FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Assignment of bachelor's thesis

Title:	Digital Persona
Student:	Lukáš Marek
Supervisor:	Ing. Jan Šedivý, CSc.
Study program:	Informatics
Branch / specialization:	Web and Software Engineering, specialization Computer Graphics
Department:	Department of Software Engineering
Validity:	until the end of summer semester 2022/2023

Instructions

Use the standard software engineering modeling techniques to design a 3D conversational photorealistic persona conversing about social topics. Follow these steps: Survey and select a conversational development platform. Second, analyze the graphical development tools and choose the most convenient ones for creating a conversational persona. Third, design a 3D person manually or use a selected character generator. Fourth, implement a conversation about at least three social topics. Assume the application will run on a high-end gaming PC with a dedicated graphics card. Finally, test the developed application.

Electronically approved by Ing. Radek Richtr, Ph.D. on 12 January 2022 in Prague.

Bachelor's thesis

DIGITAL PERSONA

Lukáš Marek

Faculty of Information Technology
Department of Software Engineering
Supervisor: Ing. Jan Šedivý, CSc.
May 11, 2023

Contents

List of Figures

List of Tables

List of code listings

I'd like to start with expressing my deep thanks to my supervisor, Ing. Jan Šedivý, CSc. Not only did he come up with the idea for this thesis, but he also recognized my potential already during my high school years and urged me to pursue higher education, something I had been previously refusing. His faith in me has made a big difference in my life and brought me all the way up to this significant milestone. Next, I must mention Mgr. Bc. Barbora Šipošová, Ph.D., who's been an incredible support. Always open to my ideas, proactive, and kind, she's been a big help in enhancing my work. Thanks to Mgr. Petr Novák, Ph.D., as well. He provided advice and expertise that aided to my understanding and application of statistical methods in this thesis. I can't forget to thank R.U.R. Postproduction and its CTO, Michal Mocňak, for letting me use their infrastructure and lending me a second laptop which allowed for concurrent testing of two participants, significantly boosting my research. Furthermore I'm grateful to PromethistAI for sponsoring my research and supporting my academic journey. Lastly, I want to thank my amazing data gathering squadron: Alena Línková, Kateřina Melicharová, Lucie Kučerová, and Kateřina Bašová. To everyone who's helped me on this journey, thank you. Your contributions have made this thesis possible.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Czech Technical University in Prague has the right to conclude a licence agreement on the utilization of this thesis as a school work pursuant of Section 60 (1) of the Act.

In Prague on May 11, 2023

Abstract

This thesis explores the potential of digital personas, virtual embodied talking beings combining computer graphics with conversational artificial intelligence. Conversational platform called Flowstorm and a character generator, Metahuman Creator, were employed in development of showcase application in which users can converse with digital personas about various social topics. Developed underlying solution powering this Windows application takes the form of an adaptable Unreal Engine plugin that can power a variety of future projects of different needs. Thesis additionally demonstrates how multi-disciplinary approaches can make the testing phase more interesting and efficient and also illustrates the role of digital personas in the research domain, as digital personas can offload tasks from human researchers. A pilot psychological study with 51 participants was conducted during usability testing to examine how secondary communication channels, such as colors, lights, and sound can affect user's perception of digital personas during conversations with them and how can these secondary communication channels be utilized intentionally. Furthermore the thesis discusses received feedback and offers insights into the development and possible future research.

Keywords application, Microsoft Windows, plugin, Unreal Engine, digital humans, conversational AI, color psychology, music psychology, emotional induction

Abstrakt

Tato práce zkoumá potenciál digitálních person, mluvících postav kombinujících počítačovou grafiku s konverzační umělou inteligencí. Konverzační platforma Flowstorm a generátor postav, Metahuman Creator, byly použity při vývoji ukázkové aplikace, v níž uživatelé mohou konverzovat s digitálními personami o různých společenských tématech. Vyvinuté řešení, které tuto Windows aplikaci pohání, má formu adjustabilního Unreal Engine pluginu, který může pohánět širokou škálu budoucích projektů různých potřeb. Práce navíc ukazuje, jak lze multidisciplinárními přístupy udělat fázi testování zajímavější a efektivnější, přičemž ilustruje roli digitálních person v oblasti výzkumu, neboť digitální persony na sebe mohou převzít úkoly lidských výzkumníků. V rámci testování použitelnosti byla provedena pilotní psychologická studie s 51 participanty. Účelem bylo zjistit, jak sekundární komunikační kanály, jako jsou barvy, světla a zvuky, mohou ovlivnit uživatelovo vnímání digitálních person během konverzací s nimi, a jak lze tyto kanály využít. Práce dále diskutuje získanou zpětnou vazbu a nabízí vhled do vývoje a možného budoucího výzkumu.

Klíčová slova aplikace, Microsoft Windows, plugin, Unreal Engine, digitální lidé, konverzační AI, psychologie barev, psychologie hudby, emoční indukce

List of abbreviations

AI	Artificial intelligence
API	Application programming interface
ASR	Automatic speech recognition
AWS	Amazon Web Services
DAW	Digital audio workstation
GUI	Graphical user interface
HUD	Heads-up display
IK	Inverse kinematics
LFS	Large file storage
LOD	Level of detail
MEL	Maya Embedded Language
MIDI	Musical Instrument Digital Interface
ML	Machine learning
NLP	Natural language processing
NLG	Natural language generation
NPC	Non-player character
NLU	Natural language understanding
SAM	Self-assessment manikin
SSML	Speech Synthesis Markup Language
STS	Speech-to-speech
STT	Speech-to-text
TTS	Text-to-speech

Introduction

Digital personas, which are to simply put it, a virtual embodied talking beings, have become an increasingly important topic. Although conversational designers have been crafting gamebooks and dialogues for video games for quite some time, interactive conversational systems have experienced limited evolution over the decades. The growing interest in the metaverse and the development of advanced conversational AI, such as ChatGPT, highlight the relevance of digital personas in contemporary discussions.

The relevance of digital personas in the context of the metaverse is underscored by both Gartner's Hype Cycle illustrated in the accompanying figure 1 and an announcement from Mark Zuckerberg, founder and CEO of Meta, who after silently burying the idea of metaverse posted on Facebook "*we'll focus on developing AI personas*" [1]. Success of the metaverse is heavily dependent on the implementation of digital personas as supported by Jonathan Goodman, a metaverse specialist and a long-term developer of one called Second Life. He believes that these virtual worlds cannot solely rely on human users and that digital personas are crucial for their existence. [2] Metaverse has experienced a surge of hype followed by a "Trough of Disillusionment", during which many have recently abandoned the concept. Ondřej Dobruský, former CEO of Victoria VR, has asserted that the creation of the metaverse is not currently feasible, citing the need for advanced AI in various fields. However, once such AI is available, he paradoxically predicts that the metaverse will be established very quickly. [3] This emphasizes the need for further research and development in the field of digital personas to progress towards the "Plateau of Productivity", ultimately making the technology as usable and effective as previously hyped. Artur Sychov, the founder and CEO of the metaverse Somnium Space, predicts that the era of usability of the metaverse will arrive in the 2030s. [4]

While the previous paragraphs focus on the metaverse, it is important to note that the ideas presented in this thesis also apply to various other virtual worlds, encompassing those aimed at social interactions, as well as those created for research purposes in different fields. For example, in scenarios such as studying user behavior in a virtual reconstruction of ancient Greece, it would be unrealistic to have empty and silent streets. Intriguingly, digital personas can serve as researchers or more of like data gatherers themselves, contributing to research efforts. That is actually demonstrated in the testing phase of this thesis.

To put it differently, specifically 3D virtual worlds, often visually pleasing, currently lack robust conversational capabilities, while in contrast, conversational interfaces, that posses robust conversational capabilities typically do not incorporate visual elements at all, usually sticking only to text based interactions within chat windows. That is another reason why digital personas, combining both are important to develop. The specific goals of this thesis will be covered in the next section.

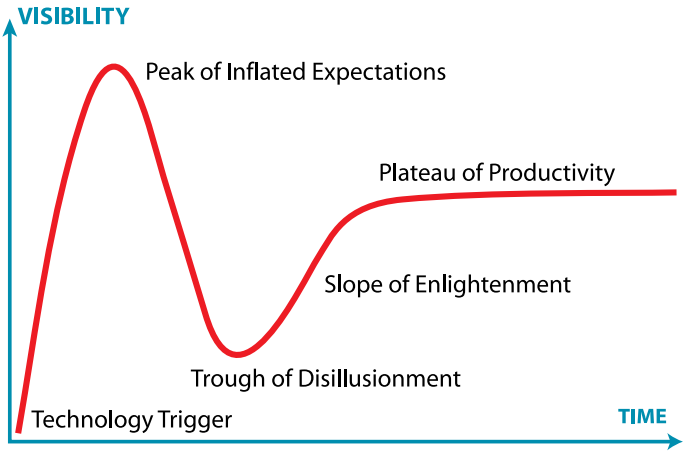


Figure 1 Gartner hype cycle [5]

In this thesis, the terms "chatbot", "voice assistant" and other similar entities are collectively referred to as "conversational interface". The purpose of using the term "conversational interface" is to encompass a broader range of technologies that facilitate communication serving various purposes between humans and computers through natural language, whether it be text-based or voice-based interactions.

0.1 Goals

This thesis primarily focuses on conversational interfaces in conjunction with computer graphics. It does not delve into the intricacies of AI itself. The goals include:

1. Creating a demo application to showcase a digital persona.
2. Developing an underlying solution utilizing a conversational platform powering this digital persona, which will be both usable and extendable for future projects. This solution is likely to take the form of a plugin.
3. Demonstrating the role of digital personas in the research domain, and exploring how the typically unpopular testing phase can be made more interesting and efficient through multidisciplinarity.

These goals aim to advance the understanding and development of digital personas, exploring their potential in various applications and contexts.

Chapter 1

3D character creation

This chapter aims to provide an understanding of the process involved in creating digital persona's bodies through the use of 3D computer graphics. An alternative approach is mentioned in Discussion.

In the realm of 3D modeling, models are often considered as individual objects, such as a pencil or a tree, while a scene represents an assembly of these objects into a complete 3D environment. This perspective mirrors the prevalent method for constructing a 3D scene, which involves building multiple models and assembling them subsequently. Each model comprises two distinct descriptions... a mathematical representation of the shape's structure and a recipe for determining the appearance of the shape under specific lighting conditions. [6]

Dr. Norman Badler from the University of Pennsylvania highlights a variety of purposes for 3D models in his article about "3D Object Modeling". [6] These purposes encompass visualizing items prior to machining, estimating costs, calculating volume and area, and determining machining durations. Additionally, 3D models are used to exercise display algorithms. Within the scope of this thesis, several other purposes he mentioned are of particular interest. These include examining the relationships between objects to understand how they fit together. Evaluating the faithfulness to real-world phenomena by analyzing how an object's surface interacts with light, as well as gauging the level of realism achievable in the model's appearance. Artistic purposes, identifying the ideal blend of real and imaginary elements to effectively convey an artist's vision and mood, while also examining the extent to which imagination can be harnessed to visualize artistic concepts. [6]

It is essential to differentiate between models and rendering in the context of 3D modeling. Models describe the object and its attributes, including shape, geometry, color, reflectivity, transmittance and more. On the other hand, the rendering algorithm is responsible for displaying and transforming the model into a screen-based view from a specific camera position. [6]

Three primary types of 3D model representations can be distinguished: analytical, polygonal, and volumetric. Analytical representation involves boundary-represented objects and relies on curves and surfaces, while polygonal representation is derived from analytically obtained surfaces (polygons, figure 1.1) and is frequently used in time-sensitive applications such as games. Volumetric representation is often obtained through measurement and is widely employed for medical purposes. However, for the purposes of this bachelor thesis, the focus will be on polygonal representation, as it is the most relevant to the discussed topic. [7]

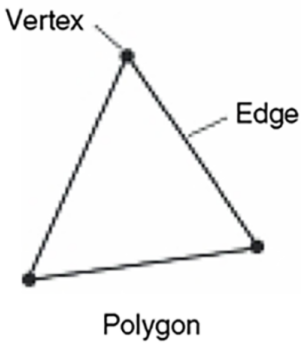


Figure 1.1 Polygon... consisting of edges and vertices

1.1 3D modeling

A variety of programs, commonly referred to as modelers, are available to assist in creating 3D models. The users of these programs are sometimes called modelers as well. These tools are discussed in the analysis chapter under the section titled 3D Tools.

When it comes to object modeling, several key issues need to be considered, such as the computational cost of the model, its effectiveness in representing the desired phenomena, and the methods utilized to obtain or create data describing the object's geometry. [6]

Two main types of approaches to 3D modeling can be identified: hard surface and organic, each tailored to different types of objects and purposes.

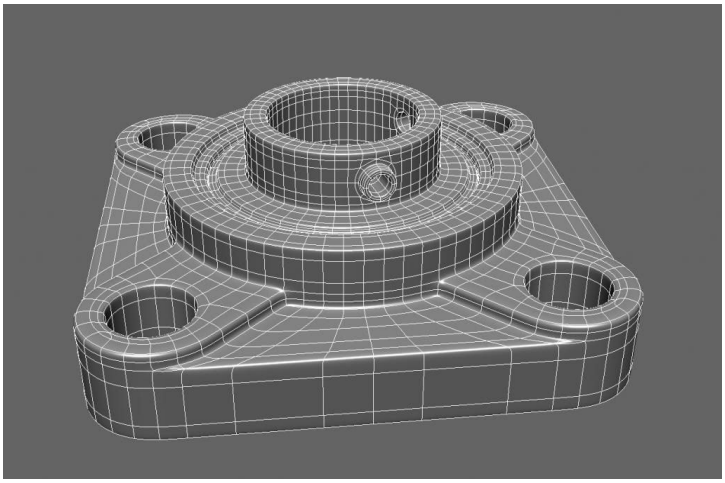
Hard surface modeling focuses on creating objects with well-defined, rigid shapes, and precise geometry. This approach is commonly used for modeling human-made objects like cars, buildings, machinery, and various other items with a manufactured or engineered appearance. The emphasis is on maintaining a high level of accuracy and control over the model's geometry.

Organic modeling, on the other hand, is more suitable for creating objects with complex, irregular shapes and forms that resemble natural or living entities, such as characters, animals, plants, and terrain. This approach mimics traditional sculpting methods, where the modeler manipulates a virtual "clay" to create the desired shape. Organic modeling often uses tools and techniques that allow for a more intuitive and free-flowing process, emphasizing the natural flow of the form and capturing the subtle details and imperfections found in living organisms or natural objects.

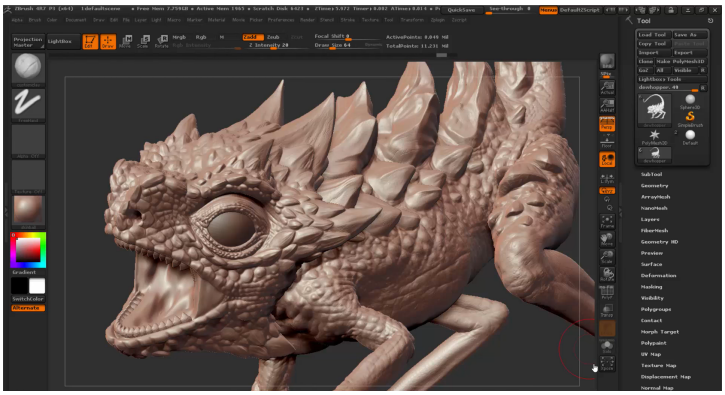
The hard surface modeling and organic modeling approaches are displayed in figures 1.2a and 1.2b, respectively. These approaches can also be combined, as demonstrated in figure 1.2c. For example, the gauntlet could have been created using hard surface modeling, with the scratches being sculpted in afterward.

One important term in the context of 3D modeling is topology. Topology refers to the arrangement and organization of vertices, edges, and faces that make up a mesh, a 3D model. It describes how these elements are connected, creating the structure and flow of the model's geometry. Good topology is crucial for a variety of reasons, particularly in hard surface modeling and animation, as it impacts the model's appearance, functionality, the ease of working with it during the creation process, and the ability to deform the model correctly without distortion.

Now the attention will transition from the shape and structure of 3D models to their recipes for determining their appearance when illuminated.



(a) Hard surface modeling [8]



(b) Organic modelling, often referred to as “sculpting” [9]

(c) Possible combination of both hard-surface and organic modelling [10]

Figure 1.2 Overview of approaches to 3D modeling

1.2 Unwrapping and texturing

The appearance of 3D models when illuminated is determined by a recipe composed of shaders, materials, and textures. This combination of elements is crucial in creating visually appealing and realistic models.

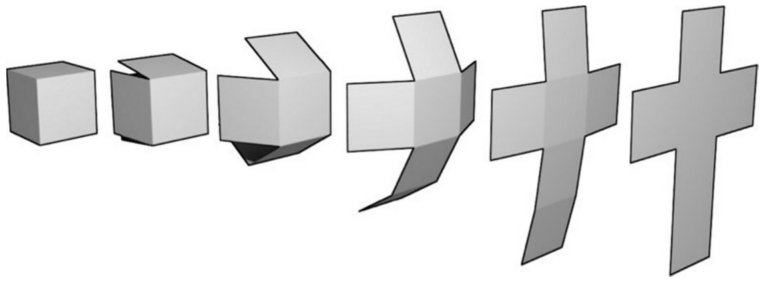
Shaders are small programs that instruct the GPU on how to render an object on the screen and perform the necessary calculations on the object. They are written in specialized scripting languages, such as HLSL. In game engines like Unreal Engine, a shader takes a material together with mesh information, and sends it to the graphics card to be rendered. It is not essential to delve into the technicalities of shaders and the rendering pipeline when it comes to the creative process of 3D modeling, which is now being described.

Materials are sets of values for various parameters that act as user-friendly interfaces for shaders. These parameters, which include color, luminosity, opacity, and others, are used for calculations within shaders. Materials allow artists to manipulate the appearance of 3D models without directly interacting with shader code. [11]

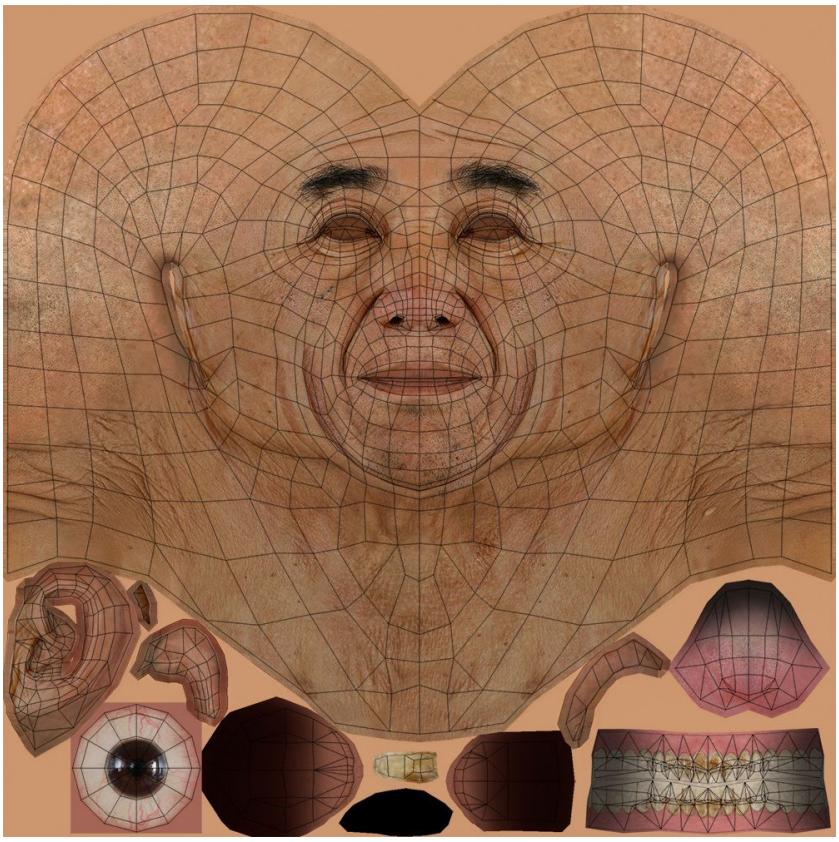
Furthermore, there are textures. Textures are 2D images that define parameter values at specific points on the 3D model, rather than affecting the entire surface uniformly. There are several types of textures that artists use to achieve more realistic and complex surface appearances:

- **Diffuse:** Defines the base color of the object.
- **Albedo:** Similar to diffuse, but represents the object's color without any lighting information, providing a more accurate base color.
- **Roughness:** Roughness of a surface determines how light is reflected and scattered, resulting in either shinier or less shiny appearance of the object.
- **Metallic:** This parameter defines the metallic properties of a material, dictating how the surface reflects light. Unlike non-metallic reflections, metallic reflections often have a color that matches the base color of the metal.
- **Normal map:** Alters the direction of surface normals to create the illusion of surface detail such as skin pores, which affects lighting calculations. It provides an efficient way to achieve complex surface details without requiring more complex geometry, thus reducing performance requirements.
- **Ambient occlusion:** Adds self-shadowing for depth and realism, without requiring real-time calculations.
- **Opacity:** Controls the transparency of the object, allowing light to pass through the surface.
- **Thickness:** Measures surface thickness for subsurface scattering effects, crucial for skin rendering where skin lets some light pass, just like wax for instance.
- **Emissive:** Indicates the areas of the object that emit light, simulating glowing surfaces.
- **Curvature:** Provides information about the convex and concave regions of a model's surface. *Useful for generating other textures.*
- **World space:** Describes the position of each point on the model in world space coordinates. *Useful for generating other textures.*

To apply these textures, 2D images, onto a 3D object, the object must be first cut into 2D pieces, similar to a paper cutout model or skinning an animal in context of organic shapes. This process is reffered to as unwrapping and is illustrated in figure 1.3.



