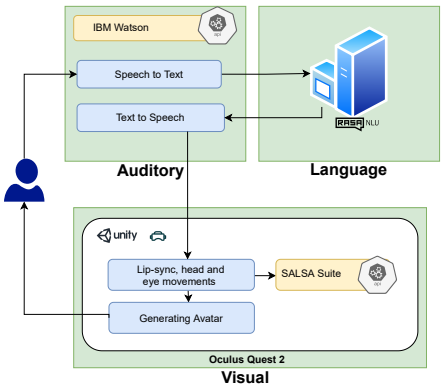

Figure 1: Architecture of the proposed child avatar.

Rasa to develop our dialogue model was motivated by the lack of available data and the desire to validate it as proof of concept for our proposed solution. We had two hundred transcripts of well-conducted training interviews from the Centre for Investigative Interviewing at Griffith University, Australia [26]. This dataset contained the conversations between a 5-7 years old child and interview trainees in a mock interview setting (CPS and police). A professional actor mimicked the child in this case.

3.3 Auditory

We tested different speech synthesis services with the main requirement to sound as child-like as possible. Based on this, we chose to use the IBM Watson services for text-to-speech[4] (TTS) and speech-to-text[5] (STT) synthesis. Watson TTS and STT are cloud service APIs that serve as communicative bridges between the language (back-end) and Visual (front-end) components. The user communicates with the front-end verbally, with the question uttered by the user sent to the IBM STT API to be transcribed with the response, then forwarded to the back-end. At the back end, the dialogue model processes this user utterance and generates an appropriate response. This response is then sent to the IBM TTS API, which then sends the generated audio response to the user at the front end.

3.4 Integration

Ngrok[6] services were used to make the back-end accessible over the internet, with tunneling used to make local servers available over the Internet. The VR front-end accessed the Rasa REST API using an HTTP Post method. The HTTP Rest APIs for the IBM Watson STT and TTS services were accessed using generated API keys and endpoint URLs for each service.

4 SYSTEM PERFORMANCE

In this section, we discuss the computational performance of our implementation. Computers with similar specifications were used at both the front-end and back-end. The system used was an HP

[1] https://rasa.com/
[2] https://github.com/umasteeringgroup/UMA
[3] https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442
[4] https://www.ibm.com/cloud/watson-text-to-speech
[5] https://www.ibm.com/no-en/cloud/watson-speech-to-text
[6] https://ngrok.com/

Pavilion Gaming Desktop TG01-0 model with an AMD Ryzen 7 3700X processor, 16 GB physical memory, an 8 GB NVIDIA GeForce RTX 2060 Super graphics processing unit (GPU), and a 512 GB SSD with Windows 10 operating system running at each end.

Front-end: The Unity3D and Oculus windows applications[7] running with Oculus Link were the primary applications run at the front-end. Utilization of resources was observed over each batch of 10 seconds. The average total CPU usage was 9.7%, and the physical memory load was 48.5% (7912 MB). The observed average GPU memory usage was 74%, GPU core load was 39%, and GPU video engine load was 29.5%. Both of these applications are mostly GPU dependent, and our off-the-shelf system is able to handle the load sufficiently. Both of these applications are mostly GPU dependent, and our system is able to handle the load of these off-the-shelf sufficiently.

Back-end: The Rasa REST API, Rasa action server and ngrok tunneling services were the primary applications running in the back-end. The utilization of resources was observed over each batch of 2-minute conversations with a dialogue model. The average total CPU usage was 1%, and the physical memory load was 35% (5723 MB). The average GPU memory usage was 4.8%, GPU core load was 4%, and GPU video engine load was 0%. The Rasa back-end was not very resource-intensive w.r.t to both CPU and GPU. The only use of the GPU occurred when the intent classification models were called using tensorflow backend.

Text to Speech: The average, max and min times from sending the text to IBM Watson TTS API to receiving the audio response were 1179 ms, 3151 ms and 429 ms, respectively.

Speech to Text: These times were calculated from sending the 10-second audio to IBM Watson STT API till receiving of the text. The SST API processes the audio snip-it as a stream while transcribing and seeks to optimize the textual response over the duration of the audio. Average, max and min times were 12 sec 544 ms, 13 sec 439 ms, and 12 sec 9 ms, respectively. These time measurements includes the length of the 10-second audio as well.

