ACIT5900

MASTER THESIS

in

Applied Computer and Information Technology (ACIT)

May 2022

Applied Artificial Intelligence

Impact of different type of Child Avatar Interactions on user Quality of Experience

Alexander William Ingvarsson Hals

Department of Computer Science

Faculty of Technology, Art and Design

OSLOMET

Abstract

Conducting an interview and communicating with children that have experienced traumatic situations can be difficult. Norway's Child Protective services received in 2017 over 58.580 reports about child maltreatment and estimated that over 118 million children in Europe are victim to abuse.

Gunn Astrid Baugerud and her team are creating an avatar in virtual reality that will work as a training avatar for police to work on their interviewing abilities. This thesis will improve this avatar by testing multiple techniques and technologies that can be applied to the avatar. The thesis will evaluate two different solutions made with Unity and one artificially generated avatar with generative adversarial network techniques. The thesis is however limited to only Unity based avatars and movement testing of movement have been restricted due to corona-19 restrictions that was applied during the thesis.

The testing of the three different solution gave indications that avatars created with generative adversarial network techniques had the best impact on realism and overall experience. The avatars created with Ready Player Me follow closely, with minor differences when it comes to appearance and experience. The worst avatar was made with Unity multipurpose avatar 2, during the questionnaire one of the avatars created with UMA2 had the worst results of all avatars.

The avatar created in the main project by Gunn Astrid Baugerud and her team has come a great way, the results of this thesis show what could be the focus points when creating the final virtual child avatar, with post-questionnaire indicating that eye contact and graphics are among the most important attributes for accomplishing realism.

Preface

This master thesis marks the end of my education at department of Computer Science at Oslo Metropolitan University. I chose the master's program in Applied Computer and Information Technology (ACIT) with focus on applied artificial intelligence at the Faculty of Technology, Art and Design due to my interests in artificial intelligence and the potential it can have on our future.

The idea for this project was put forward by Saeed Shafiee Sabet for SimulaMet. I would like to thank my supervisors at SimulaMet, Saeed Shafiee Sabet, Pegah Salehi, Syed Zohaib Hassan for your feedback, guidance, support and encouragement. I would like to give a special thanks to Saeed for providing me with data and input during all stages of my work.

Last but not least, I would also like to express my gratitude towards my supervisor at OsloMet Prof. Pål Halvorsen for valuable feedback and guidance.

Alexander Hals

Table of Contents

List of Figures

List of Tables

1. Introduction

Since 2021, OsloMet is conducting the research project “Professionals interviewing maltreated children supported via artificial avatars”, which aims to create an interview training program for child protection services and law enforcement. The training program will be digital and combine expertise in developmental psychology with artificial intelligence (AI) (Oslomet , 2022). Simula Metropolitan Center for Digital Engineering (SimulaMet) is part of this research and are working on creating the virtual reality child avatar.

This thesis will conduct tests for SimulaMet to assess the importance of avatar animations and compare the realism of different avatars, depending on the technology used to create them. The thesis will compare two Unity created avatars and one Artificial Intelligence created avatar.

1.1 Motivation

Interviewing and communicating with children that have experienced traumatic situations can be difficult. In 2017, Norway’s Child Protective services received over 58.580 reports about child maltreatment, and it is estimated that over 118 million children in Europe are victim to abuse. Of the 118 million children in Europe that experience abuse. Unfortunately, the abuse has fatal consequences for 850 children each year. However, the majority of the victims are left having to deal with psychological problems and possible altered behavior due to the traumatic experience (WHO, 2013).

There are several state issued and private actors that will try to help children that have experienced abuse, among them are Child Protective Services and law enforcement. To be effective, they need information from the victims, this is however often challenging due to children with traumatic experience are often not willing to talk with a random person. It is therefore important that the interviewer can convey the needed information on a level that the children can understand. There are even studies that show that the type of question the interviewer use is more important than the social skills and cognitive competence of the child (Simulamet). It is of importance that the interviewers can find good and reliable ways to train on how to conduct an interview with correct questions.

With recent technology like virtual reality, user can be put in an environment of their choosing without doing any actual impairment. Virtual reality could therefore be one way to

create a simulation for child interviewing. Creating a realistic environment can however be challenging, as it requires knowledge about what is important for the immersion and for the user to have a good user experience.

This paper will assess what features are important when creating a child avatar in virtual reality and conduct a user-case study to get feedback.

1.2 Problem statement

The challenge for this project is to create a Virtual Reality Avatar that can help law enforcement with training how to communicate or interview a child that has experienced something traumatic. How the Virtual Avatar is behaving during the interaction is important for the user experience, and there is therefore a need for more information on what functionalities are important for the user experience to be more realistic. When observing the avatar in the virtual reality, certain functions like movement and animations make a difference on user experience, while others are not as important. The objectives for the thesis are:

Objective 1: Get an overview of available methods for creating talking avatars.

When creating avatars there are many different solutions and techniques available to use. The first objective will therefore be to get an overview of the available technology that can be used.

Objective 2: Generate realistic talking avatars, with the identified methods from objective 1.

The methods identified in objective 1 will be used to create avatars.

Objective 3: Compare the generated avatars and examine which one creates the best user experience.

When the avatars are created, an experiment will be conducted to test which one of the avatars give the best user experience.

Objective 4: Compare the generated avatars and examine which one has the most realistic appearance.

The experiment will also test which one of the avatars has the most realistic appearance.

1.3 Scope

When searching for available techniques and technologies there are too many available solutions, as there are solutions created on numerous different game engines and environments. This thesis will therefore narrow the scope down to Unity based avatars.

The equipment used is also a limiting factor of why Unity was used instead of other solutions, such as those based on the Unreal engine. The Unreal Engine requires a lot more computable power to run its avatars than Unity, which is discussed further in chapter 5.3 Future works.

1.4 Ethical considerations

Some ethical considerations for this project are the collection of authentic information. Transcribed investigative interviews of alleged victims of abuse and maltreatment requires us to adhere to special ethical requirements for the use of sensitive data that involves vulnerable children (Simulamet). All interviews and assessment of any participants in this project will be conducted fully anonymous. The project is registered with Norsk senter for forskningsdata NSD. Any avatar images will be created with morphed real images of different children to make the face untraceable and unidentifiable to any real person. This project will also make ethical contribution by making it possible for interviewers to practice their interviewing skills on an avatar instead of a real child.

1.5 Research method

The thesis will find out if any of the objectives improve the overall realism when creating an avatar, the test results will be tested with a repeated measure ANOVA test to find any significant differences. During the questionnaire the participants will be asked to answer a survey with questions related to the user experience.

The design of the avatars in this project will be created in Unity with models gathered from UMA2, Ready Player One and animations gathered from Mixamo which is a free library with full-body animations. Lip-sync will be created with Unity and voice lines will be provided by SimulaMet in a file with pre-made computer voice that mimic authentic voices. All the

tests will use the same two voice lane, one female and one male. The questionnaire will be created in six samples to create random order for the participants.

1.6 Main Contribution

The project will during the research phase focus on two main sections, one is development of avatars, and the second part is evaluation of assessing the avatars made in the first section. The evaluation will answer the problem statements mentioned above in section 1.2. The main contribution of this paper is the following:

Objective 1: Get an overview of available methods for creating talking avatars.

The technologies used in this project are Unity and avatars created with Generative adversarial network. Mixmao was for animations and lip-sync Unity was created with Salsa and Ocolus.

Objective 2: Generate realistic talking avatars, with the identified methods from objective 1.

In Unity two different techniques were used to create avatars, first technique based on Unity Multipurpose Avatar 2 and second one based on Ready Player Me. The third avatar was created with MakeItTalk and GAN technologies. In the project a total of twelve different avatars was created, four based on Unity Multipurpose Avatar 2, four based on Ready Player Me and four based on MakeItTalk. Unity multipurpose avatar 2 and Ready Player Me used two different lip sync solutions and MakeItTalk used a third solution.

Objective 3: Compare the generated avatars and examine which one creates the best user experience.

After comparing the three avatars the overall best technology and technique used to create the best user experience based on the results was an artificially generated avatar created with MakeItTalk.

Objective 4: Compare the generated avatars and examine which one has the most realistic appearance.

After comparing the avatars, the overall most realistic avatar when it came to appearance was an artificially generated avatar created with MakeItTalk.

1.7 Thesis outline

The thesis is organized in the following order:

Chapter Two – Background: this chapter contain background information for more context. Chapter two will start with related works and what has been done so far and then go more into type of technologies that are used.

Chapter Three – Methodology: The third chapter will go into detail on the methods that was used to create the avatars. This section is divided into one part with the general design of the avatars and a second part then more in detail on how each of the avatar solutions was created. How the animation was created with Mixamo will be mentioned during the last part of chapter three.

Chapter Four – Experiments and Results: this chapter will mention how the experiment was created and conducted. Information on participants and test conditions will be mentioned and additionally, the results from the test will be presented and any findings will be presented and discussed in this chapter.

Chapter Five – Conclusion: this chapter will summarize what has been done in the thesis and suggest any future work that can improve the work of creating a child avatar.

2. Background

This chapter will provide some background context too future explain the motivation and reason of this thesis. This chapter begins with related work and any similar virtual reality chatbots and then continue into the main project of which this thesis is accommodating. The chapter will go more into detail of what is planned for the project and mention some of the progress done in reference to creating a virtual reality avatar. There will also be sections that will explain technologies that is used in the project.

2.1 Virtual chatbots

There are multiple training programs that use avatars for interview training, Virtualspeech have created a virtual environment where the user can learn how complete a job interview (Virtualspeech, 2022). HAL open science has published an article about using virtual avatars as children's companions (Elsa Thiaville, 2020). Another publishment talks about how to design an avatar-mediated system for child interview training (Johansson, 2015).

2.2 Multimodal Virtual Avatars for Investigative Interviews with Children

In August 2021, a research article with the title Multimodal Virtual Avatars for Investigative Interviews with Children was published by Gunn Astrid Baugerud and co (Gunn Astrid Baugerud, 2021). This study presented their ongoing work on how to train police officers to conduct interviews with children that have traumatic experience or have experienced abuse. Their goal is to create a training program with multimodal model that use artificial intelligence, chatbot, generation of visual content, text-to-speech, and speech-to-text to create a training program that officers can use for training purposes. The goal is that the program will be able to create unlimited scenarios in an environment that is realistic, in this environment the officers will hopefully be able to ask the avatar question and it would respond appropriately.

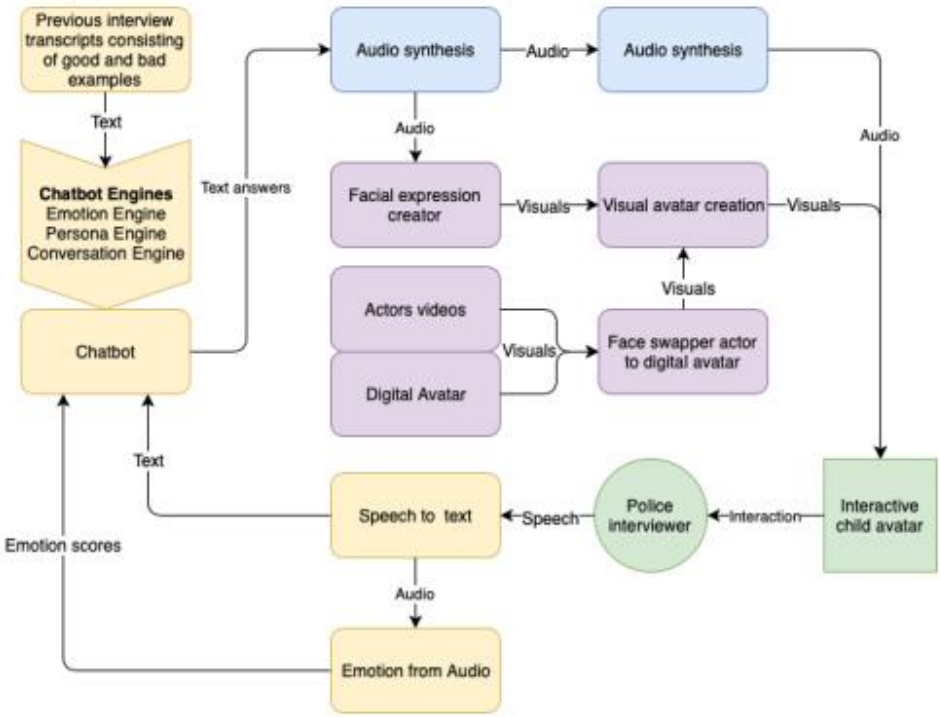


Figure 2.1 System architecture. Green blocks mark the interactive parts, yellow is text related, blue is audio, and purple marks (Gunn Astrid Baugerud, 2021)

Figure 2.1 show the system architecture that is planned for the final avatar that is going used in this project. The Australian group of the project created an interactive training system that created interviews with actors and simulated interviews with a child avatar. Participants for this training system was given training in how to choose effective questions and completed multiple interviews with virtual avatar several times a week. The results of this study showed that interview training with avatars and with regular feedbacks enhances the interviewers interview technique.

The project is still ongoing, and the paper mention that the ongoing first work have been focused on research and the design for proof of concept. Currently in the project they have experienced the use of Faceswap (Deepfakes., 2019) and ObamaNet (Rithesh Kumar, 2018). They first tried Faceswap, but due to the technology requiring full video and many images of the face to work properly it does not give the wanted realistic results. Secondly, they tried ObamaNet which is a technology that move the mouth region and synchronize it with a given audio input. The video had however some uncoordinated mouth movement

that made it less realistic but was found to be something that could be improved in future experiments.

2.3 Proposal Avatar 2.0

The Proposal Avatar 2.0 (Simulamet) is linked to 2.2 and is the proposal given to this project. This proposal has more information about how the group is working and their current progress. The proposal go more into detail on how the development of the avatar is going, the proposal mention that unlike earlier approaches a completely new avatar concept is planned. This new concept is based around a chatbot, and a visual avatar combined with iterative design and experimental prototyping, something that will hopefully produce a realistic virtual child that can be used in interviewer-training. The interactive chatbot is based on deep neural network with 1000's of hours of real recoded interview as data. This will create a chatbot that have a greater pool of answers and will be able to answer more than simple questions.

The project is divided into four work packages and some of the work packages are subdivided into tasks. These work packages and sub-tasks are linked to the figure 2.2. The first work package is to conduct a pilot study evaluation and testing to evaluate students' perceptions of the avatar's interactional characteristic. The second work package is about the research and development of chatbots and avatars, the second work package has more sub-tasks with focus on chatbot development, avatar development and combining the chatbot and avatar. The third work package will focus on two-group pretest-posttest design with a pre-v post-training measurement, the two groups conducting the test will hopefully be around 80 child protective service representatives and 80 police force representatives. The last work package will be a six-month follow-up, this work package is split into two sub-tasks where the first sub-task will do a withing-participants assessment and the second sub-

task will be to pre-, post- and follow-up training assessment.