(a) Theoretical image of the unwrapping process [12]



(b) Unwrapped and textured face [13]
(depicted with a wireframe overlay describing its topology)

Figure 1.3 Unwrapping

1.3 Rigging and animation

1.3.1 Skeletal animations

So far, the focus has been on creating static 3D objects that do not exhibit any movement. To bring a 3D character to life and transform it into a digital persona capable of moving, it is necessary to provide the model with a skeleton.

A skeleton (figure 1.4), in the context of 3D modeling and animation, is an underlying structure that mimics the joints and bones of a real-life creature. The skeleton is composed of a hierarchy of interconnected bones and joints, each responsible for controlling specific parts of the model. This hierarchy enables complex, coordinated movements by allowing bones to influence one another, much like the way bones and joints work together in a real organisms. For instance, when the thigh bone is rotated, the lower leg, ankle, and foot will follow.

To create such movements, the vertices of a 3D model are connected to the bones in the skeleton. This connection is achieved through a process called skinning or vertex weighting. Each vertex is assigned a weight for each bone it is connected to, determining the level of influence that particular bone has on the vertex. The sum of the weights for a given vertex must equal 1 to maintain the integrity of the model's shape. When a bone is moved or rotated, the associated vertices are affected according to their assigned weights, resulting in a smooth deformation of the mesh that simulates the way muscles and skin would move in response to the underlying bones.

While a skeleton provides the basic structure and movement capabilities for a 3D character, it is often insufficient for animators to create complex and nuanced motions. To address this challenge, animators use a rig (figure 1.5), which is a higher-level system built on top of the underlying skeleton. A rig consists of additional control objects, such as handles or sliders, that are designed to simplify and streamline the animation process. Rigs allow animators to manipulate multiple bones and joints[1] simultaneously, as well as to create more natural and intuitive movements. They can also include features like inverse kinematics (IK) and constraints, which enable the animators to maintain certain relationships between different parts of the character while animating. In a typical scenario, an IK system can handle the automatic reconfiguration of a character's arm parts when engaging with an object, enabling animators to focus on controlling the object's movement while the IK system adjusts movement of different sections of the arm to fit the action.

1.3.2 Morph targets

The preceding paragraphs discuss the skeletal animation approach, but there is an another method known as morph target animation. Morph target animation is employed when numerous small, per-vertex changes are needed for a model, as opposed to large-scale and usually rotating movements typically managed by skeletal animations. One suitable application of morph targets is creating realistic facial expressions for video game characters. This approach usually involves working with multiple stored mesh variations called target poses (figure 1.6) or key poses, alongside a base pose representing the neutral state of the animation. To generate different animation sequences, the position of each mesh vertex is combined with one or more target poses using a weight vector. The components of this vector correspond to a specific target pose and indicate its influence on the outcome. [14]

[1] In addition to bones and joints, morph targets described in the next section can also be utilized by rig. While a rig can move a bone connected to a shoulder, transforming the arm, a morph target can for instance simultaneously add wrinkles to the clothing worn by the character without the need for cloth simulations.

To blend between target poses, a difference mesh is created for each target pose, reflecting the per-vertex difference between the target pose and the base pose. These difference vectors serve as bases in a vector space, meaning that each vertex in the output mesh can be constructed by combining these bases using a weight vector. Specifically, for each output vertex $\mathbf{v_i}$ at time $\mathbf{t}$ in $\mathbf{N}$ morph targets with base target vertex $\mathbf{b_i}$, weight vector $\mathbf{w}$, and target pose vertex $\mathbf{p_i}$, the following relationship holds: [14]

$$\boldsymbol{v_i}(t) = \boldsymbol{b_i} + \sum_{k=0}^{N} w_k(t) \cdot \left(\boldsymbol{p_{k,i}} - \boldsymbol{b_i}\right).$$

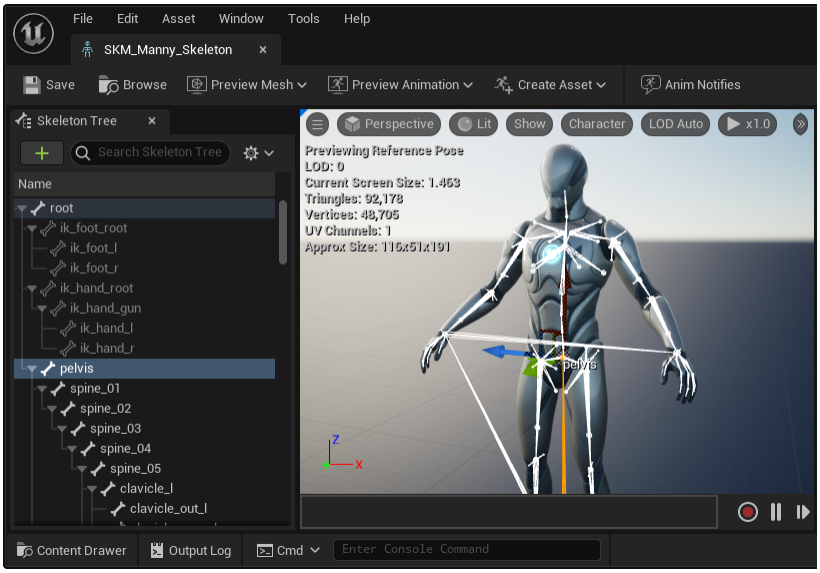


Figure 1.4 Skeleton, a bone hierarchy [15]

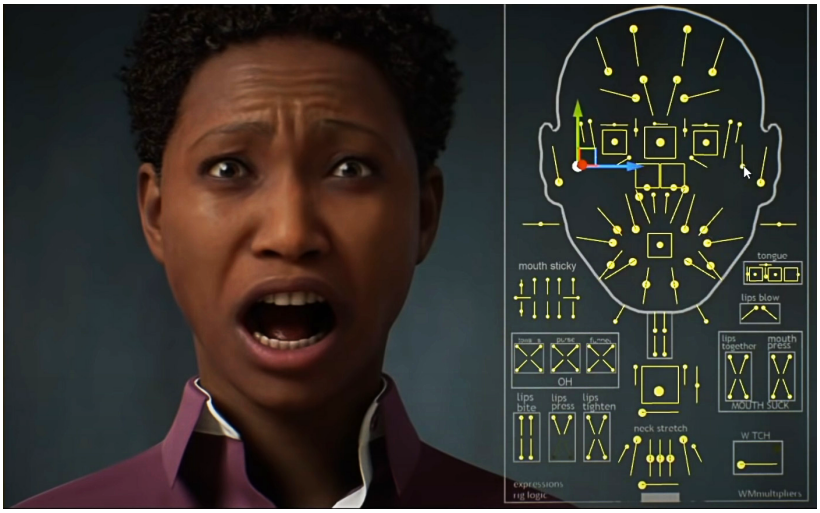


Figure 1.5 Rig of Metahuman's face [16]

Figure 1.6 Target poses [14]

Chapter 2

Colors and lighting

Careful consideration of colors and lighting can play a supporting role in enhancing the overall user experience. This chapter delves into the various aspects of color theory and light design, examining their significance during interactions as well as during the development process. It is important to acknowledge that colors have unique properties due to their characteristics stemming from physics, human biology, and the way in which they are perceived, as that differs from individual to individual. By investigating these aspects, a better understanding of their application can be achieved.

One of the main motivating factors for studying color theory and light design in the context of digital personas is the potential for color-concept-emotion associations and connotations. As stated, "*Color-concept-emotion associations also have the potential to enhance human-computer interactions in many real- and virtual-world domains, e.g., online shopping, and avatar construction in gaming environments.*" [17]. Therefore, by understanding how colors and lighting can evoke emotions and facilitate more engaging interactions, the development of digital personas can be greatly enhanced.

It is also important to consider the fact that individuality, culture and age differences in color acuity exist. Nevertheless, the implementation of carefully considered color guidelines can still greatly enhance interactions of users with digital personas and even the environments in which they oftentimes spend the majority of their time. By acknowledging and adapting to these differences, digital personas can be designed in a way that caters to most users, if not all, providing a more inclusive[1] and effective interaction experience. [18]

In this chapter, the exploration of colors and lighting will span various aspects, including history, color perception, composition, associations and connotations, storytelling, psychology, and physiology. Through this comprehensive analysis, it will become evident how important colors and lighting can be in shaping interactions with digital personas and the virtual worlds they inhabit.

[1] For the sake of simplicity, this chapter does not take into account individuals with specific conditions such as color blindness, chromesthesia (sound involuntarily evoking an experience of color) etc.

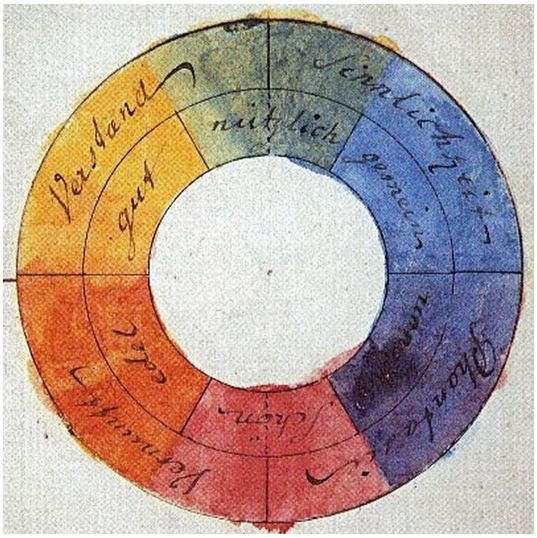


Figure 2.1 Goethe's color wheel from the book Theory of Colours released in 1810 A.D. [20]

2.1 History

The study of color theory has fascinated scholars and artists for centuries, indicating the importance of understanding the intricacies of color and its impact on human emotions. The various perspectives and models that have emerged over time reflect a desire to comprehend the essence of color and its relationships with human experiences. However, no single model can definitively claim to encompass the whole truth about color theory, which is why it is essential to examine the subject from its very foundations in the next sections.

Wolfgang von Goethe, a prominent figure in the study of color theory, developed a color wheel as can be seen in figure 2.1 that sought to associate positive and negative emotions with different colors. His pioneering work paved the way for others to build upon his ideas and delve deeper into the world of color and emotion. Even two centuries later, Claudia Cortes and Shirley Willet proposed similar models, demonstrating the everlasting relevance of color theory. [19]

In more recent times, researchers have continued to explore color theory, attempting to create models grounded in scientific understanding. For example, Niels A. Nijdam proposed a 2D grid with two axes — Pleasure and Arousal — in an article titled "Mapping emotion to color" [19]. While Nijdam's model offers a fresh perspective on the subject, it is not without its limitations, as the exact distribution of emotions on the grid remains unknown.

Given the multitude of perspectives and the lack of a universally accepted model, it is crucial to approach the study of color theory and light design from the very basics by examining the foundations of how humans perceive light as color. In the following section called Color Perception the inner workings of the human eye and the nature of light as electromagnetic waves will be described.

2.2 Light as electromagnetic waves

In the study of color perception, it is essential to understand the nature of light itself. Light is a form of electromagnetic waves. These waves are generated by the oscillation of electrically charged materials, and they travel at a speed of approximately 300,000 km/s. Thanks to this speed, light has the ability to transfer information virtually instantaneously for human perception. [7]

The portion of the electromagnetic spectrum that is perceived as light, known as visible light, spans wavelengths from 380 to 720 nanometers. However, visible light is only a small fraction of the entire electromagnetic spectrum, which also includes for example X-rays, gamma rays, infrared, and ultraviolet light and more. Images captured at different wavelengths exhibit varying properties and provide distinct information about observed objects. X-rays and gamma rays have significant applications in medicine, particularly in visualizing internal body functions. Infrared light, emitted by heated objects, is useful for detecting people and objects at night. Despite their practical uses, excessive exposure to the light of these wavelengths can be detrimental, e.g., ultraviolet or infrared light can cause the decomposition of proteins and damage to the lenses in the eyes. [7]

When it comes to perceiving color, we primarily see light that has been reflected off objects. When light encounters an object, certain parts of the spectrum are absorbed, while others are reflected. The reflected portion is what we perceive as the color of the object. Several factors influence the perception of color, including hue, intensity, saturation, and brightness. [7]

- **Hue:** refers to the dominant frequency of the perceived light.
- **Intensity:** describes how light or dark it appears.
- **Saturation:** denotes the tightness of the spectrum around a given wavelength.
- **Brightness:** indicates the amount of achromatic (white) light present.

These aspects of light and their effect on human perception will be explored in detail in the subsequent sections. [7]

2.3 Eye

Humans rely on the eye, a spherical organ, as their primary means of perceiving light and color. Light enters the eye through the lens and is projected onto the retina. It is worth noting that the lens absorbs approximately 8 % of the incoming light, but during the typical aging process, the human eye's lens becomes more yellow, rigid, and disperses light to a greater extent, causing age-dependent variations in the perception of light and the presence of optical errors. [21]

The retina functions as a sensor comprised of rods and cones, with an estimated 120 million rods and 8 million cones present in the human eye. This means that for every cone, there are about 15 rods. The distribution of rods and cones across the retina is not uniform. The majority of rods are concentrated in the fovea, which is situated at the center of the optical axis where most light is focused. [7]

Rods are about 10 times more sensitive to light than cones and play a more significant role in peripheral vision. Cones, on the other hand, are responsible for color perception and are classified into three categories based on their photopigment. The distribution of cones by photopigment is not uniform either: 64 % of cones are red, primarily sensitive to a wavelength of 575 nm (closer to yellow but with some overlap); 32 % are green, with peak sensitivity at 535 nm; and a mere

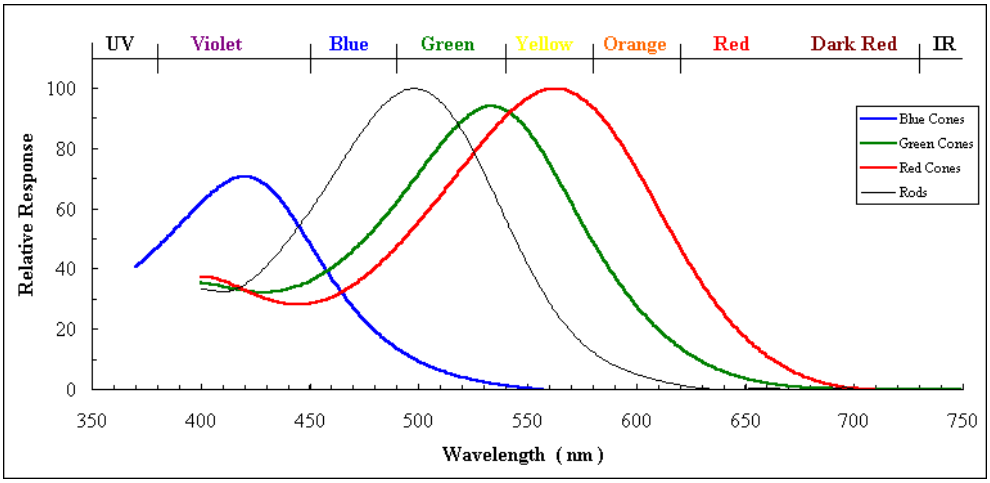


Figure 2.2 Sensitivity of rods and cones to different parts of the visible spectrum [24]

2 % are blue, sensitive to 520 nm. It is important to clarify that cones are not referred to as red, green, or blue because they possess these colors, but rather because they are sensitive to these specific colors. [7]

The characteristics and applications of separate red, blue, and green light will be discussed in more detail in the following paragraphs:

- **Red light:** is well-suited for certain applications, in part because it has a non-dazzling nature that can be beneficial in some contexts. A key distinction between rods and cones lies in their respective sensitivity and response to various light conditions. Rods are primarily responsible for vision in dim light and are more sensitive to overstimulation, while cones are responsible for color vision and are less sensitive in such conditions. In particular, extreme red light is detected by red cones and has the most significant wavelength difference compared to the light perceived by rods, as seen in figure 2.2. This difference ensures that the rods are not overstimulated. Rods recover from overstimulation, which has tendency to cause dazzle or discomfort, much slower compared to the relatively quicker recovery of cones. [7], [22]

  Due to these properties, red light has found practical use in various settings, such as cockpits during nighttime missions. Properties of red light ensure that pilots can maintain their night vision while still being able to read instruments and navigate effectively. [23]

- **Green light:** specifically green-yellow, is the color to which the eyes are most sensitive. This heightened sensitivity is due to the overlapping perceived spectrums of green and red cones in this color range. Additionally, green light is affected the least by chromatic aberration in the eyes, as can be seen again in figure 2.3. This contributes to the overall sharpness and clarity of the vision when viewing green light.

  As a result, green light has become the preferred choice for various applications, particularly those requiring optimal visual clarity. One such example are night vision goggles, which predominantly utilize green light to ensure the highest possible image quality. This preference for green is supported by an article about night vision goggles that states "*since green focuses on the retina more readily, it has been chosen as the color for most systems*" [25]

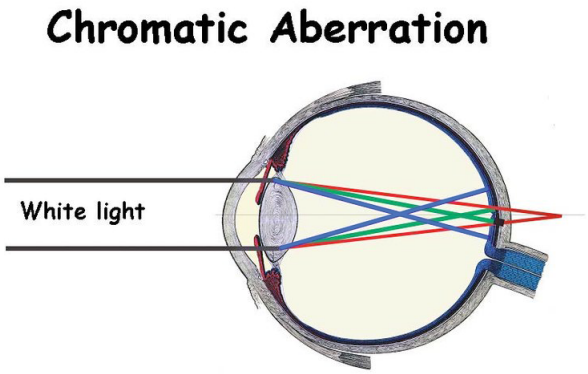


Figure 2.3 Chromatic aberration in the human eye [26]

- **Blue light:** chromatic aberration is an optical phenomenon that occurs when a lens is unable to bring all wavelengths of light to the same focal point. This results in the usually noticeable fringes of colors appearing along the boundaries of objects in an image. In the human eye, green and red light tend to be focused relatively well on the retina. However, "*The difference of refraction between red and blue might be up to 3 diopters.*" [26] as depicted in figure 2.3

  The brain compensates for chromatic aberration in most situations by processing and adjusting the perceived image. However, this compensation may not be as effective when no other color is present. Considering these optical properties, blue light may not be the best choice for certain applications, such as concert lighting. Audience members positioned further away from the stage may struggle to focus their sight on the performers, resulting in a blurred vision of the scene.

Apart from fovea responsible for sharp and detailed color vision, one such notable aspect of the eye's structure is the presence of a blind spot. The blind spot occurs because the nerves collecting information from photoreceptors run in front of them, and all of them connect at a specific location. This design flaw is attributed to evolution. Interestingly, octopuses, for example, do not have a blind spot, as their eyes evolved in such a way that the nerves are routed behind the photoreceptors. Despite this evolutionary flaw, the brain's image-fixing capabilities compensate for the blind spot, making it unnoticeable in everyday life. [27]

Another fascinating aspect of color processing in the eye involves the recombination of color information. The red and green stimuli are combined into a single stimulus defined as a ratio expressing their relative proportions. A similar process occurs for blue and yellow stimuli. This is the reason why certain color combinations, such as yellowish-blue or reddish-green, are imperceptible to the human eye. [7]

Moreover, the human eye can perceive an extensive range of light intensities but cannot process all of them simultaneously. Instead, the eye adapts to the current lighting conditions, enabling the adjustment of vision to accommodate various levels of intensity. [7]

2.4 Composition

Striking the right balance of light and color is essential for creating visually appealing and engaging imagery. This section will state some of the mistakes and compositional rules that have been observed as efficient and tested thoroughly through history of art. As stated in the previous section, even though eyes can adapt "*Too much light offends our gaze, and too little weakens it*" [28]. Therefore, the most pleasant image is often found somewhere in the middle of these extremes.