Overall, the system performance results show that an average PC can run the components for the current iteration. Although the prototype is designed with the game engine, it is not as sensitive to time delay as games where a delay of around 200 ms can destroy the user experience [30]. In the case of interviewing a child, delays up to even a few seconds can be considered normal due to the child's hesitation. In the next section, we discuss the system's responsiveness to evaluate this aspect in more detail.

5 INTERACTIVE USER STUDY

This study was an interactive test paradigm based on ITU-T Rec. P.809 which suggests that participants should interact with the system in a designed scenario using a specific assessment method. In this study, users were asked to interact with the system by interviewing the child avatar. This study was designed to evaluate the user's QoE as well as to investigate how effectively the system enhanced the learning experience and the acquisition of knowledge and skills about communicating with abused children. In summary, the main research questions of this study are the following:

- **RQ1:** How is the experts' QoE from using the system?
- **RQ2:** Can such a system enhance the knowledge and skills in interviewers maltreated children?
- **RQ3:** What are the most important quality aspects in the success of such a system?

5.1 Experiment Design

The invited participants were all experienced in the field of child protection because users with no-prior experience do not have the competency either to ask the right questions or to judge the quality of the interview. In this study, users (CPS workers and child welfare students) interacted with the child avatar prototype by wearing a head-mounted display (Oculus Quest 2) and engaging with the avatar in a VR environment in which the child avatar performed spoken conversations using a STT and TTS system as discussed in Section 3. A screenshot of the subject wearing the VR headset and interacting with the child avatar[8] is shown in Figure 2.

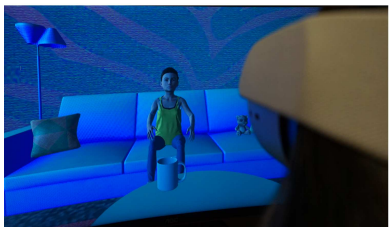

Figure 2: A screenshot of the VR user-view of the child avatar.

5.1.1 Demographic of the participants. In this study, 11 experienced subjects, including seven females, three males, and one transgender participated. They were first asked to rate their experience with CPS and VR using a 10-point scale ranging from not experienced to expert. The mean (M) participant's experience in CPS was 6.18 (standard deviation (SD)=2.67). However, these experts in the field of CPS were inexperienced users of VR (M=1.18, SD=.45), i.e., most of the participants were experiencing VR for the first time. For that reason and for the sake of training, the participants interviewed the child avatar twice in the VR environment. Fortunately, none of the participants reported experiencing any motion sickness after using the prototype. Two users expressed confusion in the first interview, but claimed that they could adapt to the situation.

5.1.2 Questionnaire. After each interview, the experts were asked to fill out a post-experience questionnaire as well as to answer a few open-ended questions targeting their opinion about the use of the virtual avatar for training. The questionnaire was designed to measure overall user experience, the interactivity of the system (Responsiveness from GIPS [34] as discussed in [31]), flow items in iGEQ [10] as well as four questions proposed by social science experts regarding learning effects. The first items in the questionnaire used a 7-point continuous scale, known as Extended Continuous-ACR scale as was recommended in ITU-T Rec. P.809, and the rest of the questions used 5-point Likert scale. However, the EC-ACR is also based on a Likert scale, which uses five labels ranging from "strongly disagree" to "strongly agree". In order to transfer it back

[7] https://www.oculus.com/setup/

[8] https://www.dropbox.com/s/k4yokutx27btt2b/AvatarDemo.mp4?dl=0

to a 5-point ACR, the scale transformation [15] was used. Table 1 provides an overview of the items in the questionnaire.

5.1.3 Story. To setup the story-line for this experiment, transcripts of interviews were clustered based on different personas. Availability of multiple numbers of different transcripts and enough details to generate conversations of reasonable length was used to select the persona for this study. The child avatar in this story is named Theresa, and she is five years old. She visits an adult friend, named Jerry, every Tuesday, and there is suspicion of sexual harassment or abuse on her last visit. Her mother has brought her to the child protection agency to be interviewed about the suspected abuse. Participants of the interactive study interviewed the child avatar verbally with the goal to elicit her account of the incident.

