

Synthesizing a Talking Child Avatar to Train Interviewers Working with Maltreated Children

Author

Salehi, Pegah, Hassan, Syed Zohaib, Lammerse, Myrthe, Sabet, Saeed Shafiee, Riiser, Ingvild, Roed, Ragnhild Klingenberg, Johnson, Miriam S, Thambawita, Vajira, Hicks, Steven A, Powell, Martine, Lamb, Michael E, Baugerud, Gunn Astrid, Halvorsen, Pal, Riegler, Michael A

Published

2022

Journal Title

Big Data and Cognitive Computing

Version

Version of Record (VoR)

DOI

[10.3390/bdcc6020062](https://doi.org/10.3390/bdcc6020062)

Rights statement

© 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Downloaded from

<http://hdl.handle.net/10072/428523>

Griffith Research Online

<https://research-repository.griffith.edu.au>



Article

Synthesizing a Talking Child Avatar to Train Interviewers Working with Maltreated Children

Pegah Salehi ^{1,*}, Syed Zohaib Hassan ¹, Myrthe Lammerse ², Saeed Shafiee Sabet ¹, Ingvild Riiser ², Ragnhild Klingenberg Røed ², Miriam S. Johnson ², Vajira Thambawita ¹, Steven A. Hicks ¹, Martine Powell ³, Michael E. Lamb ⁴, Gunn Astrid Baugerud ², Pål Halvorsen ^{1,2}, and Michael A. Riegler ^{1,5}

- ¹ SimulaMet, 0167 Oslo, Norway; syed@simula.no (S.Z.H.); saeed@simula.no (S.S.S.); vajira@simula.no (V.T.); steven@simula.no (S.A.H.); paalh@simula.no (P.H.); michael@simula.no (M.A.R.)
 - ² Department of Computer Science, Oslo Metropolitan University, 0130 Oslo, Norway; myrthela@oslomet.no (M.L.); ingvildr@oslomet.no (I.R.); rar@oslomet.no (R.K.R.); mirsin@oslomet.no (M.S.J.); gunnba@oslomet.no (G.A.B.)
 - ³ Centre for Investigative Interviewing, Griffith Criminology Institute, Griffith University, Brisbane 4122, Australia; martine.powell@griffith.edu.au
 - ⁴ Department of Psychology, University of Cambridge, Cambridge CB2 3RQ, UK; mel37@cam.ac.uk
 - ⁵ Department of Computer Science, University of Tromsø, 9037 Tromsø, Norway
- * Correspondence: pegah@simula.no; Tel.: +47-92097694



Citation: Salehi, P.; Hassan, S.Z.; Lammerse, M.; Sabet, S.S.; Riiser, I.; Røed, R.K.; Johnson, M.S.; Thambawita, V.; Hicks, S.A.; Powell, M.; et al. Synthesizing a Talking Child Avatar to Train Interviewers Working with Maltreated Children. *Big Data Cogn. Comput.* **2022**, *6*, 62. <https://doi.org/10.3390/bdcc6020062>

Academic Editor: Moulay A. Akhloufi

Received: 30 April 2022

Accepted: 21 May 2022

Published: 1 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: When responding to allegations of child sexual, physical, and psychological abuse, Child Protection Service (CPS) workers and police personnel need to elicit detailed and accurate accounts of the abuse to assist in decision-making and prosecution. Current research emphasizes the importance of the interviewer's ability to follow empirically based guidelines. In doing so, it is essential to implement economical and scientific training courses for interviewers. Due to recent advances in artificial intelligence, we propose to generate a realistic and interactive child avatar, aiming to mimic a child. Our ongoing research involves the integration and interaction of different components with each other, including how to handle the language, auditory, emotional, and visual components of the avatar. This paper presents three subjective studies that investigate and compare various state-of-the-art methods for implementing multiple aspects of the child avatar. The first user study evaluates the whole system and shows that the system is well received by the expert and highlights the importance of its realism. The second user study investigates the emotional component and how it can be integrated with video and audio, and the third user study investigates realism in the auditory and visual components of the avatar created by different methods. The insights and feedback from these studies have contributed to the refined and improved architecture of the child avatar system which we present here.

Keywords: Child Protection Services (CPS); interview training; virtual child avatar; generative pre-trained transformer 3 (GPT-3); generative adversarial networks (GANs)

1. Introduction

Child sexual abuse (CSA), violence, and neglect are major global public health issues that have far-reaching immediate and long-term implications for the children involved, as well as for society in general [1]. Child abuse represents a risk to the child's existence and development. Research proves that it has major consequences for the child, leading to cognitive, behavioral, and social problems, as well as substance abuse, serious mental health problems, and death [2]. CSA alone is identified by the World Health Organization (WHO) as substantively contributing to the global burden of disease [3]. To investigate this problem, child protective services (CPS) and law enforcement personnel must interview the children concerned when planning responses to such abuse; the interviews thus play

important roles in safeguarding children to promote their welfare and protect them from harm [4]. Children are often interviewed as both victims and key witnesses to the abusive incidents [5]. As there is often a lack of corroborative evidence in these cases, informative interviews with child complainants play a crucial role in their investigation [6].

Most abused children do not have any physical signs of abuse [7]. This means that the progress of the investigation depends upon the informativeness of the child's account of the incident. Therefore, the interviewer's ability to ask good questions that maximize the quality and quantity of the information provided by the child is crucial. The investigative interview is thus a vital component of a comprehensive child abuse investigation and a prime opportunity for investigators to elicit accurate and detailed information from alleged victims of child abuse [8]. A large body of research on children's cognitive and social development, supplemented by field studies on children's ability to describe experienced stressful and traumatic events, have identified best practices for how investigative interviewing with maltreated children should be conducted [9]. These best-practice investigative interview guidelines provide interviewers with clear instructions regarding how child witnesses should be questioned and supported in a non-suggestive way during the interview to maximize the value of their testimony [10,11]. In particular, this includes offering the children open-ended prompts while avoiding forced-choice and suggestive questions because free recall questions encourage more accurate and longer responses from children [12,13] (e.g., "What happened next?" where the child is provided with no information that could influence their answers). Interviewers should also avoid suggestive or leading questions (e.g., "the person touched your private part. Is it true?" or any other question that might lead the child to tell a specific story). However, unfortunately most interviewers do not adhere to such best-practice guidelines [14].

Powell et al. [15] found that interactive computer-based learning activities can improve the effectiveness and productivity of investigative interviewers. Moreover, the area of Artificial Intelligence in Education (AIE) facilitates the learning process and transfers knowledge through humanoid robots or virtual avatars. In the same vein, there has been significant progress in recent years towards synthesizing realistic digital humans, avatars, characters, and agents [16]. The approach we intend to follow is to develop a realistic child avatar that involves the integration and interaction of different components with each other, including dialogue models, auditory, emotional and visual components of the avatar. Our initial prototypes [17,18] demonstrated the potential of such an avatar, and following the process of improving the interactive child avatar, we have employed cutting-edge technologies to synthesize a realistic child avatar; for example, we have used the RASA and GPT-3 for dialogues, the IBM Watson service for auditory, the GPT-3 and BART for emotions, and GANs and Unity game engine for the visual appearance of the avatar. In this paper, we systematically discuss the results obtained using these various tools and techniques.

In addition, this article compares and contrasts various state-of-the-art ways of incorporating multiple characteristics of the child avatar in three user studies. The first user study assessed the entire system and found that it was well received by experts who emphasized the importance of realism; the second user study looked at the emotional component and how it could be integrated with video and audio features; and the third user study looked at the realism in the auditory and visual components of the avatar created using various methods. Based on the findings and user comments from these three investigations, the paper describes the architecture of our child avatar system. In summary, the main contributions of our work are:

- An investigation of the potential learning effects and user experience with the system.
- An investigation of the realism of the synthetic voices compared to natural voices.
- An examination of emotion extraction with different models based on children's answers in investigative interviews.
- An investigation of the realism of several methods for generating the appearance of the talking avatar.

- An investigation of the system architecture regarding the integration and interaction of various system components.

The rest of the paper is organized as follows: Section 2 provides an overview of the state-of-the-art with respect to different system components and highlights the importance of the child avatar interview training system. Section 3 describes the system in overview and discusses the material and methods used to develop various components of the system. Section 4 evaluates the performance of various system components and discusses the results of our three user studies conducted using the system components, and Section 5 discusses the results and their limitations along with suggestions for future research. Finally, Section 6 concludes the paper and highlights the main findings.

2. Related Work

This section reviews related work on each of our research questions separately.

2.1. Investigative Interview Training

A number of training approaches have been developed to provide essential information about how to conduct these interviews in compliance with best-practice guidelines. One of the most prevalent methods of imparting knowledge to police officers and CPS workers involves traditional classroom-based teaching. However, a large body of research from all over the world of investigative interviewing of children [9] has shown that attempts to improve the quality of investigative interviewing using such traditional classroom-based instructions have not been productive [14]. Professionals tend not to follow recommended questioning strategies. Instead, they tend to ask many risky option-posing and suggestive questions rather than the open-ended questions that are preferred, resulting in unreliable children's reports about their experiences [19–22]. Following most training programs, participants are required to participate in mock-interview activities involving an instructor/professional actor portraying an abused child, with feedback provided by a trainer. Recently, the use of mechanical avatars in interviewer training have been shown to be advantageous, especially when integrated with feedback [9,15,23–26]. By training interviewers in the adoption of recommended best-practice interview strategies using a dynamic avatar, including question types that encourage children to make the fullest possible use of their cognitive and communicative abilities while avoiding strategies like suggestive questions, has the potential to be beneficial. Conducting simulated interviews may be particularly suitable for basic training, as well as practice and refresher training, which are known to be especially important for the maintenance of skills. Using dynamic avatars in training may be a very efficient way of acquiring and training practical skills. In addition, AI-based interview training is less expensive, can be used over an extended period of time and may be much more accessible to users than traditional in-person-training [15,27].