To control the viewer's attention and to highlight key elements in an image, such as digital persona or perhaps some interactable objects. Aerial perspective can be used. By simulating the atmospheric effects that occur over distance, such as shifts in hue and reduced contrast, can provide depth and dimension, usually in the background, and may also contribute to a visually appealing composition. The strategic use of light and color can help an object stand out from the rest of the scene, creating a sense of hierarchy and focus. [28]

One has to be careful with colors, as the harmonious combination of a few colors can be pleasing to the eye, while an excessive use of many colors may be overwhelming and distracting. Furthermore, local colors can appear in various forms, influenced by reflections, light sources, and indirect lighting. [28] So an object and its corresponding material created outside of the scene may look differently when placed in some, and that is something creators should be vary of. Additionally the scene itself can alter the way we perceive the color not as color but as something more. We associate colors with different meanings as they usually appear in different contexts in different environments. Moreover not only the meaning of the color can be affected according to what context it appears in, but we can also use it other way around. To bestow the meaning of color on the context it appears in.

2.5 Color associations and connotations

Colors and their associations play a vital role in human perception and experience. Different colors can evoke various emotions and connotations, both consciously and unconsciously. However, these associations can be influenced by factors such as cultural differences, personal experiences, and societal norms. Despite that, there are colors that share common meanings across cultures, it is essential to consider these variations during the development process. [29]

For example, an article called "*CLex: A Lexicon for Exploring Color, Concept and Emotion Associations in Language*" [17] highlights the difference in color associations for joy between Indian and US annotators, with yellow being connected to joy for Indian annotators and pink for US annotators. Despite these differences, there is a considerable overlap in color associations. Lexicons such as CLex mentioned in an article above or EmoLex and others based on statistics can provide insights into these connotations and help navigate cultural and personal differences. [17]

To state some of the associations, colors such as red and brown are strongly linked with anger, green with anticipation, black with fear, pink with joy, black, brown, and gray with sadness, yellow and orange with surprise, and blue and brown with trust. [17] It's a custom to use red for enemies in video games. Some of these associations can be logically traced back to historical or environmental influences. On the other hand, other associations, like red evoking hunger may require context for accurate interpretation. Some mobile operators are using red in their logo and hopefully no one ever found an appetite for SIM cards.

Considering the diverse range of color associations, it is crucial to emphasize the importance of context in determining the meaning of a color. Colors can have both positive and negative meanings[2], and understanding the context is key to accurately interpreting their significance. While some colors may evoke certain emotions or responses in specific situations, colors alone do not inherently possess these associations. For example, an article exploring the relationship between colors of face masks and perceived emotions of people wearing them found no significant correlation between color and perceived emotion. [30] This could be explained by the missing context to establish such connection. However, this does not mean that colors alone have no effect, this will be discussed in a section about physiology later.

Probably the most outstanding and very useful aspect of color associations is the ability to create them intentionally. Designers and storytellers can strategically use colors to evoke specific emotions or establish connections with the audience. The director of Toy Story 3 Lee Unkrich described the importance of blue in his movie "*We came up with the concept of blue connoting safety and home. At the beginning of the film, Andy... his bedroom is blue, the sky is blue, his T-shirt is blue. He is in blue jeans, he's got a blue car — these are not accidents. These are conscious choices. Everything in the movie is there for a reason. We are making a commitment to say that blue will connote safety and trying to avoid that in situations where we don't want the audience to feel safe.*" [31] This example demonstrates the power of color in storytelling and the potential to influence emotions and perceptions through conscious color choices.

2.6 Storytelling

The power of colors and lighting in storytelling is further exemplified through the intentional use of them in creating visually engaging narratives. The emotional experience of a viewer can be shaped by the journey through scenes that unfolds in colors and lighting of each. From the warm, comforting light in the introduction of a story to the darkness that sneaks in as tension rises, and the eventual warmth of resolution, the strategic use of lighting and color plays an important role in storytelling. [31]

One powerful example of such journey of colors can be found in the film Up. "*After Ellie's death, Carl's world becomes dark and gloomy. Color is slowly reintroduced in the form of the balloons that lift him into the sky, Russell and his rainbow sash of Wilderness Explorer merit badges, and Kevin, the colorful bird of Paradise Falls. Magenta, which is Ellie's color, remains absent from the film until Carl reaches Paradise Falls.*" [31], this adds a layer of emotional depth to the story. Similarly, as the example from Toy Story 3 in the last section, this helps to create associations with the color and break the film into separate parts with different meanings to support the story.

Color can be used to organize and group different elements or scenes, providing a sense of cohesion and structure to the narrative. This property is shared with sound and music, which will be discussed in a later chapter Composition and uses. In the context of digital personas, this shared property of both visual and auditory perception could be used to better differentiate between various dialogs and create a sense of familiarity with the topic after recurring encounters with them.

Regarding the lighting design itself, the way a face is lit can affect the character's appearance, recognizability, and emotional appeal. "*The characters may be saying one thing, but if the color and lighting make the scene feel gloomy, or if the music is unsettling, the audience knows something else is going on—the character's dialogue can't be taken at face value.*". [31] Lighting styles such as low-key lighting, which often employs a single light source, create strong emotional

[2] or even none, context is crucial

reactions by obscuring or distorting facial features. For instance, underlighting, which places the light source below the subject, is unnatural and often used to highlight criminal or evil characters in crime and horror films. Similarly, lighting directly from above or silhouetting can obscure the subject's eyes, increasing the sense of intimidation and unease. However, there are exceptions, such as 90-degree side lighting, which may still provide sufficient information about a person's intentions and emotional state. This is explained by the fact that the face is symmetrical. [32], [33]

Conversely, lighting that reveals facial features more naturally can evoke more positive reactions from viewers. For example, a study found that lighting setups that exposed facial features in a manner typical of natural settings (e.g. such as from sun) led to more positive responses. This demonstrates the importance of lighting in guiding the viewer's emotional response and understanding of a character's intentions and emotional state. [32], [33]

2.7 Psychology

Colors and lighting can have an impact on psychological well-being, with various studies and theories exploring their influence on emotions, perceptions, and mental states. The book *Color Psychology And Color Therapy; A Factual Study Of The Influence of Color On Human Life* [29] by Faber Birren is a great overview of many studies dedicated to colors from various different points of view. Despite the fact that some statements like personal preference of color can say something about one's personality (athletic people prefer red, intellectuals blue and so on) is purely a speculation added only as a "fanciful note" other parts can bring a great insight.

For example, in his work, Birren suggests that "*normal persons who are or attempt to be well adjusted to the world, and hence "outwardly integrated," like color in general and warm colors in particular. "Inwardly integrated" persons may favor cool colors and be none too enthusiastic about them*"[29]. This is supported and extended by Mahnke's statement that it is essential for designers to understand these psychological preferences and avoid creating environments that conflict with individuals' personalities. There is a fallacious belief that prescribing passive environments for extroverted temperaments or active environments for introverted personalities will lead to positive outcomes, but people are unlikely to be happier in surroundings that clash with their inherent dispositions. [34]

Mental disorders can also affect the way we perceive colors, as the visual process involves both neural and optical components. Research has demonstrated that abnormalities in color sensitivity and perception in the retina may be indicative of mental disorders. Additionally, color therapy, which involves patients drawing with different colors, is another example of how color can suggest a psychological state. Victims of abuse or trauma may use colors like red and black more often. [29], [35]

Understanding these influences is crucial for designers, therapists, and researchers alike. Despite skepticism within the medical profession, color therapy has the potential to benefit individuals if studied rationally and impartially. Some rare conditions, such as urticaria solare, provide evidence for the biological effects of visible light on the human body. In this case, exposure to visible blue and violet light causes photosensitivity and skin afflictions. Next section, will explore how colors and lights can affect human physiology, including blood pressure and other bodily functions. [29]

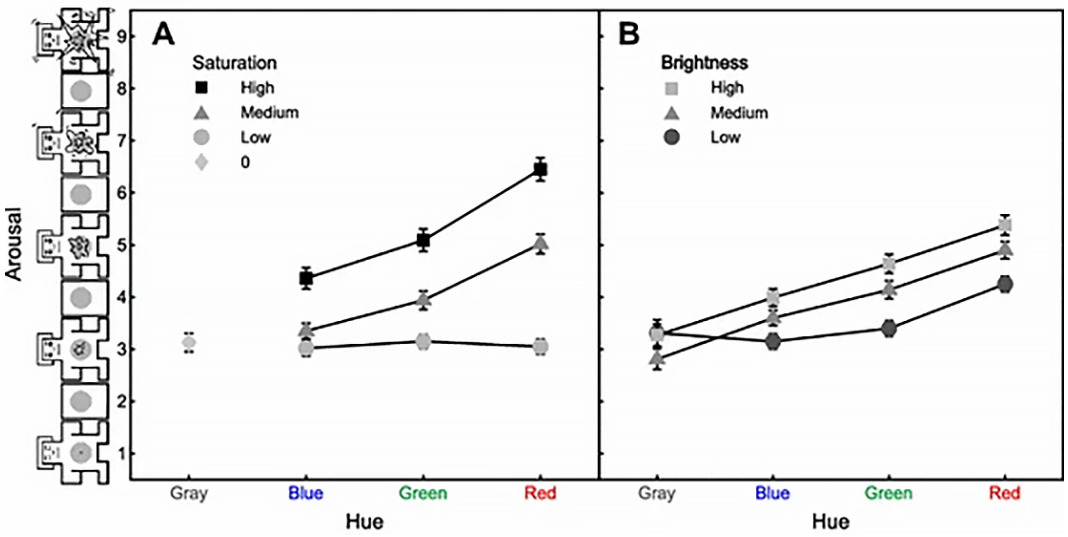


Figure 2.4 Average arousal ratings of different light wavelengths (hue) in relation to saturation and brightness [36]

2.8 Physiology

Colors and lighting, apart from impact on psychology, can have physiological effects on human bodies, influencing various bodily functions and responses. Across cultures and religions, the color red has been associated with a wide range of feelings and concepts, from happiness and bravery to sin and mourning, also extrovertness, but apart from these volatile associations and its corresponding psychological impacts, red has been found to have physiological effects, stimulating brain-wave activity, increasing heart rate, and raising blood pressure. [31]

Interestingly, research conducted by Charles Féré already in the 1880s suggested that red light had the most stimulating effect, while violet light was the most calming. Modern studies have confirmed this. In this study called *Color and emotion: effects of hue, saturation, and brightness* [36], electrodermal (skin conductance response; SCR) and cardiovascular (heart rate) parameters were measured as they are considered correlates of arousal. Participants in these studies viewed large LED panels displaying different colors before a black background, revealing that the effects of color on emotion depends on the combination of hue, saturation, and brightness[3]. It is interesting to note that arousal rises as the wavelength of the light rises as can be seen in figure 2.4. [18], [36]

One could draw a conclusion that blue light is calming. Although this could be said, it's a bit more complicated than that as blue monochromatic light has been shown to be more effective than longer-wavelength light in enhancing alertness, potentially due to its impact on melatonin[4] production. [37]

[3]To avoid misunderstanding, it is important to note that there might be some confusion in terminology, as in this article the term *brightness* is used to refer to *intensity*. These terms were explained in a section *Light as electromagnetic waves*

[4]hormone responsible for regulating sleep-wake cycles and promoting relaxation

Another interesting physiological aspect related to light and color is the potential impact of light on the eyes and gaze. The contrast between the sclera and iris in the human eye allows for the transmission of information through gaze. When shadows obscure this contrast in the eyes, as described in the section about *storytelling*, it may hinder the ability to send visual signals, potentially causing unease in others. It is thought that the human visual system likely developed before the widespread use of artificial lighting sources emitting light from below, such as firelight. [33], [38]

2.9 Conclusion

In conclusion, it is challenging to assign universal meanings to colors such as "blue represents good", since interpretations vary across cultures, individuals and most importantly contexts in which they appear. Lexicons may aid in determining associations, but their usefulness and applicability remains debatable. What remains unquestionable is the ability to create these color-concept-emotion associations to achieve specific outcomes. When selecting colors for particular associations, one should consider the properties of light, rooted in physics and human biology, as they have been shown to influence physiology and psychology. This comprehensive understanding of colors and lighting will be beneficial during the development and testing phase, ensuring that the final outcome is both visually appealing and effectively communicates the intended message.

Chapter 3

Sound

Sound, as a vital aspect of storytelling and user experience, plays an indispensable role in films, games, and applications. As Steven Spielberg once said "*Sound and music make up more than half of communicating a story, greater even than what you're seeing*" [39]. Sound functions as a secondary medium for conveying a wide range of information to the user, including details about:

- **Character states**: conveying levels of exhaustion, emotions, injuries, or progress in leveling up in video games, or if they are a friend or foe
- **Environmental cues**: providing information about the setting and character's orientation through wind, crunching snow, or footsteps on various surfaces
- **Meta information**: generating tension and fostering anticipation of upcoming events, or establishing a feeling of a particular time of day
- **Overall theme**: communicating the experience's genre and mood, such as epic or mystical

In an ideal scenario, sound should originate from objects that are in motion or of significant importance, as well as from elements that cannot be conveyed using other senses, but are still important for the overall experience. [40]

In addition to providing explicit information, sound serves to evoke specific moods, emphasize chosen themes, and accentuate key activities. This aspect of sound design contributes to the emotional depth and richness of the experience, engaging users on a deeper level. In the following section, the relationship between sound and color will be explored to further understand the intricate connections that exist between these sensory experiences.

3.1 Sound in relation to color

The connections between our senses, including sight and hearing, are physiologically intertwined, and they can have a mutual impact on each other. This section will primarily focus on the interplay between sight and hearing, as the ways to artificially trigger other senses on command are not widely established. This interplay between senses can lead to suppression or amplification of sensations. For example, popcorn may not taste as flavorful when consumed while riding a roller coaster. Conversely, when trying to hear something in a quiet environment, such as at night in the woods, the effort to listen can seem to enhance both sight and hearing. [29]

Sergey Kravkov, a Russian psychologist and psychophysiologist did found that the pitch of a sounds can shift the appearance of colors, their hue. Low pitch tends to shift the appearance of light towards shorter wavelengths while high pitch tends to do the opposite, shift the appearance to longer wavelengths. This effect can cause red to appear deeper or more bluish, orange to become reddish, yellow to turn brownish or reddish, green to appear bluer, and blue to resemble violet. Furthermore studies have also shown that rod vision (intensity) might be reduced during sound stimulation, whereas cone vision (color) could be enhanced, particularly for green light. It has been suggested that high-pitched sounds may help sharpen perception to some extent. The findings of Kravkov have been confirmed and expanded upon by researchers Frank Allen and Manuel Schwartz. Additionally, researchers Karwoski and Odbert found associations between slow music and blue color, fast music and red color, high notes and light colors, and deep notes and dark colors. [29]

Music has numerous parameters, including contour, rhythmic content, pitch content, length, mode, orchestration, texture, register, tempo, harmonic progression, harmonic function, and contrapuntal framework. This section will not delve into the details of each parameter, but will highlight tempo and mode. Studies have shown that quicker tempo, which refers to the speed of the music, is associated with more saturated colors and higher color intensity, and that for both minor and major modes[1]. Notably, quicker tempo and more saturated colors[2] both increase arousal, suggesting a connection between these two aspects. Sound can affect physiological responses too. For instance, listening to sad sounding music can cause a decrease in heart rate and skin conductance level but an increase in blood pressure, while happy sounding music can lead to a decrease in respiration depth. [41], [42]

According to the Yerkes-Dodson law, the effect of arousal on performance is characterized by an inverted U-shaped curve, with the best cognitive performance at moderate levels of arousal and lower performance at both low and very high levels. Mood induced by music affects performance as well, with positive moods improving performance on tasks such as categorization, complex decision making, creative problem solving, sorting, and heuristics. Conversely, "*boredom or negative moods can lead to poor performance*" [42].

In terms of music parameters, mode affects mood, while tempo influences arousal. Major/minor mode does not impact arousal that much, only the scale of the result. Conversely, mood is affected primarily by major/minor mode, but not tempo really, it's a similar story to how arousal is affected, but the other way around. Overall, music is perceived as more enjoyable when major mode pieces are faster and minor mode pieces are slower, possibly because these combinations complement each other well. Importantly, tempo manipulations affect arousal, while mode manipulations affect mood. All of this is shown in figure 3.1 and figure 3.2. [42]

To clarify... "*The scientific investigation of emotion has been hampered by the fact that many distinct phenomena are often studied under the same linguistic rubric. Several different words, such as emotion, mood, or feeling, have been used interchangeably by researchers, who often work in different sub-disciplines of the field, to refer to the same basic phenomena. Conversely, many research programmes investigate the same affective phenomenon, but use quite different terminology. This has led to an inevitable degree of confusion*". [43] In this thesis, the term "mood" will primarily be utilized, as it was also the term utilized in the articles from which this research was derived. Mood typically refers to relatively long-lasting emotions, which may have stronger consequences for cognition[3]. Arousal, on the other hand, refers to the degree of physiological activation or the intensity of an emotional response. The following section will discuss when and what type of sound should be played to achieve desired results. [42]

[1] major mode sounds happy or uplifting while minor mode sounds more sad or melancholic, this is achieved through a selection of notes that are used

[2] how color affects arousal is discussed in a previous section *physiology* in the context of color perception

[3] "cognition" refers to the mental processes involved in acquiring, processing, and understanding information. This includes functions such as perception, memory, learning, problem-solving, and decision-making

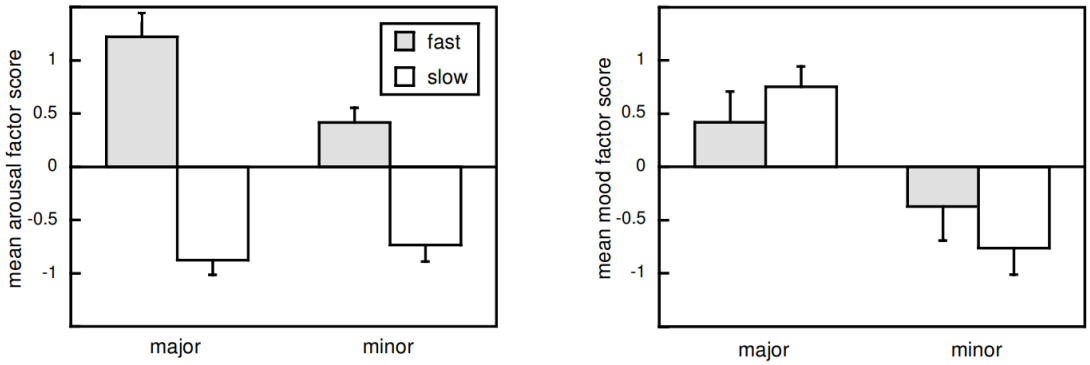


Figure 3.1 Mean arousal and mood ratings of tempo and mode combinations [42]

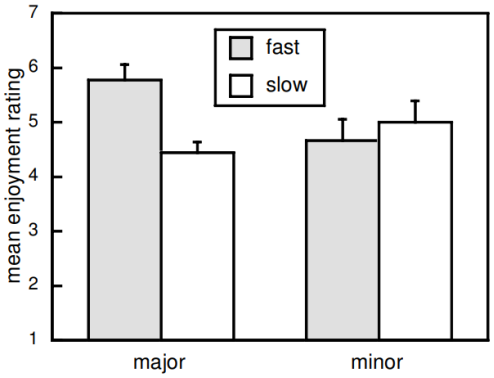


Figure 3.2 Mean enjoyment ratings of tempo and mode combinations [42]

3.2 Composition and uses

Selecting appropriate music is important. When choosing the right music for a particular scene or situation, the intended outcome becomes more evident to the user. For example, aggressive music may encourage confrontation, while ambient tunes could inspire a more peaceful approach. Familiarity with recurring music cues not only helps users anticipate actions but also supports learning and decision-making in future encounters with similar cues. When combined with visual[4], audio effectively contributes to the creation of memorable experiences. [39]

The concept of leitmotif plays a crucial role in creating a sense of familiarity and association in the context of sound. Leitmotif, often translated as "leading motive" is a recurring musical theme associated with a specific character, object, or idea. As these themes recur throughout a piece, they become more meaningful and contribute to a larger musical structure. [44]

Theme is vaguely defined as anything heard more than once, and the repetition of these themes strengthens the connection with the related subject. By employing leitmotifs, creators can enhance the association between sound and visual elements, similar to the way colors establish associations. This shared property of sound and color can be further utilized to amplify the desired effect on the audience. [44]

The process of establishing associations works in the following steps:

1. "*Recognizing familiar elements.*"
2. "*Recalling other music or schemata that make use of those elements.*"
3. "*Perceiving the associations that follow from the primary associations.*"
4. "*Noticing what is new and how familiar elements are changed.*"
5. "*Interpreting what all this means.*" [44]

Sound associations, much like color associations, are of course again influenced by factors such as culture, race, gender, and religion, but these culturally dependent associations are then further filtered through an individual's unique experiences. For instance, Western listeners may associate Zlatý Kolovrat composed by Antonín Dvořák with a specific meaning, but the personal connection to the piece depends on whether the listener has engaged with the related story written by Karel Jaromír Erben. Another example... a piece of music might remind one person of young love if heard during a magical prom night, while evoking corporate drudgery for another if it was part of an office playlist. Music can create so called autobiographical memories, a "*meanings cemented onto music colored with emotion from past experience*"[44]. This concept of music being "colored" by emotion further emphasizes the connection between sound, color and concept-emotion associations.