Table 1: List of questionnaire items [34], [10]

Overall QoE
QoE: How would you rate your overall quality of experience?
Responsiveness
RE1: I noticed delay between my actions and the outcomes.
RE2: The responsiveness of my inputs was as I expected.
RE3: My inputs were applied smoothly.
Flow
FL1: I forgot everything around me.
FL2: I felt completely absorbed.
Learning Effects
Communication: It improves knowledge and skills in communication.
Self-efficacy: It can enhance my self-efficacy
Learning Experience: It would enhance my learning experience.
Pedagogical Benefits: It has pedagogical benefits.

5.2 Evaluation and Results

This section discusses the results of the conducted experiment. The results are discussed based on various aspects, namely the user experience, learning effect, expert comments, and importance of the quality aspects.

5.2.1 RQ1: User Experience Aspects. Figure 3 shows the user ratings of QoE, Flow, and Responsiveness in their first and second interviews. Overall, we can see a trend towards improvement in the user experience after training with the system. For the overall QoE, the first interview had a $M = 3.62$, $SD = 1.25$, and the second interview created a Mean Opinion Score (MOS) of $M = 3.92$, $SD = .71$, which implies a good QoE on the ACR scale. However, a paired-sample t-test showed that there was no significant difference between the MOSs of the QoE in the first and second interviews $t(10) = -1.47$, $p = .17$. The results were similar for Flow with the experts feeling more immersed in the second interview ($M = 3.45$, $SD = .57$) than in the first ($M = 3.0$, $SD = .89$), but the difference was not significant ($t(10) = -1.66$, $p = .12$). However, the user responsiveness significantly improved from the first interview ($M = 2.78$, $SD = .42$) to the second interview ($M = 3.18$, $SD = .45$, $t(10) = -3.13, p = .01*$) because the participants had to press a button each time they finished a question. This contrasts with natural speaking, and participants needed time to adapt to the system. In addition, because of the button pressing, the participants felt there was a delay between their actions and the outcome of the system, underlining the need to improve the responsiveness of the system. A button-less experience should be implemented as a future improvement.

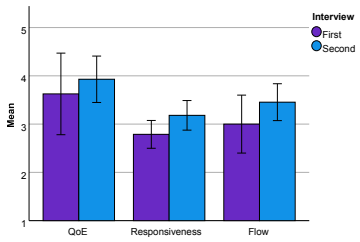

Figure 3: Bar-plot (95% confidence interval) of means showing quality assessments in the first and second rounds of testing.

5.2.2 RQ2: Learning Effects. Figure 4 shows the bar-plot of user ratings for all items regarding learning effects in the questionnaire. One can observe that, in all the aspects, experts agreed that the virtual avatar could help them to improve themselves. 72% of the participants stated that such an AI-based child avatar could help them to acquire knowledge and skills in communication, 81% opined that the avatar could enhance their self-efficacy, and all the participants agreed that the avatar could improve their learning experience.

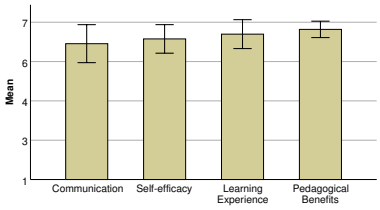

Figure 4: Barplot (95% confidence interval) of means for learning effects.

5.2.3 RQ3: Importance of the Quality Aspects. The experts were asked about their opinions on the importance of each quality aspect. This could help us to further improve the system with respect to the requirements and expectations of actual users of the child avatar system. Experts were asked to sort the importance of the aspects *interactivity, impassiveness, realism, presence, escapism, and investigative* from 1 to 6 (least important to most important). Figure 5 shows the importance of each aspect based on the experts' judgment. Interactivity, realism, and presence were rated as more important, identifying our next steps for developing the child avatar system.