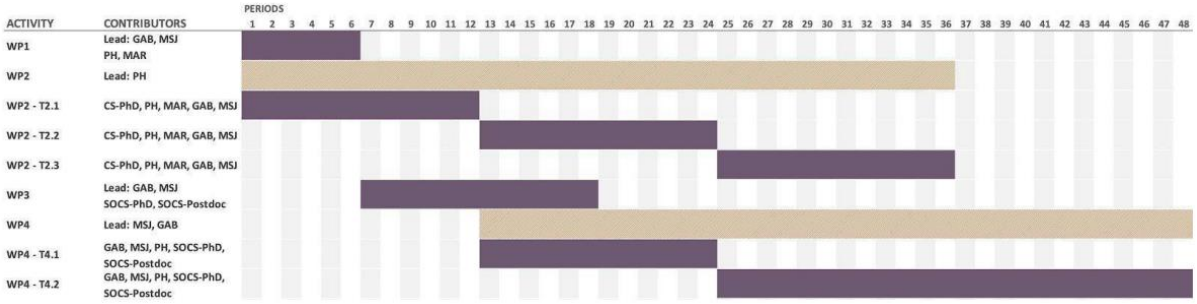


Figure 2.2 Gantt chart of project plan with allocated time and personnel, WP is short for work packages and T is short for sub-tasks

For the different work packages different teams will lead the package, for work package one OsloMet will be leading and guiding this task. For the second work package SimulaMet will lead the technical work, for the third package OsloMet research will have responsibility with close collaboration with international partners, CPS and the police.

The proposal also mentions the main project group as follows. Gunn Astrid Baugerud, Associate professor at Oslo Metropolitan University will lead the project. Gunn is in expert in child maltreatment and has a degree in cognitive development psychology (PhD), child welfare (M.A) and psychology (M.Sc). Miriam S. Johnson, Associate professor at Oslo Metropolitan University, is a clinical psychologist with degrees in witness psychology (PhD), social and community psychology (M.Sc), and clinical psychology(cand.psych). Michael Alexander Riegler, Chief research Scientist, SimulaMet. Michael has a degree in computer science from the University of Oslo (UiO) and Klagenfurt University. Pål Halvorsen, Professor and Chief Rsearch Scientist SimulaMet, also holds degree in computer science from the University of Oslo. The Project also have two PhD positions that will together with Pål Halvorsen and Michael Alexander Riegler have as primary responsibility to develop the technical part. This thesis is part of the work that the PhD students is conducting.

2.4 Unity

In this thesis the avatar will be created with the use of Unity, Unity is a cross-platform game engine that is often used to create simulations and video games for computers, consoles and mobiles. Unity was announced in 2005 and was only working for OS X but has since been expanding and is now usable by over 27 platforms (Freecodecamp, 2020). Unity uses an all-purpose game engine for 2D and 3D graphics and has support for C# scripting. The interface

on Unity is very user friendly, most of its functions are drag and drop and there are countless guides and videos that can help developers with creating their own game, simulation or environment.

2.5 Virtual reality

The avatar that is going to be created in this thesis is planned to be created to work in virtual reality. This section will quickly look on what virtual reality is and what the challenge with this technology. Virtual reality is a technology that gives the user the experience to simulate anything a developer has created as if the person was there. This is achieved by having virtual reality headsets that in current iteration have base stations around the room or sensors on the headset that it synchronizes with so that the headset knows where the person is in the room. The user is free to move around in the real world, but the virtual reality headset can project something different for the user. Virtual reality has the technology to transfer a person from their living room into a tennis match or skiing downhill, and by adding sound the experience becomes even more real and can sometimes make people lose balance and even hurt them self-due when reacting to something inside the virtual environment. To interact with the virtual reality each user has special controllers that function like your hands would in real life, they are projected into the virtual headset and have real-time tracking and often vibration to make the immersion even more real.

There are multiple manufacturers that have different virtual reality headsets and controllers, but the main difference is often just the controller and graphic standard. This project is currently using Oculus Rift as the virtual headset see figure 2.3.

Figure 2.3 Example of current virtual reality with Oculus Rift and what the person can see is displayed on the monitor (Hjetland, 2021)

2.6 Generative Adversarial Network

Mentioned previously in 2.2 and 2.3 the current Avatar development using DeepFake technologies. In order to create realistic video content while using a person's facial expression generative adversarial network (GAN) will be integrated into the DeepFake system (Simulamet). GAN is not widely known so this section will in short words explain what GAN is and what GAN is used for. GAN was created in 2014 by Ian Goodfellow and colleagues and is a class of machine learning framework (Wikipedia , 2022). GAN can read a training set data and from the training set create a plausible picture, GAN can make a photo into a emojis, generate photographs of human faces and face aging, Figure 2.4 show the progression that GAN has had in its capability since 2014 till 2017 (Brownlee, 2019).

Figure 2.4 Example of the progression in the Capabilities of GAN from 2014 to 2017 (Brownlee, 2019)

2.7 Summary

In this chapter related work and projects that have close relation with the thesis has been mentioned and explained. The main project of creating a Multimodal Virutal Avatar has been explained in more detail and the plan for the project and progress has been mentioned in the Proposal 2.0.

Unity, Virtual Reality and generative adversarial network technologies that are relevant for the thesis has been explained shortly.

3. Methodology

In this chapter the methods used to create the different avatars will be presented, the chapter will start with what the original plan for the thesis were and what has changed since the beginning of the thesis. The chapter will go more into detail on how UMA2, Ready Player Me and MakeItTalk was used and will end with details on how the animations was managed with Mixamo.

3.1 Design of Avatars

During the planning phase of creating the child avatar that would be evaluated in this thesis the plan was to create a fully working child avatar in virtual reality. However, due to the corona situation and restrictions it became apparent that doing a physical test would not be possible. Therefore, the test was changed to be planned around making a pre-made video questionnaire and finding out which one of the avatars would be best suited for virtual reality and find out what users prioritized when interacting with avatars in games.

The avatars in this project are all created with Unity, but the avatars are created with different libraries and technologies. The first library that was used was UMA2 chosen due to SimulaMet already had created earlier solutions UMA2, more about UMA2 at 3.2. The second avatar was created with the use of Ready Player Me more at 3.3 and the third solution was created with GAN technology more in section 3.4.

When the three different avatars where ready they were standardized by removing the background and any furniture that was present on some of the avatar solutions. The background and furniture were removed due to making the test as similar as possible without any outside factors making an impact on realism and the answers from participants. After the avatars were ready it was time to record and do the cropping of the videos and uploading to YouTube, this had unfortunate complications due to the size of the video recorded was so small that it created shorts on YouTube instead of normal video and Google Sheets does not accept shorts. This was fixed by forcing and tricking the URL that was used in Google Sheet to look like a normal URL.

3.2 Unity Multipurpose Avatar 2

The first avatar is created with Unity Multipurpose Avatar 2, this is a library/asset that developers can find on unity asset store (UMA Steering Group, 2021). UMA2 is a big

package having everything needed to create an avatar, of the available premade avatars one was chosen to start creating animation and getting to know how to use Unity see figure 3.1.

Figure 3.1 Snippet of one avatar made with UMA

A prototype avatar was already created by SimulaMet with UMA2, so main work with the UMA2 avatar was to create animations and lip sync with the use of SALSA lipSync Suite (crazyD, 2022). SimulaMet had created an earlier environment that could also be used see figure 3.2.

Figure 3.2 Avatar interaction room created for the avatars by SimulaMet

The first avatar was made with UMA2 due to SimulaMet already having set up avatars with UMA2, this made it easier to learn the basics of Unity with the help of the PHD students.

3.3 Ready Player Me

The second avatar was created with Ready Player Me, which is an avatar first created on a webpage (Ready Player Me, u.d.) and download with Ready Player Me SDK. On the webpage the user can freely customize the avatar to their linking with multiple pre-made options. There are options to change skin color, face shape, nose shape, eye shape, mouth shape, hair style, eye color and everything needed to create an avatar to own preference see figure 3.3 for reference.

Figure 3.3 Snippet of readyplayer.me/avatar, the bar below shows options of what the user can change. This is one of the avatars used for ready player me in the thesis

To download the avatar from the website Ready Player Me Unity Avatar SDK is required, this can be downloaded from readyplayer.me (Ready Player Me, 2022). Once the SDK is downloaded into Unity one can download the avatar from the webpage with a. glb url. The .glb url is copy pasted into Ready Player Me SDK and one can choose to have pre-set eye animations and voice animation see figure 3.4 for reference.

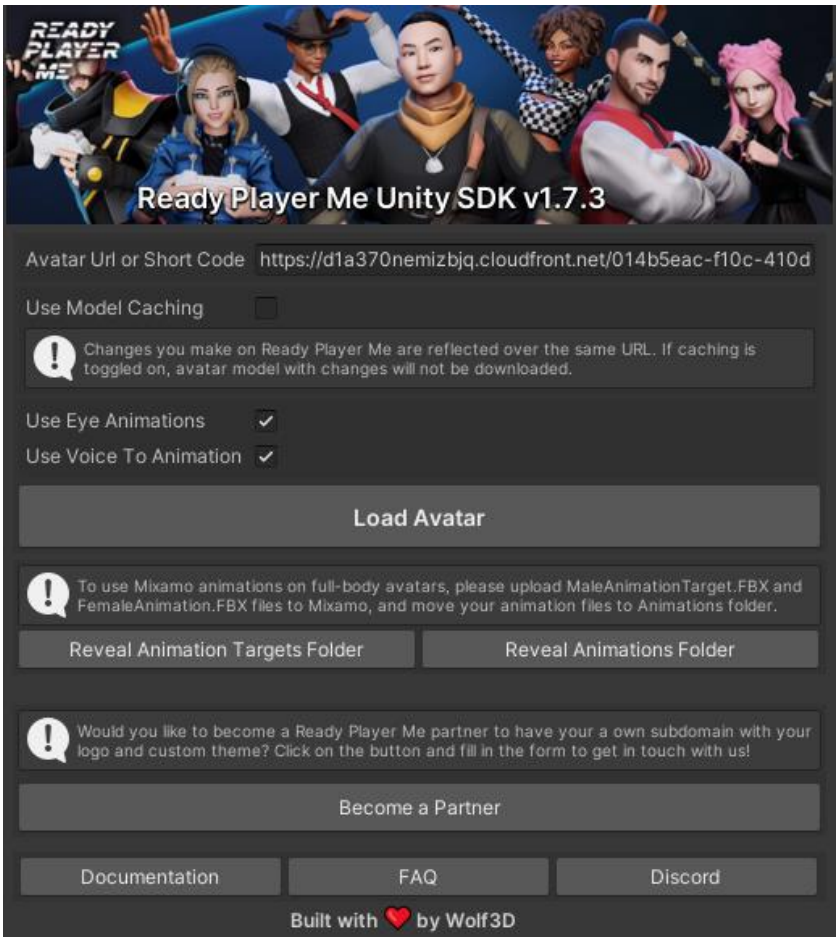


Figure 3.4 Snippet of ready player me Unity SDK, used for downloading avatars from readyplayer.me

Once the avatar is loaded one can start with animations and how to use the voice animation. The option to use voice to animation as seen in figure 3.4, give the avatars that is downloaded a Voice Handler in the inspector see figure 3.5 for reference. Ready Player Me use Ocolus lip sync (Unity - oculus, 2021). The Voice Handler take either an audio clip or microphone input and animates voice movement on the avatar. The animation is however not perfect and does not make different mouth movement depended on how loud the audio clip is or how loud one speaks.

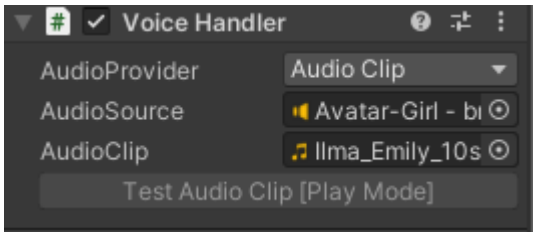


Figure 3.5 Snippet of the voice handler found in the inspector when selecting the avatar

The reason to use Ready Player Me as one of the avatars is due to the voice handler, trying to create a good and working lip sync is a challenge. When creating lip synchronize it is hard to synchronize the face movement and lips to believable animations. And as one of the problem statements of this thesis is to find out what avatar is the most realistic it became a challenge to find a good lip sync and a need to evaluate more than one lip sync solution became apparent.

3.4 MakeItTalk

The third avatar was created with MakeItTalk. MakeItTalk uses GAN technology to generate state-of-the-art animated faces with PC-AVS (Hang Zhou, 2021) and StyleGan (Tero Karras S. L., 2019) (Tero Karras S. L., 2020). PC-AVS takes a facial image and generates a talking-head which is controlled by another facial image trained with VoXCeleb2 (Zisserman, 2018) that generate expressive animations for the face based on the sound input. So, for instance figure 3.6 show the process work, first you have the facial image on the left and the PC-AVS facial image in the middle, the middle image is controlled by the facial expression made on the right picture and the output used for the videos is the middle image. This was done on four different images, two male and two female.

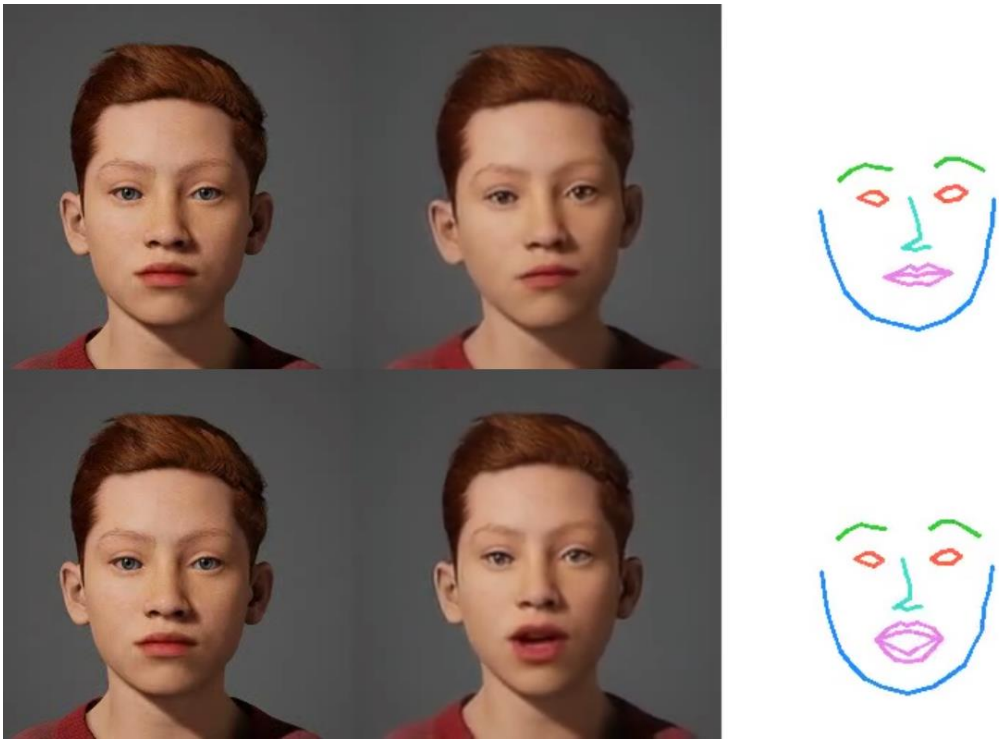


Figure 3.6 Snippet of MakeItTalk generation with two different facial expressions

3.5 Mixamo

The animation applied to UMA2 avatars and Ready Player Me avatars a library named Mixamo was used (Mixamo, 2022). This webpage has pre-made animations that can be imported and fixed to the skeleton used by UMA2 and Ready Player Me, see figure 3.7 for examples of animations on Mixamo.