The association between music and visual is bidirectional. Not only can music evoke specific images or scenarios in our minds, such as a particular movie scene or a feeling bound to it, but the opposite is also true. When exposed to a specific visual element, we may be reminded of a piece of music or a melody connected to that visual. This reciprocal relationship highlights the profound connection between sound and visuals in our minds. [44]

[4] and possibly other sensory stimuli

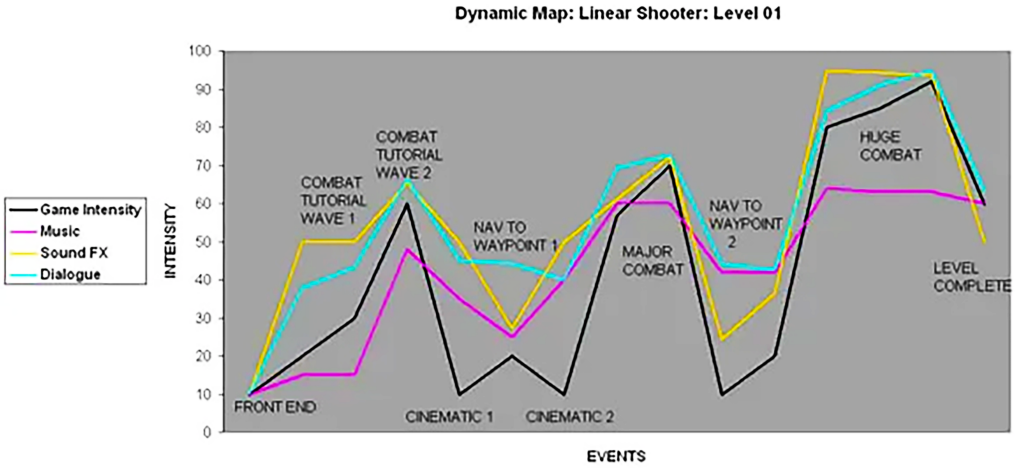


Figure 3.3 Sound adapting to and reflecting occurring events [39]

Furthermore, the power of associations is not limited to situations where music and visuals are interlinked or complementary. Even when music and context appears mismatched, such as a comedic theme played over a serious scene[5], the association between them can still evoke a potent emotional response, which may be even stronger than that in conventional situations. This incongruity may lead to humor, shock, disgust or discomfort, showcasing the strength of musical associations. [44]

Although utilizing such contrasts, even though powerful[6], may not always be safe or practical for some purposes. Deliberate use of dichotomy has to be approached with caution, but that applies for every intentional use of associations as well. Introspection should be avoided.

In order to prevent user fatigue or boredom, it is essential that sound does not play constantly, but rather reflects the occurring events within the application. That is shown in figure 3.3. A dynamic change in music and sound in real-time is more achievable by focusing on more rhythmic compositions instead of melodic ones. This approach allows for smoother and quicker transitions between tracks, reducing the likelihood of distorted or unpleasant audio artifacts caused by overlapping melodies for a prolonged time. In any case, these transitions between different audio tracks should not be linear. A variety of functions, other than linear, can be employed to avoid hearable abrupt starts and ends in the transitions. For instance, in racing or sports games, licensed music is often used, as it provides a more real-life like context. However, blending licensed music tracks is challenging, limiting real-time adaptation to the occurring events. [39]

One of the very useful benefits of employing sound in games is the ability to segment applications into distinct sections. This concept links back to the idea that choosing the right music for a particular scene or situation helps make the intended outcome more evident to the user. It also highlights another similarity between the use of sound and color, as both can serve to segment different parts of an application. Although, when transitioning between distinct segments, it is crucial to avoid playing all sounds simultaneously, as this would create a “wall of noise” and potentially overwhelm the user. [39], [40]

[5] for eternal damnation of your soul please visit this site `http://bennyhillthis.com` and insert a link to the Schindler’s List trailer, you’ll hardly find any better example of such dichotomy and its strong effect [44]

[6] according to this `http://youtu.be/i8HePfa7WYs` Youtube video such concept of dichotomy is used by Pixar to make an audience cry, by putting a happy sounding music over a sad scene [45]

In some cases, music does not need to play all the time or at all. The necessity of sound in an application depends on factors such as platform, user's familiarity with the application and genre in case of a video game. For example, the game called Myst was designed without music, relying instead heavily on environmental sounds to set the mood, create familiarity, and establish the right atmosphere. This approach resulted in a "sound-centric" application, where the focus is primarily on sound effects. Similarly, "music-centric" applications primarily revolve around music. An example of such an application could be TikTok, though this distinction is not widely established and may be debated. For this explanatory purpose though, terms "sound-centric" and "music-centric" serve well. What widely established, and often used is, is the term "vococentric" coined by Michel Chion [46], which indicates that the application or film revolves around dialogues. It does not imply the absence of music, in fact, the use of music as dialogue underscoring often outweighs other uses, at least when measured in screen time for movies. [47], [48], [49]

[40] Sound can be divided into categories such as:

- **DX:** dialogues, any verbal speech
- **MX:** music, usually non-diegetic (see the following list for an explanation of terms diegetic and non-diegetic)
- **SFX:** sound effects, usually sounds played for a short period of time, can originate from objects in the scene, clicking on GUI, signaling changes and so on
- **FOL:** foley, sounds produced by the user or its avatar
- **BG:** background sounds or noise, even diegetic music

[40] Furthermore, sound can be categorized as:

- **Diegetic:** originating from within the scene and perceivable by the characters living in that scene
- **Non-diegetic:** not originating from the scene and heard only by the user/player/viewer

Game Myst used only diegetic sounds to create a realistic lively environment. However, both diegetic and non-diegetic sounds can be used together, and mixed together, which is often utilized. However we can separate these two concepts and utilize it as another tool for segmentation. The movie Casablanca employed this technique, using non-diegetic music during scenes of Nazi-occupied streets and soothing diegetic music during indoor scenes played on a piano by a friend of the main character. This approach showcases the diverse ways sound can be utilized to create distinct atmospheres and segment applications into different parts. [47], [49]

3.3 Conclusion

Sound is a valuable asset for transferring messages from media to observers through various means. It has the capacity to evoke emotional responses, immerse users, and create a sense of familiarity. When choosing music, inspiration can be for instance drawn from this [50] emotion-annotated dataset of 80s-styled game music, but it is essential to remember that sound and music are still an aesthetic choices of metaphors too. [39]

All this knowledge of sound's impact and potential will prove beneficial during both the testing as well as development phases. As Ralph Eggleston, an animator by training who introduced colorscripting at Pixar, once remarked that "*He will often choose a piece of music to listen to while drawing the colorscript — a composition unrelated to the film except that it captures the moods he's trying to convey.*" [31]

Chapter 4

Analysis

After acquiring a foundation of theoretical knowledge, the next step involves investigating the tools available to achieve the established goals. Following this Analysis chapter, the focus will shift to the design and implementation of the solution followed by throughout testing in which the final solution will be examined.

4.1 Conversational platforms

A conversational platform is a system that facilitates conversational design by employing natural language understanding and other components to enable human-like interactions between user and a computer[1], typically through text or voice. To better understand the dynamics of such interactions, it is essential to first examine the concept of conversation itself. The terminology of such conversational systems includes:

- **Conversation:** An interaction between a human and a conversational interface, or between humans, although that is not the focus here, consisting of a series of turns.
- **Turn:** A single exchange within a conversation, including the user input (utterance) and the system's response.
- **Utterance:** The user's input in a conversation, representing what the user says or types to convey their intent.

The key components of a conversational platforms are:

- **Automatic Speech Recognition (ASR)[2]:** Transforms spoken language input into text.
- **Natural Language Understanding (NLU):** Processes and interprets the user's input, identifying intents and extracting entities.
  - **Intent:** The meaning or purpose behind a user's input, representing what the user aims to achieve during the interaction. By understanding the user's intent, the conversational platform can respond accordingly.

[1] in this case a digital persona

[2] also known as Speech-to-Text (STT), but in the context of conversational platforms more often referred to as ASR

  - **Entity:** Pieces of data extracted from user input, such as names, dates, places, or phone numbers, that the conversational interface can remember and utilize in the conversation, e.g., addressing the user by name.

- **Dialogue Management:** Handles the flow and structure of the conversation.
- **Natural Language Generation (NLG):** Crafts responses for the system to deliver to the user.

In addition to these primary components, conversational systems also include parts that communicate with the external world, typically through API connectors or, in the case of Flowstorm, commands and actions.

- **Command:** Metadata originating from the conversational platform, not part of the direct communication, but capable of triggering actions such as animations or other effects.
- **Action:** Events or operations triggered by commands, which may include animations, sounds, or visual elements like QR codes appearing on the screen.

4.1.1 Rasa

Rasa is a powerful open source conversational AI platform that enables developers to build conversational interfaces. Rasa is unique in that it allows for complete customization of your conversational interface, making it an ideal platform for businesses or developers who want to create a unique conversational experience for their users. Rasa even advertises "*It's Not a Bot, It's Your Brand*" [51].

However, this customizability comes at a price of Rasa's one of the main disadvantages, its complexity. Being a Python based framework requiring a local installation calls for a significant amount of technical knowledge and expertise to set up and configure properly. Rasa's lack of a graphical user interface (GUI) means that developers need to work with code to build and train the conversational interface. Additionally, Rasa requires hosting and deployment on a server, which can be challenging for those who are new to server administration. However, for those who prefer a managed hosting solution, Rasa-as-a-Service is also available to completely take care of the hosting and deployment of the conversational interface. [51]

Of course Rasa-as-a-Service is not free, same as other versions of Rasa, except for the open-source one. Rasa is available in three different versions, catering to different needs. The open-source version is free and provides the core features for building and training conversational interfaces. However, for those who need additional functionalities and premium support, Rasa offers a Pro version and an X/Enterprise version. These versions come with features such as collaborative low-code GUI or analytic tools. Rasa Pro also offers an AI-powered voice solution provided by AudioCodes VoiceAI Connect, which enables communication with the conversational interface through voice. [51]

Rasa's customizability, flexibility, and community makes it a popular choice among developers looking to build complex and highly customizable conversational interfaces. However, the learning curve and technical expertise required to use it effectively may pose a challenge for those new to development in this field. [3]

[3] Another framework similar to Rasa is DeepPavlov, which is an open-source conversational AI framework as well. However, for the purposes of this text, it is not necessary to delve into DeepPavlov as the insight would not offer any additional significant information.

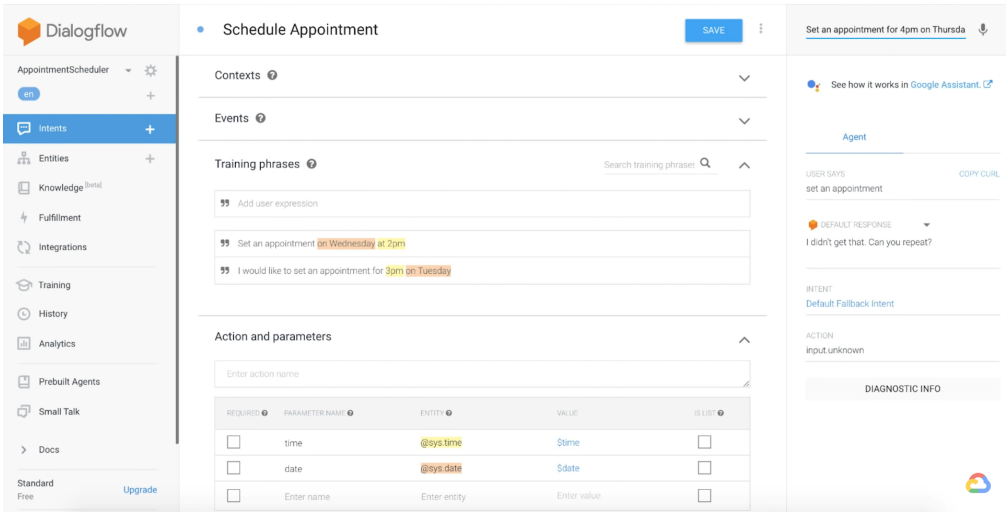


Figure 4.1 Dialogflow's GUI [53]

4.1.2 Dialogflow

Google's Dialogflow is an AI-powered development platform that allows businesses and developers to create natural language processing capable conversational interfaces for various purposes, from customer support to sales and marketing.

One of the biggest advantages of Dialogflow is its strong entity recognition ability. This means that it can recognize and extract important information from user input, such as dates, times, locations, and other types of data. This makes it easier for developers to create conversational interfaces that can understand and respond to users' requests more accurately.

Dialogflow also integrates with Google Assistant, allowing businesses to create conversational interfaces that can interact with users on a variety of Google-enabled devices. Google Assistant is available on over 500 million devices, including smartphones, tablets, smart speakers, smart displays, and even cars and watches. This means that businesses can create conversational interfaces that can reach a wide audience and provide customer support and other services across multiple platforms. This integration also enables businesses to leverage other Google services, such as Google Calendar and Google Chat, to enhance its functionality. [52]

For larger companies with larger customer base, more complex and demanding requests handling more intricate service offerings comes a Dialogflow CX that provides a more advanced solution. Dialogflow CX offers features such as versioning and collaborative options, which allow teams to work together on conversational interface's development and management.

One major advantage of Dialogflow is that it is a hosted online platform with a graphical user interface. This means that businesses don't need to worry about hosting or server management, as Google takes care of all the hosting and maintenance. All this makes it easier for businesses and developers to get started with development, as they can simplify building their conversational interface without worrying that much about the technical details.

4.1.3 Amazon Lex

Amazon Lex is a versatile platform for creating conversational interfaces integrating deep learning technologies that power Amazon Alexa. The platform enables developers to build conversational interfaces in a cost-effective and scalable manner. By providing a few example phrases, Amazon Lex constructs a natural language model that understands related phrases, making it straightforward for developers to create conversational interfaces in just minutes without upfront costs. [54]

Similar to Google Dialogflow, Amazon Lex offers a comprehensive solution for building both voice and text based conversational interfaces. However, it integrates seamlessly with Amazon Web Services (AWS) infrastructure, such as AWS Lambda, which allows developers to build highly scalable and flexible applications on the cloud that can respond to events in real-time. [54] On the other hand, entity recognition in Amazon Lex may not be as robust as in Google Dialogflow.

A notable feature of Amazon Lex is the "Automated Chatbot Designer", which uses machine learning to analyze conversation transcripts between callers and agents, semantically clustering them around common intents and related information. This feature can assist in designing conversational interfaces more efficiently, reducing manual effort and human error, and consequently improving the end-user experience. This functionality could be particularly useful for call centers and larger businesses with existing recorded traffic. [55]

4.1.4 Voiceflow

Voiceflow is a conversational platform that stands out for its exceptional GUI and comprehensive documentation. While some other platforms do offer GUIs, Voiceflow provides dialogue trees as graphs with nodes that showcase the flow of conversations, providing a clearer overview of the conversation flow compared to other options.

One of the advantages of Voiceflow is its ability to import and export NLU data directly from and to other conversational platforms. This feature allows users to leverage their existing NLU data, such as intents, utterances and entities to kick-start their designs more efficiently.

Furthermore, Voiceflow offers various types of nodes for different functions, such as making API calls to other services or displaying a menu with buttons in the chat for faster responses. However, for a more seamless experience in sending commands without resorting to workarounds in response messages, additional types of nodes may be required.

4.1.5 Flowstorm

Flowstorm is a conversational platform that, while similar, differs from VoiceFlow in several aspects. In Flowstorm, there is no need to create and save new intents in a list. Instead, designers can simply write potential user inputs directly in the node. This approach, apart from being smoother and quicker, is particularly suitable for social topics where user inputs are often unique and cannot be reused. In spite of that, of course that pre-prepared intents for common inputs such as yes/no responses are available.

Flowstorm is well-suited for designing large and complex dialogues, offering the option to create a hierarchy through subdialogues. To further enhance clarity and facilitate navigation in complex dialogues, Flowstorm utilizes color-coded nodes.

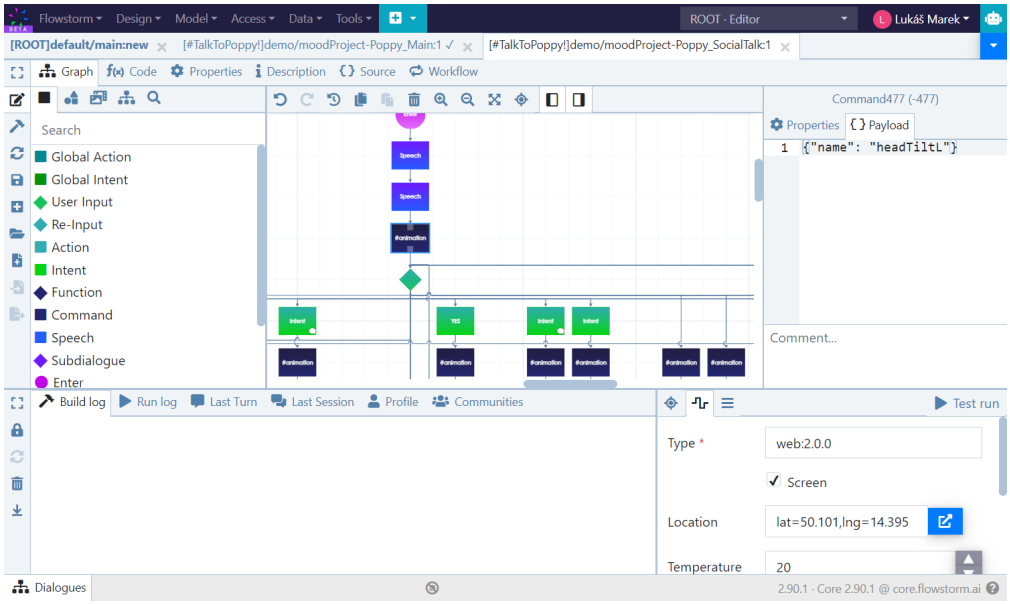


Figure 4.2 Flowstorm's GUI

Similar to VoiceFlow, it can interact with external systems and use functions, but it uses Kotlin instead of JavaScript. Additionally, Flowstorm provides a dedicated node type for sending metadata, such as commands for controlling animations in the contex of digital personas, eliminating the need for workarounds.

One notable distinction between Flowstorm and other platforms is that it is completely free, even though it demonstrates several advantages over competing platforms in certain aspects as shown in figure 4.2

	Flowstorm	Rasa	DeepPavlov	Amazon ASK	Google Dialogflow	Voiceflow
SaaS	✓			✓	✓	✓
Open-source	✓	✓	✓			*p*
Visual editor	✓				*p*	*p*
Custom code	✓	✓				
ML-based NLU	✓	✓	✓	✓	✓	✓
Sub-dialogues	✓					
Sharing assets	✓					✓
Built-in NRG	✓					
Analytic tools	✓			✓	✓	✓

Figure 4.3 Flowstorm compared to other options [56]

4.1.6 ChatGPT

ChatGPT, despite its powerful capabilities and current popularity, is not a conversational platform in the traditional sense. It lacks the ability to control its responses, and there is no option for sending commands to manage animations or other functionalities. This limitation makes it unsuitable for use in domains such as marketing, HR, therapeutic psychology and many other where professionals often require precise control over their conversational interfaces.