5.2.4 Expert Comments and Suggestions. In addition to the questionnaire, experts were asked to write a paragraph about the child avatar system. The written texts were gathered and coded to extract quantifiable results. Six of 11 participants mentioned that the virtual avatar was helpful/useful and could assist them, six experts stated that the virtual avatar is very important/valuable, 3 experts stated that the child avatar provided realistic training, and all the participants wrote that training/practicing with the system would enhance their knowledge/insights. Interestingly, one of the experts said, "I already learned in this first conversation how child conversations can be improved." However, this study aimed only to evaluate the first version of the system and gather comments from the experts rather than to teach them. In addition to the written texts, users' feedback during the experiment was collected when they were thinking out loud. Some of the participants expressed confusion due to the need to press the button each time after finishing their questions. A couple of participants expressed the urge to take notes, which was impossible due to the VR head-mounted display constraints.

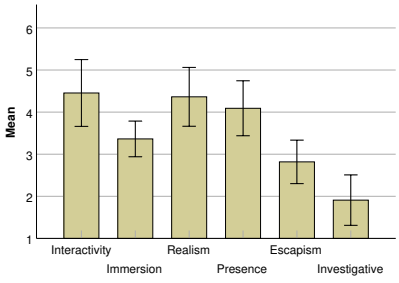

Figure 5: Bar-plot (95% confidence interval) of mean of the importance of aspects.

6 DISCUSSION AND FUTURE WORK

This paper proposed an AI-driven talking avatar in VR that could understand and respond to human language to emulate an investigative interview with maltreated children. As a prof of concept, we presented the system architecture of the prototype, and using a subjective study with expert participants. In addition, the aim was to have a qualitative study with experts where they are able to express their opinion with open questions rather than a quantitative study. The study showed that the system can improve the knowledge and skills of the CPS and law enforcement personnel. In addition, it was shown that having the system in the VR can create a higher level of QoE and helps to increase the realism and presence that were rated important by experts. We targeted a specific group of participants who either have active or passive experience in interacting with abused children and are familiar with current teaching/training programs. The low number of participant is attributed to the fact that it is hard to manage such professionals to participate in our research as they had their professional commitments as well. However, high agreement between the experts indicates that even higher number of such professionals would have produced similar results.

The current implementation of our system is meant to serve as a baseline for our future work in this research area. We are currently working to develop a more generalizable and flexible dialogue model using GPT-3, coupled with deterministic models built on annotated conversational data for feedback mechanism and prompt generation. Based on the experts' feedback, the system's responsiveness has to be improved, and a button-less experience is planned. Moreover, note-taking during the interviews was pointed out as a critical activity an interviewer often perform during the interview, which we plan to address in future work. Emotional recognition of child responses can be exploited to alter the features of child audio like tone or pitch to give a more realistic experience and visual expression of these sentiments. Recent research has shown that there is a high correlation between non-verbal and verbal expressions of emotions [14]. Contrast to that, experienced experts are of the opinion that maltreated children show and deal with emotions differently, and some traumatized children do not show any emotions at all. Therefore, displaying emotion during the conversation should be within the context of children with different personas. To create realistic personas, there is a need for a video dataset of such interviews, which is not available to us at the time, as the data is very sensitive so it's hard to get hold of it. We are collaborating with our partners in CPS and police from two different countries to get access to this data, and to be able to add emotions to audio/video into the system in future works. Furthermore, we will also investigate the use of realistic talking avatars in more detail and whether the uncanny valley effect [19, 21] exists in the context of such child interviews, meaning that users feel a sense of anxiety or discomfort in response to a highly realistic avatars resulting in low QoE.

7 CONCLUSION

In this paper, an implementation of our proposed VR-based interview training system was described. It is developed using the Unity game engine and a dialogue model developed using RASA. Audio communication with system was synthesised using IBM text-to-speech and speech-to-text services. The game system was discussed and evaluated objectively in terms of system performance and subjectively using an interactive user study. The participants of this study were CPS experts, and the results show that the child avatar system is well received by the experts and can help them obtain better communication skills for interviewing abused children. In addition, our findings identified *interactivity*, *realism* and *presence* as the important quality aspects to target a successful training program. Using the results of this study, a second prototype will be designed to create more presence and realism for users, which validates our initial hypothesis of using VR over other interfaces [32]. Finally, we conclude, from the observed system performance of our first avatar version, that a standard PC can handle the system's processing requirements.