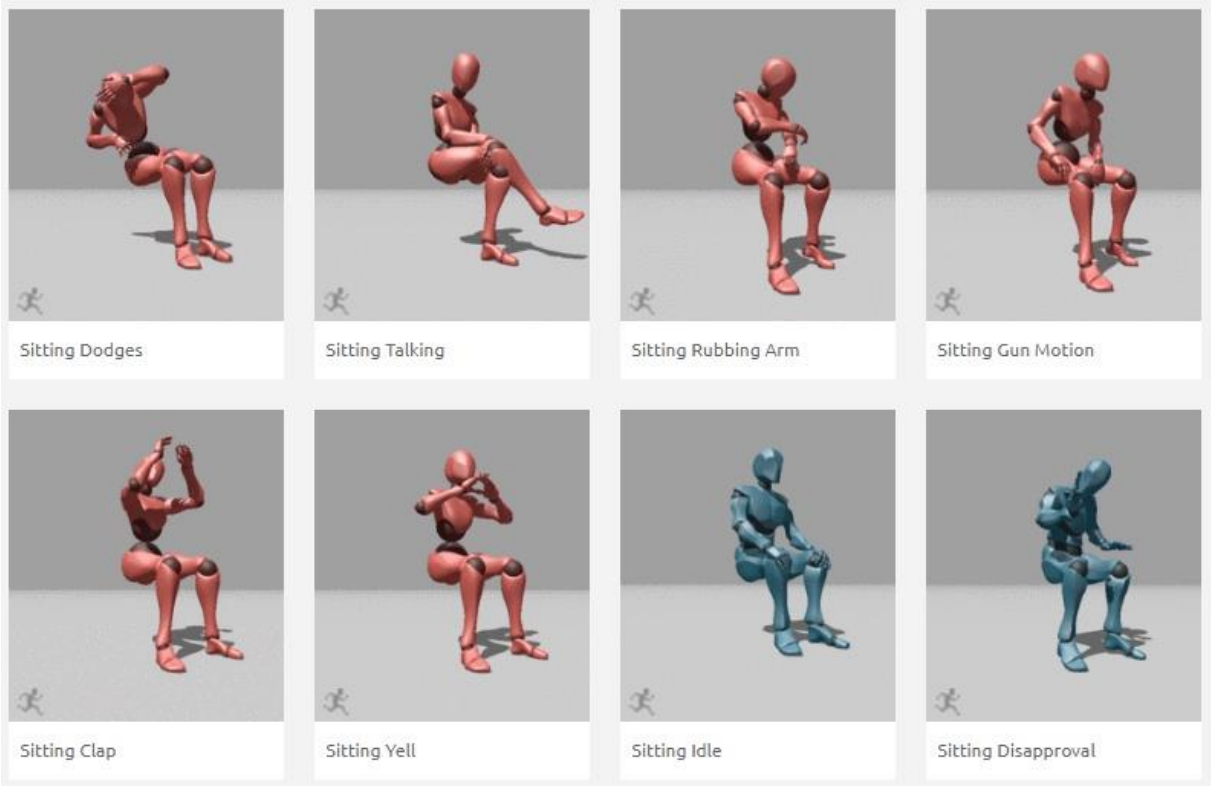


Figure 3.7 Snippet of mixamo.com and a preview of some available sitting animations (Mixamo, 2022)

Each avatar in Unity has an animator that control their animations see figure 3.8 for example of the animator. The animator is a simple drag and drop function with priorities and conditions being set by the user. Example of this is the arrows seen on figure 3.8, the arrows indicate where the animations should proceed after finishing their previous animation, in this case the avatar start its animation by sitting idle and then can be prompted to do certain animations by for example a C# script.

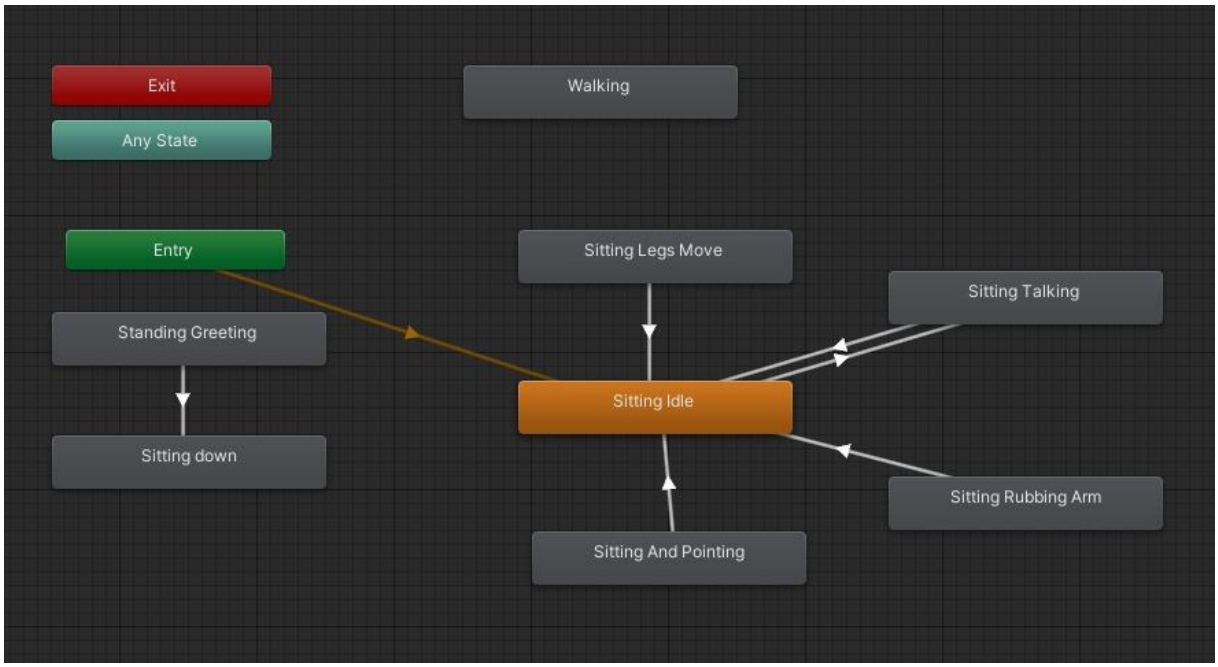


Figure 3.8 Snippet of animator for one of the Ready Player Me avatars

To have the animations work at given times a C# script was created; this was created having the animations play when the key assigned to their animation was pressed while the game was running in Unity see figure 3.9 for the code.

```
using System.Collections;
using System.Collections.Generic;
using UnityEngine;

public class AnimControl : MonoBehaviour
{

    public GameObject theAvatar;

    void Update()
    {
        if (Input.GetButtonDown("1key"))
        {
            theAvatar.GetComponent<Animator>().Play("Sitting Idle");

        }
        if (Input.GetButtonDown("2key"))
        {
            theAvatar.GetComponent<Animator>().Play("Sitting Talking");
        }
        if (Input.GetButtonDown("3key"))
        {
            theAvatar.GetComponent<Animator>().Play("Sitting Legs Move");
        }
        if (Input.GetButtonDown("4key"))
        {
            theAvatar.GetComponent<Animator>().Play("Sitting And Pointing");
        }
        if (Input.GetButtonDown("5key"))
        {
            theAvatar.GetComponent<Animator>().Play("Sitting Rubbing Arm");
        }

    }
}
```

Figure 3.9 C# code for making animations play from the animator

The choice to use Mixamo was due to convenience and the ease of use. Instead of creating the animation from the ground up it was easier to tweak the animations to fit the

characters made with UMA2 and Ready Player Me, especially when not having any prior experience with Unity and how to create animations.

3.6 Summary

This chapter has gone into detail on how the design of the different avatars was planned process of creating the avatars based on the plan. The chapter started with basis information on how the design of avatars should be then went into more detail on each solution.

The first solution with UMA2 was created with Unity and UMA2 use SALSA LipSync Suite for lip sync. During the UMA2 section the planned background were presented. The second solution was created with Ready Player Me, which uses an external website to create the avatar then import into Unity. Ready Player Me avatar uses Ocolus for lip sync. The third solution is MakeItTalk which is a GAN created avatar and use VoXCeleb2 for lip sync.

The last section goes into detail on how the animation handler works on Unity, this is for creating animation for UMA2 and Ready Player Me avatars.

4. Experiments and Results

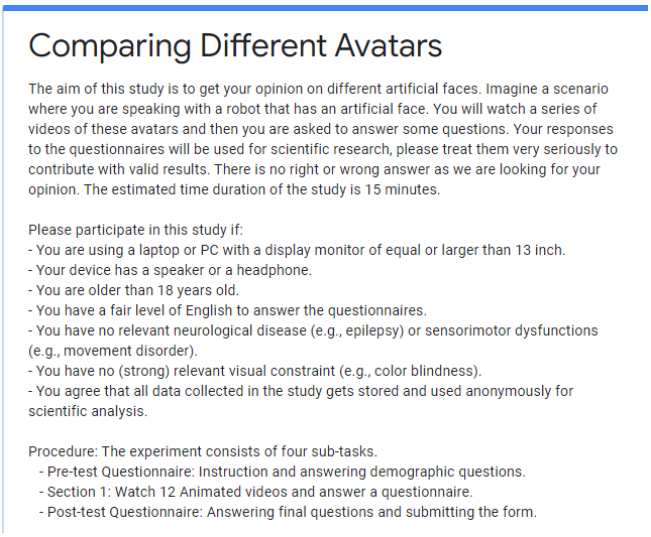
In this chapter the experiment will be presented. The first part will mention the test design and how the questionnaire was created and the layout of the questionnaire. Then the next section will go more into detail on the participants for the experiment and followed by a section about test questions.

The results will then be presented in five sections with the first three sections comparing UMA, Ready Player Me and MakeItTalk within each group. The fourth section will compare the three best avatars, one from each solution. And the last section will present the post-questionnaire results and the results will then be discussed.

4.1 Test Design

The questionnaire was created using google forms and were split into four different sections. Section one had a general overview of the questionnaire and what it contained, section one also had some requirements written down so that the participants would have more standardized equipment and equal prerequisites see figure 4.1. The prerequisite had the following information:

- You are using a laptop of PC with a display monitor of equal or larger than 13 inches
- Your device has a speaker or a headphone
- You are older than 18 years old
- You have a fair level of English to answer the questionnaire
- You have no relevant nourological disease (e.g epilepsy) or sensorimotor dysfunctions (e.g., movement disorder).
- You have no (strong) relevant visual constraint (e.g color blindness).
- You agree that all data collected in the study gets stores and used anonymously for scientific analysis



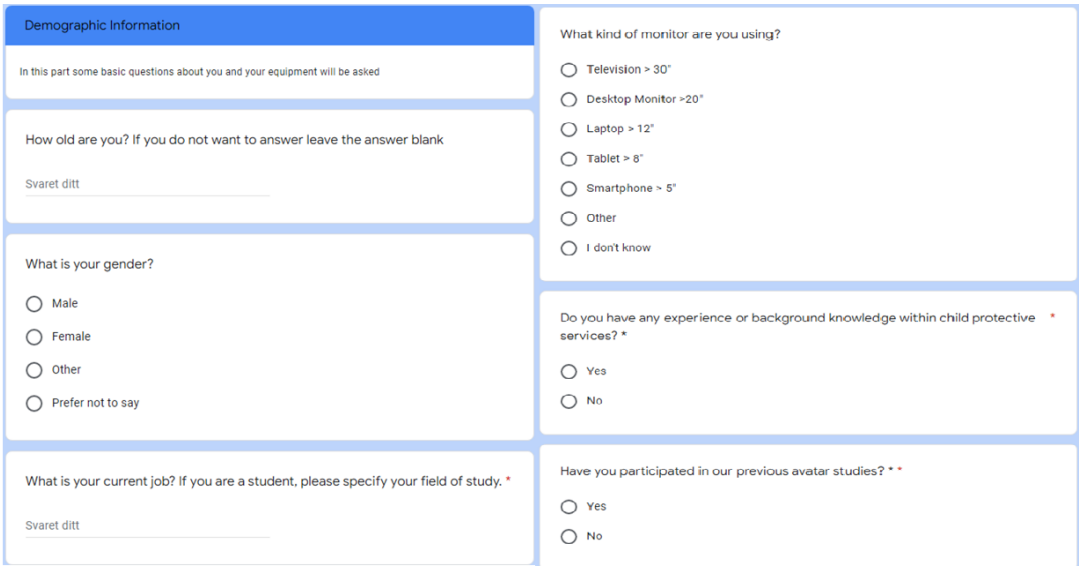
Comparing Different Avatars

The aim of this study is to get your opinion on different artificial faces. Imagine a scenario where you are speaking with a robot that has an artificial face. You will watch a series of videos of these avatars and then you are asked to answer some questions. Your responses to the questionnaires will be used for scientific research, please treat them very seriously to contribute with valid results. There is no right or wrong answer as we are looking for your opinion. The estimated time duration of the study is 15 minutes.

Please participate in this study if:
- You are using a laptop or PC with a display monitor of equal or larger than 13 inch.
- Your device has a speaker or a headphone.
- You are older than 18 years old.
- You have a fair level of English to answer the questionnaires.
- You have no relevant neurological disease (e.g., epilepsy) or sensorimotor dysfunctions (e.g., movement disorder).
- You have no (strong) relevant visual constraint (e.g., color blindness).
- You agree that all data collected in the study gets stored and used anonymously for scientific analysis.

Procedure: The experiment consists of four sub-tasks.
- Pre-test Questionnaire: Instruction and answering demographic questions.
- Section 1: Watch 12 Animated videos and answer a questionnaire.
- Post-test Questionnaire: Answering final questions and submitting the form.

Figure 4.1 Snippet of the first section of the questionnaire for bigger picture see appendices figure A.1



Demographic Information

In this part some basic questions about you and your equipment will be asked

How old are you? If you do not want to answer leave the answer blank

Svaret ditt

What is your gender?

- Male
- Female
- Other
- Prefer not to say

What is your current job? If you are a student, please specify your field of study. *

Svaret ditt

What kind of monitor are you using?

- Television > 30"
- Desktop Monitor >20"
- Laptop > 12"
- Tablet > 8"
- Smartphone > 5"
- Other
- I don't know

Do you have any experience or background knowledge within child protective services? *

- Yes
- No

Have you participated in our previous avatar studies? * *

- Yes
- No

Figure 4.2 Snippet of the second section of the questionnaire, with demographic questions. For bigger picture see appendices figure A.2 and figure A.3

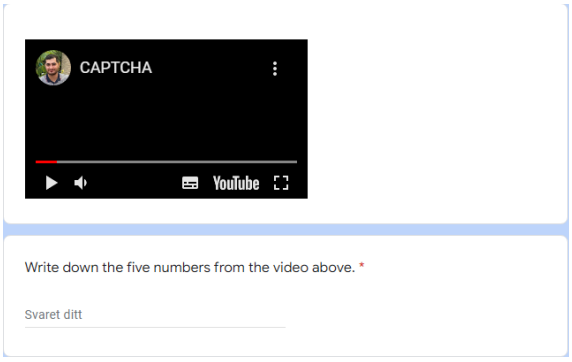


Figure 4.3 Snippet of the second section of the questionnaire, with a homemade CAPTCHA video. For bigger picture see appendices figure A.4

Section two asked about the demographic information about the participants for reference see figure 4.2. This section asked for how old the person was, this question was optional. The second question was the gender of the person, this was separated into male, female, other or prefer not to say and this was also an optional question. The first required question in this part is the field of work or study, this was an input question, so the user must write in their answer. The third question asked about what type of monitor the participant is using, this was split into seven radio buttons with the following alternatives: Television > 30”, Desktop monitor >20”, Laptop >12”, Tablet > 8”, Smartphone >5”, other or I don’t know.