However, ChatGPT can be integrated within conversational platforms. Both Voiceflow and Flowstorm, for example, employ ChatGPT to enhance their capabilities. By leveraging ChatGPT, conversational designers can generate a larger variety of utterances and speech outputs fast. This added robustness and variability helps prevent repetition and can contribute to a more natural and engaging user experience. In this way, ChatGPT can help these conversational platforms stand out from their competitors by improving the quality of their conversational design.

4.2 3D tools

The following 3D tools will be explored and considered for creating the digital persona's body and overall physical appearance.

4.2.1 Maya

Maya, initially developed by Alias Systems Corporation in 1998 and later acquired by Autodesk in 2005, is a widely-used 3D computer graphics software, primarily known for its exceptional animation capabilities. It has become a popular choice among professionals in the fields of animation, visual effects, and game development. Maya is highly versatile, offering customization and extensibility options through built-in scripting languages such as MEL (Maya Embedded Language) and Python. [57]

While Maya provides a variety of tools for tasks like modeling, sculpting, and texturing, its true strength lies in animation. It has an advanced rigging, scripting and skinning system, as well as extensive animation tools and features, including inverse kinematics. These features allow animators to work efficiently and achieve high-quality results.

Maya's popularity extends beyond just animation, as it is also a widely-used tool for creating visual effects in blockbuster movies, TV shows, and video games. In order to achieve high-quality, photorealistic or stylized visuals, Maya supports various rendering engines, such as Arnold, RenderMan, and V-Ray. These rendering engines are not designed for real-time rendering and are computationally intensive, often taking advantage of rendering farms to produce final output. However, they offer the ability to achieve high-quality results, utilizing advanced techniques like ray tracing to simulate realistic lighting and reflections. [58]

4.2.2 Blender

Blender is a versatile, open-source 3D software known for its all-encompassing capabilities and extensive community support. As a free program, Blender offers an accessible alternative to many commercial applications, providing a wide range of tools for modeling, sculpting, texturing, rigging, animation, and rendering. However, its jack-of-all-trades nature means that although it can perform a wide array of tasks, it may not always be the top choice for every specific task when compared to specialized software. Blender is not as widely used in larger studios as some other specialized software options. This can be attributed to the software's ever-changing nature, which can make it difficult for large teams to maintain consistency and compatibility.

However, in recent years, Blender has seen significant improvements and increased interest from larger organizations, such as Ubisoft, which have invested in its development. These investments have helped to enhance Blender's capabilities and make it more appealing to professional artists and large studios. Nonetheless, Blender's primary appeal remains its status as a free, open-source, and versatile tool that enables artists to explore and utilize a wide range of 3D tasks without needing to invest in multiple specialized software packages. [59]

One of Blender's key advantages is its large and active community, which contributes to the software's development and provides a wealth of resources, including tutorials, forums, and add-ons. This collaborative atmosphere fosters rapid innovation and frequent updates, ensuring that Blender remains up-to-date with the latest industry trends.

4.2.3 ZBrush

ZBrush, developed by Pixologic, is a digital sculpting software known for its remarkable ability to handle millions of polygons with ease, making it an industry favorite for high-detail sculpting. Its powerful sculpting tools and features are arguably the best in the market, enabling artists to create intricate and detailed models with unparalleled precision.

One remarkable feature of ZBrush is its ability to work with multiple levels of detail (LODs) already during the modeling process. This feature simplifies the creation of final LODs for the model, streamlining the workflow and reducing potential issues. Though primarily a sculpting tool, ZBrush also offers some rigging and hand-painting texture options, making it a versatile choice for certain 3D tasks. Its GUI can be seen back in figure 1.2b depicting organic modelling.

4.2.4 Marvelous Designer

Marvelous Designer is a groundbreaking 3D modeling software that has become the industry standard for creating realistic clothing in films, video games, and other digital media. In the past, everything, including human faces, was hard surface modeled due to the limitations posed by computing power. As technology advanced, computers became capable of handling more polygons, allowing artists to simulate clay and sculpt intricate details. Today, with even more powerful computers, artists no longer need to sculpt clothing. Marvelous Designer allows them to design patterns and cuts of fabric, which can then be sewn together and draped over characters using cloth simulations. This workflow is similar to that of real tailors. This innovative approach to clothing design has been widely adopted in Hollywood and other major entertainment industries[4], resulting in remarkably lifelike garments that enhance the overall visual appeal and realism of digital characters.

4.2.5 Substance 3D Painter

Substance 3D Painter, now owned by Adobe, was originally developed by Allegorithmic under the name Substance Painter. In the past, texturing was predominantly carried out on a 2D canvas, where an unwrapped topology of the 3D model was overlaid and texture artists painted and placed images accordingly. However, this method became increasingly impractical as the complexity of models and the number of unwrapped islands grew.

Adobe Photoshop has been used for texturing, but the emergence of tools like Quixel SUITE[5], which allowed for more realistic texturing and 3D painting directly onto the 3D model, began to challenge its dominance. However, being a Photoshop plugin, Quixel SUITE developed a reputation of being buggy. This paved the way for Substance 3D Painter to become the leading choice as a standalone software in the texturing realm.

Substance 3D Painter offers procedural texture generation based on the model's properties, utilizing textures such as curvature, ambient occlusion, thickness, and world position. It also employs "smart materials", which can be created in Substance 3D Designer – another software in the same family of texturing tools. The combination of these features has made Substance 3D Painter a powerful and versatile solution for artists seeking to create realistic and detailed textures for their 3D models.

[4] but also by fashion designers outside the entertainment industry
[5] and many before it

4.2.6 Unity

Unity developed by Unity Technologies is a widely popular game engine, while robust and versatile, its simplicity and ease of use has made it a go-to choice for game developers across various platforms, ranging from mobile devices to consoles and PCs. It provides a comprehensive platform for creating both 2D and 3D games, as well as other interactive content such as simulations and both virtuall and augmented reality experiences. One of the key factors contributing to Unity's success is its choice of programming language: C#. This modern, object-oriented language is easy to learn and offers a powerful, flexible foundation for game development.

Another crucial aspect of Unity's appeal is its extensive community of developers. As the largest and most active community in the game engine world, Unity users benefit from a wealth of resources, including tutorials, forums, and Unity Asset Store. This strong network of support empowers developers to learn, collaborate, and troubleshoot more effectively, thereby speeding up the development process and improving the overall quality of their projects. [60]

4.2.7 Unreal Engine

Unreal Engine, developed by Epic Games, is a cutting-edge game engine with a strong reputation for its exceptional visual capabilities. While it has a smaller community compared to Unity, it still boasts a substantial following thanks to its outstanding performance and features. Unreal Engine has found applications beyond the gaming industry, including architectural visualizations, automotive industry and even the movie industry, such as in "Rogue One: A Star Wars Story" for real-time rendering and "The Mandalorian" for StageCraft technology. StageCraft is a groundbreaking virtual production tool which combines physical sets with real-time rendered digital backgrounds on LED walls to create immersive environments for actors and filmmakers. [61], [62]

Epic Games are known for investing in innovative technologies related to Unreal Engine that push the boundaries of 3D graphics. Examples include Lumen, a new real-time global illumination rendering engine. Nanite, a virtualized geometry system[6] for managing LODs of high-polygon models and the Metahuman Creator, which allows the generation of high-quality, 3D scan-like human characters.

However, Unreal Engine's impressive capabilities come with a level of complexity. It relies on C++, a more intricate programming language than Unity's C#, as well as its own visual scripting language called Blueprints. Blueprints can both increase and decrease the complexity for developers, depending on their role, task, familiarity with the language and the specific requirements of their projects, making development process both more accessible and challenging at the same time. Despite its complexities, Unreal Engine remains a highly sought-after choice for those looking to create visually stunning, cutting-edge projects across various industries.

4.2.8 Metahuman Creator

Metahuman Creator is an innovative tool developed by Epic Games for Unreal Engine that allows users to generate highly realistic, 3D scan-like quality human characters for use in their projects. One of its standout features is that it is streamed, meaning that it does not require powerful hardware to run, making it accessible to a wide range of users. However, it's important to note that while Metahuman Creator is free to use, it is exclusive to Unreal Engine.

[6] it intelligently streams and renders only the necessary polygons and also "*it intelligently does work on only the detail that can be perceived and no more*" [63]

4.2.9 Character Creator

Character Creator is a tool developed by Reallusion for generating digital characters. While it may not produce results as visually realistic as those from Metahuman Creator, it offers more customization options and a less computationally demanding output, making it suitable for a wider range of projects.

One of the key differences between Character Creator and Metahuman Creator is the hair. While Metahuman Creator uses a groom system that allows for individual hair strands to be rendered, Character Creator relies on more traditional hair cards. Hair cards are essentially planes with opacity maps applied to them, resulting in a more optimized and performance-friendly hair representation, although not as realistic as the groom system.

Character Creator is not a free tool, and users are required to pay for access to its features and content. Additionally, unlike Metahuman Creator, which runs on the cloud, Character Creator requires developers to have their own hardware to run the software.

4.2.10 iClone

iClone by Reallusion, like Character Creator, is primarily an animation tool that streamlines the animation workflow, making it easier for users to create high-quality animations. While some argue that the tool is not as powerful as other options, its extensive animation store and user-friendly features make it a popular choice for many developers.

One of iClone's strengths is its ability to post-process and clean motion capture data, which can be invaluable when working with raw motion capture data[7]. Additionally, the software includes a feature that can generate lip sync animations from an audio file with transcribed text. If the transcribed text is not provided, iClone can perform speech-to-text conversion, although this may result in some errors that need to be corrected by the developer using it.

The generated lip sync animation can be further refined by the developer, but it's important to note that this process is not real-time and can take a significant amount of time to complete. Furthermore, integrating this feature into a real-time application may require some workarounds.

4.2.11 Omniverse Audio2Face

Nvidia's Omniverse Audio2Face is an AI-powered solution that delivers unparalleled lip sync quality, surpassing any other available technology. It not only generates highly accurate lip sync animations but also incorporates emotions, creating corresponding facial expressions that enhance the realism. Despite its impressive capabilities, Omniverse Audio2Face has some drawbacks. The technology is not yet production-ready and demands significant computational resources.

[7] which is essentially a collection of 1D signals, similar to multi-channel audio. This means that similar techniques can be used for reducing noise and other processing. The analogy between sound and animation is certainly interesting.

4.2.12 Oculus Lipsync

Oculus Lipsync is a real-time lip sync solution originally developed by Oculus for use in VR social interactions between users. It is designed to generate lip-sync animations directly from microphone input, making it an excellent choice for real-time applications.

In addition to real-time lip syncing, Oculus Lipsync can produce higher-quality lip sync animations from audio files, although not in real-time. To achieve this in real-time, a workaround must be implemented to override the microphone input and use data from an audio file instead. Despite this minor limitation, Oculus Lipsync is a lightweight and production-ready solution.

4.3 Sound tools, sources and services

This section presents an overview of various sound tools, sources, and services that can be employed. The overview includes an exploration of methods for obtaining suitable background music, sound libraries, and essential text-to-speech and speech-to-text services. The aim is to provide a comprehensive understanding of the available options and their potential use.

4.3.1 Digital audio workstations (DAWs)

Digital Audio Workstations (DAWs) are software environments for music production that have become indispensable in the creation of commercially released music since the early 2000s. They are widely used by professional recording engineers in studios and by bedroom producers alike. The rise of DAWs can be attributed to the rapid increases in computing power, which has revolutionized the ability to handle digital audio, and the unique creative possibilities it offers that have significantly transformed music practices. DAWs offer opportunities impacting both the conception and organization of musical ideas that, once understood, can be harnessed to great effect. They have become instruments in their own right, much like the piano or guitar in previous eras of songwriting. [64]

DAW typically provides a visual environment represented graphically on a computer screen, allowing users to manipulate two main forms of information:

- **MIDI (Musical Instrument Digital Interface):**
  - Protocol for communication between digital musical instruments and computers
  - MIDI data contains:
    - Note on/off messages (pitch, velocity, and duration)
    - Control changes (volume, modulation, etc.)
    - Program changes (instrument selection)
    - System messages (timing, synchronization)
- **Audio (Audio data):**
  - Representation of sound in a digital format
  - Audio data contains:
    - Waveform data (sampled amplitude values of sound waves)
    - Sample rate (number of samples per second)
    - Bit depth (accuracy of each sample, typically 16 or 24 bits)
    - Channels (mono, stereo, or surround sound) [64], [65]

The design of the DAW interface significantly affects creative decision-making and workflow. These software platforms often share common elements, such as the main sequencer interface, mixer, “piano roll” for MIDI editing, waveform display, and traditional score. DAWs often model electronic music technology of the past, including samplers, drum machines, synthesizers, signal processors, and effects. Some well-known DAWs are Apple’s Logic Pro, Avid’s Pro-Tools, Propellerhead’s Reason, Ableton’s Live and many more. [64]

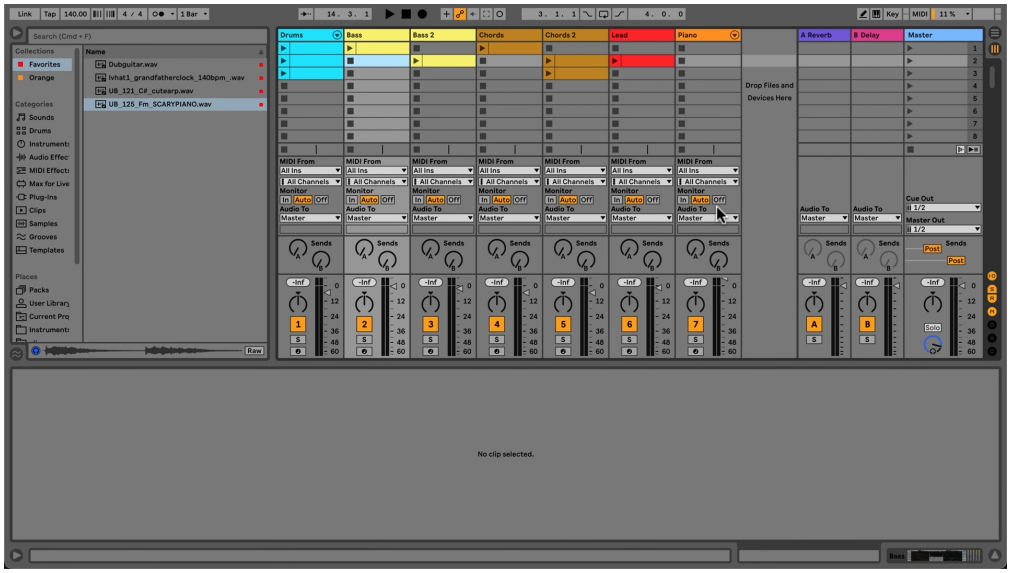


Figure 4.4 Ableton Live session view, no waveform display, just colorful playback buttons [66]

4.3.1.1 Ableton Live

Ableton Live is state of the art DAW that has gained popularity among musicians, producers, and performers due to its unique features and capabilities that cater to various aspects of music creation and performance. This section will discuss the distinguishing characteristics of Ableton Live, highlighting its innovative design and integration possibilities.

A distinguishing feature of Ableton Live is the "Session View" shown in figure 4.4, which differs from traditional linear DAWs by not relying on a single timeline. Instead, it allows users to experiment with ideas, mix and match them, and control live performances. Each clip can be started and stopped independently while remaining in sync, providing a conducive environment for idea generation, song structure development, and live performance management. Of course, in addition to the Session View, Ableton Live also incorporates the "Arrangement View", which offers a more conventional workspace for recording audio and MIDI takes, splicing them together, and organizing ideas on a timeline. This view enables musicians and producers to refine their performances and create structured compositions. [67] [66]

Ableton Live features an extensive range of sample-based instruments, synthesizers, and effects for creative sound design. The software includes packs from renowned sound designers, such as Spitfire Audio, as well as Live's Curated Collection series, providing users with diverse options to expand their sonic palette. Apart from that, Ableton Live can function as the central hub for various music setups, facilitating the recording of audio sources like guitars, vocals, and drums. Ableton Live can even dynamically adjust its tempo based on incoming audio, effectively integrating itself into the performance. Users can perform with clips in Session View, manipulate sounds, and remix compositions in real-time. The software also allows users to record and loop audio on-the-fly, providing opportunities for improvisation and unique performances. [68]

It's integration capabilities are worth noting. "Max for Live", an integrated feature of Ableton Live, lets you customize your experience by creating your own devices and even controlling physical objects like lights and motors. This opens up a world of possibilities for live performances, installations, and experimental projects that go beyond the realms of traditional music production. [67]

4.3.1.2 Audacity

Audacity is an open-source audio editor that offers a range of features for recording, editing, and processing audio files. Although its capabilities as a digital audio workstation (DAW) are debatable due to its limited multitrack and MIDI functionality, Audacity remains a valuable tool for various audio-related tasks.

One of Audacity's primary function is the ability to record live audio while providing basic editing tools for manipulating audio files, such as cutting, copying, pasting, deleting, duplicating, and splitting. These tools allow users to refine their recordings, remove unwanted segments, or rearrange sections to create cohesive compositions. Users can modify the speed, pitch, or volume of recordings and apply several built-in effects that can be applied to any part of the audio, enabling users to enhance their recordings or achieve specific sonic results. These effects range from simple equalization and compression to more complex reverberation or distortion. [69]

While Audacity has limited MIDI capabilities, it does support the importation of MIDI files as "Note Tracks". Users can listen to these tracks using a basic electronic piano sound and convert them into audio, albeit through a slow process. This functionality offers minimal MIDI support but may be sufficient for some user's needs. [70]

Audacity, although not widely considered a full-fledged DAW, offers a range of features and capabilities that are valuable for various audio-related tasks. Its functionality in recording, editing, and processing audio has made it a popular choice for podcast creators and those in need of a straightforward audio editing solution.

4.3.2 Audio libraries

In the Conclusion of sound section, the possibility of using an emotion-annotated dataset of 80s style music as an inspiration was mentioned. However, composing original music might not be the most efficient approach. Various sources offer quality music, some even for free. The following platforms are worth noting:

- **SoundCloud:** lacks ambient tracks and primarily features vocals, making it less suitable for dialogue underscoring
- **Free Music Archive:** offers limited number of genre categories and no mood filter
- **PremiumBeat:** provides a mood filter but has a small selection of free music
- **YouTube Audio Library:** features free music with genre and mood filters
- **Pixabay:** contains free music and an extensive range of filters

Comparing Pixabay and YouTube Audio Library, Pixabay offers fewer tracks but more accurate search results. On the other hand, YouTube Audio Library provides a larger selection, which might require more time to choose from but offers more variety. The mood filters are particularly useful since the focus, in this case, is on the mood evoked rather than the genre of music.

A potentially efficient approach to acquiring music could involve selecting tracks from these audio libraries and modifying them only as needed, such as looping them to create an infinite ambience.

4.3.3 Text-to-speech and speech-to-text

Incorporating text-to-speech (TTS) and speech-to-text (STT) services in sound tools is crucial for the development of digital personas. These services enable digital personas to hear user speech input and respond using synthesized speech, which is particularly important for conversational platforms that rely on text or utilize these or similar services. It is worth noting that conversational interfaces that communicate with users using voice tend to have a more distinct personality compared to those that communicate through text only. [71] Several notable options for the design phase are summarized below. They all share some common features: support for viseme[8] output, the possibility to create custom voices (although this is not a simple process and is often experimental), and the use of SSML[9]. Differences between these options mainly lie in their pricing models.

Amazon:

- **Transcribe:** offers 60 minutes of STT services for 12 months. [72]
- **Polly:** provides a free tier with 5 million characters per month for speech or speech marks requests in the first 12 months, and 1 million characters per month for Neural voices. Viseme output and custom voice support are available. [73]

Google:

- **Speech-to-Text:** provides 60 minutes of free transcribing and analyzing audio per month. [74]
- **Text-to-Speech:** allows 1 million characters for WaveNet voices and 4 million characters for Standard voices each month for free. Supports viseme output and offers an experimental beta version for training custom voices. [75]

Microsoft Azure:

- **Speech-to-Text:** offers 5 free audio hours per month. Supports the creation of custom speech models for improved performance in noisy environments or for domain-specific jargon, which is particularly useful for police officers and other emergency first responders. [76]
- **Text-to-Speech:** provides 0.5 million characters for free per month. Supports viseme output (currently available only for en-US neural voices) and custom voice training for responsible use. [77]

By comparing the services based on price, Amazon Transcribe and Google Speech-to-Text both provide 60 minutes of free service per month, while Microsoft Azure Speech-to-Text offers 5 free audio hours per month. For TTS services, Google Text-to-Speech provides the same number of free characters per month as Amazon Polly (5 million, Google divides it into 1 million characters for WaveNet voices and 4 million characters for Standard voices) followed by and Microsoft Azure Text-to-Speech (0.5 million).