2.2. Emotions

Emotional intelligence is the ability to perceive, use, understand, and manage emotions [28]. The level of emotional intelligence that an investigative interviewer portrays influences the execution and performance of the emotional labor [29]. Which refers to jobs where employees are expected to recognize emotions and act accordingly. These jobs have guidelines and rules that need to be followed to ensure the quality of the work completed [30]. The results of an investigative interview with an adult depend on how officers handle the emotions of the interviewee. The outcome, the interviewees' well being, and therapeutic jurisprudence are all positively influenced if the interview is conducted in an emotionally intelligent way [31]. Research addressing emotional intelligence in child interviews is limited, but Albaek et al. [32] showed a need to address professionals' emotional distress in child abuse cases in one qualitative meta-synthesis. To train CPS workers and police officers to conduct investigative interviews effectively, it is important that the training module emphasizes the right kind of emotionally expressive responses to children's emotions. By having the avatar express different emotions, just like real children

would portray them, the interviewer can practice handling emotions in different scenarios by, for example, offering non-suggestive emotional support when interviewing children.

Research shows that by classifying emotions through the identification of facial cues, we can distinguish seven universal emotions [33–35]. These seven emotions are *joy, sadness, anger, disgust, contempt, fear, and surprise*. These emotions can be expressed both verbally as well as non-verbally. Katz and Hunter [36] and Karni-Visel et al. [37] classified non-verbal emotions based on real-life interviews and discovered that non-verbal expressions are 10 times more commonly expressed than verbal emotional expressions. However, traumatized children may not show any emotion, either positive or negative [38]. This phenomenon is called numbing. The paradoxical combination of expressing more non-verbal emotions and numbing creates a difficult landscape for the talking child avatar. Therefore, it is important to portray these emotions during the conversation with different degrees of expression based on different child avatar personas. We must pay attention to the extent to which the different personas express different emotions, as this can vary considerably between children. The expression of the emotions will be noticeable in both visual and auditory outputs with emotions, for example, changing facial expressions or the pitch of the voice.

2.3. Chatbot

Typical applications of chatbots are found in call centers, e-commerce customer services, and internet gaming. Chatbots in these sectors employ different machine learning algorithms to conduct auditory or textual conversations with end users [39]. Chatbots that mimic an allegedly abused child are much more complex to realize than simple question-answering or even open-ended social chatbots [40]. Social chatbots have become more realistic and advanced in the last few years. The goal of a social chatbot is to establish an emotional connection with the person who is using it. A social chatbot is not the same as an investigative training module, and it also must be able to recognize emotions and track emotional changes during a conversation. Verbal or non-verbal emotions are an important part of the conversation and can convey a lot of required information. However, there are some implementations of chatbots that can converse with a certain emotional component [41]. XiaoIce is a social chatbot developed by Microsoft [42] which detects the sentiment reflected in the input before responding. However, identifying, understanding and displaying emotions, and modeling the impact of emotions on the conversation quality are challenging tasks that we are currently addressing. In addition to the emotions expressed in the conversation, there is a need to design and generate different personas that have the ability not only to show emotions differently, but also to react to emotionally charged conversations. Li et al. [43] showed how a neural-based approach can model the different personas. Parametrization of emotions and personas into a chatbot could really improve realism and immersiveness for the end-users, i.e., in our case, CPS trainees. Therefore, this social chatbot aims to recognize emotions and present a consistent personality. This personality would involve acting according to a given age, gender, language, speaking style, attitude, level of knowledge, etc. We have started working on these individual aspects of the chatbots, and we discuss them in the sections below.

2.4. Auditory

A conversation is a two-way flow of information. For a human to communicate with a computer program using speech, one must find ways to translate between the textual and auditory domains. This requires so-called speech-to-text (STT) and text-to-speech (TTS) methods that work at low latencies to ensure a natural conversational flow. With the recent advances in function of CNNs [44] and recurrent neural networks (RNNs) [45], STT methods are approaching human parity [46]. Benchmarks such as LibriSpeech [47] and the 2000 NIST Speaker Recognition Evaluation [48] have shown that current state-of-the-art deep learning methods can recognize speech with very few mistakes [49,50]. As auditory speech is more information-rich than pure text, we may also extract paralinguistic information

such as emotions. Databases, such as IEMOCAP [51], can be used to train speech emotion classifiers. Examples of such classifiers include CNN long short-term memory (LSTM) networks [52] and CNNs [53], which take audio and/or (log-mel) spectrograms as input. TTS research has also developed considerably in recent years. Models such as Wavenet [54,55] and Waveglow [55] have shown immense promise in using generative models to produce very realistic speech from input transcripts. These methods are state-of-the-art and work efficiently, which makes them apposite for the system presented here [56].

2.5. Visual

Due to the free access to large-scale datasets and the rapid development of deep learning technologies, especially Generative Adversarial Network (GAN) [57] and its variants, unprecedented successes have been achieved in the generation of realistic multimedia content.

Among the recent innovations in humanoid avatars, the Uneeq (<https://digitalhumans.com/>, accessed on 24 May 2022) has designed a conversational AI system based on video and audio components with natural voice and highly realistic facial expressions, and the Soul Machines (<https://www.soulmachines.com/>, accessed on 24 May 2022) has developed a digital brain that emulates human cognitive processes.

In the following, we provide a basic classification for understanding the differences between various facial manipulation techniques regarding the amount of manipulation (as shown in Figure 1). A brief explanation of each of them is given below.

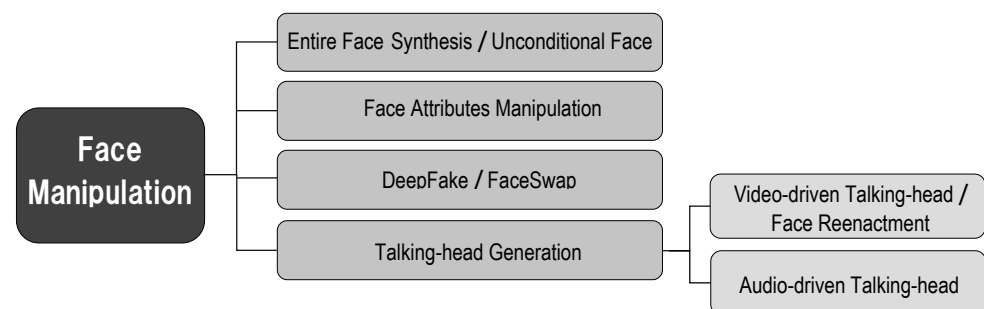


Figure 1. A comprehensive category of face manipulation techniques.

Entire Face Synthesis (or Unconditional Face Generation) means the production of non-existent portrait images in the real world, e.g., PGGAN [58] and StyleGAN (<https://thispersondoesnotexist.com/>, accessed on 24 May 2022) [59] fall into this manipulation, which generates ultra-realistic photos of humans who do not actually exist; even humans have difficulty assessing their realism.

Face Attributes Manipulation consists of editing some attributes of a face such as hair, skin color, eyeglasses, gender, and age as when using StarGAN [60]. Moreover, FaceApp (<https://www.faceapp.com/>, accessed on 24 May 2022) has popularized facial attribute editing as an application.

DeepFake (or FaceSwap) commonly replaces one person's face in an image or video with another person's. In this respect, the combination of source videos results in a fake video that shows an action or an event that never occurred in reality, for example, creating fake news [61]. In the same way, a digital avatar has been developed to swap faces in video chats (<https://blog.siggraph.org/2021/01/ai-avatars-virtual-assistants-and-deepfakes-a-real-time-look.html/>, accessed on 24 May 2022). The DeepFaceLab (<https://github.com/iperov/DeepFaceLab>, accessed on 24 May 2022) [62] is also an open sourced software package to build high fidelity face-swapping videos.

Talking-head Generation can synthesize precise lip synchronization, head pose motion, and natural facial expressions from a specific person by capturing one or more input signals. Depending on the input data type, the existing approaches to talking-head generation can be divided into two broad categories: audio-driven and video-driven. Many

audio-driven methods avoid straight mapping from audio to image, but instead first map audio to an intermediate step like 2D facial landmarks [63–67] or 3D face shapes [68–72], and then render photorealistic videos. Audio-driven talking-head generation is inherently difficult because it generates hand and mouth movements based only on an audio signal. Thus, some of the proposed works control the motions of a subject with one or more additional videos as input categorized as **video-driven** methods (a.k.a **face reenactment**) [73–78]. Similarly, Face2Face [73] enables real-time facial reenactment of a target video sequence, i.e., animating the facial expressions of the target video using a source actor and re-rendering the manipulated output video in a photorealistic fashion. Some models also focus on training their model using a specific person [79,80]. Audio-driven methods have received more attention in recent research because they aim to synthesize a universal talking-head model for various subjects and applications.

2.6. Child Interview Training Avatars

Empowering Interviewer Training (EIT) [27] is an investigative interview training program. Child responses are pre-defined in the system and responses are selected using a rule-based algorithm. Based on the selected response, prerecorded videos of children showing different emotions are chosen by a human operator and shown to the user. Pomedda et al. conducted multiple studies using this system to analyze the training effects after having multiple sessions of these mock interviews, the effects of feedback and reflection on improving the quality of the investigative interviews [24,81–83]. In Sweden, Linnæus university and AvBIT Labs have also introduced an online interview training system. They also use prerecorded audio responses and videos of a child avatar and human operator. The user is shown an appropriate video response with suitable emotions controlled using Wirecast software controls via the Skype interface [84,85]. Even though development and testing of these systems have successfully transferred the investigative skills needed to interview abused children, these systems are not dynamic in the response generation and have human input during the response selection phase. This makes them rigid and harder to operate.

We propose an AI-driven training system that can dynamically respond to the questions and provide a higher realism during the training interviews. In addition, we propose a system that would be completely independent of human input.