The next two questions ask the participants if they have any experience or background knowledge within child protective services and if they have taken part in any previous avatar studies. Section two ends with a captcha video to confirm that the user have working headphones or speakers for part three.

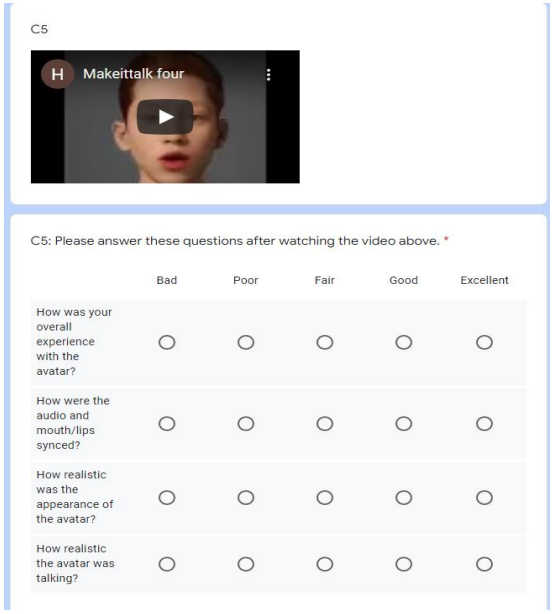
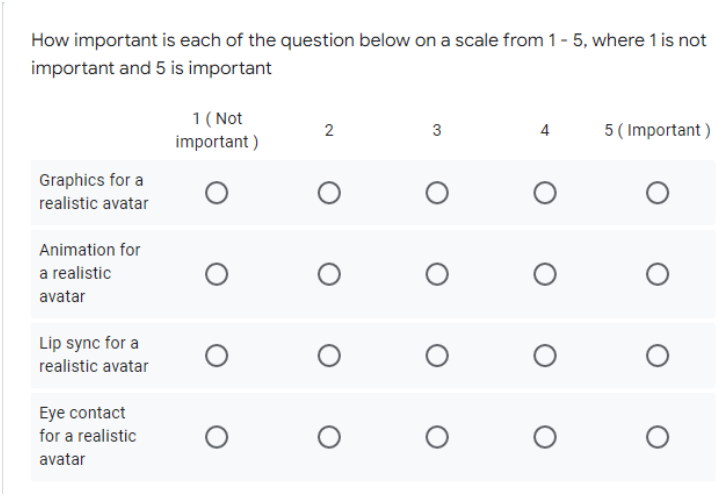


Figure 4.4 Snippet of part three of the questionnaire, for a bigger picture see appendices A.10

Section three has all the avatar videos with questions see figure 4.3 for reference. Here each video had four questions related to how the user experience was when watching the video. Each video was cut to be around ten seconds long and all male avatars used one of two different male voices, the same was done for the female avatars that also used one of two different female voices. The avatars were split into three groups with four in each group, so four UMA avatars, four Ready Player Me avatars and four MakeItTalk avatars. The

avatars were given a number from C1-C12 and to randomize the order of questions there was made six different questionnaires in a random order that was made with a random number generator. All of the different version started with an anchor avatar named C0 see appendices figure A.5 for reference.


How important is each of the question below on a scale from 1 - 5, where 1 is not important and 5 is important

	1 (Not important)	2	3	4	5 (Important)
Graphics for a realistic avatar	○	○	○	○	○
Animation for a realistic avatar	○	○	○	○	○
Lip sync for a realistic avatar	○	○	○	○	○
Eye contact for a realistic avatar	○	○	○	○	○

Figure 4.5 Snippet of part four of the questionnaire, for bigger picture see appendices figure A18

The last section of the questionnaire had four questions about the importance of different technologies and aspects for a realistic avatar see figure 4.5. The questions were created to find what the focus point for future avatars should be and what the participants viewed as most important when thinking of a realistic avatar.

4.2 Participants

In this questionnaire 23 participants participated consisting of 3 females and 20 males, the age ranges from 22-64 years. Most of the participants worked or studied something IT related, one worked as a paramedic, one as a teacher and two as a retail worker see figure 4.6 for reference. The questionnaire was available on different monitors, 14 of the participants used a desktop monitor above 20 inches, 4 used a smartphone around 5 inches, 3 used a laptop above 12 inches and 2 used a television monitor above 30 inches see figure 4.7 for reference. None of the participants had any prior experience with child protective service or taken part in any earlier avatar studies.

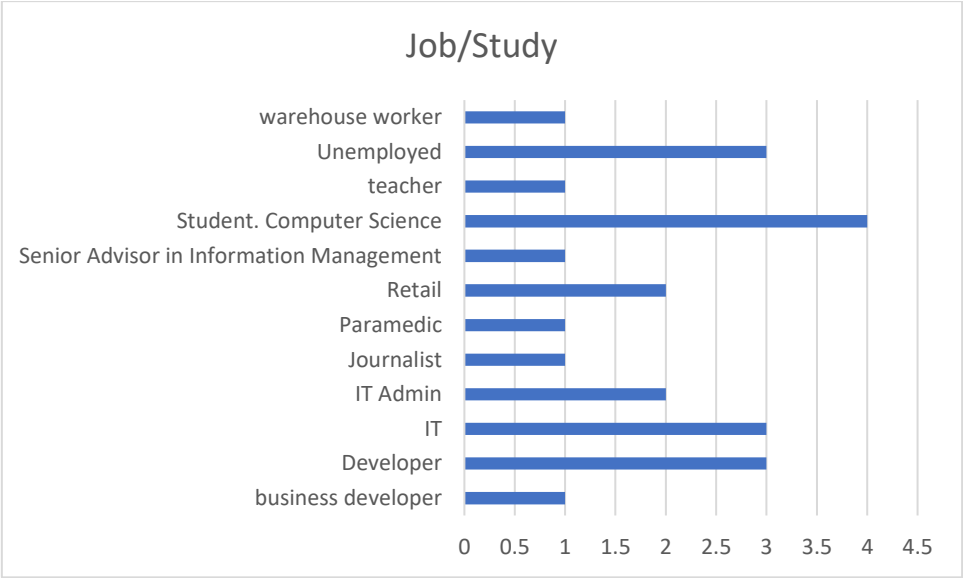


Figure 4.6 Results of participants and their field of work

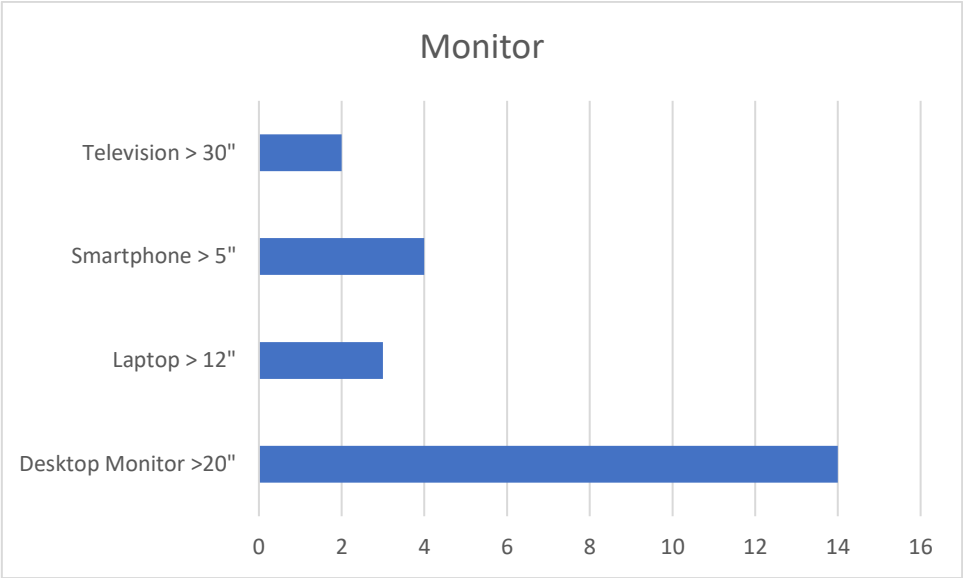


Figure 4.7 Results of participants and their monitor size

4.3 Test questions in the questionnaire

Questions P2	Label
Q1	How old are you? If you do not want to answer leave the answer blank
Q2	What is your gender?
Q3	What is your current job? If you are a student, please specify your field of study.
Q4	What kind of monitor are you using?

Q5	Do you have any experience or background knowledge within child protective services?
Q6	Have you taken part in our previous avatar studies?

Table 4.1 List of items in demographic section of the questionnaire

Questions P3	Label
Q1	How was your overall experience with the avatar?
Q2	How were the audio and mouth/lips synced?
Q3	How realistic was the appearance of the avatar?
Q4	How realistic the avatar was talking?

Table 4.2 List of items in post-experience in the questionnaire

Questions P4	How important is each of the questions below on a scale from 1-5
Q1	Graphics for a realistic avatar
Q2	Animations for a realistic avatar
Q3	Lip sync for a realistic avatar
Q4	Eye contact for a realistic avatar

Table 4.3 List of items in the post-test questionnaire to find what is most important for the participants when it comes to a realistic avatar

All videos are set to be around ten seconds long for normalization. All videos have a background that should not interfere with the experience, see figure 3.1 for an example.

4.4 Results

The results will be presented in five sub-sections, first UMA, Ready Player Me and MakeItTalk will be evaluated and compared against their own equal avatars and then in the fourth section the best avatar from each solution will be evaluated against each other. The fifth section will present the post-questionnaire results.

4.4.1 UMA created avatars

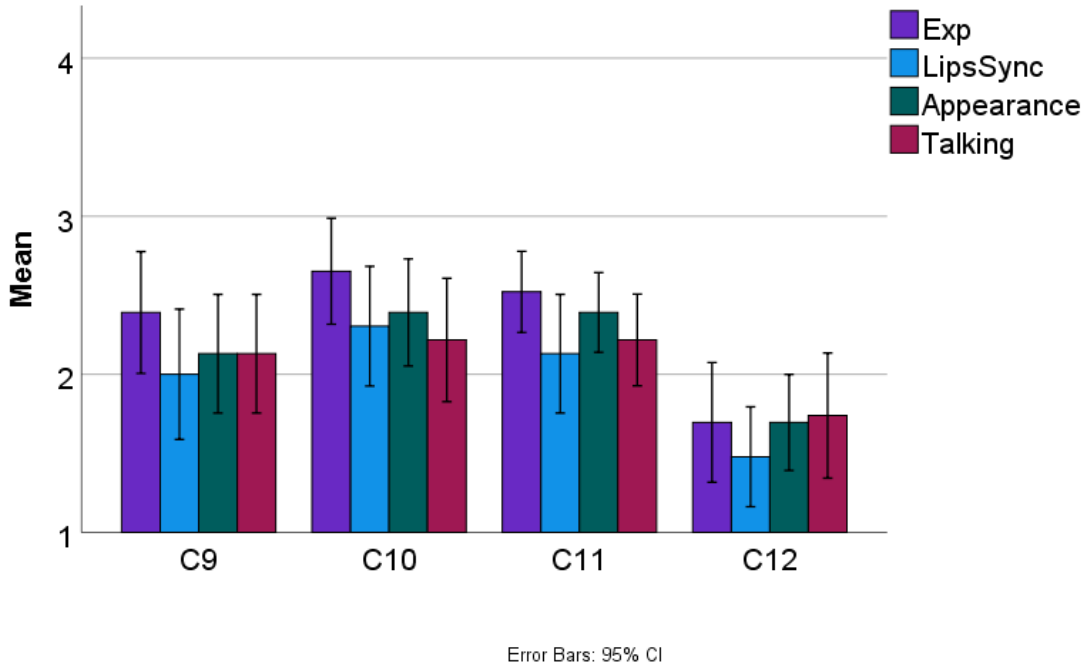


Figure 4.8 Results of comparison of UMA2 Avatars

Avatar ID	Mean	Std. Deviation
C9	2,39	.89
C10	2.65	.78
C11	2.52	.59
C12	1.70	.88

Table 4.4 Descriptive statistic table of UMA avatars overall experience

Avatar ID	Mean	Std. Deviation
C9	2,13	.87
C10	2.39	.78
C11	2.39	.58
C12	1.70	.70

Table 4.5 Descriptive statistic table of UMA avatars appearance

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.
Appearance	.515	13.753	5	.017

Table 4.6 Mauchly's test of sphericity on UMA avatars appearance

Source Factor	Type III sum of Squares	df	Mean square	F	Sig.
Sphericity Assumed	7,435	3	2,478	5,440	0.002

Table 4.7 Within-subjects effects on UMA avatars Appearance

Factor (I)	Factor (J)	Sig.
C9	C10	1.000
C9	C11	1.000
C9	C12	.088
C10	C11	1.000
C10	C12	.021
C11	C12	.006

Table 4.8 Pairwise comparisons of UMA avatars appearance

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.
experience	.625	9.729	5	.084

Table 4.9 Mauchly's test of sphericity on UMA avatars overall experience

Source Factor	Type III sum of Squares	df	Mean square	F	Sig.
Sphericity Assumed	12.554	3	4.185	7.047	.000

Table 4.10 Within-subjects effects on UMA avatars overall experience

Factor (I)	Factor (J)	Sig.
C9	C10	1.000
C9	C11	1.000
C9	C12	.006
C10	C11	1.000
C10	C12	.007
C11	C12	.014

Table 4.11 Pairwise Comparisons of UMA avatars overall experience

Figure 4.8 show the quality rating for the UMA animated avatars. To find the best avatar for UMA a ANOVA repeated measure was conducted for the mean quality aspects overall experience, and experience. Since all the avatars within UMA used the same lip sync and sound for talking lip sync quality and talking is not investigated in this section but will be done when comparing the best avatars from each section.

The results for comparing UMA appearance are presented in table 4.5, table 4.6 and table 4.7. A repeated measure ANOVA was performed to compare the effect of appearance on realistic avatar. There was a statistically significant difference in realistic avatar between at least two groups F(3.66) = 5.44, p = .002. The post-hoc pairwise comparing with Bonferroni correction is reported in table 4.8. Results show that avatar C9 is statistically significant different from C12 but is not statistically significant different from C10 and C11 in appearance experience. C10 is also statistically significant different from C12 in appearance experience and the same is true when comparing C11 and C12.

The results for comparing UMA overall experience are presented in table 4.4, table 4.9 and table 4.10. A repeated measure ANOVA was performed to compare the overall experience of the avatar. There was a statistically significant difference in overall experience between at least two groups s F(3.66) = 7.047, p = .000. The post-hoc pairwise comparing with Bonferroni correction is reported in table 4.11. Results show that C9 is statistically significant different from C12 but is not statistically significant different from C10 and C11 in overall experience. C10 is also statistically significant different from C12 in overall experience, and the same is true when comparing C11 and C12.

To summarize, based on the results we can see that there is a significant difference between the four avatars made with UMA for both overall experience and appearance. The post-hoc show that C12 is significantly worse than the rest of the avatars when it comes to appearance and the same is true for overall experience. Of the remaining three avatars the best UMA avatar is C10 and will be used when comparing the best avatars in section 4.4.4.