ACKNOWLEDGMENTS

This research is sponsored by the Research Council of Norway, project number #314690.

REFERENCES

[1] Gunn-Astrid Baugerud, Miriam S Johnson, Helle BG Hansen, Svein Magnussen, and Michael E Lamb. 2020. Forensic interviews with preschool children: An analysis of extended interviews in Norway (2015–2017). *Applied Cognitive Psychology* 34, 3 (2020), 654–663.

[2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[3] Nathalie Carrick, Jodi A Quas, and Thomas Lyon. 2010. Maltreated and nonmaltreated children's evaluations of emotional fantasy. *Child Abuse & Neglect* 34, 2 (2010), 129.

[4] Ann-Christin Cederborg, Yael Orbach, Kathleen J Sternberg, and Michael E Lamb. 2000. Investigative interviews of child witnesses in Sweden. *Child abuse & neglect* 24, 10 (2000), 1355–1361.

[5] Dante Cicchetti and Sheree L Toth. 2005. Child maltreatment. *Annu. Rev. Clin. Psychol.* 1 (2005), 409–438.

[6] Kevin Charles Dalli. 2021. Technological Acceptance of an Avatar Based Interview Training Application: The development and technological acceptance study of the AvBIT application.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Pierre Gander. 1999. *Two myths about immersion in new storytelling media.* Lund University.

[9] Sebastian Hobert. 2019. Say hello to 'coding tutor'! design and evaluation of a chatbot-based learning system supporting students to learn to program. (2019).

[10] Wijnand A IJsselsteijn, Yvonne AW de Kort, and Karolien Poels. 2013. The game experience questionnaire. *Eindhoven: Technische Universiteit Eindhoven* 46, 1 (2013).

[11] ITU-T Recommendation P.809. 2018. *Subjective Evaluation Methods for Gaming Quality.* International Telecommunication Union, Geneva.

[12] David Johansson. 2015. Design and evaluation of an avatar-mediated system for child interview training.

[13] Miriam Johnson, Svein Magnussen, Christian Thoresen, Kyrre Lønnum, Lisa Victoria Burrell, and Annika Melinder. 2015. Best practice recommendations still fail to result in action: A national 10-year follow-up study of investigative interviews in CSA cases. *Applied Cognitive Psychology* 29, 5 (2015), 661–668.

[14] Yael Karni-Visel, Irit Hershkowitz, Michael E Lamb, and Uri Blasbalg. 2021. Nonverbal emotions while disclosing child abuse: the role of interviewer support. *Child maltreatment* (2021), 10775595211063497.

[15] Friedemann Köster, Dennis Guse, Marcel Wältermann, and Sebastian Möller. 2015. Comparison between the discrete ACR scale and an extended continuous scale for the quality assessment of transmitted speech. *Fortschritte der Akustik, DAGA* (2015).

[16] Michael E Lamb. 2016. Difficulties translating research on forensic interview practices to practitioners: Finding water, leading horses, but can we get them to drink? *American psychologist* 71, 8 (2016), 710.

[17] Michael E Lamb, Deirdre A Brown, Irit Hershkowitz, Yael Orbach, and Phillip W Esplin. 2018. *Tell me what happened: Questioning children about abuse.* John Wiley & Sons.

[18] Michael E Lamb, Yael Orbach, Irit Hershkowitz, Phillip W Esplin, and Dvora Horowitz. 2007. A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the NICHD Investigative Interview Protocol. *Child abuse & neglect* 31, 11-12 (2007), 1201–1231.

[19] Karl F MacDorman, Robert D Green, Chin-Chang Ho, and Clinton T Koch. 2009. Too real for comfort? Uncanny responses to computer generated faces. *Computers in human behavior* 25, 3 (2009), 695–710.

[20] Fernando A Mikic-Fonte, Martín Llamas-Nistal, and Manuel Caeiro-Rodríguez. 2018. Using a Chatterbot as a FAQ Assistant in a Course about Computers Architecture. In *2018 IEEE Frontiers in Education Conference (FIE)*. IEEE, 1–4.