3. Materials and Methods

Figure 2 provides an overview of the entire architecture of the child investigative interview avatar. Our proposed idea for the interaction flow between different system components and the flow of data/information between them is well traceable in the architecture. Yellow parts mark the language model, where GPT-3 is trained on real and mock interviews, green parts are the emotional engine that extracts emotion from the generated texts and provides the input to be applied in the avatar's visual and auditory output, blue parts are the auditory system, which uses IBM Watson API to convert the text to audio and audio to text, and purple parts show the visual sections, where the generated audio and visual aspects of the avatar are combined to create a talking face.

Although all of these modules can be developed in isolation, integrating all the components would still be needed. This section discusses future research and the work in progress for each component.

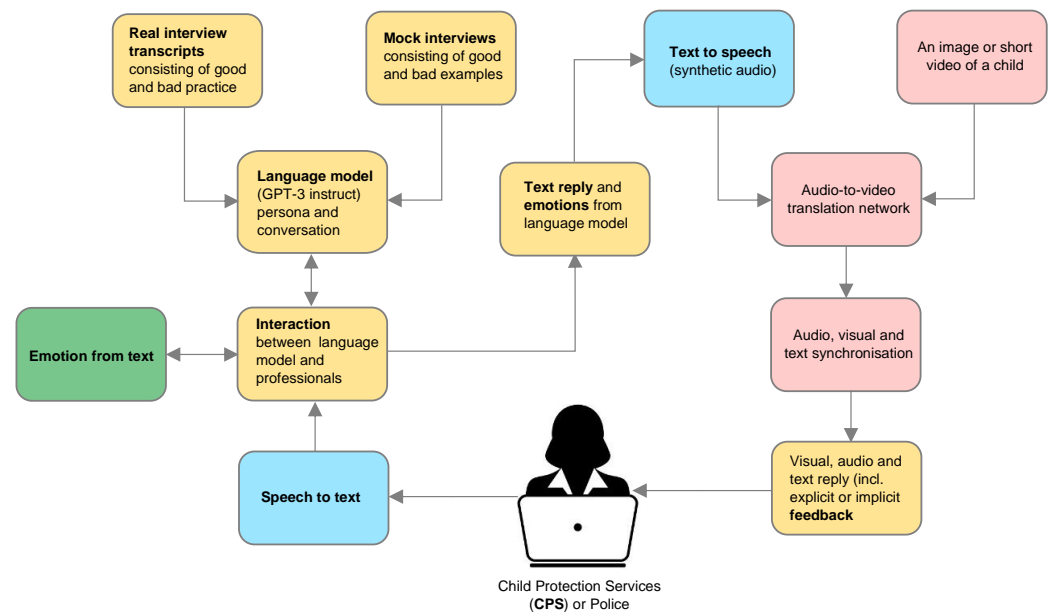


Figure 2. System architecture. Green blocks denote the interactive parts, yellow blocks are language-related, blue audio-related, and pink the parts of the system related to visualization.

3.1. Language

There are two sources of data, mock and real interviews, which will provide the training data for our chatbot. The Centre for Investigative Interviewing at Griffith University, Australia [15] provided 1000 transcripts of the mock interviews conducted as part of their investigative training for social workers, police officers and psychologists. In these mock interviews, a trained actor mimics an allegedly abused child. Real-life child investigative interviews will be added to the system at a later stage since we want to initially have a more rigid and controllable version of the avatar.

The current chatbot was developed using Rasa (<https://rasa.com/>, accessed on 24 May 2022). It provides an open source framework to develop the automated solutions for text-based conversations. The choice of RASA at this stage was motivated by the fact that we lack enough data to develop a solution from scratch and aim to develop a prototype for proof of concept. Additionally, with the small amount of data, RASA provides an environment to control the flow of conversation. At the time of developing the first chatbot, we had two hundred transcripts of well-conducted training interviews. The dataset we created comprises conversations between a child aged 5–7 years and an interviewer. RASA is powered by the TensorFlow [86] back-end framework. RASA has multiple modules which employ different deep learning models to develop a complete dialogue model. Each module takes in the training data in a module specific format. The NLU module employs the Dual Intent and Entity Transformer (DIET) [87] for training to predict intent and entities in the utterances jointly.

We extracted personas from the available transcripts by manually clustering them based on scenario and chose one of them to develop our RASA chatbot. We aim to build a chatbot that can show different personas and we are working with GPT-3 to achieve that. We plan to use fine-tuned GPT-3 on our transcripts to allow us to capture the behavior of children with different personas dynamically. We classify questions asked by interviewer into 15 different categories and child responses as productive and non-productive. We plan to use these data to develop deterministic models that can be employed for feedback mechanism and regulating GPT-3 during the conversation to alter the behavior of the child bot based on the type of question asked.

We conducted an interactive study that followed ITU-T Rec. P.809 [88] test paradigm which suggests that participants should be evaluated using a specified method after they interact with the system in a specific scenario. In this study, we invited participants from CPS agencies who had some experience interviewing abused children [17]. They were asked to interview the child avatar mimicking a six years old child allegedly sexually abused. The goal of this study was to evaluate the user's quality of experience and assess the system's capability to enhance the learning experience and the acquisition of knowledge and skills for communicating with abused children.

3.2. Auditory

We tested different speech synthesis services to see which one sounded the most childlike. We chose to use IBM Watson services for text-to-speech (<https://www.ibm.com/cloud/watson-text-to-speech>, accessed on 24 May 2022) (TTS) and speech-to-text (<https://www.ibm.com/no-en/cloud/watson-speech-to-text>, accessed on 24 May 2022) (STT) synthesis because of the range of synthetic voices and in-built options for pace and pitch adjustment. Watson TTS and STT are cloud service APIs that serve as communicative bridges between 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. The dialogue model processes the user utterance at the back end and generates the appropriate response. This response is then sent to the IBM TTS API, sending the audio response to the user at the front end.

3.3. Emotions

Unfortunately, it is difficult to obtain audiovisual field interviews due to privacy concerns. Therefore, the prediction of the emotions is purely based on written text rather than a combination of both textual, audio, and visual input from transcripts of mock interviews [15]. Due to the lack of annotated data, we started our experiments with a pre-trained transformer model for sequence classification. From the HuggingFace library (<https://huggingface.co/>, accessed on 24 May 2022) we used the zero-shot classification pipeline in combination with the BART large model. The BART large model [89] is an autoencoder for pretraining sequence-to-sequence models. This model is made specifically for GLUE tasks [90], with sentiment classification as a sub-task. We present the model with our set of labels. The model then returns the class labels with their corresponding probabilities. We also experimented with the use of GPT-3 [91] and with different approaches to predict and choose the correct sentiment from the seven options. The emotions were predicted using only one sentence, the whole story up until the current sentence, and using a sliding window with and without a threshold to restart the window. The window size experiments varied with window sizes of 3, 5, 7, 10, and 15. If a threshold is set and the single sentence prediction has a higher probability than a certain threshold, it restarts the window. The choice for including such a threshold was made to be able to recognize the potential for sudden significant changes in emotions. If a child, for example, starts telling about how they started crying, the sentiment turns very sad very quickly, and the context is thus of less influence.

We commenced using the before-mentioned seven universal emotions. However, we also performed the same experiments using only the four basic emotions of joy, sadness, anger, and fear [92] instead of the seven. The reason to change the subclass of emotions from four to seven was due to the low Intra-class Correlation Coefficient score on the seven dimension experiments, as it only scored 0.537 on average measures.

The importance of the emotion classification part becomes cognizable when it is integrated into the complete system. It plays an integral part in the visual and audio output by altering the output based on the specified emotion. We created an emotion pipeline that predicts both the emotional valence of the interviewer's input and the chatbot's output. Due to the closed-off RASA environment, it is not possible to directly classify the bot's

responses in the environment. However, it is possible to classify the input in the RASA environment. Using the requests package (<https://github.com/psf/requests>, accessed on 24 May 2022) in Python connected to the localhost address where the RASA bot is running, we are able to receive and classify both the human input as well as the RASA output. This is necessary because classifying the RASA response is of great importance for our system. The emotion pipeline receives these texts as input and then provides the input for the audio and visual part.

3.4. Visual

Our early efforts are dedicated to research and proof-of-concept development. To discover the right approach to generate a visual child avatar, we went through a trial-and-error procedure that involved three prototypes: **Faceswap**, **Unity Multipurpose Avatar (UMA)** and **Talking-head Generation**. The following is a description of each.

In the first prototype, we narrowed the scope to a system that can lip-synch a specified audio stream by manipulating a video or image of a person. Similarly, we chose Faceswap [93], which allows us to swap two people's faces, an open source deepfake software, that is based on two autoencoders with a shared encoder in which the encoder learns the common features of the source and the target faces, while the two decoders learn to generate the source and target faces. The method requires a full video and many images of the face to be transferred and each new face requires a separate neural network. In the second approach, the child avatars were developed using the open source project Unity Multipurpose Avatar (UMA) (<https://github.com/umasteeringroup/UMA>, accessed on 24 May 2022) system. We customized the characters by merging meshes and textures, and the audios were synchronized with the avatar. Using the Unity game engine asset SALSA LipSync Suite (<https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442>, accessed on 24 May 2022), we generated eye, head, and mouth movements in sync with audio.

After reading a number of publications [94–96], we turned to audio-to-video translation methods using GANs for realistic avatar synthesis which were popularized when used to generate a fake video of Obama (<https://www.youtube.com/watch?v=cQ54GDm1eL0>, accessed on 24 May 2022) [63,97]. Following an examination of the two approaches of ObamaNet [97] and the method proposed by Suwajanakorn et al. [63] in our previous work [98], we experimented with the ICface [74], a lightweight model for face animators that is driven by human interpretable control signals. This method belongs to video-driven talking-head generation (a.k.a., face reenactment) because it uses another video as input to regulate expression, pose, mouth, eye, and eyebrows movements.

We next used two contemporary state-of-the-art procedures to obtain more results after collecting the necessary understanding regarding talking-head methods, including: PC-AVS [75] and MakeItTalk [65]. The same general approach was adopted for all methods: an audio stream generated by IBM Watson TTS was given to the network as input to synthesize lip-synced videos.