4.4.2 Ready Player Me created avatars

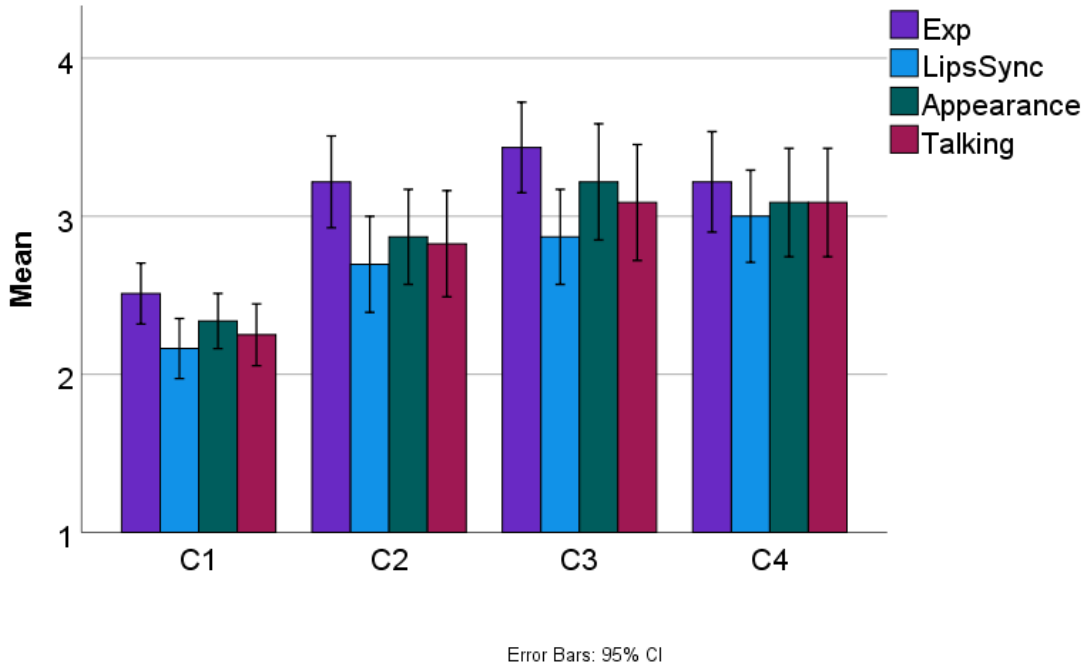


Figure 4.9 Comparison results of Ready Player Me Avatars

Avatar ID	Mean	Std. Deviation
C1	3.17	.83
C2	3.21	.67
C3	3.43	.66
C4	3.21	.74

Table 4.12 Descriptive statistic table of Ready Player Me avatars overall experience

Avatar ID	Mean	Std. Deviation
C1	2.87	.87
C2	2.87	.70
C3	3.22	.85
C4	3.09	.79

Table 4.13 Descriptive statistic table of Ready Player Me avatars appearance

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.
Appearance	.519	13.609	5	.018

Table 4.14 Mauchly's test of sphericity on Ready Player Me avatars appearance

Source Factor	Type III sum of Squares	df	Mean square	F	Sig.
Sphericity Assumed	2.033	3	.678	2.327	0.083

Table 4.15 Within-subjects effects on Ready Player Me avatars appearance

Factor (I)	Factor (J)	Sig.
C1	C2	1.000
C1	C3	.015
C1	C4	1.000
C2	C3	.256
C2	C4	.808
C3	C4	1.000

Table 4.16 Pairwise Comparisons of Ready Player Me avatars appearance

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.
Experience	.644	9.113	5	.105

Table 4.17 Mauchly's test of sphericity on Ready Player Me avatars overall experience

Source Factor	Type III sum of Squares	df	Mean square	F	Sig.
Sphericity Assumed	.957	3	.319	1.312	.278

Table 4.18 Within-subjects effects on Ready Player Me avatars overall experience

Factor (I)	Factor (J)	Sig.
C1	C2	1.000
C1	C3	.820
C1	C4	1.000
C2	C3	.577
C2	C4	1.000
C3	C4	.340

Table 4.19 Pairwise Comparisons of Ready Player Me avatars overall experience

Figure 4.9 show the quality rating for the Ready Player Me animated avatars. To find the best avatar for Ready Player Me an ANOVA repeated measure was conducted for the mean quality aspects overall experience, and experience. Since all the avatars within Ready Player Me used the same lip sync and sound for talking lip sync quality and talking is not

investigated in this section but will be done when comparing the best avatars from each section.

The results for comparing Ready Player Me appearance are presented in table 4.13, table 4.14 and table 4.15. A repeated measure ANOVA was performed to compare the effect of appearance on realistic avatar. There was a statistically significant difference in realistic avatar between at least two groups F(3.66) = 2.327 p = .083. The post-hoc pairwise comparing with Bonferroni correction is reported in table 4.16. Results show that avatar C1 is statistically significant different from C3 but is not statistically significant different from C2 and C4. C2 is statistically significant different from C3 and C4. C3 is however, not statistically significant different with C4.

The results for comparing Ready Player Me overall experience are presented in table 4.12, 4.17 and 4.18. A repeated measure ANOVA was performed to compare the overall experience of the avatar. There was a statistically significant difference in overall experience between at least two groups F(3.66) = 1.312, p = .278. The post-hoc pairwise comparing with Bonferroni correction is reported in table 4.19. Results show that C1 is statistically significant different from C3 but is not statistically significant different from C2 and C4. C2 is also statistically significant different from C3, and the same is true when comparing C3 to C4.

To summarize, based on the results we can see that there is a significant difference between the four avatars made with Ready Player Me for both overall experience and appearance. The post-hoc show that C3 is significantly better than C1 and C2 when it comes to appearance, but when it comes to overall experience C3 is significantly better than all the other three avatars. With these results C3 is the best avatar from Ready Player Me and will be used when comparing the best avatars in section 4.4.4.

4.4.3 MakeItTalk created avatars

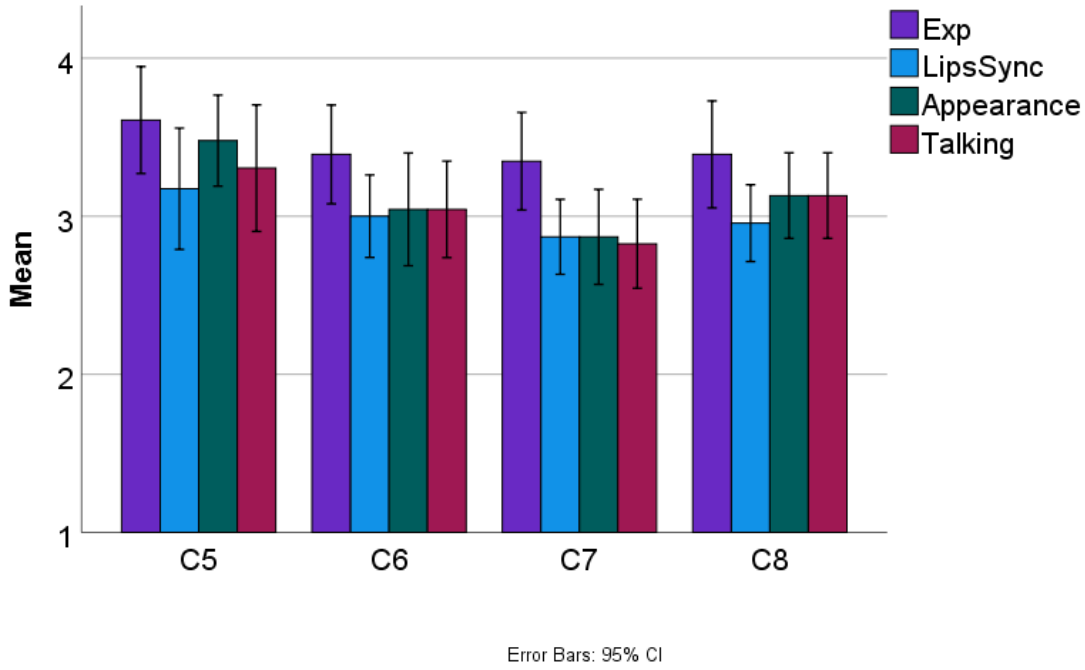


Figure 4.10 Comparison results of MakeItTalk avatars

Avatar ID	Mean	Std. Deviation
C5	3.60	.78
C6	3.39	.72
C7	3.35	.71
C8	3.39	.78

Table 4.20 Descriptive statistic table of MakeItTalk avatars overall experience

Avatar ID	Mean	Std. Deviation
C5	3.48	.67
C6	3.04	.82
C7	2.87	.69
C8	3.13	.63

Table 4.21 Descriptive statistic table of MakeItTalk avatars appearance

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.
Appearance	.959	.867	5	.973

Table 4.22 Mauchly's test of sphericity on MakeItTalk avatars appearance

Source Factor	Type III sum of Squares	df	Mean square	F	Sig.
Sphericity Assumed	4.522	3	1.507	3.620	.045

Table 4.23 Within-subjects effects on MakeItTalk avatars appearance

Factor (I)	Factor (J)	Sig.
C5	C6	.283
C5	C7	.013
C5	C8	.436
C6	C7	1.000
C6	C8	1.000
C7	C8	.971

Table 4.24 Pairwise Comparisons of MakeItTalk avatars appearance

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.
Experience	.670	8.309	5	.140

Table 4.25 Mauchly's test of sphericity on MakeItTalk avatars overall experience

Source Factor	Type III sum of Squares	df	Mean square	F	Sig.
Sphericity Assumed	.957	3	.319	.875	.459

Table 4.26 Within-subjects effects on MakeItTalk avatars overall experience

Factor (I)	Factor (J)	Sig.
C5	C6	1.000
C5	C7	.662
C5	C8	1.000
C6	C7	1.000
C6	C8	1.000
C7	C8	1.000

Table 4.27 Pairwise Comparisons of MakeItTalk avatars experience

Figure 4.10 show the quality rating for the MakeItTalk animated avatars. To find the best avatar for MakeItTalk an ANOVA repeated measure was conducted for the mean quality aspects overall experience, and experience. Since all the avatars within MakeItTalk used the same lip sync and sound for talking lip sync quality and talking is not investigated in this section but will be done when comparing the best avatars from each section.

The results for comparing MakeItTalk appearance are presented in table 4.21, 4.22 and 4.23. A repeated measure ANOVA was performed to compare the effect of appearance on realistic avatar. There was a statistically significant difference in realistic avatar between at least two groups F(1.22) = 4.520 p = .045. The post-hoc pairwise comparing with Bonferroni correction is reported in table 4.24. Results show that avatar C5 is statistically significant different from C6, C7 and C8. C6 is not statistically significant different from C7 and C8 and C7 is not statistically significant different with C8.

The results for comparing MakeItTalk overall experience are presented in table 4.20, 4.25 and 4.26. A repeated measure ANOVA was performed to compare the overall experience of the avatar. There was a statistically significant difference in overall experience between at least two groups F(3.66) = .875 p = .459. The post-hoc pairwise comparing with Bonferroni correction is reported in table 4.27. Results show that C5 is statistically significant different from C7 but is not statistically significant different from C6 and C8. C6 is not statistically significant different from any of the other avatars. C7 is not statistically significant different from C8.

To summarize, based on the results we can see that there is a significant difference between the four avatars made with MakeItTalk for both overall experience and appearance. The post-hoc show that C5 is significantly better than the other three when it comes to appearance, but not when it comes to overall experience. On overall experience C5 is not statistically better than C6 and C8 but is still the best avatar from MakeItTalk and will be used when comparing the best avatars in section 4.4.4.

4.4.4 Comparing best avatars

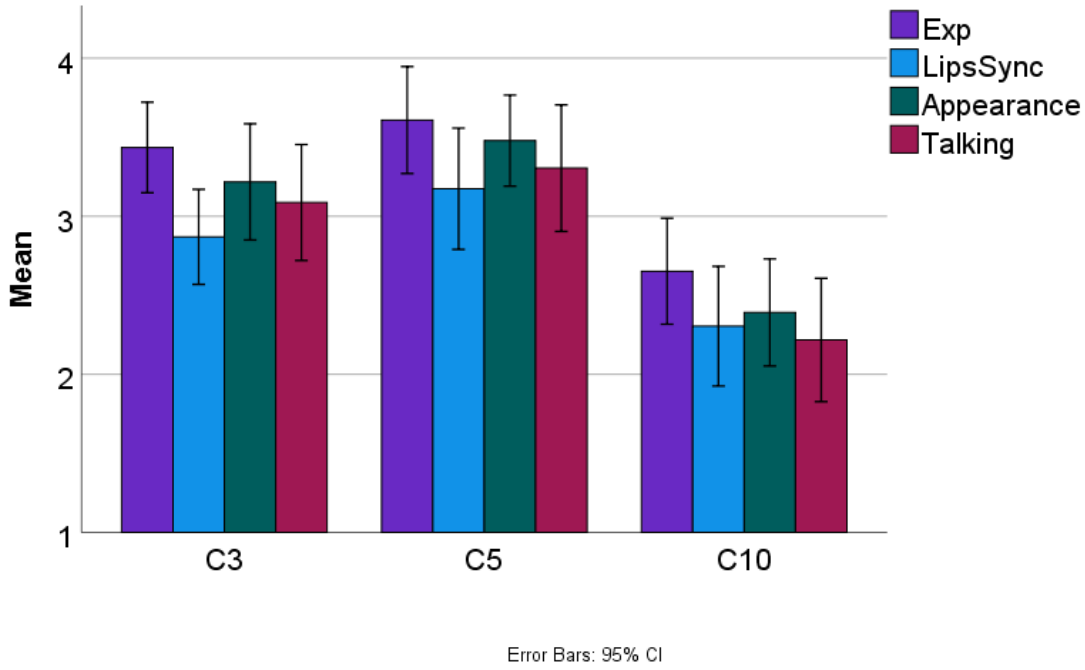


Figure 4.11 Comparison results of the best avatars from UMA2, Ready Player Me and MakeItTalk

Avatar ID	Mean	Std. Deviation
C3	3.43	.66
C5	3.61	.78
C10	2.65	.78

Table 4.28 Descriptive statistic table of when comparing the three best avatars overall experience

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.
Experience	.575	11.634	2	.003

Table 4.29 Mauchly's test of sphericity on three best avatars overall experience

Source Factor	Type III sum of Squares	df	Mean square	F	Sig.
Sphericity Assumed	.957	3	.319	.875	.000

Table 4.30 Within-subjects effects on three best avatars overall experience

Factor (I)	Factor (J)	Sig.
C3	C5	.486
C3	C10	.005
C5	C10	.002

Table 4.31 Pairwise Comparisons on three best avatars overall experience

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.
Talking experience	.791	4.936	2	.085

Table 4.32 Mauchly's test of sphericity on three best avatars talking experience

Source Factor	Type III sum of Squares	df	Mean square	F	Sig.
Sphericity Assumed	15.217	2	7.609	13.155	.000

Table 4.33 within-subjects effects on three best avatars talking experience

Factor (I)	Factor (J)	Sig.
C3	C5	.610
C3	C10	.007
C5	C10	.001

Table 4.34 Pairwise Comparisons on three best avatars talking experience

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.
Lip sync	.912	1.944	2	.378

Table 4.35 Mauchly's test of sphericity on three best avatars lip sync experience

Source Factor	Type III sum of Squares	df	Mean square	F	Sig.
Sphericity Assumed	8.957	2	4.478	8.311	.001

Table 4.36 Within-subjects effects on three best avatars talking lip sync experience

Factor (I)	Factor (J)	Sig.
C3	C5	.329
C3	C10	.060
C5	C10	.004

Table 4.37 Pairwise Comparisons on three best avatars lip sync experience

Figure 4.11 show the quality rating for comparing the three best avatars. To find the best avatar among C3, C5 and C10 an ANOVA repeated measure was conducted for the mean quality aspects overall experience, and experience. In this section lip sync and talking test results will also be reported.