[21] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.

[22] Chris Newlin, Linda Cordisco Steele, Andra Chamberlin, Jennifer Anderson, Julie Kenniston, Amy Russell, Heather Stewart, and Viola Vaughan-Eden. 2015. *Child forensic interviewing: Best practices.* US Department of Justice, Office of Justice Programs, Office of Juvenile ….

[23] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[24] Andrew Perkis, Christian Timmerer, Sabina Baraković, Jasmina Baraković Husić, Søren Bech, Sebastian Bosse, Jean Botev, Kjell Brunnström, Luis Cruz, Katrien De Moor, et al. 2020. QUALINET white paper on definitions of immersive media experience (IMEx). *arXiv preprint arXiv:2007.07032* (2020).

[25] Francesco Pompedda, Angelo Zappalà, and Pekka Santtila. 2015. Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychology, Crime & Law* 21, 1 (2015), 28–52.

[26] Martine B Powell, Belinda Guadagno, and Mairi Benson. 2016. Improving child investigative interviewer performance through computer-based learning activities. *Policing and Society* 26, 4 (2016), 365–374.

[27] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[29] Marie-Laure Ryan. 1999. Immersion vs. interactivity: Virtual reality and literary theory. *SubStance* 28, 2 (1999), 110–137.

[30] Saeed Shafiee Sabet, Steven Schmidt, Saman Zadtootaghaj, Carsten Griwodz, and Sebastian Möller. 2020. Delay sensitivity classification of cloud gaming content. In *Proceedings of the 12th ACM International Workshop on Immersive Mixed and Virtual Environment Systems*. 25–30.

[31] Saeed Shafiee Sabet, Steven Schmidt, Saman Zadtootaghaj, Babak Naderi, Carsten Griwodz, and Sebastian Möller. 2020. A latency compensation technique based on game characteristics to mitigate the influence of delay on cloud gaming quality of experience. In *Proceedings of the 11th ACM Multimedia Systems Conference*. 15–25.

[32] Pegah Salehi, Syed Zohaib Hassan, Saeed Shafiee Sabet, Gunn Astrid Baugerud, Miriam Sinkerud Johnson, Michael A. Riegler, and Pål Halvorsen. 2022. Is More Realistic Better? A Comparison of Game Engine and GAN-based Avatars for Investigative Interviews of Children. *Workshop on Intelligent Cross-Data Analysis and Retrieval* (2022).

[33] Karen J Saywitz, Thomas D Lyon, and Gail S Goodman. 2017. 19 When Interviewing Children: A Review and Update. *The APSAC handbook on child maltreatment* (2017), 310.

[34] Steven Schmidt. 2021. *Assessing the Quality of Experience of Cloud Gaming Services.* P.h.D. thesis, Technische Universität Berlin.

[35] Jack P Shonkoff. 2016. Capitalizing on advances in science to reduce the health consequences of early childhood adversity. *JAMA pediatrics* 170, 10 (2016), 1003–1007.

[36] Sharob Sinha, Shyanka Basak, Yajushi Dey, and Anupam Mondal. 2020. An educational Chatbot for answering queries. In *Emerging Technology in Modelling and Graphics*. Springer, 55–60.

[37] Kathleen J Sternberg, Michael E Lamb, Yael Orbach, Phillip W Esplin, and Susanne Mitchell. 2001. Use of a structured investigative protocol enhances young children's responses to free-recall prompts in the course of forensic interviews. *Journal of applied psychology* 86, 5 (2001), 997.

[38] Sheree L Toth and Dante Cicchetti. 2013. A developmental psychopathology perspective on child maltreatment. *Child maltreatment* 18, 3 (2013), 135–139.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[40] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* (2015).

[41] WHO. 2020. *Child maltreatment.* https://www.who.int/news-room/fact-sheets/detail/child-maltreatment

[42] Mai Xu, Chen Li, Zhenzhong Chen, Zulin Wang, and Zhenyu Guan. 2018. Assessing visual quality of omnidirectional videos. *IEEE transactions on circuits and systems for video technology* 29, 12 (2018), 3516–3530.

[43] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.