Another service to consider is the vosk-language-server, which operates offline and provides speech-to-text capabilities. However, this service must be run alongside the final application or hosted separately, and while it offers the advantage of offline functionality, the accuracy of its results may not be as high, compared to other cloud-based services. [78]

[8] Viseme is a facial expression representing movement, or rather the shape of the mouth when making a particular sound or phoneme, in our case implemented as morph targets. TTS services can return the names of these visemes with a time tag.

[9] by using SSML tags in the text, developers can add pauses, emphasize specific words or phrases, and modify intonation in TTS services outputs

4.4 Proposed tool-set

The optimal pipeline for creating a 3D character would involve using ZBrush for sculpting the face and body, Marvelous Designer for clothing, Substance 3D Painter for texturing, and Maya for rigging and animation. However, the author, who has experience in hard surface modeling, is not proficient in human anatomy required for sculpting humans, nor is he a rigger or animator. Therefore, a more efficient tool-set is proposed.

The author opted to use a character generator, which consistently produces high-quality results and handles the rigging process. Metahuman Creator was chosen over Character Creator due to its free access, more user-friendly interface, and more realistic results. The limited customizability of Metahuman Creator is not a significant drawback in this case, as the showcase application does not require a large number of visually unique digital personas. This choice also implies the selection of Unreal Engine over Unity, as Metahumans are exclusive to Unreal Engine. Unreal Engine offers superior visuals compared to Unity. Metahumans also utilize Unreal Engine's standardized skeleton, which allows for easy use of animations from the Unreal Engine Marketplace without the need for extensive animation retargeting. Moreover, Unreal Engine provides some animation capabilities itself for further modification if needed.

The digital persona KAI (Konversational AI) is a Metahuman resculpted in Blender and retextured in Substance 3D Painter to convey an artificial, robot-like appearance. This design choice helps to distinguish him from other personas and highlights the use of a generative model developed at PromethistAI, as opposed to traditional dialogue trees. All visemes for all personas required for lip sync were created as morph targets in Maya. Additionally, the author created the wave animation used by Poppy and Seb to greet users.

Regarding music, the author lacks musical education and relied on free music libraries to obtain suitable tracks for the research described later. Tracks were chosen by the author based on their perceived appropriateness, and potential users were then asked to rate them in an online survey. The tracks with higher median scores were selected, although no statistically significant difference was found between them.

Employing this particular tool-set and workflow made the development process faster and more efficient without requiring expertise in multiple distinct professions. This method eliminated the need for reinventing the wheel and optimized the author's work. The content creation process has been presented now, and the inner-workings, the way in which different elements come together and other design choices within the final solution will be discussed in future chapters.

4.5 Existing solutions

In this section, situated between the analysis and design chapters, the aim is to explore existing solutions in order to understand their functionality and identify potential improvements for the proposed solution.

The approach to dialogues in games, and other interactive applications, has remained largely unchanged since the early days of gamebooks. For instance, gamebook called "An Examination of the Work of Herbert Quain" dates way back to the 1941. This gamebook contained a three-part story with two branch points, resulting in nine possible readings. Gamebooks function by offering readers multiple paths through the narrative, allowing for non-linear exploration. For example, if a reader encounters a choice between a right or left path at page 88, they can turn to page 238 for the right path or page 165 for the left path. Reading pages chronologically does not reveal a coherent narrative. [79]

Given that the Unreal Engine appears to be the most suitable game engine for this task, the analysis will focus on dialogue solutions in the context of Unreal Engine. However, it is worth noting that these solutions do not significantly differ from those found in Unity or any other game engine. Several existing solutions found on the Unreal Engine Marketplace are listed below:

- "Ascent Dialogue System - C++ Visual Tool for Branched Dialogues" [80]
- "Dialogue Component" [81]
- "Simple Dialogue System" [82]
- "Dialogue Plugin" [83]
- "Dialogue System X" [84]

The solutions listed, and dozens more, offer various features and usability aspects, but their core functionality remains the same. Some solutions offer a visual tool for creating dialogues built directly into Unreal Engine. Many are implemented as components, which is a favorable approach, as the author's solution will likely be developed as a component as well[10]. Certain solutions support variables in text (e.g. using a user's name saved in a variable) and voiceovers, while others provide a gameplay-friendly event and condition system. This functionality is essential for controlling animations and thus non-verbal communication[11]. Random replies in dialogues found in some solutions are also desired, as they contribute to more natural and less repetitive reoccurring conversations[12].

However, as already stated the core concept of these solutions has not changed significantly since the gamebooks era, relying on pre-prepared replies (figure 4.5). Although this approach can be engaging in story-based adventure games, it may not be suitable for non-gaming applications such as therapeutic sessions, casual conversations, or simulations where capturing real user input is crucial for realism. For applications, beyond gaming, where understanding user intent is vital, a shift towards more realistic and adaptive conversational systems is necessary.

There are existing solutions that attempt to incorporate natural language processing into dialogue systems, aiming to enhance their capabilities and create more natural interactions. However, some of these solutions suffer from high interdependence between various modules outside the project itself or are too complex to use and might require some knowledge of machine learning.

One such project is "FANTASIA", a "Framework for Advanced Natural Tools and Applications with Social Interactive Agents". Developed at the University of Padua and the University of Naples Federico II in Italy. A collection of tools designed for use with Embodied Conversational Agents. Despite its academic origins and potential usefulness, the complexity of FANTASIA may pose challenges for those without a background in machine learning or related fields. [85]

Another project that incorporates natural language processing is "VR-Vendor", which uses Amazon Lex for natural language understanding and processing. It offers support for Unreal Engine 4 and some limited for Amazon's Lumberyard[13]. While VR-Vendor demonstrates the potential of integrating natural language processing into dialogue systems its unusual reliance on external modules makes it unsuitable for any real use. [86]

[10] the concept of a component will be elaborated upon in the implementation chapter

[11] which is why Flowstorm appears to be the most suitable conversational platform for the proposed solution, as it allows for command-based event triggering

[12] a feature commonly found in conversational platforms

[13] Lumberyard is a game engine that has its origins in Crytek's in-house game engine, CryEngine. Although CryEngine was released to the public some time ago and offers impressive graphical capabilities, it has not managed to attract a large community.

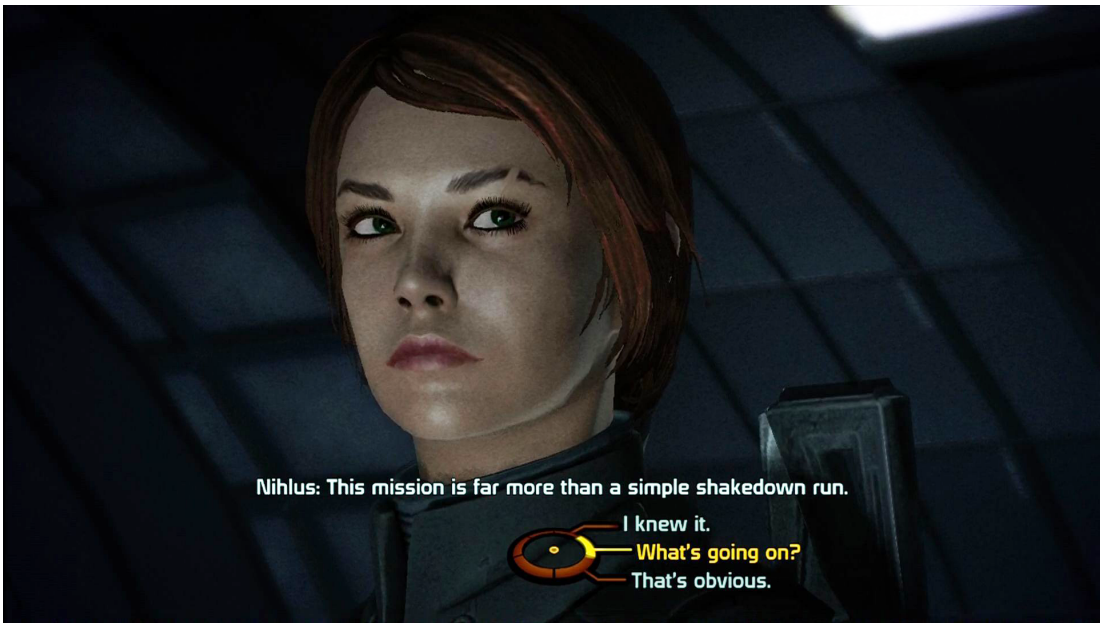


Figure 4.5 Pre-prepared replies in the video game called Mass Effect [87]

Chapter 5

Design

In this chapter, the design phase of the solution is presented, building upon the knowledge gained from the analysis phase. The main focus of this section is to provide a comprehensive understanding of the overall structure and functioning of the solution, while the implementation details will be covered in the subsequent chapter. The design phase is structured to cover several aspects, including functional and non-functional requirements, component diagram, use case model, business process model, and domain model.

The following functional and non-functional requirements have been identified for the solution:

- **Functional Requirements:**
  - The solution must facilitate conversations between users and digital personas in an Unreal Engine.
  - The solution must be able to handle multiple users concurrently.
  - The user should have the option to choose the topic and persona with which they want to converse.
  - The solution must support voice-based conversational exchange, utilizing speech-to-text and text-to-speech services.
  - The solution must be able to trigger various events, such as animations, sound effects and so on based on the conversation.
- **Non-Functional Requirements:**
  - **Modularity:** The solution must be adaptable to various configurations, allowing developers to choose whether to include features such as lipsync, speech-to-text services, or some specific event triggers.
  - **Simplicity:** The solution should be simplistic, enabling developers to modify or expand its functionality easily as required for their specific use case.

These requirements describe the resulting application and ensure that the solution is capable of catering to a wide range of use cases in the future. The application's core component, the Flowstorm Link plugin serving as a connector, is designed to be minimalistic yet versatile and easily adaptable, enabling seamless communication with the Flowstorm conversational platform and facilitating various types of interactions within the Unreal Engine.

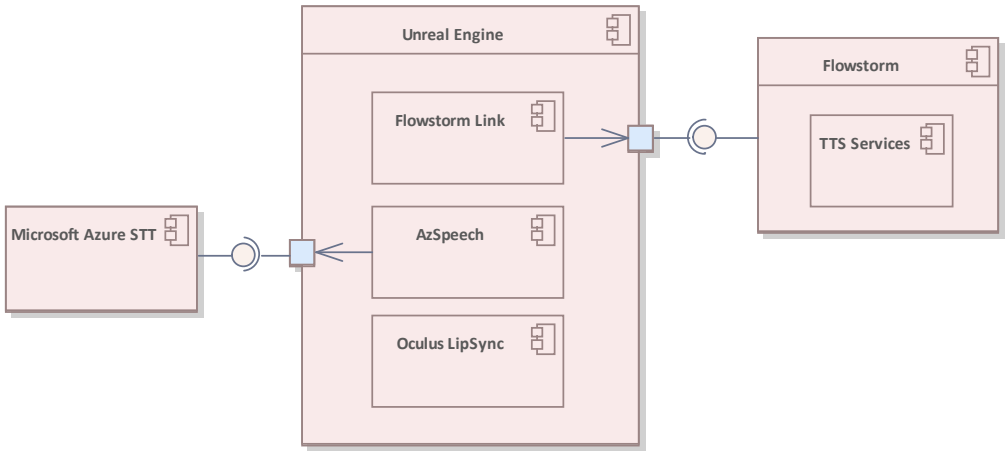


Figure 5.1 Component diagram

5.1 Components

The architecture of the proposed solution is based on a component diagram, which illustrates the interactions between different systems and subsystems that work together to create a comprehensive application. The core component of this solution is Unreal Engine, a game engine that serves as an environment for the digital persona, encompassing graphics, code, sound, and other multimedia elements. Unreal Engine is where the digital persona exists.

Unreal Engine communicates with Flowstorm, a conversational platform, through an internal subsystem called Flowstorm Link. Flowstorm Link, a plugin, and the baseline of this thesis, enables the integration of Flowstorm as a next-generation conversational system which utilizes natural language processing into Unreal Engine. Flowstorm also incorporates Text-to-Speech services, simplifying the process of providing voices for digital personas.

Another external system employed is Microsoft Azure Speech-to-Text, which transcribes user voice input. The main advantage of using Azure STT lies in its ease of integration within Unreal Engine through an internal subsystem, the AzSpeech plugin. AzSpeech provides functionality for capturing user voice, detecting silence, and transcribing speech. This silence detection feature facilitates a more natural user experience, as users can simply stop speaking, and the digital persona will respond accordingly, eliminating the need for a push-to-talk approach in which the user has to have a button pressed while speaking.

The final noteworthy and key component is the real-time LipSync plugin by Oculus, an internal subsystem responsible for generating lip-sync animation data in real-time based on the audio files of the digital personas' speech. Manually animating every spoken line is impractical if not impossible, making real-time animation generation essential. An alternative to this plugin is the external system, Audio2Face by NVIDIA, which can be connected to Unreal Engine through built in Live Link. Even though Audio2Face provides a more natural-looking lip-sync animations, it significantly increases performance requirements and presents challenges related to hosting, scaling, and communication. Therefore, the Oculus Lipsync plugin is chosen for its production-ready capabilities and lower performance overhead.

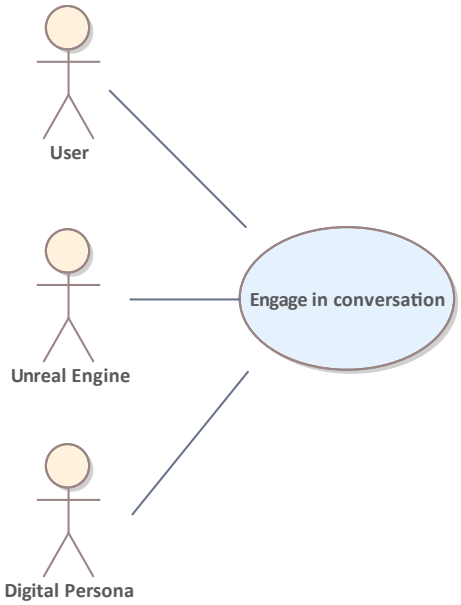


Figure 5.2 Use case model

5.2 Use case and business process model

Use case and business process models help to describe the functionality and flow of the application, as well as the roles and interactions of the different actors involved. Application consists of a single use case: "Engage in a conversation". This use case involves three actors: the User, the Unreal Engine, and the Digital Persona. Although the Unreal Engine and Digital Persona could be combined into a single "Application" actor, they are separated in these models to better distinguish the roles and responsibilities of the Digital Persona as a 3D talking avatar and the Unreal Engine, which manages technical aspects of the application in the background. Digital persona is also more humanized this way and that's also the reason why this finer granularity was chosen.

The application is divided into two main types of levels: the main menu level and the conversation levels. The main menu level serves as the starting point for the user, offering three digital personas to choose from, the user can exit the application from there. Upon selecting a persona, the main menu level is unloaded, and the conversation level with the chosen digital persona is loaded. Conversation level begins with the digital persona greeting the user and initiating a conversation. The user can respond, leading to a turn-based conversation, or remain silent, in which case the digital persona will react accordingly after a while. At any point, the user can return to the main menu level, either to select a different digital persona or to exit the application altogether.

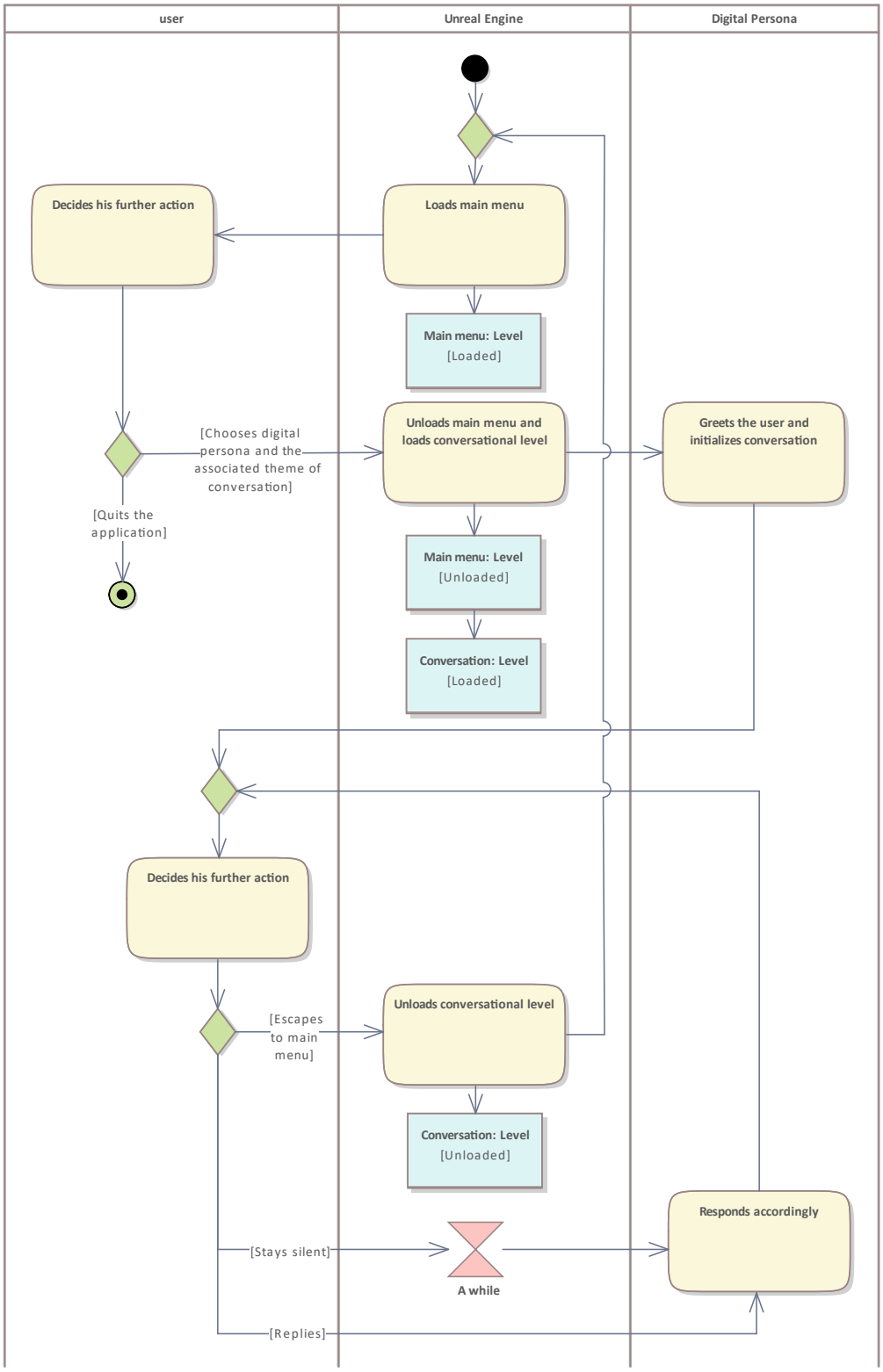


Figure 5.3 Business process model

5.3 Domain model

In the context of the Unreal Engine, a level refers to a distinct environment or scene within the application. Levels play an essential role in organizing the different elements and aspects of the application, providing a structured space for various actors and interactions to take place. Note that the term "actor" here refers to "objects" within the Unreal Engine levels and should not be confused with actors from use case diagrams. The arrangement of these actors helps create a distinct atmosphere and setting for the user experience.

Atmosphere in a level can be influenced by various types of lights, such as directional light, point light, spot light, rect light, and skylight. These are going to be explained later. It is important to note that a level can have any number of lights or none at all, depending on the desired atmosphere.

In addition to lights, atmosphere is also achieved through post-process volumes, which allow for the adjustment of visual effects and rendering settings within a specific area of the level. Like lights, post-process volumes can be added or omitted from a level as needed.