The results for comparing top three avatars on overall experience are presented in table 4.28, 4.29 and 4.30. A repeated measure ANOVA was performed to compare the overall experience between the top three avatars. There was a statistically significant difference in overall experience between at least two groups $F(2.44) = 13.098$ $p = .000$. The

post-hoc pairwise comparing with Bonferroni corrections is reported in table 4.31. Results show that avatar C3 is statistically significant different from C5 and C10. C5 is also statistically significant different from C10.

The results for comparing top three avatars on talking experience are presented in table 4.32 and 4.33. A repeated measure ANOVA was performed to compare the overall talking experience between the top three avatars. There was a statistically significant difference in overall talking experience between at least two groups $F(2.44) = 13.155$ $p = .000$. The post-hoc pairwise comparing with Bonferroni corrections is reported in table 4.34. Results show that avatar C3 is statistically significant different from C5 and C10. C5 is also statistically significant different from C10.

The results for comparing top three avatars on lip sync experience are presented in table 4.35 and table 4.36. A repeated measure ANOVA was performed to compare the overall lip sync experience between the top three avatars. There was a statistically significant difference in overall lip sync experience between at least two groups $F(2.44) = 8.311$ $p = .001$. The post-hoc pairwise comparing with Bonferroni corrections is reported in table 4.37. Results show that avatar C3 is statistically significant different from C5 and C10. C5 is also statistically significant different from C10.

To summarize, based on the results we can see that there is a significant difference between the three methods UMA, Ready Player Me and MakeItTalk for overall experience, lip sync and talking. The post-hoc results show that talking for C10(UMA) is significantly lower than the other two, but there is no difference between C5(MakeItTalk) and C(Ready Player Me), this was also true for lip sync and overall experience and based on the results the best avatar is C5 (MakeItTalk).

4.4.5 Post-Questionnaire results

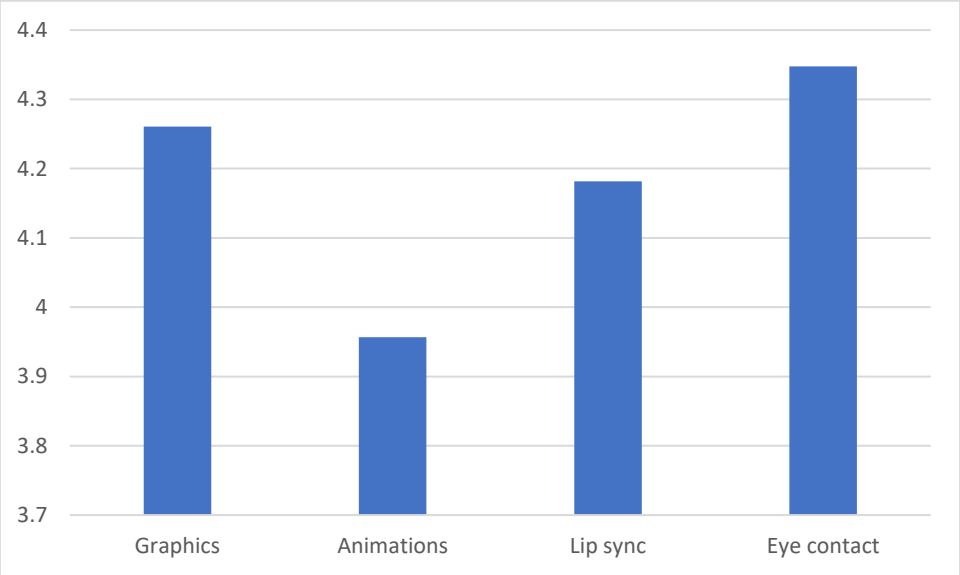


Figure 4.12 Results of post-questionnaire with the average answer from 1-5

The Post-test questionnaire was created to figure out what the participants found as the most important feature when creating a realistic avatar. The results shown in figure 4.12 show the importance of each question from 1 to 5 where 5 is very important. The results indicate that eye contact is more important than lip sync, graphics and animations when creating a realistic avatar. The least important feature was animations, this can be due to the questionnaire not having fully implemented animations for all avatars and the videos being zoomed in. Going forward this could indicate that focusing on eye contact is going to be important when creating future solutions.

4.5 Limitations

When creating a lip sync for avatars there are multiple other animations to take into consideration, among them are facial movement and body language. There were also limitations to how the test could be conducted, the first plan was to use Virtual Reality and evaluate the participants in a VR environment, but due to restrictions the early work was shifted towards video recorded avatars instead. The test was therefore done non-interactive and was instead a passive test. Other limitations are time constraint on a short thesis, there was therefore not time to create more than three working avatars in time for testing.

4.6 Discussion

When comparing the avatars, it became clear that the best avatar was created using artificial intelligence and GAN technologies. This could however be due to the more realistic appearance of the pictures used to create the avatars in MakeItTalk. However, there are limitations to MakeItTalk, it is unable to process the pictures and create a talking avatar in real time. So even though MakeItTalk had the best avatar it is going to be hard to use it when creating an avatar that should be able to interact with the user in real time. For real time talking the avatar created on Ready Player Me has the best results. Unfortunately, the plan to implement and test with a fourth solution using Unreal Engine and MetaHuman (Epic Games, 2022). MetaHuman has avatars using over 8k pixels and other implemented solutions to create lip sync, I will mention more about Unreal Engine and MetaHuman in future works.

When creating the avatars all four groups had individual lip-sync solutions, there could therefore be interesting to know if a UMA2 avatar could perform better on the test results with oculus lip-sync instead of Salsa. When looking at the results UMA2 and Salsa lip sync is almost one point lower than Ready Player Me and Oculus lip-sync.

In UMA2 and Ready Player Me there are one avatar in each comparison that has a statistically lower score, for the UMA avatar it can be due to the lip sync not working properly and the resolution might not process the lime green color that good making the shading and overall look a little rough. For the Ready Player Me avatar I find it interesting that the participants experience was so different when all four of the avatars is based on the same motion, lip sync and created from the same base template. The only thing that is different from two of the avatars is the female voice line. I however don't see how that can impact the results so drastically, as the voice is computer generated and should not perceive any sociolect differences.

When comparing the three best avatars the results indicate that MakeItTalk and Ready Player Me is superior to the UMA avatar, and MakeItTalk being slightly better than the best Ready player Me avatar. Taking the post-questionnaire results into account I believe that MakeItTalk and Ready Player Me both having better eye contact and graphics compared to UMA made an impact on how the participants experienced the different avatars. The

post-questionnaire ranked eye contact as a the most important part when creating a realistic avatar, this might be due to eye contacting being important for understanding and reading other people. Eye contact is one of the best non-verbal ways for humans to communicate and therefore it makes sense that when creating a realistic avatar some of the most prominent features of human communication should be included. As written in (Koike, 2019) eye contact activates limbic mirror system, this system makes your neurons and the person you have eye contact to sync up and fire the same neurons. So, when creating an avatar that should convey the user that they are sad or have trouble eye contact could enhance the experience.

Some unforeseen challenges when creating the avatars was the complexity of using Unity and learning everything from scratch. There were also some challenges when trying to expand the solutions with the use of Unreal Engine, but there my own computer did not have enough process power and the required graphical card was beyond my equipment. Due to changing the test from a virtual reality to video recorded avatars the need for animation was less important. The videos had to be similar so that neither avatar solution had a big edge, the background where therefore made similar to MakeItTalk which had a neutral background. A lot of the animation made for Ready Player Me was cut and the camera moved closer to only include the chest and up. This made most of the animation made for Ready Player Me useless and could be why MakeItTalk have an edge over Ready Player Me.

4.7 Summary

In this chapter the test design is presented, some general information of the participants and test conditions.

Based on the results we can see that there is a significant difference between the four avatars made with UMA, Ready Player Me and MakeItTalk for both overall experience and appearance.

Also based on the results we can see that there is a significant difference between the three methods UMA, Ready Player Me and MakeItTalk for overall experience, lip sync and talking. The post-hoc results show that talking for C10(UMA) is significantly lower than the other two, but there is no difference between C5(MakeItTalk) and C(Ready Player Me),

this was also true for lip sync and overall experience and based on the results the best overall avatar is C5 (MakeItTalk).

Post-test questionnaire indicates that when creating a realistic avatar, the focus should be more on eye contact than animations. It is, however, also important to have good graphics and lip sync.

5. Conclusion

5.1 Summary

Creating a realistic avatar that should be able to communicate with the users in real-time is not an easy task. In this thesis different techniques and technologies have been used to future develop and create avatars. During the thesis three types of lip sync solutions and three different appearances have been evaluated.

The thesis has gone into detail on how the design of the different avatars was planned and how the different technologies were used. Detailed information on how UMA2, Ready Player Me and MakeItTalk was used in the project and the difference between the lip sync solutions.

An evaluation was conducted to evaluate the different avatars against each other, and ANOVA repeated measure was conducted to find if any significant differences between the avatars. The results showed that there are significant differences between the three avatar solutions. The results indicate that artificially created avatar had the best overall avatar, this could be due to the graphics on the artificially created avatar being better, something that was supported by the post-questionnaire results having eye contact and graphics as the most important aspects of a realistic avatar. The results also indicate that going forward on should focus on eye contact, graphics, and lip sync more than animations and that going forward using artificially influenced avatars could create the best user experience.

5.2 Contributions

The research completed in the thesis had four objectives to try to find answer:

Objective 1: Get an overview of available methods for creating talking avatars.

The technologies used in this project are Unity and avatars created with Generative adversarial network. Mixmao was for animations and lip-sync Unity was created with Salsa and Ocolus.

Objective 2: Generate realistic talking avatars, with the identified methods from objective 1.

In Unity two different techniques were used to create avatars, first technique based on Unity Multipurpose Avatar 2 and second one based on Ready Player Me. The third avatar was created with MakeItTalk and GAN technologies. In the project a total of twelve different avatars was created, four based on Unity Multipurpose Avatar 2, four based on Ready Player Me and four based on MakeItTalk. Unity multipurpose avatar 2 and Ready Player Me used two different lip sync solutions and MakeItTalk used a third solution.

Objective 3: Compare the generated avatars and examine which one creates the best user experience.

After comparing the three avatars the overall best technology and technique used to create the best user experience based on the results was an artificially generated avatar created with MakeItTalk.

Objective 4: Compare the generated avatars and examine which one has the most realistic appearance.

After comparing the avatars, the overall most realistic avatar when it came to appearance was an artificially generated avatar created with MakeItTalk.

5.3 Future Work

A fourth solutions made with MetaHuman (Epic Games, 2022) and Unreal Engine was also planned to be implemented, but due to unforeseen challenges with both CPU, GPU and RAM Unreal Engine and MetaHuman was unfortunately not used in this thesis. For future solutions implementing an avatar created with MetaHuman would probably be the best free to use technology available. MetaHuman also have a support to use a new lip-sync animation handler called Acculips (Reallusion Inc. , 2022). Acculips works like MakeItTalk where the input is processed by Acculips and then animations are preset with facial expressions helping to enhance the lip sync experience. So instead of only the mouth moving the whole face with eyebrows, cheeks, nose, ears are working together to be as realistic as possible. This could however look like it should have the same issues that MakeItTalk had with not being compatible with real-time avatars, but there are other ways to use Acculips where you can have a camera tracking your facial expression in real-time and implementing to the MetaHuman avatar in Unreal Engine. This could be a place to start when creating a new avatar.

Going forward it could be important to keep in mind the results from the post-questionnaire that indicated that eye contact and graphics are important to create a realistic avatar. It could therefore also be interesting to find out more about what type of technologies can be implemented to increase eye contact with the user when inside a virtual environment. Graphics is something that can be solved using MetaHuman with the possibilities of up to 8k resolution something that is a graphic standard that is a lot higher than any of the solutions in this thesis.

6. References

Brownlee, J. (2019, july 12). *machinelearningmastery.* Retrieved from machinelearningmastery - 18 Impressive Applications of Generative Adversarial Networks (GANs): https://machinelearningmastery.com/impressive-applications-of-generative-adversarial-networks/

crazyD, c. (2022, February 23). *crazyminnowstudio What is SALSA LipSync?* Retrieved from crazyminnowstudio What is SALSA LipSync?: https://crazyminnowstudio.com/docs/salsa-lip-sync/modules/overview/

Deepfakes. (2019). *Faceswap.* Retrieved from Faceswap: https://github.com/deepfakes/faceswap

Elsa Thiaville, J.-M. N. (2020, December 15). *HAL.* Retrieved from HAL open science: https://hal.archives-ouvertes.fr/hal-03068405/document

Epic Games. (2022, May 15). *Unreal Engine*. Retrieved from Unreal Engine - MetaHuman: https://www.unrealengine.com/en-US/metahuman-creator

Freecodecamp. (2020, February 13). *freecodecamp - Unity Game Engine Guide: How to Get Started with the Most Popular Game Engine Out There.* Retrieved from freecodecamp: https://www.freecodecamp.org/news/unity-game-engine-guide-how-to-get-started-with-the-most-popular-game-engine-out-there/

Gunn Astrid Baugerud, M. S. (2021). *Multimodal Virtual Avatars for Investigative Interviews with Children.* New York, NY, USA: Association for Computing Machinery.

Hang Zhou, Y. S. (2021). Pose-controllable talking face generation by implicitly modularized audio-visual representation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 4176–4186).

Hjetland, G. B. (2021, October 10). *NRK.* Retrieved from NRK - Norsk politi seier nei til banebrytande overgreps-forsking: https://www.nrk.no/norge/norsk-politi-seier-nei-til-banebrytande-overgreps-forsking-1.15681084

Johansson, D. (2015, Spring). Retrieved from http://www.diva-portal.org/smash/get/diva2:787845/FULLTEXT01.pdf

Koike, T. S. (2019, February 28). *What Makes Eye Contact Special? Neural Substrates of On-Line Mutual Eye-Gaze: A Hyperscanning fMRI Study.* https://doi.org/10.1523/ENEURO.0284-18.2019. Retrieved from https://doi.org/10.1523/ENEURO.0284-18.2019

Mixamo. (2022, May 5). *Mixamo part of Adobe Family*. Retrieved from Mixamo - Get animated: https://www.mixamo.com/

Oslomet . (2022, May 1). *Oslomet*. Retrieved from Oslomet - Research - Projects: https://www.oslomet.no/en/research/research-projects/professionals-interviewing-maltreated-children-supported-artificial-avatars

Ready Player Me. (2022, May 04). *Ready Player Me - SDK*. Retrieved from Ready Player Me - SDK: https://docs.readyplayer.me/ready-player-me/integration-guides/unity

Ready Player Me. (n.d.). *Ready Player Me*. Retrieved from Ready Player Me - Create avatar: https://readyplayer.me/hub

Reallusion Inc. . (2022, May 15). *Reallusion* . Retrieved from Reallusion - Lipsync animation: https://www.reallusion.com/iclone/lipsync-animation.html

Rithesh Kumar, J. S. (2018). *ObamaNet: Photo-realistic lip-sync from tex.* Cornell University.

Simulamet. (n.d.). *Proposal Avatar 2.0.*

Tero Karras, S. L. (2019). A style-based generator architecture for generative adversarial networks. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 4401–4410).

Tero Karras, S. L. (2020). Analyzing and improving the image quality of stylegan. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 8110–811).