MakeItTalk was trained on the VoxCeleb2 [99] dataset, which contains video segments from a variety of speakers and can synthesis expressive talking-head videos with an audio stream and a portrait face image as the only inputs. It also generalizes well to unseen facial images. MakeItTalk presents a self-attention based LSTM able to disentangle content and style in audio and lead to the generation of a speaker-aware talking-head. PC-AVS takes a single facial image as an input and generates a talking-head whose poses are controlled by the pose on another source video. The method implicitly devises a posture code that is free of mouth shape or identification, then modularizes audio-visual representations into spaces for audio content, head movement, and identity without relying on any intermediate information such as landmarks and 3D face shapes.

We conducted a user study to assess the two promising Unity game engine and talking-head techniques. The findings are briefly addressed in the next section.

4. Results

Eventually, users learning to conduct expert investigative interviews will use the talking child avatars directly. Therefore, subjective evaluation of humans is an essential step. In this section, the extracted results are described for each language, auditory, emotional, and visual components.

4.1. Language

The first interactive study involved expert users interacting with a six-years-old sexually abused child to validate our first prototype as a proof of concept for the proposed training system for investigative interviewer training. The core of the system used RASA as the dialogue model. Results of this user study showed that it was well received by the CPS workers. 72% opined that it could help them obtain knowledge and skills for communicating with abused children and 81% stated that it could improve their self-efficacy. The current chatbot was developed specifically to practice an investigative interviewing methodology following best-practice guidelines and inspired by the research on interview methodology conducted by researchers at the US National Institute of Child Health and Human Development (NICHD) [11]. Although the dialogue model can answer questions about the child's life that might not be relevant to the alleged incident it is still not capable of having a generic conversation with the child. It is hard to model small talk with RASA, as it leads to many intent definitions that negatively affect the performance of the intent classification model. We believe we can deal with the shortcomings of RASA chatbot using GPT-3 to keep the story coherent while also being able to answer generic small talk. RASA provided more control over how we wanted the conversation to proceed, but it is not a scalable solution.

4.2. Auditory

The user study had synthetic voices generated by IBM Watson TTS and recorded voices of natural speech, each of them for two different genders, as test conditions. This section investigated whether computer-generated voices can be as realistic as human voices and the appropriateness of the voices associated with each character. Figure 3 shows the results of the user study that compared the voices generated by a computer and by a human. A factorial ANOVA compared animated and GAN-generated avatars, where each had both synthetic and natural voices. The factorial ANOVA shows no significant main effect of type of voices ($F(1, 35) = 1.39, p = 0.24$), meaning that the computer generated voices were rated equivalent to human voices and there was no significant difference between them.

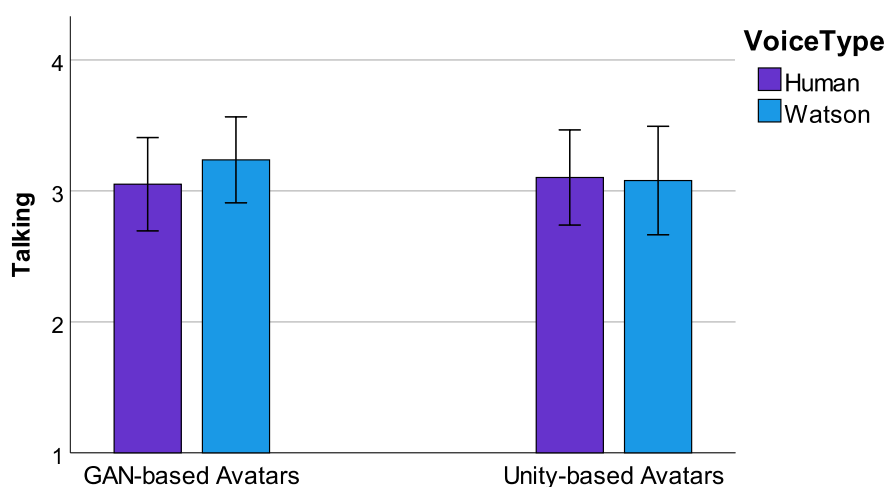


Figure 3. A comparison between natural and synthetic voices in animated unity-based and GAN-based avatars.

4.3. Emotions

As explained in Section 3, we started with both GPT-3 and BART predicting one sentence at a time. Comparing this with the results of a user study yielded promising results. However, these results do not always make sense in a larger context. Take the sentence “I was on the playground” as an example. This sentence is innocent and suggests enjoyment when viewed on its own, but it may no longer have that positive connotation in the context of abuse. The single-sentence prediction misses the importance of context. The opposite occurs when predicting the emotion based on the whole story. At some point, the model is not able to pick up subtle emotional changes since there is too much context. Thus, the single-sentence prediction is less complicated but not accurate since it misses important context. By contrast, the whole story has too much context, resulting in results that are not accurate either. The sliding window appeared to offer the perfect solution to this problem. Using a window consisting of between 3 and 7 sentences was expected to be ideal. However, the threshold did not work as expected. The hypothesis was that it would help spot sudden substantial emotional changes in the story, for example, if a child started crying. However, a restart of the window often occurred with sentences classified as enjoyment. Consequently, the model predictions got worse due to the implementation of the threshold.

It is challenging for humans to reach a consensus when classifying data excerpts with seven emotions as the options. There was unanimous agreement about only one of the twenty-one questions, that one being a single sentence excerpt. Although there was a clear winner for eleven questions, the Intraclass Correlation Coefficient (ICC) based on average measures was low on seven dimensions, scoring only 0.537 on average. The ICC describes how strongly elements in the same group resemble each other. An ICC between 0.40 and 0.59 is a moderate score. However, it is preferable to have an ICC score of at least 0.75 [100]. When we reduced the number of emotions to four, the participants were better at agreeing on the best fitting class. Reducing the number of categories resulted in more clear winners than before. The ICC also increased from 0.537 to 0.788: it was significantly higher with four dimensions than with seven.

A comparison between GPT-3 and BART was based on human annotations from a survey of 21 participants. 52.4% of the participants identified as female and 47.6% as male. Most of the participants were between the age of 26 and 35, but people between 18 and 25, 36 and 45, and 56 and 65 years also participated. The survey included eight single-sentence excerpts and 12 excerpts for each window size. The window sizes that we used were 3, 5, and 7.

The results showed that BART could predict more single sentences correctly, whereas GPT-3 was better at predicting sentences in context. One of the sentences in the user study was “We watched a movie and then we got some ice cream and then we went to bed.”. The human consensus, the GPT-3 model, and BART, classified this as enjoyment. It becomes more difficult to classify sentences such as “It really hurt”, which BART and the human participants classified as sadness, while GPT-3 considered it to be anger.

Sometimes both the models and the user study results were in agreement. This happened for the excerpt shown in Figure 4. It was unanimously considered to belong to the class fear.

However, there are also instances where both models made different classifications than human participants. This happened with the conversation shown in Figure 5. Both models only saw the responses that the child gave us and thus, classified this excerpt as enjoyment, whereas the participants saw the whole conversation and classified it as fear instead.

There is not always a reason why one or both models are wrong. Sometimes, it is a matter of interpretation. As mentioned above, GPT-3 is better when the context is involved, as was also the case in Figure 6, where both models predicted a different emotion based on the text given to them.

Interviewer	Child
Hm-hm.	And I really needed to go to the toilet.
Okay, and then what happened next?	I went to the toilet by myself because mum couldn't leave my baby brother by himself.
Ah-ha.	And then I went to the toilet by myself.
Okay, and tell me everything that happened from when you went to the toilet.	I went into the toilet and I didn't lock the door because I didn't want to get stuck in.
Hm-hm.	And then the bad man came inside.

Figure 4. Excerpt from the user study with window size 5 where both models are in agreement with the human opinion.

Interviewer	Child
Okay, so no one else was home. Where was everyone else?	Work.
At work, okay, and why was Mark there?	To play.
To play, okay. Okay, so you've told me quite a bit. Let me just doublecheck that I've got all of that right. So Mark was with you and you were playing Nintendo and you stood up and broke the controller. Then he got angry at you for that so you went to the garden and got a stick and then he hit you with it lots of times and then after that, he just left and he told you not to tell anyone.	Yes.

Figure 5. Excerpt from the user study with window size 3 for which both models were not in agreement with the human raters.

Interviewer	Child
Tell me more about the part where he put his hand in your undies.	He wrapped his hand around my doodle and then I said I didn't like it.
Hm-hm.	And then he laughed and took it out.
Okay, what happens next?	And then I went home.
Hm-hm, and was anyone else with you at uncle George?	No.
Hm-hm, and can you tell me more about the big boys game? Have you played it more than one time?	Yes.

Figure 6. Excerpt from the user study with window size 5 for which both GPT-3 and the human raters agreed that this should be classified as fear, while the BART model classified its as anger.

4.4. Visual

Here, we report on the findings and lessons learned using the tools introduced in Section 3.4. First, we investigated the deepfake models and why they are unsuitable for creating realistic avatars. Then, using a subjective study, we evaluated the realism of the talking-head techniques and avatars created by the game engine as the solution to be integrated into the child avatar system.

Faceswap [93] is the leading free and open source multi-platform Deepfakes software that uses computer graphics and visualization techniques. Faceswap is used to swap the faces of two people. According to the results of published research [98], when a man's visage is used on a child, Faceswap does not produce a realistic outcome. Although two people who are very similar in appearance could create a better result, we cannot rely on close likeness for our use case. As a result, Faceswap appears not to be suitable for creating a realistic-looking avatar.

ICface [74] can transfer the expressions from a driving video to a source face image (as shown in Figure 7) video-driven talking-head (refer to Section 2.5). Although ICface can be applied to arbitrary face images, the results are not as visually plausible. Moreover, manipulating expressions in images using human interpretable control signals such as head movement angles and Action Units (AUs) is tedious, time consuming, and causes repetitive facial expressions. Therefore, this method is not efficient enough to employ to generate a realistic avatar.