Each level can have its own set of functionality and logic, defining the high-level behavior of actors, as well as the interactions and events that can occur within the level, such as through user interface elements, and any other items relevant to the application. This high-level logic is mediated through a "level blueprint" which will be explained further in the implementation chapter. Each level has one level blueprint.

In Unreal Engine, an actor represents any object that can be placed or spawned in a level, including characters, meshes, lights, sound, cameras, and many other types of objects. Actors are the base class for any interactive object within a level, as they can have internal logic.

A digital persona is a specialized case of an actor, inheriting properties from the base actor class. It contains both internal logic as well as external functionality brought in by components. Actors in the Unreal Engine can be seen as small scenes of their own, containing a local coordinate system, where different objects can be placed and driven by the logic inside the actor. The digital persona has several components, including:

- **Flowstorm Link:** An actor component responsible for the main functionality of the digital persona. *A conversational brain*
- **Oculus Lipsync:** Another actor component, a plugin, providing the ability for the digital persona to open its mouth when speaking
- **Media Sound:** Responsible for playing sound, specifically the digital persona's speech that is downloaded from the conversational platform realtime
- **Subtitles HUD:** A reference to the Heads-Up Display (HUD) used to display a status bar when it is ready for the user to speak.
- **Skeletal Mesh Component:** Provides the digital persona with a body.
- **Camera**: A camera component is included with the digital persona to simplify the process of setting up cameras and their settings in each level.

The digital persona has the ability to perform actions, usually animations, which are received from Flowstorm through the Flowstorm Link plugin. Derived from this "digital persona" parent actor are digital personas named Poppy, Seb, and KAI (Konversational AI), with the only difference being their body, the skeletal mesh component and their topic of conversation. No changes are made to the underlying logic.

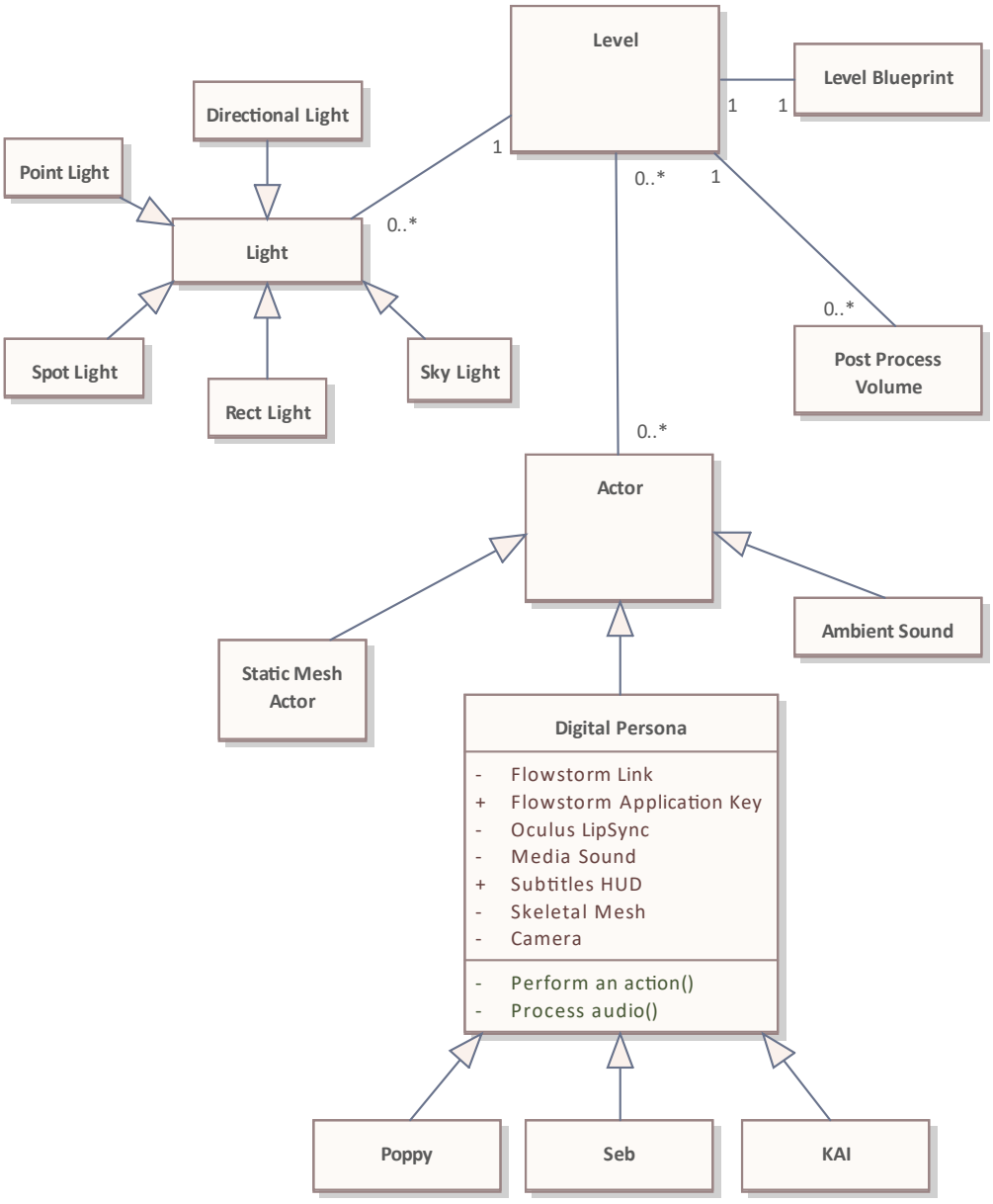


Figure 5.4 Domain model

Chapter 6

Implementation

In this chapter, the detailed implementation of the concepts discussed in the design chapter will be explored. However, before delving into the specifics, it is essential to provide a brief introduction to programming in Unreal Engine.

Unreal Engine offers two primary methods for programming: C++ and Blueprints. C++ provides the foundation for creating various functions and classes, while Blueprints, a visual scripting language, enable the execution of these functions. Blueprints are somewhat similar to the business process model in UML, as they outline the flow of activities within the application. This visual scripting language allows for a better understanding of the processes taking place within the application, making it easier to design and troubleshoot complex interactions.

It is important to note that some functions cannot be triggered directly from C++ because they only exist within Blueprints. In certain cases, it might not even make sense to initiate these functions from code. However, if necessary, an event dispatcher can be implemented to achieve this functionality. Event dispatchers will be explained in the next section, which focuses on the Flowstorm Link component.

Blueprints are particularly popular among 3D artists and other non-programmers due to their relative ease of use. The visual scripting language enables these individuals to implement simple logic themselves. To put it differently, coders can prepare functions for non-programming team members, such as game designers and 3D artists. This enables these individuals to easily integrate the prepared functions into their workflow, enhancing collaboration and streamlining the overall development process.

Components, specifically actor components, play a crucial role in the functionality of Unreal Engine actors. By adding actor components to an actor, new features and capabilities can be introduced, extending the functionality of the blueprint associated with that actor. In this way, actor components allow for the modular addition of functionality to an actor without directly modifying the actor's base class. Concrete components used in digital persona are shown in figure 6.1.

Flowstorm Link, a plugin, the subject of the next section, is implemented as such actor component. By integrating Flowstorm Link as a component, the digital persona, or any other actor, can easily take advantage of the features and functionality provided by the plugin. In the following section, the implementation details of the Flowstorm Link component will be discussed in-depth, shedding light on how the various aspects of the design are brought to life within the application.

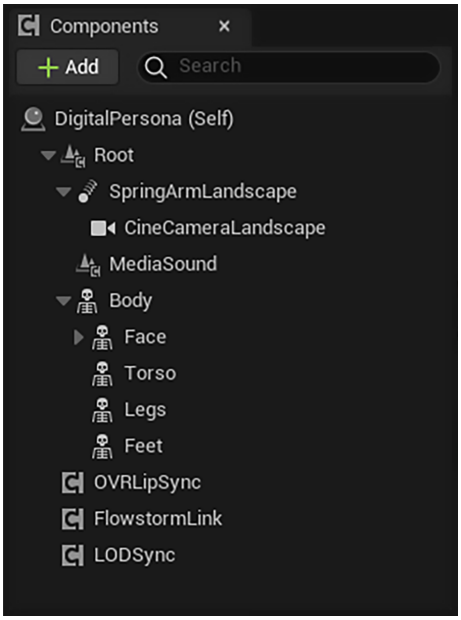


Figure 6.1 Digital persona's components

6.1 Flowstorm Link Plugin

6.1.1 Blueprints side

To provide a better understanding of the overall flow that takes place in Flowstorm Link, let's first explore the Blueprints side of Flowstorm Link before delving into its C++ side. The flow goes from left to right, and to make it readable, the graph in figure 6.2 has been split in half. But before discussing the flow further, it is essential to understand the different types of nodes.

Event dispatchers, hereafter referred to only as "events", in Flowstorm Link are represented by red nodes within a graph. These red nodes function differently from blue nodes, which are functions that execute immediately when the flow reaches them (similar to receiving a token in UML). In contrast, red nodes representing events wait for a signal. Upon receiving the signal, they generate a token. A signal can be sent from either C++ or other Blueprints. In addition to red and blue nodes, there are black nodes, which represent macros. Macros are sub-graphs that have been collapsed into a single node for clarity.

Nodes can have input parameters coming to the left side and output parameters coming from the right side. Class instances, or objects, are represented by small blue nodes attached to functions on the left side [1]. Pink nodes represent strings, while that one green node present represents an enumeration.

- **Red:** Events
- **Blue:** Functions / objects
- **Black:** Macros (sub-graphs)
- **Pink:** Strings
- **Green:** Enumerations

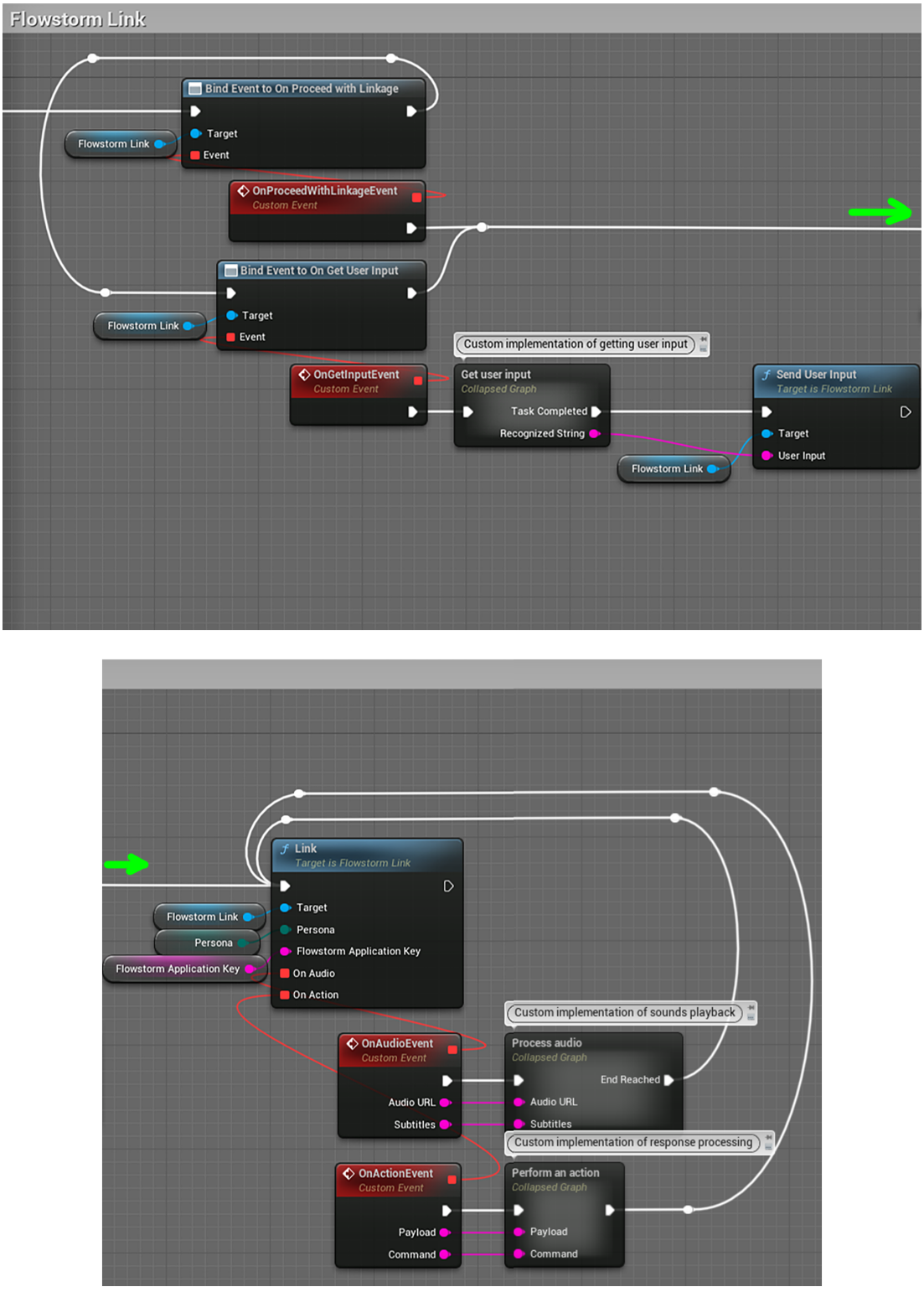


Figure 6.2 Flowstorm Link Blueprints

After binding[2] the two events “On Proceed with Linkage” and “On Get User Input”, as displayed in the upper part of the figure 6.2 of the graph, the flow proceeds to the “Link” function in the lower part of figure 6.2.

Link takes an enumeration input parameter, allowing users to choose from one of three pre-prepared dialogues (Flowstorm Application Keys): Poppy, Seb, and KAI. This option is mainly for testing purposes aimed at first-time users new to the plugin and Flowstorm. However, if a custom Flowstorm Application Key is provided as a string, it overwrites the enumeration, enabling access to a specific dialogue within Flowstorm.

The Link function serves as the primary function for the entire plugin, it operates in a loop, requesting user input when necessary and sending the responses from Flowstorm for further processing within the digital persona. For that, it has two events attached: “On Audio Event” and “On Action Event”. These events function as return mechanisms for the Link function, allowing for the return of different parameters and directing the flow to various parts of the graph as needed. All the events stated are triggered from the C++ side.

The modularity and simplicity of the plugin will become evident with the showcase of three macros in the graph: ”Get User Input”, ”Process Audio”, and ”Perform an Action”. These macros can be replaced with custom solutions tailored to the specific needs of any project. In the case of digital persona:

- **Get User Input:** This macro leverages the AzSpeech plugin, allowing the persona to ”hear” what the user says through Microsoft Azure STT. This process obtains the user input string, which is then sent to Flowstorm.
- **Process Audio:** This macro manages the download and playback of the persona’s speech. It also handles the integration of Oculus LipSync to synchronize the persona’s mouth movements with the speech. Additionally, the transcription is displayed as subtitles on the screen.
- **Perform an Action:** This macro triggers animations, enabling the persona to move, make various expressions, and possibly perform other actions.

Consider a scenario where instead of digital persona the goal is to create a talking vending machine that communicates with the user through an old-school terminal. In this scenario, several relatively simple modifications can be made:

- ***Get User Input:** The “Get User Input” macro can be modified to replace the TTS with a text-based input method in Unreal Engine, as the user will type their requests into the terminal.
- ***Process Audio:** The audio output can be retained, but instead of displaying subtitles on the screen, the output can be redirected to the vending machine’s terminal or omitted entirely along with LipSync.
- ***Perform an Action:** As the vending machine will not have extensive movements, the “Perform an Action” macro can be adjusted to change textures or spawn particle effects, rather than using skeletal animations.

On the Flowstorm side, it is only a matter of customizing the content being sent. For example, change the type of command from “gesture” to “dispense” and replace the payload with “energy bar” instead of “pleasured”. This would allow for comprehensive spawning an “energy bar” actor or any other desired reaction that could be implemented in Unreal Engine. The two figures 6.3 6.4 illustrate the necessary changes on the Flowstorm side.

[1]this arrangement is similar to notation of calling a class function as “ClassInstance.myFunction()”
[2]this enables the event to “listen” for a signal, event without prior binding is “deaf”

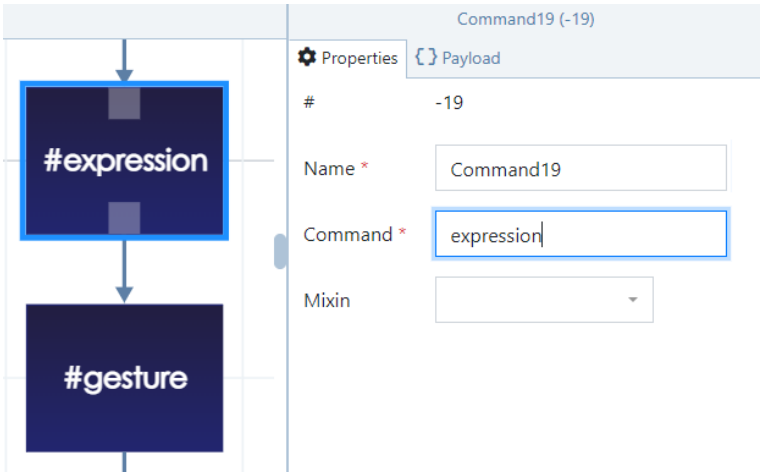


Figure 6.3 Flowstorm command

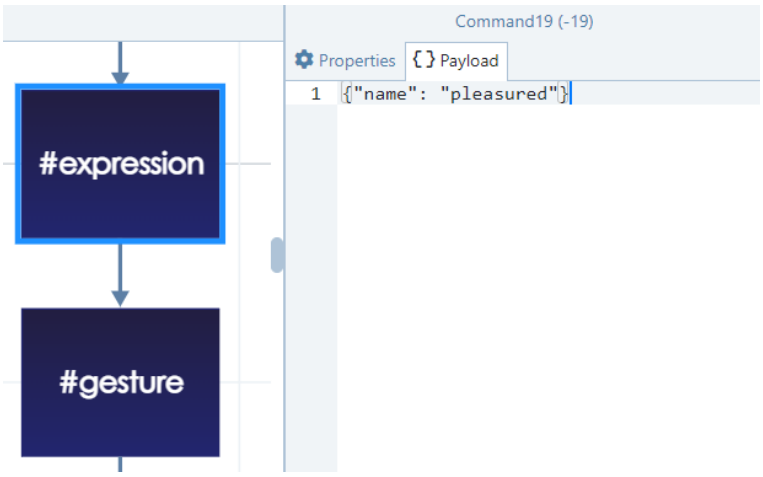


Figure 6.4 Flowstorm payload

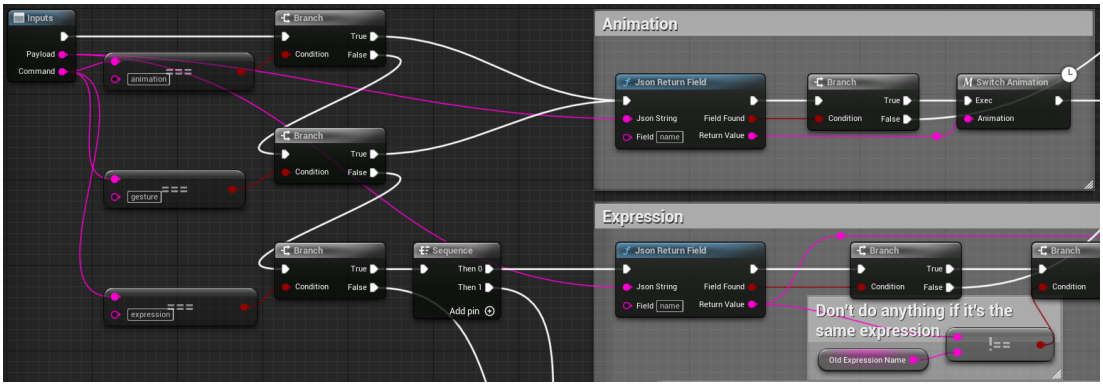


Figure 6.5 Part of the insides of "Perform an action" macro

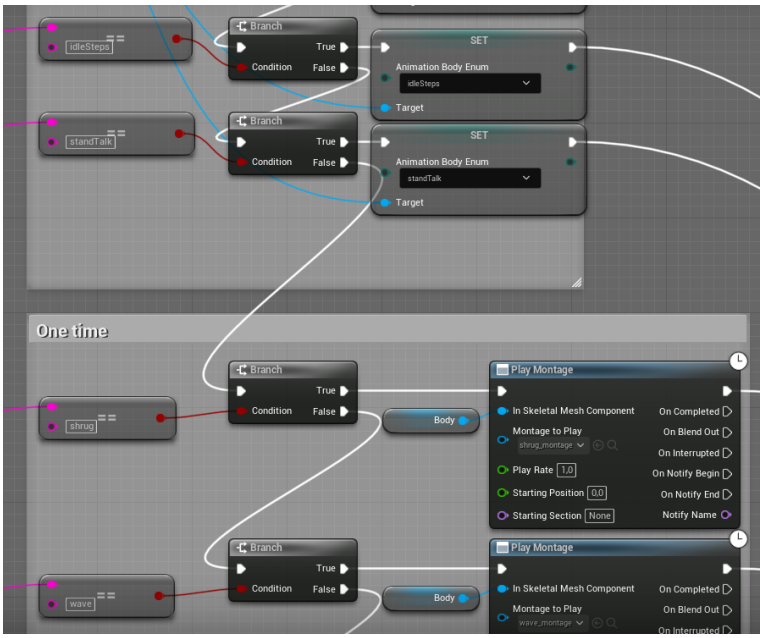


Figure 6.6 "Switch animation" macro found in top right corner of figure 6.5

The utilization of Flowstorm Link is ultimately up to the developers and their specific project requirements. To provide an insight into how the "Perform an action" macro is implemented in the digital persona, the following figures 6.5 6.6 are presented. Initially, the flow is directed to the "Perform an action" macro upon receiving it from Flowstorm. Within the "Perform an action" macro, the type of action is identified, such as changing an animation, performing a gesture, or altering an expression. This determination is made using a "switch" comprised of branching nodes performing string comparisons. Once the action type is decided, it is executed. For example, if an animation action is received, the flow enters the "Switch animation" macro, where the appropriate animation is played. Afterward, the flow returns to the Link function to continue processing further actions or audio, or to request user input.