UMA Steering Group. (2021, April 19). *Unity Asset Store*. Retrieved from Unity Asset Store - UMA 2 - Unity Multipurpose Avatar: https://assetstore.unity.com/packages/3d/characters/uma-2-unity-multipurpose-avatar-35611

Unity - oculus. (2021, May 24). *Oculus Lipsync Unity*. Retrieved from Oculus Lipsync Unity - Oculus Lipsync Unity Integration v29: https://developer.oculus.com/downloads/package/oculus-lipsync-unity/

Virtualspeech. (2022, March 13). *virtualspeech*. Retrieved from virtualspeech - Job Interview Preparation: https://virtualspeech.com/courses/interview-vr

WHO. (2013). *WHO*. Retrieved from World Health Organization: WHO.com

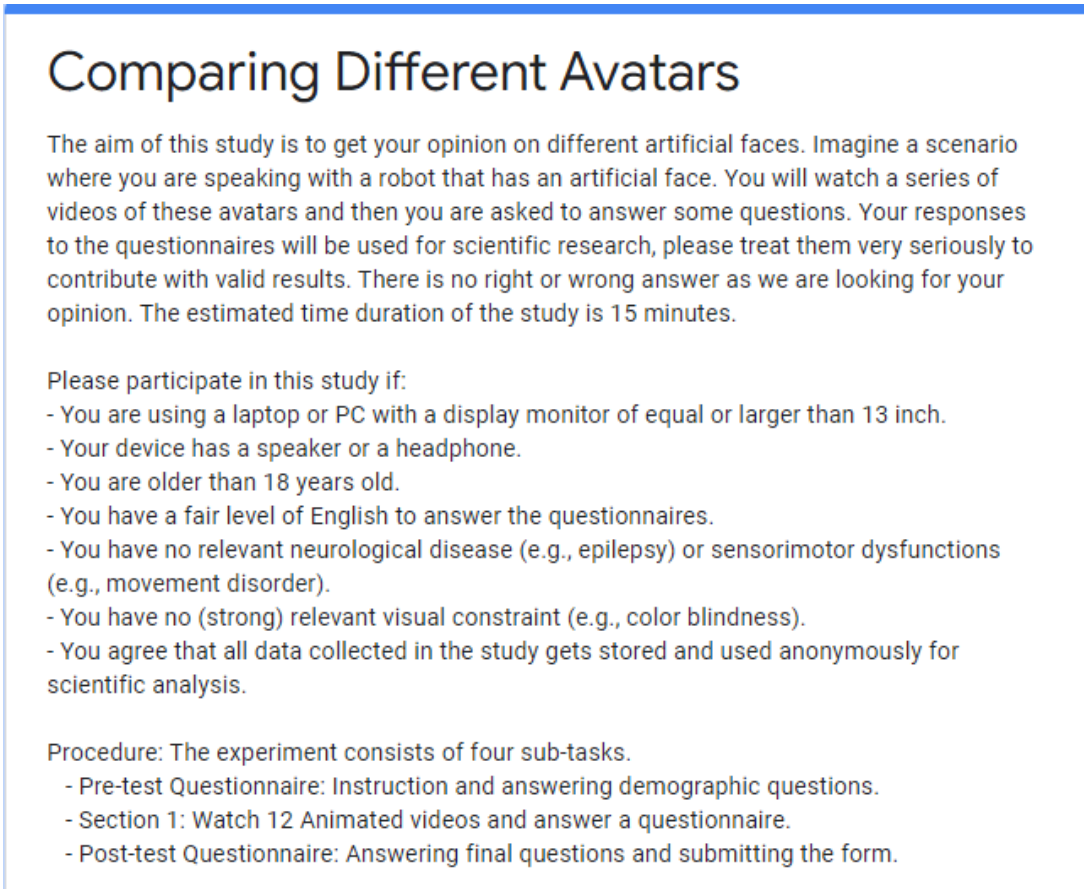
Wikipedia . (2022, April 24). *Wikipedia - Generative adversarial network.* Retrieved from Wikipedia - GAN: https://en.wikipedia.org/wiki/Generative_adversarial_network

Zisserman, J. S. (2018). VoxCeleb2: Deep Speaker Recognition. *Proc. Interspeech 2019*, (pp. 1086-1090).

Appendices A

A.1 Google Forms Questionnaire

In this appendix the full questionnaire used will be presented with screenshots:



Comparing Different Avatars

The aim of this study is to get your opinion on different artificial faces. Imagine a scenario where you are speaking with a robot that has an artificial face. You will watch a series of videos of these avatars and then you are asked to answer some questions. Your responses to the questionnaires will be used for scientific research, please treat them very seriously to contribute with valid results. There is no right or wrong answer as we are looking for your opinion. The estimated time duration of the study is 15 minutes.

Please participate in this study if:
- You are using a laptop or PC with a display monitor of equal or larger than 13 inch.
- Your device has a speaker or a headphone.
- You are older than 18 years old.
- You have a fair level of English to answer the questionnaires.
- You have no relevant neurological disease (e.g., epilepsy) or sensorimotor dysfunctions (e.g., movement disorder).
- You have no (strong) relevant visual constraint (e.g., color blindness).
- You agree that all data collected in the study gets stored and used anonymously for scientific analysis.

Procedure: The experiment consists of four sub-tasks.
- Pre-test Questionnaire: Instruction and answering demographic questions.
- Section 1: Watch 12 Animated videos and answer a questionnaire.
- Post-test Questionnaire: Answering final questions and submitting the form.

Figure A.1 Pre-questionnaire information from the questionnaire used in the thesis

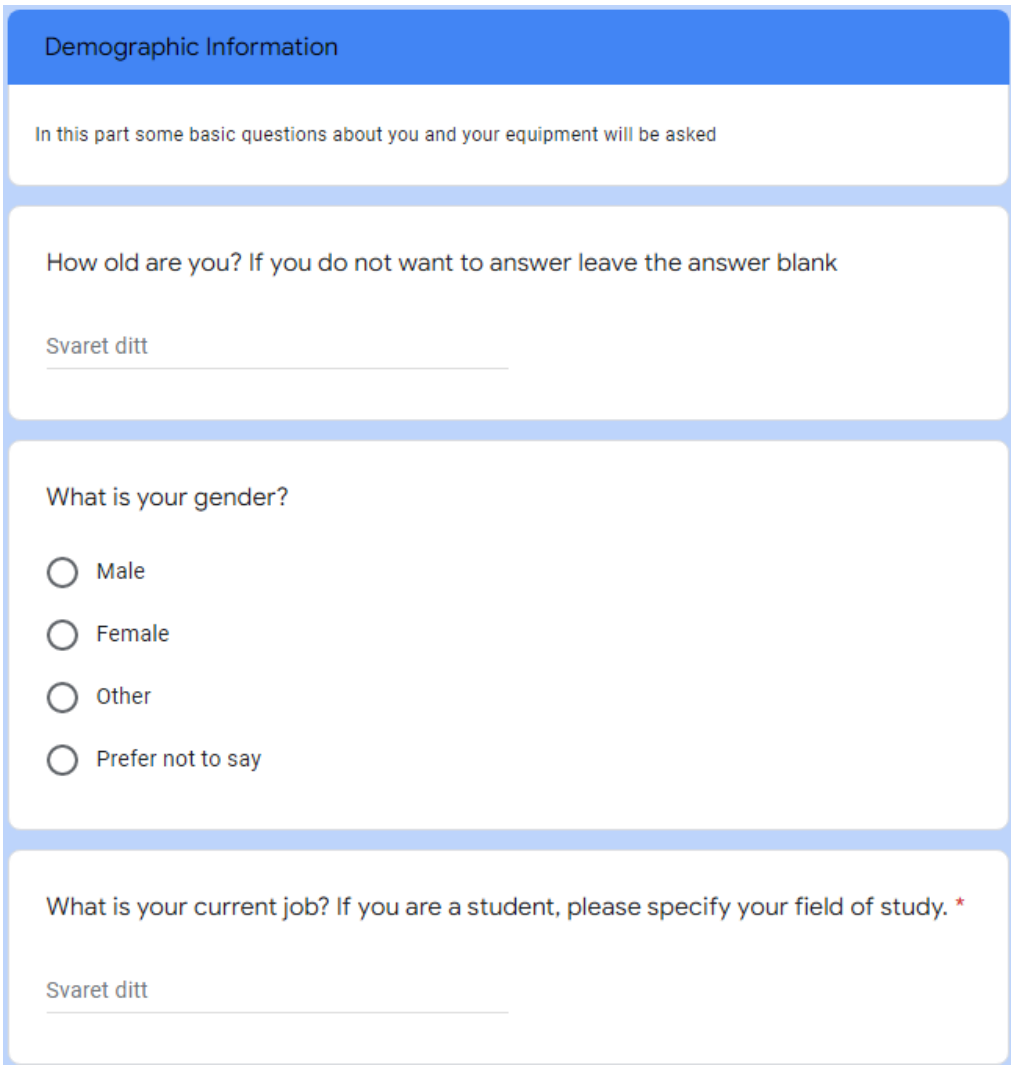


Figure A.2 First part of the demographic questions used in the questionnaire for the thesis

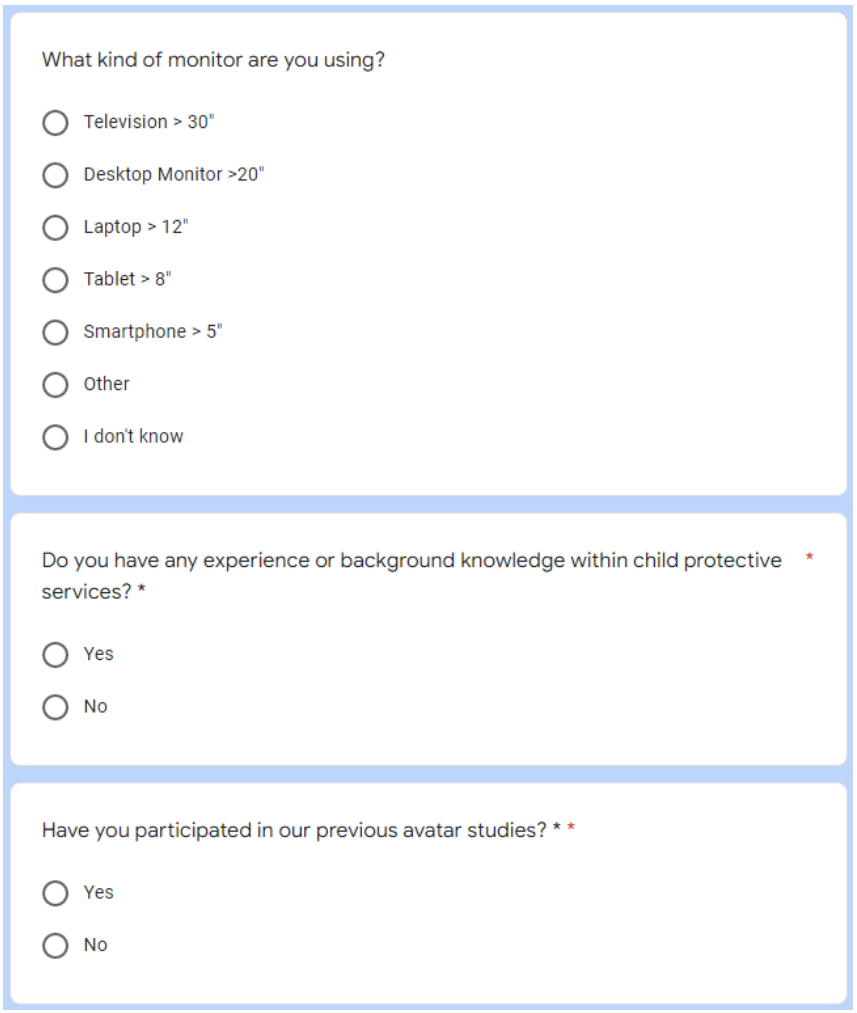


Figure A.3 Second part of the demographic questions used in the questionnaire for the thesis.

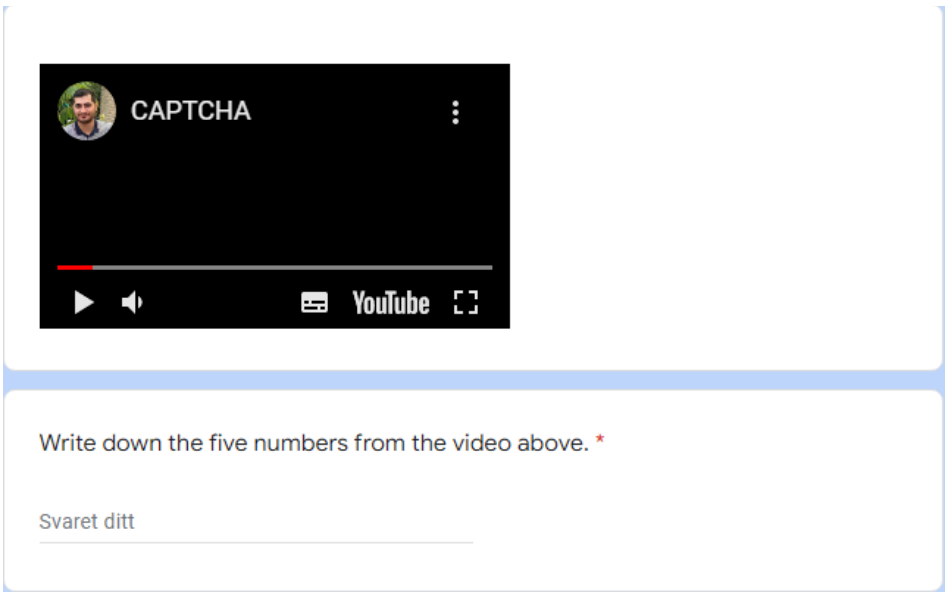


Figure A.4 Third and last part of the demographic questions used in the questionnaire for the thesis

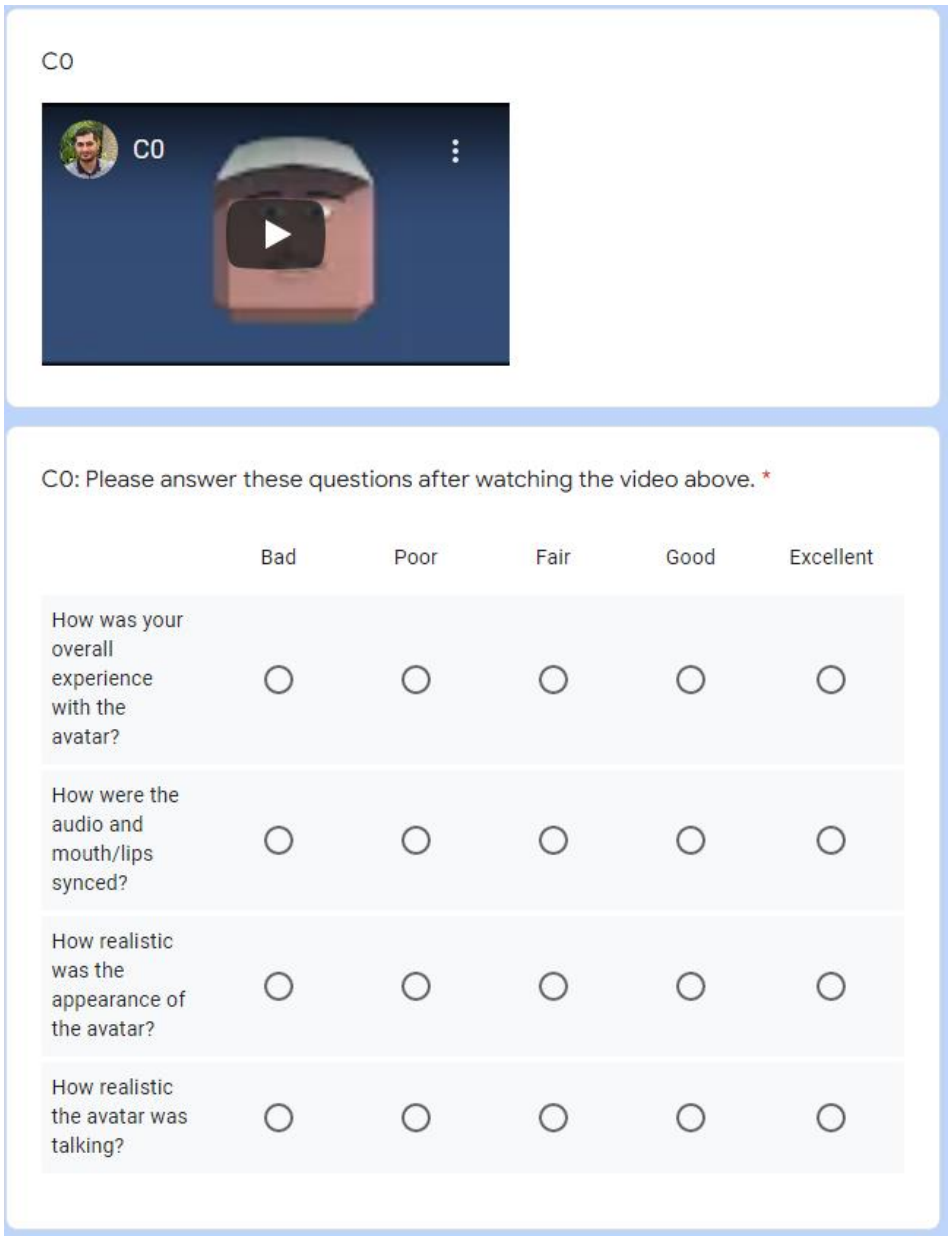


Figure A.5 First video in the third part of the questionnaire, this is C0 that was used as an anchor rated from bad to excellent on four different questions.