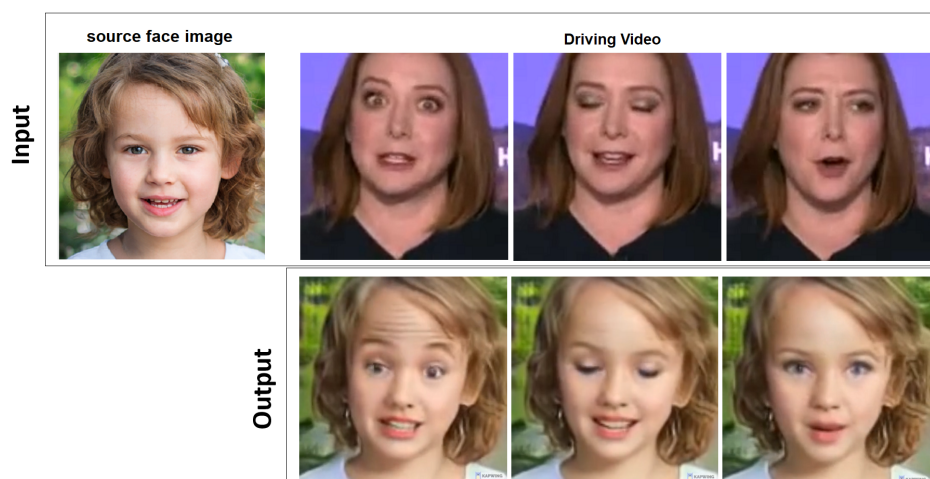


Figure 7. Given an arbitrary source face image generated by styleGAN [59,101] and a driving video, ICface [74] has generated the talking-head of a child.

We investigated the efficacy of the two approaches to talking-head generation using GANs and Unity game engine-based generation by conducting a crowdsourced user study. In both cases the audio stream used as input was obtained from IBM Watson (Introduced in Section 3.2) and recorded natural human voices. We generated 18 avatars using the two open source talking-head methods MakeItTalk [65] and PC-AVS [75], which can generalize well on any desired facial images. Here, StyleGAN [59,101] was utilized to create some child portrait images. Input facial images and several frames of video images thus generated are shown Figure 8. We also developed ten animated avatars using the Unity game engine. Finally, 28 ten-second video clips of the created avatars were provided using methods based on either GANs or the Unity game engine.

The study was conducted through crowdsourcing and Microworker (<https://www.microworkers.com/>, accessed on 24 May 2022) was used as a recruiting environment, with users referred to a questionnaire tool hosted on a separate server that contained the videos. To ensure the validity and reliability of the collected data, only high-performing crowdworkers who performed the best on the test were invited to participate.

Overall, 39 crowdworkers provided valid results in this study, including 10 women, 27 men, and two of other genders. The crowd workers aged 19 to 54 years (Median = 28, Mean = 29.58, SD = 8.36) were geographically evenly distributed between Europe, Asia and America (North and South).

Participants were asked to evaluate each video with regard to three statements: 'How realistic was the talking avatar?', 'How realistic was the avatar's appearance?', 'How was your overall experience with the avatar?' and 'How were the audio and mouth/lips synced?' on a scale between 1 to 5 (5-strongly agree, 4-agree, 3-neither agree nor disagree, 2-disagree, 1-strongly disagree). More details about the dataset and the study design are given in [18].

We first examine the difference between the MakeItTalk [65] and PC-AVS [75] as audio-driven models from different perspectives. From our point of view, PC-AVS can generate correct lip-synching, but it can not maintain the identity and resolution of the input facial image, and it also generates talking-heads with no blink motions. MakeItTalk, on the other

hand, can perform blink actions and maintain the resolution of the input facial image by depending on exact facial landmarks. Still, it does not accurately synchronize the shape of the mouth. Additionally, natural head movement is one of the most important aspects of effective communication in the real world. MakeItTalk can generate subtle head motions, unlike most existing approaches that generate face animation with a fixed head pose. These motions involve many repeating behaviors and movement patterns that only minimally swing around. In contrast, PC-AVS takes a brief target video and an audio stream as input to transfer the head position, resulting in more varied rhythmic head motions based on the input video clip. However, because this strategy depends on another short video clip, it is not particularly useful in our situation.

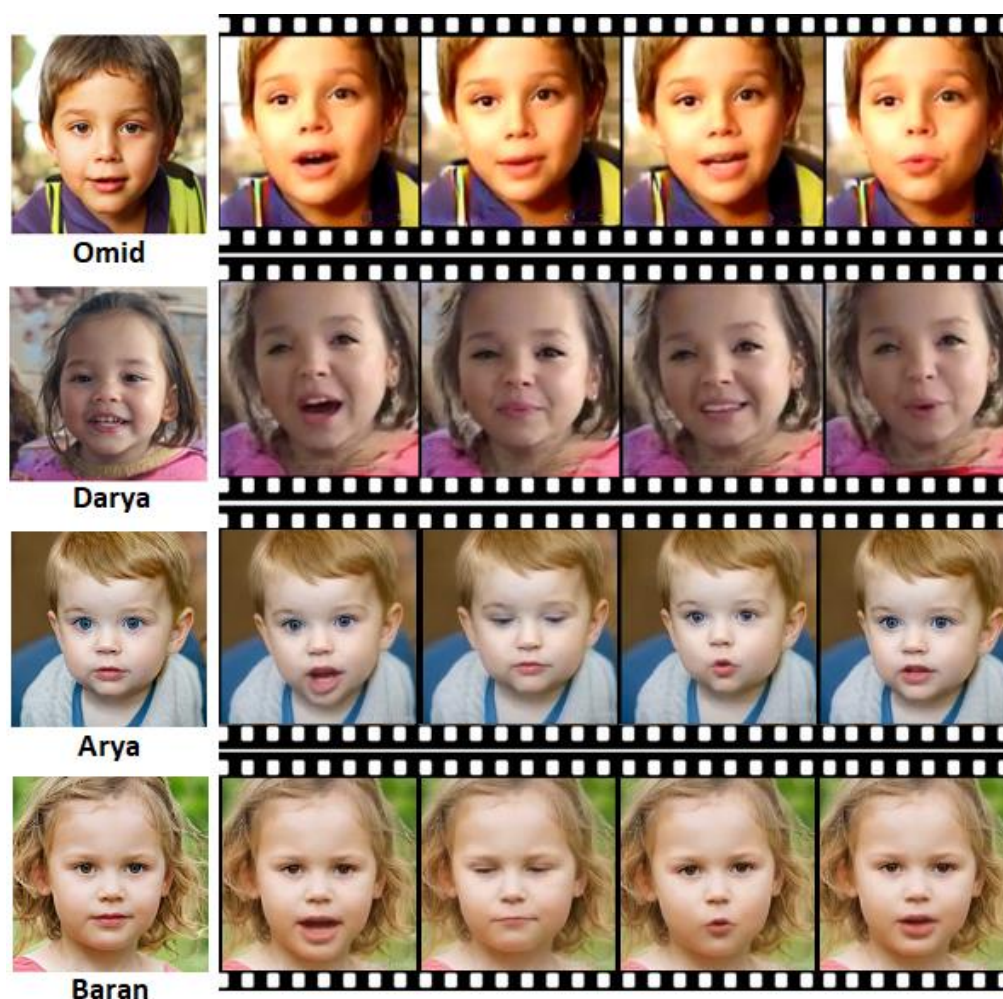


Figure 8. Illustration of a talking-head video generated using two methods, PCAVS [75] and MakeItTalk [65]. The input is an image generated using styleGAN and an audio generated using IBM Watson. The first two rows: PCAVS and the second two rows: MakeItTalk.

The bar-plots in Figure 9 depict the user ratings showing no significant difference between the MakeItTalk and PC-AVS models for four different characters. We expected MakeItTalk to appear more realistic from the user's perspective, but it did not. Four repeated measure factorial ANOVAs were used to compare the main effects of models and characters and their interactions on all four quality dimensions. According to all four quality criteria talking, appearance, experience and lips sync, there was no significant difference between the two models (MakeItTalk and PC-AVS), and the characters (the selected facial image). For the realism of *Talking*, there were no significant main effects of the model (p -value (p) more than <0.05) ($p < 0.6$), or characters ($p < 0.5$) and no significant interaction between model and characters ($p < 0.06$). For *Appearance*, no main effects were

observed for the model ($p < 0.09$), or characters ($p < 0.21$) and there was no significant interaction between model and characters ($p < 0.09$). For *Overall experience*, there was no main effect of model ($p < 0.86$), no main effect of characters ($p < 0.29$) and no interaction between model and characters ($p < 0.38$). And for lip-sync, there was no main effect of model ($p < 0.52$), no main effect of characters ($p < 0.17$) and no interaction between model and characters ($p < 0.07$).

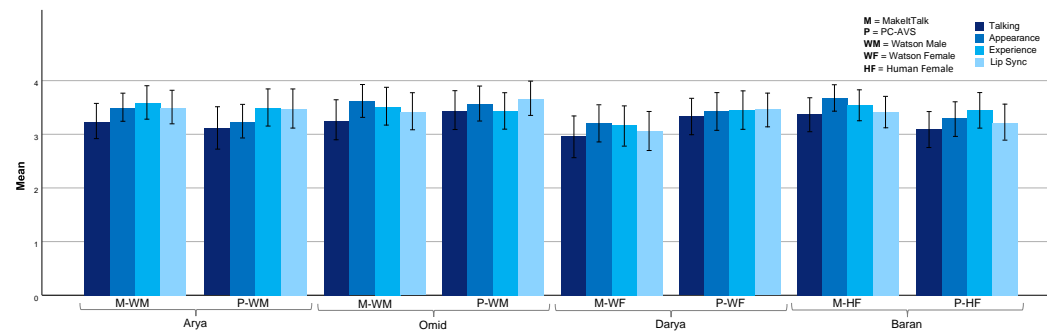


Figure 9. Bar-plot (95% confidence interval) for comparison of MakeItTalk [65] and PC-AVS [75].

Interaction with the virtual avatar necessitates techniques to generate high-fidelity talking-head frames while maintaining human observers' trust and empathy. According to Uncanny Valley theory [102,103], if an avatar is humanoid but imperfect, its non-human features instill confusion and even fear in the audience. Thus, we developed animated facial avatars using the Unity game engine to see whether the uncanny valley exists in the context of a child interview.

To investigate the avatars created by the game engine, a one-way repeated measure analysis of variance (ANOVA) for the mean of overall experience and appearance was performed. The lip-sync quality was not explored in the data analyses because all of the animated characters used the same technique for lip-sync generation utilizing the Salsa component in the Unity game engine. The animated avatars' quality ratings are shown in Figure 10.

Next, the best-rated avatars created with GAN were compared with the animated avatars created by the Unity game engine. The best avatars in GAN were Omid and Darya from model PC-AVS, which were selected based on how realistic their *Appearance* and *Talking* were rated. These two avatars are compared with Sarah and Liam's animated avatars created by the Unity game. Figure 10 shows the quality ratings for these four avatars. Sarah's character was rated slightly lower than the others, which might be due to weaker lip synchronization, but there was no statistically significant difference. A factorial ANOVA showed that although there was a general trend for more realism to be perceived in the videos created by GANs, there was no significant main effect of realism for *Talking* $F(1,37) = 2.54, p = 0.12$, *Appearance* $F(1,37) = 2.66, p = 0.11$ and *Overall experience* $F(1,37) = 1.09, p = 0.30$. Overall, the results showed that although the avatars generated by GANs were slightly more realistic, they did not necessarily create a higher quality of experience and realism for the user.

In addition to the quality aspects, we asked participants an open question about their reasons for preferring one character over the other at the end of the questionnaire. Almost all of them agreed that the avatars generated by GANs were more realistic. Four participants stated that lip-synching in GAN-generated avatars was better than in animated avatars. One participant even stated, "The GAN-generated avatars could trick me into thinking that an actual person is talking." Three participants unequivocally stated that the talking faces appeared frightening, and one participant mentioned that something was off-putting and felt incorrect. This is because GAN-generated avatars are innately unlikable unless they are perfect and natural, as explainable by the uncanny valley theory.

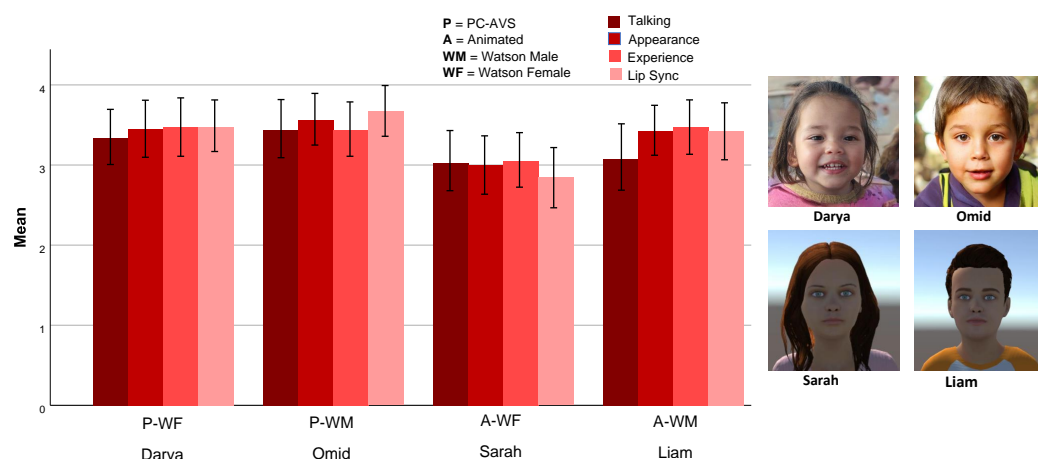


Figure 10. Bar-plot (95% confidence interval) to show results of the user study for the evaluation of two of the best female and male characters created for both the GAN-Based and game engine-based approaches.

5. Discussions and Future Work

Enhancing the disclosure of child abuse can facilitate earlier protection, prevention, and prosecution when children are victims of sexual, physical, and emotional abuse, thereby possibly having a huge impact on individuals’ mental health and the future of society in general. The paper describes how an avatar-assisted mode of training might help CPS workers and police officers learn how to effectively interview sexually and physically abused children. Thanks to recent developments in AI, we proposed synthesizing a realistic digital child avatar, attempting to emulate an abused child. Our ongoing work focuses on integrating various components, including the language, auditory, emotional, and visual components of the avatar. The proposed system can dynamically respond to the questions and provide a higher level of realism during the training interviews and would be completely independent of human input. Therefore, unlike previous systems that were too rigid in their response generation, lacked generality, and required human input, our proposed system can generate dynamic responses without human operation. Such a system would be more dynamic and cost-effective because it lowers expenditures on human resources. This study discussed lessons and outcomes achieved by adopting various techniques. Moreover, three user studies were conducted to comprehensively evaluate a number of these methods.

The first user study showed that, although RASA provides an environment to manage the flow of dialogue within brief mock interviews, it is not easy to model small talk with RASA because it leads to many intent definitions, negatively affecting classification of the model’s intent. We believe that by using GPT-3, we can address the shortcomings of the RASA chatbot, which will maintain the coherence of the story while also answering generic small talk. RASA gave us greater control over how the conversation flows, but it is not a robust solution as it has not been trained on open-domain data to sensibly answer the questions that are not in transcripts used during training.

There were also promising initial results concerning the emotional component. However, it is essential that more annotated data get analyzed and compared to the output of the existing models. We can incorporate the right model into our emotion pipeline. After that, we can start implementing the pipeline in both the auditory and visual components, following which users can evaluate what does and does not work. However, at this stage the biggest improvements would be made if there were annotated text data available in combination with the corresponding videos. It is also worth researching whether better results would be obtained if the interviewer’s questions were also classified. Figure 5 shows this could be a positive influence. However, it may negatively affect other currently correctly classified examples, so it is worth investigating.

Another user study was conducted to explore the realism of the avatar's auditory and visual components. For the audio component, the avatars were tested using synthetic voices generated by computer and natural human voices. The results showed that synthetic voices generated by IBM Watson TTS can be as realistic as natural voices. Based on those results, we decided to use the synthetic voices for the avatar.

For the visual component, facial manipulations using the FaceSwap and video-driven talking-head methods were ineffective. We therefore turned to the Audio-driven talking-head using GANs and Unity game engine and then conducted a user study to compare them. Contrary to our predictions, GAN-generated and Unity game engine avatars were rated similarly. The best generated avatars using the two approaches were not rated differently. The results are even more startling because more than half of the participants claimed to prefer conversing with avatars developed using the Unity game engine, even though the avatar's appearance and overall talking were less realistic. A few participants noted that the GAN-generated avatars appeared to be frightening in response to an open question. This could be evidence of the presence of an uncanny valley [102,103]: if an avatar is humanoid but flawed, the viewer may experience weird feelings of uneasiness and even revulsion as a result of the avatar's non-human traits. We aim to improve the realism of GAN-generated avatars in the future so that fewer artifacts are produced. Furthermore, we will experiment with combining the two methodologies, with the talking-head formed using an animated facial image as input. Further, future works will explore the human influencing factors and their impact on the user experience. The human influencing factors include demographic characteristics such as age, gender and socio-economic background, physical and mental constitution, and the user's emotional state [104], all of which can have an impact on the user experience.

In addition, we intend to consider the versatile and novel LSP model [67] in future work. LSP introduces a real-time system that generates talking heads with natural head movement, lip-sync, and eye blinks using only audio signals at more than 30 frames per second. Furthermore, the system requires training on a several-minute video of the desired character. We can provide this video from a real human generated using GANs or an existing talking animation. Thus, LSP appears practical and ideal for developing a digital child avatar due to its high-fidelity video creation and real-time capabilities.

6. Conclusions

In forensic interviews, obtaining valuable information from a neglected or abused child requires the interviewer to have good interviewing skills. Our ongoing work aims to synthesize an interactive virtual child avatar in real-time, mimicking a child so that law enforcement and CPS workers can efficiently be trained to learn and improve these skills. This paper presented the interactive child avatar system, mimicking an abused child. The system is designed using different artificial intelligence-based technologies such as the avatar's language, auditory, emotional, and visual components. Furthermore, using three subjective studies, various system components were investigated. The results of the first user study showed that participants believed the child avatar system could effectively improve conversational skills and was well received by the CPS workers. Furthermore, the second study examined emotion extraction using different models and discussed how emotion could be integrated with auditory and visual components. The third user study first showed that the synthetic voices generated by computers could be as realistic as natural voices, and then investigated the realism of various techniques for generating child avatars. It was shown that the GAN-based and game engine-based avatars could create the most realistic avatars. Using the insights provided by these three user studies, the refined and improved architecture of the child avatar system was presented, and the integration and interaction of various components were discussed.