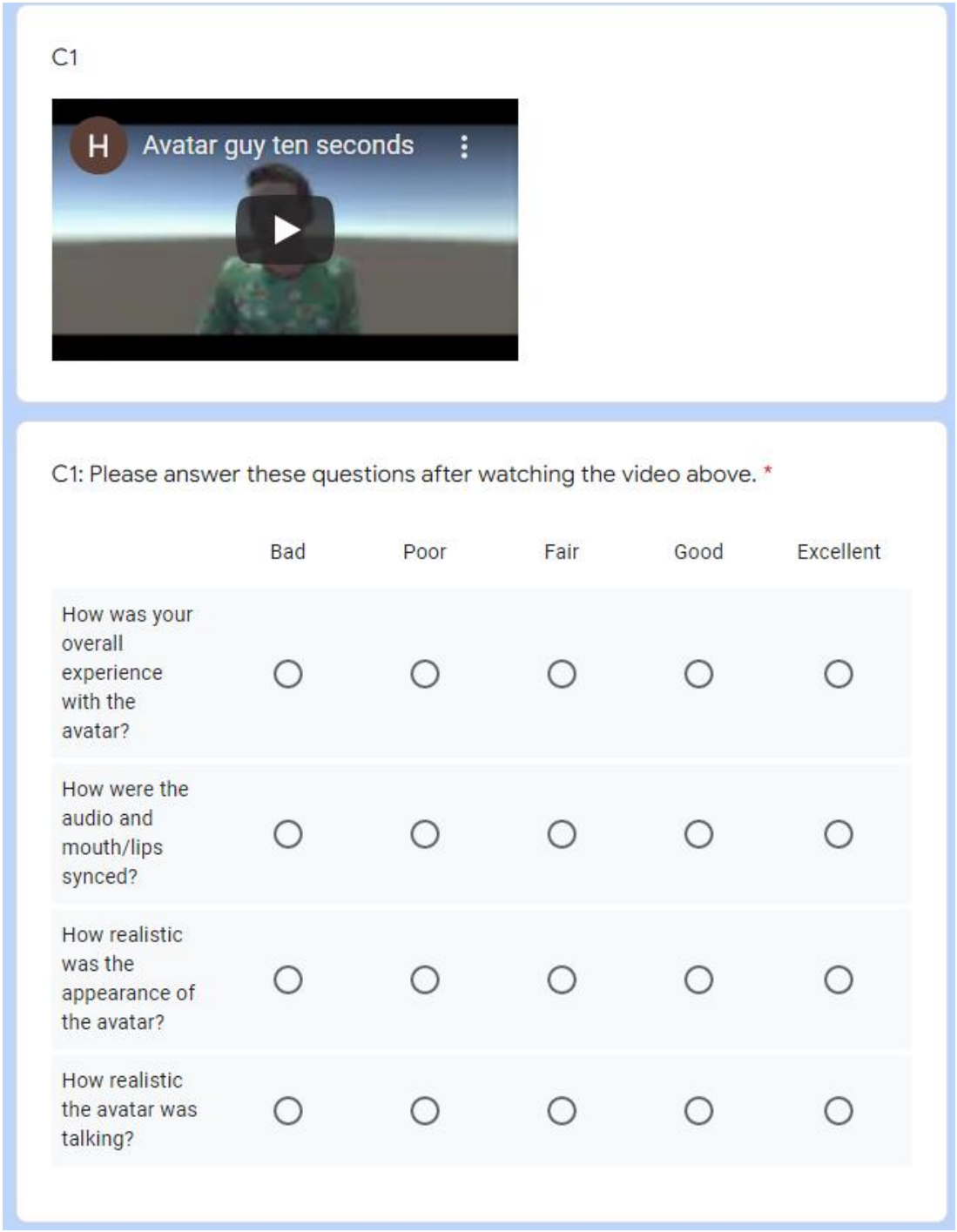


Figure A.6 First Ready Player Me video in the third part of the questionnaire, this is avatar C1 rated from bad to excellent on four different questions.

C2

C2: Please answer these questions after watching the video above. *

	Bad	Poor	Fair	Good	Excellent
How was your overall experience with the avatar?	○	○	○	○	○
How were the audio and mouth/lips synced?	○	○	○	○	○
How realistic was the appearance of the avatar?	○	○	○	○	○
How realistic the avatar was talking?	○	○	○	○	○

Figure A.7 Second Ready Player Me video in the third part of the questionnaire, this is avatar C2 rated from bad to excellent on four different questions.

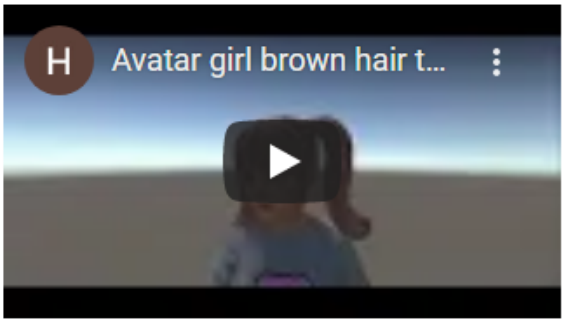
C3

C3: Please answer these questions after watching the video above. *

	Bad	Poor	Fair	Good	Excellent
How was your overall experience with the avatar?	○	○	○	○	○
How were the audio and mouth/lips synced?	○	○	○	○	○
How realistic was the appearance of the avatar?	○	○	○	○	○
How realistic the avatar was talking?	○	○	○	○	○

Figure A.8 Third Ready Player Me video in the third part of the questionnaire, this is avatar C3 rated from bad to excellent on four different questions.

C4



C4: Please answer these questions after watching the video above. *

	Bad	Poor	Fair	Good	Excellent
How was your overall experience with the avatar?	○	○	○	○	○
How were the audio and mouth/lips synced?	○	○	○	○	○
How realistic was the appearance of the avatar?	○	○	○	○	○
How realistic the avatar was talking?	○	○	○	○	○

Figure A.9 Fourth Ready Player Me video in the third part of the questionnaire, this is avatar C4 rated from bad to excellent on four different questions.

C5

C5: Please answer these questions after watching the video above. *

	Bad	Poor	Fair	Good	Excellent
How was your overall experience with the avatar?	○	○	○	○	○
How were the audio and mouth/lips synced?	○	○	○	○	○
How realistic was the appearance of the avatar?	○	○	○	○	○
How realistic the avatar was talking?	○	○	○	○	○

Figure A.10 First MakeItTalk video in the third part of the questionnaire, this is avatar C5 rated from bad to excellent on four different questions.

C6

C6: Please answer these questions after watching the video above. *

	Bad	Poor	Fair	Good	Excellent
How was your overall experience with the avatar?	○	○	○	○	○
How were the audio and mouth/lips synced?	○	○	○	○	○
How realistic was the appearance of the avatar?	○	○	○	○	○
How realistic the avatar was talking?	○	○	○	○	○

Figure A.11 Second MakeItTalk video in the third part of the questionnaire, this is avatar C6 rated from bad to excellent on four different questions.

C7

C7: Please answer these questions after watching the video above. *

	Bad	Poor	Fair	Good	Excellent
How was your overall experience with the avatar?	○	○	○	○	○
How were the audio and mouth/lips synced?	○	○	○	○	○
How realistic was the appearance of the avatar?	○	○	○	○	○
How realistic the avatar was talking?	○	○	○	○	○

Figure A.12 Third MakeItTalk video in the third part of the questionnaire, this is avatar C7 rated from bad to excellent on four different questions.

C8

C8: Please answer these questions after watching the video above. *

	Bad	Poor	Fair	Good	Excellent
How was your overall experience with the avatar?	○	○	○	○	○
How were the audio and mouth/lips synced?	○	○	○	○	○
How realistic was the appearance of the avatar?	○	○	○	○	○
How realistic the avatar was talking?	○	○	○	○	○

Figure A.13 Fourth MakeItTalk video in the third part of the questionnaire, this is avatar C8 rated from bad to excellent on four different questions.

C9.

C9: Please answer these questions after watching the video above. *

	Bad	Poor	Fair	Good	Excellent
How was your overall experience with the avatar?	○	○	○	○	○
How were the audio and mouth/lips synced?	○	○	○	○	○
How realistic was the appearance of the avatar?	○	○	○	○	○
How realistic the avatar was talking?	○	○	○	○	○

Figure A.14 First UMA2 video in the third part of the questionnaire, this is avatar C9 rated from bad to excellent on four different questions.

C10

C10: Please answer these questions after watching the video above. *

	Bad	Poor	Fair	Good	Excellent
How was your overall experience with the avatar?	○	○	○	○	○
How were the audio and mouth/lips synced?	○	○	○	○	○
How realistic was the appearance of the avatar?	○	○	○	○	○
How realistic the avatar was talking?	○	○	○	○	○

Figure A.15 Second UMA2 video in the third part of the questionnaire, this is avatar C10 rated from bad to excellent on four different questions.

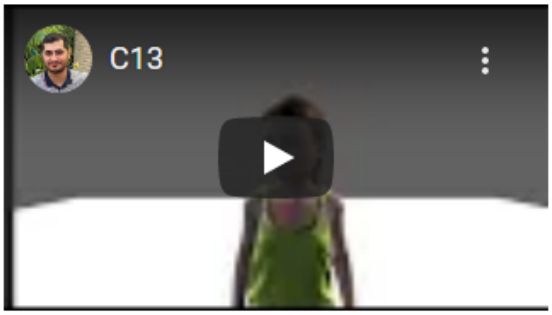
C11

C11: Please answer these questions after watching the video above. *

	Bad	Poor	Fair	Good	Excellent
How was your overall experience with the avatar?	○	○	○	○	○
How were the audio and mouth/lips synced?	○	○	○	○	○
How realistic was the appearance of the avatar?	○	○	○	○	○
How realistic the avatar was talking?	○	○	○	○	○

Figure A.16 Third UMA2 video in the third part of the questionnaire, this is avatar C11 rated from bad to excellent on four different questions.

C12



C12: Please answer these questions after watching the video above. *

	Bad	Poor	Fair	Good	Excellent
How was your overall experience with the avatar?	○	○	○	○	○
How were the audio and mouth/lips synced?	○	○	○	○	○
How realistic was the appearance of the avatar?	○	○	○	○	○
How realistic the avatar was talking?	○	○	○	○	○

Figure A.17 Fourth UMA2 video in the third part of the questionnaire, this is avatar C12 rated from bad to excellent on four different questions.

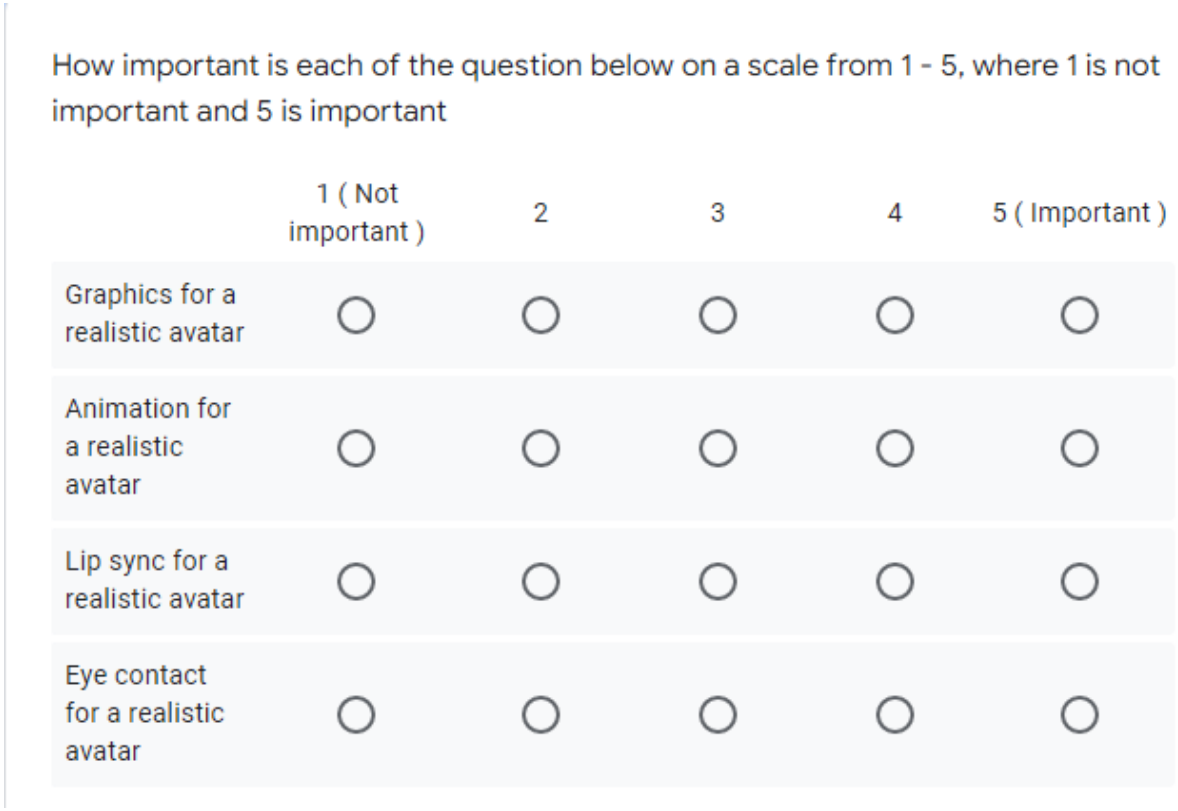


Figure A.18 Post-questionnaire questions, where the participants were asked to rate each question from 1-5