Author Contributions: Investigation, P.S., S.Z.H., M.L., S.S.S. and I.R.; Methodology, P.S., S.Z.H., M.L. and S.S.S.; Project administration, S.S.S., G.A.B., P.H. and M.A.R.; Writing—original draft, P.S., S.Z.H. and M.L.; Writing—review & editing, P.S., S.Z.H., M.L., S.S.S., I.R., R.K.R., M.S.J., V.T., S.A.H.,

M.P., M.E.L., G.A.B., P.H. and M.A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research is sponsored by the Research Council of Norway, project number #314690 (“Interview training of child-welfare and law-enforcement professionals interviewing maltreated children supported via artificial avatars”).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Sethi, D.; Bellis, M.; Hughes, K.; Gilbert, R.; Mitis, F.; Galea, G. *European Report on Preventing Child Maltreatment*; World Health Organization, Regional Office for Europe: Geneva, Switzerland, 2013.
- Widom, C.S. Longterm consequences of child maltreatment. In *Handbook of Child Maltreatment*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 225–247.
- World Health Organization. *Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks*; World Health Organization: Geneva, Switzerland, 2009.
- Dixon, L.; Perkins, D.F.; Hamilton-Giachritsis, C.; Craig, L.A. *The Wiley Handbook of What Works in Child Maltreatment: An Evidence-Based Approach to Assessment and Intervention in Child Protection*; John Wiley & Sons: Hoboken, NJ, USA, 2017.
- Brown, D.; Lamb, M. Forks in the road, routes chosen, and journeys that beckon: A selective review of scholarship on childrens testimony. *Appl. Cogn. Psychol.* **2019**, *33*, 480–488. [[CrossRef](#)]
- Lamb, M.E.; La Rooy, D.J.; Malloy, L.C.; Katz, C. *Children’s Testimony: A Handbook of Psychological Research and Forensic Practice*; John Wiley & Sons: Hoboken, NJ, USA, 2011; Volume 53.
- Adams, J.A.; Farst, K.; Kellogg, N.D. Interpretation of Medical Findings in Suspected Child Sexual Abuse: An Update for 2018. *J. Pediatr. Adolesc. Gynecol.* **2018**, *31*, 225–231. [[CrossRef](#)] [[PubMed](#)]
- Newlin, C.; Steele, L.C.; Chamberlin, A.; Anderson, J.; Kenniston, J.; Russell, A.; Stewart, H.; Vaughan-Eden, V. *Child Forensic Interviewing: Best Practices*; US Department of Justice, Office of Justice Programs, Office of Juvenile: Washington, DC, USA, 2015.
- Lamb, M.E.; Brown, D.A.; Hershkowitz, I.; Orbach, Y.; Esplin, P.W. *Tell Me What Happened: Questioning Children about Abuse*; John Wiley & Sons: Hoboken, NJ, USA, 2018.
- Lamb, M.E.; Orbach, Y.; Hershkowitz, I.; Esplin, P.W.; Horowitz, D. 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 Abus. Negl.* **2007**, *31*, 1201–1231. [[CrossRef](#)] [[PubMed](#)]
- Powell, M.B.; Brubacher, S.P. The origin, experimental basis, and application of the standard interview method: An information-gathering framework. *Aust. Psychol.* **2020**, *55*, 645–659. [[CrossRef](#)]
- Lyon, T.D. Interviewing children. *Annu. Rev. Law Soc. Sci.* **2014**, *10*, 73–89. [[CrossRef](#)]
- Powell, M.B.; Hughes-Scholes, C.H.; Smith, R.; Sharman, S.J. The relationship between investigative interviewing experience and open-ended question usage. *Police Pract. Res.* **2014**, *15*, 283–292. [[CrossRef](#)]
- Lamb, M. Difficulties translating research on forensic interview practices to practitioners: Finding water, leading horses, but can we get them to drink? *Am. Psychol.* **2016**, *71*, 710–718. [[CrossRef](#)] [[PubMed](#)]
- Powell, M.B.; Guadagno, B.; Benson, M. Improving child investigative interviewer performance through computer-based learning activities. *Polic. Soc.* **2016**, *26*, 365–374. [[CrossRef](#)]
- Seymour, M.; Riemer, K.; Kay, J. Actors, avatars and agents: Potentials and implications of natural face technology for the creation of realistic visual presence. *J. Assoc. Inf. Syst.* **2018**, *19*, 4. [[CrossRef](#)]
- Hassan, S.Z.; Salehi, P.; Røed, R.K.; Halvorsen, P.; Baugerud, G.A.; Johnson, M.S.; Lison, P.; Riegler, M.; Lamb, M.E.; Griwodz, C.; et al. Towards an AI-Driven Talking Avatar in Virtual Reality for Investigative Interviews of Children. In Proceedings of the 2nd Edition of the Game Systems Workshop (GameSys ’22), Athlone, Ireland, 14–17 June 2022.
- Salehi, P.; Hassan, S.Z.; Sabet, S.S.; Baugerud, G.A.; Johnson, M.S.; Riegler, M.; Halvorsen, P. Is More Realistic Better? A Comparison of Game Engine and GAN-based Avatars for Investigative Interviews of Children. In Proceedings of the ICDAR Workshop, ACM ICMR 2022, Newark, NJ, USA, 27–30 June 2022.
- Cederborg, A.C.; Orbach, Y.; Sternberg, K.J.; Lamb, M.E. Investigative interviews of child witnesses in Sweden. *Child Abus. Negl.* **2000**, *24*, 1355–1361. [[CrossRef](#)]
- Baugerud, G.A.; Johnson, M.S.; Hansen, H.B.; Magnussen, S.; Lamb, M.E. Forensic interviews with preschool children: An analysis of extended interviews in Norway (2015–2017). *Appl. Cogn. Psychol.* **2020**, *34*, 654–663. [[CrossRef](#)]
- Korkman, J.; Santtila, P.; Sandnabba, N.K. Dynamics of verbal interaction between interviewer and child in interviews with alleged victims of child sexual abuse. *Scand. J. Psychol.* **2006**, *47*, 109–119. [[CrossRef](#)] [[PubMed](#)]

22. Lamb, M.E.; Orbach, Y.; Sternberg, K.J.; Aldridge, J.; Pearson, S.; Stewart, H.L.; Esplin, P.W.; Bowler, L. Use of a structured investigative protocol enhances the quality of investigative interviews with alleged victims of child sexual abuse in Britain. *Appl. Cogn. Psychol. Off. J. Soc. Appl. Res. Mem. Cogn.* **2009**, *23*, 449–467. [[CrossRef](#)]
23. Brubacher, S.P.; Shulman, E.P.; Bearman, M.J.; Powell, M.B. Teaching child investigative interviewing skills: Long-term retention requires cumulative training. *Psychol. Public Policy Law* **2021**, *28*, 123–136. [[CrossRef](#)]
24. Krause, N.; Pompèdda, F.; Antfolk, J.; Zappala, A.; Santtila, P. The Effects of Feedback and Reflection on the Questioning Style of Untrained Interviewers in Simulated Child Sexual Abuse Interviews. *Appl. Cogn. Psychol.* **2017**, *31*, 187–198. [[CrossRef](#)]
25. Haginoya, S.; Yamamoto, S.; Pompèdda, F.; Naka, M.; Antfolk, J.; Santtila, P. Online simulation training of child sexual abuse interviews with feedback improves interview quality in Japanese university students. *Front. Psychol.* **2020**, *11*, 998. [[CrossRef](#)]
26. Haginoya, S.; Yamamoto, S.; Santtila, P. The combination of feedback and modeling in online simulation training of child sexual abuse interviews improves interview quality in clinical psychologists. *Child Abus. Negl.* **2021**, *115*, 105013. [[CrossRef](#)]
27. Pompèdda, F.; Zappalà, A.; Santtila, P. Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality. *Psychol. Crime Law* **2015**, *21*, 28–52. [[CrossRef](#)]
28. Mayer, J. D.; Salovey, P. What is emotional intelligence? In *Emotional Development and Emotional Intelligence: Educational Implications*; Basic Books: New York, NY, USA, 1997; pp. 3–33.
29. Joseph, D.L.; Newman, D.A. Emotional intelligence: An integrative meta-analysis and cascading model. *J. Appl. Psychol.* **2010**, *95*, 54–78. [[CrossRef](#)]
30. Hochschild, A.R. *The Managed Heart: Commercialization of Human Feeling*; University of California Press: Berkeley, CA, USA; London, UK, 2012.
31. Risan, P.; Binder, P.E.; Milne, R.J. Emotional Intelligence in Police Interviews—Approach, Training and the Usefulness of the Concept. *J. Forensic Psychol. Pract.* **2016**, *16*, 410–424. [[CrossRef](#)]
32. Albaek, A.U.; Kinn, L.G.; Milde, A.M. Walking Children Through a Minefield: How Professionals Experience Exploring Adverse Childhood Experiences. *Qual. Health Res.* **2018**, *28*, 231–244. [[CrossRef](#)] [[PubMed](#)]
33. Ekman, P.; Friesen, W.V. A new pan-cultural facial expression of emotion. *Motiv. Emot.* **1986**, *10*, 159–168. [[CrossRef](#)]
34. Ekman, P.; Heider, K.G. The universality of a contempt expression: A replication. *Motiv. Emot.* **1988**, *12*, 303–308. [[CrossRef](#)]
35. Matsumoto, D. More evidence for the universality of a contempt expression. *Motiv. Emot.* **1992**, *16*, 363–368. [[CrossRef](#)]
36. Katz, L.F.; Hunter, E.C. Maternal meta-emotion philosophy and adolescent depressive symptomatology. *Soc. Dev.* **2007**, *16*, 343–360. [[CrossRef](#)]
37. Karni-Visel, Y.; Hershkowitz, I.; Lamb, M.E.; Blasbalg, U. Nonverbal Emotions While Disclosing Child Abuse: The Role of Interviewer Support. *Child Maltreatment* **2021**, *29*, 10775595211063497. [[CrossRef](#)]
38. Kerig, P.K.; Bennett, D.C.; Chaplo, S.D.; Modrowski, C.A.; McGee, A.B. Numbing of Positive, Negative, and General Emotions: Associations with Trauma Exposure, Posttraumatic Stress, and Depressive Symptoms Among Justice-Involved Youth: Numbing of Positive, Negative, or General Emotions. *J. Trauma. Stress* **2016**, *29*, 111–119. [[CrossRef](#)]
39. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *CIM* **2018**, *13*, 55–75.
40. Vinyals, O.; Le, Q. A neural conversational model. *arXiv* **2015**, arXiv:1506.05869.
41. Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; Liu, B. Emotional chatting machine: Emotional conversation generation with internal and external memory. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
42. Zhou, L.; Gao, J.; Li, D.; Shum, H.Y. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *arXiv* **2019**, arXiv:1812.08989.
43. Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.P.; Gao, J.; Dolan, B. A persona-based neural conversation model. *arXiv* **2016**, arXiv:1603.06155
44. Tachibana, H.; Uenoyama, K.; Aihara, S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 4784–4788.
45. Amberkar, A.; Awasarmol, P.; Deshmukh, G.; Dave, P. Speech recognition using recurrent neural networks. In Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, India, 1–3 March 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–4.
46. Xiong, W.; Droppo, J.; Huang, X.; Seide, F.; Seltzer, M.L.; Stolcke, A.; Yu, D.; Zweig, G. Toward Human Parity in Conversational Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2017**, *25*, 2410–2423. [[CrossRef](#)]
47. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A simple data augmentation method for automatic speech recognition. *arXiv* **2019**, arXiv:1904.08779.
48. Sadjadi, O.; Greenberg, C.; Singer, E.; Mason, L.; Reynolds, D. *NIST 2021 Speaker Recognition Evaluation Plan*; NIST: Gaithersburg, MD, USA, 2021.
49. Zhang, Y.; Qin, J.; Park, D.S.; Han, W.; Chiu, C.C.; Pang, R.; Le, Q.V.; Wu, Y. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv* **2020**, arXiv:2010.10504.
50. Chung, Y.A.; Zhang, Y.; Han, W.; Chiu, C.C.; Qin, J.; Pang, R.; Wu, Y. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *arXiv* **2021**, arXiv:2108.06209.

51. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower Provost, E.; Kim, S.; Chang, J.; Lee, S.; Narayanan, S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [[CrossRef](#)]
52. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323. [[CrossRef](#)]
53. Fayek, H.; Lech, M.; Cavedon, L. Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Netw.* **2017**, *92*, 60–68. [[CrossRef](#)]
54. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 4779–4783.
55. Prenger, R.; Valle, R.; Catanzaro, B. Waveglow: A flow-based generative network for speech synthesis. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–19 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 3617–3621.
56. Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; Oord, A.V.D.; Dieleman, S.; Kavukcuoglu, K. Efficient Neural Audio Synthesis. *arXiv* **2018**, arXiv:1802.08435.
57. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
58. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
59. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8110–8119.
60. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8789–8797.
61. Agarwal, S.; Farid, H.; Gu, Y.; He, M.; Nagano, K.; Li, H. Protecting World Leaders Against Deep Fakes. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–20 June 2019; Volume 1.
62. Perov, I.; Gao, D.; Chervoniy, N.; Liu, K.; Marangonda, S.; Umé, C.; Dpfks, M.; Facenheim, C.S.; RP, L.; Jiang, J.; et al. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv* **2020**, arXiv:2005.05535.
63. Suwajanakorn, S.; Seitz, S.M.; Kemelmacher-Shlizerman, I. Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph. ToG* **2017**, *36*, 1–13. [[CrossRef](#)]
64. Chen, L.; Maddox, R.K.; Duan, Z.; Xu, C. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7832–7841.
65. Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; Li, D. Makelttalk: Speaker-aware talking-head animation. *ACM Trans. Graph. TOG* **2020**, *39*, 1–15. [[CrossRef](#)]
66. Meshry, M.; Suri, S.; Davis, L.S.; Shrivastava, A. Learned Spatial Representations for Few-shot Talking-Head Synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13829–13838.
67. Lu, Y.; Chai, J.; Cao, X. Live speech portraits: Real-time photorealistic talking-head animation. *ACM Trans. Graph. TOG* **2021**, *40*, 1–17. [[CrossRef](#)]
68. Yi, R.; Ye, Z.; Zhang, J.; Bao, H.; Liu, Y.J. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv* **2020**, arXiv:2002.10137.
69. Thies, J.; Elgharib, M.; Tewari, A.; Theobalt, C.; Nießner, M. Neural voice puppetry: Audio-driven facial reenactment. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 716–731.
70. Chen, L.; Cui, G.; Liu, C.; Li, Z.; Kou, Z.; Xu, Y.; Xu, C. Talking-head generation with rhythmic head motion. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 35–51.
71. Richard, A.; Lea, C.; Ma, S.; Gall, J.; De la Torre, F.; Sheikh, Y. Audio-and gaze-driven facial animation of codec avatars. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 11 August 2021; pp. 41–50.
72. Song, L.; Wu, W.; Qian, C.; He, R.; Loy, C.C. Everybody’s talkin’: Let me talk as you want. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 585–598. [[CrossRef](#)]
73. Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2387–2395.
74. Tripathy, S.; Kannala, J.; Rahtu, E. Icfac: Interpretable and controllable face reenactment using gans. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2020; pp. 3385–3394.

75. Zhou, H.; Sun, Y.; Wu, W.; Loy, C.C.; Wang, X.; Liu, Z. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4176–4186.
76. Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; Wang, X. Talking face generation by adversarially disentangled audio-visual representation. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 9299–9306.
77. Wiles, O.; Koepke, A.; Zisserman, A. X2face: A network for controlling face generation using images, audio, and pose codes. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 670–686.
78. Ha, S.; Kersner, M.; Kim, B.; Seo, S.; Kim, D. Marionette: Few-shot face reenactment preserving identity of unseen targets. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10893–10900.
79. Bansal, A.; Ma, S.; Ramanan, D.; Sheikh, Y. Recycle-gan: Unsupervised video retargeting. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 119–135.
80. Kim, H.; Garrido, P.; Tewari, A.; Xu, W.; Thies, J.; Niessner, M.; Pérez, P.; Richardt, C.; Zollhöfer, M.; Theobalt, C. Deep video portraits. *ACM Trans. Graph. TOG* **2018**, *37*, 1–14. [[CrossRef](#)]
81. Pompèdda, F.; Antfolk, J.; Zappalà, A.; Santtila, P. A combination of outcome and process feedback enhances performance in simulations of child sexual abuse interviews using avatars. *Front. Psychol.* **2017**, *8*, 1474. [[CrossRef](#)]
82. Pompèdda, F.; Palu, A.; Kask, K.; Schiff, K.; Soveri, A.; Antfolk, J.; Santtila, P. Transfer of simulated interview training effects into interviews with children exposed to a mock event. *Nordic Psychol.* **2020**, *73*, 43–67. [[CrossRef](#)]
83. Pompèdda, F.; Zhang, Y.; Haginoya, S.; Santtila, P. A Mega-Analysis of the Effects of Feedback on the Quality of Simulated Child Sexual Abuse Interviews with Avatars. *J. Police Crim. Psychol.* **2022**, 1–14. [[CrossRef](#)]
84. Dalli, K.C. Technological Acceptance of an Avatar Based Interview Training Application: The Development and Technological Acceptance Study of the AvBIT Application, Master’s Thesis. Linnaeus University, Växjö, Sweden, 2021.
85. Johansson, D. Design and Evaluation of an Avatar-Mediated System for Child Interview Training. Master’s Thesis, Line University, Kanagawa, Japan, 2015.
86. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M. and Ghemawat, S. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv* **2016**, arXiv:1603.04467.
87. Bunk, T.; Varshneya, D.; Vlasov, V.; Nichol, A. Diet: Lightweight language understanding for dialogue systems. *arXiv* **2020**, arXiv:2004.09936.
88. ITU-T Recommendation P.809. *Subjective Evaluation Methods for Gaming Quality*; International Telecommunication Union: Geneva, Switzerland, 2018.
89. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv* **2019**, arXiv:1910.13461. <https://doi.org/10.48550/ARXIV.1910.13461>.
90. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels, Belgium, 1 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 353–355. [[CrossRef](#)]
91. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
92. Jack, R.E.; Garrod, O.G.B.; Schyns, P.G. Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time. *Curr. Biol.* **2014**, *24*, 187–192. [[CrossRef](#)]
93. Deepfakes. github. 2019. Available online: <https://github.com/deepfakes/faceswap> (accessed on 20 May 2022).
94. Sha, T.; Zhang, W.; Shen, T.; Li, Z.; Mei, T. Deep Person Generation: A Survey from the Perspective of Face, Pose and Cloth Synthesis. *arXiv* **2021**, arxiv:2109.02081.
95. Zhu, H.; Luo, M.; Wang, R.; Zheng, A.; He, R. Deep audio-visual learning: A survey. *Int. J. Autom. Comput.* **2021**, *18*, 351–376. [[CrossRef](#)]
96. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion* **2020**, *64*, 131–148. [[CrossRef](#)]
97. Kumar, R.; Sotelo, J.; Kumar, K.; de Brébisson, A.; Bengio, Y. ObamaNet: Photo-realistic lip-sync from text. *arXiv* **2018**, arXiv:1801.01442.
98. Bangerud, G.A.; Johnson, M.S.; Klingenberg Røed, R.; Lamb, M.E.; Powell, M.; Thambawita, V.; Hicks, S.A.; Salehi, P.; Hassan, S.Z.; Halvorsen, P.; et al. Multimodal virtual avatars for investigative interviews with children. In Proceedings of the 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval, Taipei, Taiwan, 21 August 2021; pp. 2–8.
99. Chung, J.S.; Nagrani, A.; Zisserman, A. Voxceleb2: Deep speaker recognition. *arXiv* **2018**, arXiv:1806.05622.
100. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [[CrossRef](#)] [[PubMed](#)]
101. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–19 June 2019; pp. 4401–4410.
102. Mori, M.; MacDorman, K.F.; Kageki, N. The uncanny valley [from the field]. *IEEE Robot. Autom. Mag.* **2012**, *19*, 98–100. [[CrossRef](#)]

-
103. MacDorman, K.F.; Green, R.D.; Ho, C.C.; Koch, C.T. Too real for comfort? Uncanny responses to computer generated faces. *Comput. Hum. Behav.* **2009**, *25*, 695–710. [[CrossRef](#)]
 104. Brunnström, K.; Beker, S.A.; De Moor, K.; Dooms, A.; Egger, S.; Garcia, M.N.; Hossfeld, T.; Jumisko-Pyykkö, S.; Keimel, C.; Larabi, M.C.; et al. *Qualinet white Paper on Definitions of Quality of Experience*; HAL: Bengaluru, India, 2013.