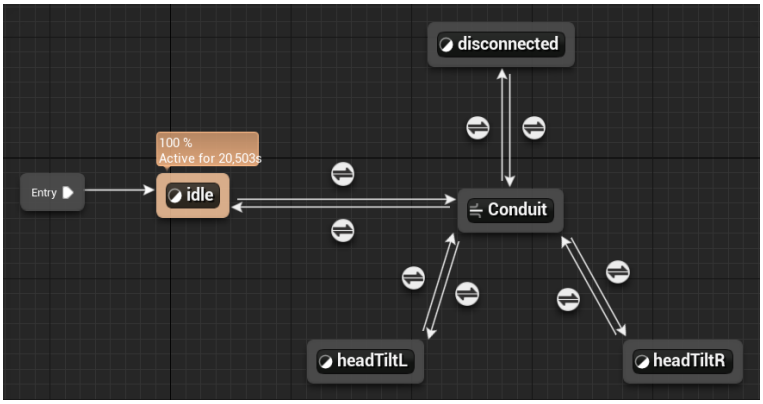


Figure 6.7 State machine responsible for head positions

In this context, the difference between "animation" and "gesture" is that animations involve state transitions through a state machine, while gestures are more like one-time animations. However, both are skeletal animations at their core. On the other hand, "expression" is implemented using morph targets, which allows for more nuanced adjustments to a character's appearance.

In the context of animations and gestures, it is important to notice that animations involve changing enumerations, while gestures play montages. Enumeration changes are reflected in the state machine shown in figure 6.7. State machines facilitate smooth transitions between different animations. On the other hand, montages function similarly to sequences, which can be thought of as animation tracks. Montages, when compared to sequences, offer additional advantages, such as allowing for smooth blending between different animations, just like state machines. However, they also enable triggering of various events and provide the option to either continue with the flow immediately or wait until an animation finishes. This makes montages more suitable for managing shorter animations, in contrast to state machines that are better suited for longer and looping animations.

Other macros like "Get User Input" and "Process Audio" won't be shown, as they may not provide any further significant information and could on the contrary create a visual clutter. Blueprints are typically quite wide, making them challenging to present clearly and effectively in this format.

Following actions were implemented:

- **Animations and Gestures:**
  - **Accept:** nods head as a signal of approval
  - **Reject:** shakes head as a signal of denial
  - **Nod:** bows down head to signal understanding
  - **Shrug:** shrugs shoulders to signal not knowing something or not caring about something
  - **Wave:** waves hand as a greeting
  - **Quote:** makes imaginary quotes in the air with its fingers to signal irony
  - **Idle:** default... standing
  - **Looking Around:** more interesting idle with some head/body rotations as if the avatar was looking around
  - **Disconnected:** looking downwards as a signal that the avatar is not listening
- **Expressions:**
  - **Empathy:** head closer to its shoulder
  - **Neutral:** neutral expression
  - **Angry:** angry expression
  - **Concerning:** concerning expression
  - **Doubt:** doubtful expression
  - **Listening:** more pleasant than neutral expression
  - **Pleasured:** smiling/happy expression
  - **Sad:** sad expression
  - **Surprised:** surprised expression
  - **Thinking:** thinking expression
  - **Wink:** eye wink
  - **Blink:** eyes blink

6.1.2 C++ side

Now, a detailed examination of how Flowstorm Link works under the hood in C++ is provided. Flowstorm Link is a class inheriting its functionality from UActorComponent, making it an actor component. Flowstorm Link utilizes several events and functions, which are introduced in the following list:

- **Events**[3]**:**
  - **On Proceed With Linkage:** Allows the Link function to call itself, enabling looping, which is not possible in the traditional sense. Just calling Link() at the end is not possible due to events playing roles of parameters, so this workaround is necessary.
  - **On GetUser Input:** Called by Link when all audio and actions received from Flowstorm have been processed.
  - **On Audio:** Event serving as a return for the Link function regarding audio and speech.
  - **On Action:** Event serving as a return for the Link function regarding actions received from Flowstorm.
- **Public:**
  - **Link:** A primary function that initializes conversations, decides on the next step during the flow, and returns outputs through associated events.
  - **Send User Input:** Makes an HTTP request to Flowstorm with the provided user input string.
  - **Json Return Field:** A helper function that finds a field in a simple non-nested Json and returns its value. It enables more structured commands on the Flowstorm Link.
  - **Get Tmp Directory:** Finds a path to the operating system's temporary files folder, which is automatically cleaned by the OS, making it an excellent choice for downloading speech audios.
  - **Get Session ID:** A simple getter that retrieves the session ID of the current conversation, used for naming the downloaded audio files.
- **Private:**
  - **On Response Received:** Receives responses from the HTTP request made by "SendUserInput" and saves them in a queue.
  - **Process Response:** Processes the queue of saved response strings for the Link function, which can then return them and direct the further flow.
- **Private data:**
  - **Flowstorm Output Queue:** Stores response strings.
  - **Initialized:** Boolean value indicating whether the conversation has been initialized.
  - **Session ID:** Unique ID of the ongoing conversation, used in "Send User Input" to allow Flowstorm to recognize which conversation instance to reply.
  - **Flowstorm Application Key:** Identification for the conversation (application) to communicate with, found in Flowstorm after creating a dialogue.

[3]Events are distinct from standard functions, which is why they are separated from the other groups. The first two events are defined in public, but all events are declared outside the class altogether.

6.1.2.1 Send User Input

The examination of the functions will now focus on SendUserInput to gain a better understanding of the roles played by Flowstorm Application Keys and Session IDs.

Flowstorm Application Key can be found in Flowstorm under ”Access - Applications” upon creating a dialogue. This application key is used in the SendUserInput function. The function sends the user's utterance as input through an HTTP request to Flowstorm. To identify which Flowstorm-powered conversation, referred to as an application in Flowstorm, to communicate with, the application key must be provided. However, the application key refers only to the “type” of conversation, and as that conversation can handle multiple concurrent users or one user can re-enter the conversation, the Session ID is used to identify which instance of that conversation to communicate with. The Session ID is provided just below the Application Key in the SendUserInput function on lines 6 and 7 respectively presented in listing 1. In the case of proposed application, the Session ID is randomly generated at the start of each conversation, however, if continuing where a previous conversation session left off is desired, the related Session ID could be provided.

```
1  void UFlowstormLinkComponent::SendUserInput(const FString& UserInput)
2  {
3      FHttpRequestRef Request = FHttpModule::Get().CreateRequest();
4      Request->SetVerb("PUT");
5      Request->SetURL("https://core.flowstorm.ai/client?key="
6          + m_FlowstormApplicationKey + "&deviceId=your-mom");
7      Request->SetHeader("Cookie", "flowstorm-session-id=" + m_SessionID);
8      Request->AppendToHeader("Content-Type", "text/plain");
9      Request->AppendToHeader("X-TtsFileType", "wav");
10     Request->SetContentAsString(UserInput);
11     Request->OnProcessRequestComplete().BindUObject(this,
12         &UFlowstormLinkComponent::OnResponseReceived);
13     Request->ProcessRequest();
14 }
```

Code listing 1 Creating HTTP request for sending user input

In listing 1, lines 8 and 9 specify that communication should occur in plain text and that TTS audio of the digital persona's speeches should be received in .wav format. The .wav format is preferred for both Unreal Engine and Oculus LipSync. At the end, lines 11 and 12 bind the OnResponseReceived function, which behaves like an event, and waits for a signal. Specifically, it waits for an HTTP request response, which will then add into a queue, and triggers the OnProceedWithLinkage event to activate the Link function. The Link function then processes the queue and directs the flow according to its contents. The queue is used because, although speech comes one by one, multiple subsequent commands are sent from Flowstorm at once, and they need to be executed sequentially.

6.1.2.2 Link

The often-mentioned Link function is presented in two parts in listings **??** **??** for better readability. The division is logical, as the first part focuses on the initialization phase, which is triggered only once, while the second part manages the overall flow. Although these could be two separate functions, they have been combined into a single function to simplify the blueprint side of the implementation.

```
1  void UFlowstormLinkComponent::Link(TEnumAsByte<EPersona> Persona,
2                                     const FString& FlowstormApplicationKey,
3                                     const FOnAudio& OnAudio,
4                                     const FOnAction& OnAction)
5  {
6      //initialize new session
7      if (m_bInitialized == false)
8      {
9          //set application key
10         if (FlowstormApplicationKey == "")
11         {
12             switch (Persona)
13             {
14                 case EPersona::Poppy:
15                     m_FlowstormApplicationKey = "62cd6a3495cbb14d83646cc9";
16                     break;
17                 case EPersona::Seb:
18                     m_FlowstormApplicationKey = "62c6b1804fd06e134b63a986";
19                     break;
20                 case EPersona::Kai:
21                     m_FlowstormApplicationKey = "62c6b1c795cbb14d8397177d";
22                     break;
23             }
24         }
25         else
26         {
27             m_FlowstormApplicationKey = FlowstormApplicationKey;
28         }
29
30         //set session ID
31         m_SessionID = FGuid::NewGuid().ToString();
32         m_bInitialized = true;
33
34         //kick off the convo
35         SendUserInput("#init");
36
37         UE_LOG(LogTemp, Warning, TEXT("Initialized now!"));
38         return;
39     }
```

Code listing 2 Initialization phase of Link

In the first 28 lines of listing 2, the Flowstorm application key is set, using either one from the predefined enumerations or a custom key if provided. A new Session ID is generated on line 31, and on line 35, Flowstorm is signaled to initialize the conversation with the digital persona (the user input can be any value in this case, “#init” or anything else). After this step, a response from Flowstorm is received, the Link function is triggered again and the initialization phase is skipped. This time the response is processed in the second part of Link. This initialization phase occurs only once, during the first time the flow reaches the Link function.

```
    //create new request if no response is available for processing
    if (m_FlowstormOutputQueue.IsEmpty() == true)
    {
        if (OnGetUserInput.IsBound())
        {
            OnGetUserInput.Broadcast();
        }
    }
    //otherwise process response
    else
    {
        TTuple<bool, FString, FString> ProcessedResponse = ProcessResponse();
        if(ProcessedResponse.Get<0>() == true)
        {
            if (OnAudio.IsBound())
            {
                OnAudio.Execute(ProcessedResponse.Get<1>(),
                                ProcessedResponse.Get<2>());
            }
        }
        //stop if end of dialogue signalled by dot is reached
        else if (ProcessedResponse.Get<0>() == false
                && ProcessedResponse.Get<2>() != "")
        {
            if (OnAction.IsBound())
            {
                OnAction.Execute(ProcessedResponse.Get<1>(),
                                ProcessedResponse.Get<2>());
            }
        }
    }
}
```

Code listing 3 Second part of Link

In the first eight lines of listing 3, the code checks if there are any responses to be processed. If there are none, the user is prompted for input. However, if there are responses from previous HTTP requests, the top response in the queue is processed. A helper function “Process Response” assists in extracting important information from the response string, storing it in a tuple with three elements: a boolean and two strings.

If the boolean is true, an audio file is associated with the response. The first string in the tuple contains the URL of the .wav format audio file, which can be downloaded and played back, while the second string contains subtitles for the digital persona’s speech, that can be displayed on screen.

If the boolean is false and the strings are empty (checking one empty string is sufficient), this indicates that the conversation has reached its end. If the strings are not empty, an action has been received, and the payload and command set in Flowstorm are returned through the On Action event.

6.2 Modified Oculus LipSync Plugin

Although Oculus LipSync offers production-level capabilities, it was not immediately suitable for the required use case. By default, it provides functionality for pre-generating files containing animation data from audio files, but this cannot be done in real time. Alternatively, it can capture audio data from a microphone and process it into animations in real time. This use case was designed for virtual lobbies, where users can meet in virtual reality, and the mouths of their avatars would move as they speak. However, the desired functionality was a combination of these two use cases: using audio files instead of a microphone and generating animations in real time. To achieve this, the functionality of capturing data from the microphone was replaced with loading data from a file. The header of the WAV file, as shown in listing 4, was read to determine its bitrate, and then an index was moved around the file's PCM[4] data, which were sequentially fed into LipSync in given time intervals.

```
struct TWAVheader
{
    uint8 Chunk[4];
    uint32 ChunkSize;
    uint8 format[4];
    uint8 Sub_chunk1ID[4];
    uint32 Sub_chunk1Size;
    uint16 AudioFormat;
    uint16 NumChannels;
    uint32 SampleRate;
    uint32 ByteRate;
    uint16 BlockAlign;
    uint16 BitsPerSample;
    uint8 Sub_chunk2ID[4];
    uint32 Sub_chunk2Size;
};
```

Code listing 4 WAV header

6.3 Atmosphere

Lighting is essential for establishing atmosphere and enhancing the perception of digital personas, as discussed in the chapter about colors and lighting. That is why modified top quality lighting presets by award-winning cinematographer Greig Fraser, famous for his work on Batman, Dune, and Rogue One: A Star Wars Story, are used. Fraser finds that virtual lights offer unique opportunities for rapid iteration and experimentation with different setups. Additionally, virtual lighting enables creative techniques, such as placing a light bulb directly in front of a camera while rendering it invisible, which would be impossible in reality. However, virtual lighting also presents challenges in certain scenarios, including the potential to make the skin of digital personas look artificial if used incorrectly. [88]

[4]PCM (Pulse Code Modulation) data represent the raw audio samples in a WAV file. A WAV file consists of a header, which contains metadata such as the bitrate or the number of audio channels, followed by the PCM data that stores the actual audio information

The following types of lights are employed:

- **Directional light:** simulates sunlight and illuminates the entire scene
- **Point light:** emits light in all directions from a single point
- **Spot light:** provides cone-shaped illumination
- **Rectangular light:** creates a rectangular-shaped light source, ideal for simulating soft-boxes
- **Skylight:** used to simulate ambient light from the sky

6.4 Conversational design

It is important to note that the dialogues presented in this thesis were not designed by the author himself. Instead, they were crafted by real conversational designers at PromethistAI, the company responsible for developing the conversational platform called Flowstorm used in the proposed solution. These designers, often linguists or psychologists, ensure a higher quality of conversation compared to those created by individuals without specialized expertise. While the author did oversee the use of command nodes and modified one of the conversations for the testing phase, more details concerning it will be provided in the next chapter.

6.5 Result

The final showcase application can be observed in the following figures[5]: 6.8 and 6.9. In figure 6.8, a main menu is displayed, where users can choose a digital persona and its associated conversation topic. Upon selecting one in the main menu, users are directed to a conversational level, as shown in figure 6.9, where the digital persona greets the user and initiates the conversation.

[5] more can be found in the appendix

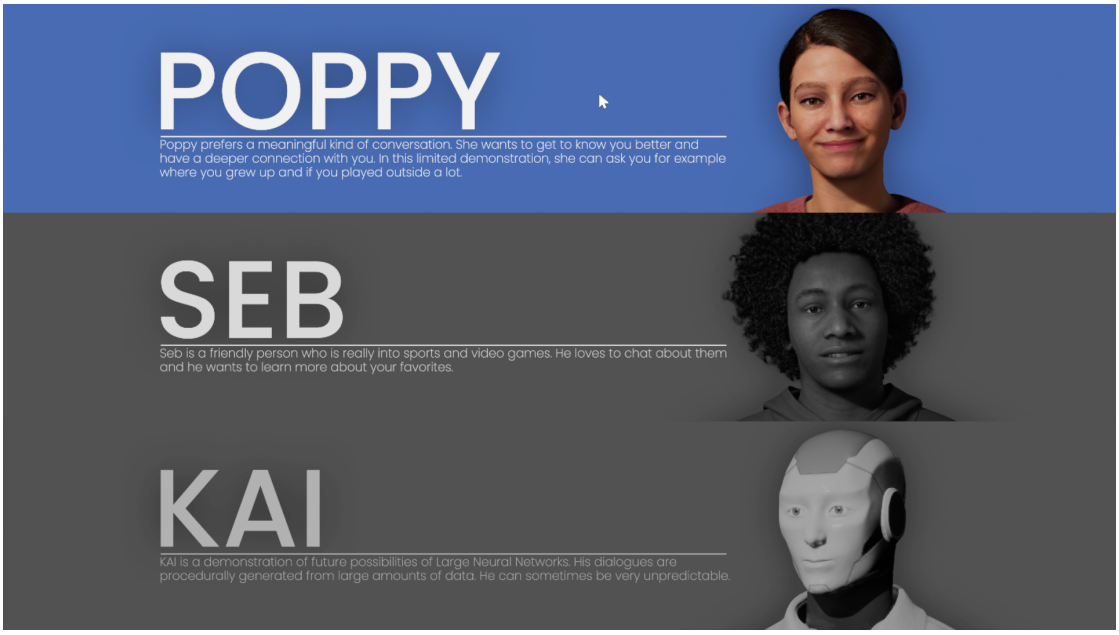


Figure 6.8 Main menu level

Figure 6.9 Conversational level (Seb)

Chapter 7

Testing

In the testing phase, the third goal is demonstrated, which focuses on illustrating the role of digital personas in the research domain and exploring the potential for making the typically unpopular testing phase more interesting and efficient through multidisciplinarity. Psychology was chosen as the additional discipline, as it can be employed in various interaction scenarios between human-human and human-computer, with digital personas being on the verge of human and computer. The increased efficiency of proposed approach lies in the ratio between the amount of information gained and the additional time and work required. Additionally, the digital persona is utilized to offload some tasks from human researchers.

The application used for testing purposes was derived from the showcase one and adapted for research purposes by incorporating music and making color and lighting adjustments during the conversation. This enabled the investigation of how secondary information channels can be utilized to influence the mood of users while conducting usability testing, thereby obtaining more information from each participant. This approach relates to the theoretical sections of the thesis, particularly the chapters on Colors and Lighting and Sound.

Instead of providing three different personas with their topics, the testing application employed only a single persona "Poppy" whose dialogue was modified and shortened for this purpose. These adjustments allowed for faster testing and ensured a uniform experience for all 51 participants. The app presented three testing levels in the main menu, labeled just "A", "B", and "C".

In level "A", color and lighting manipulations were employed to create a dim, less saturated and cooler environment, which is expected to cause the least arousal in users. Slow-paced music accompanied the dialogue, as it is also known to induce minimal arousal. In contrast, level "C" featured bright lights and highly saturated warm colors, along with fast-paced music, also associated with the vivid color palette, aiming to induce maximum arousal in users. Level "B" served as a baseline to observe the effects of these modifications compared to a neutral setting without color, light, or sound manipulations. This level used green color of medium brightness and medium saturation. The decision to use green as the base hue was based on its position between the calming blue and the highly arousing red. However, due to the post-processing of visual information done by the human eye and the combination of red and green stimuli into a single stimulus, there is no reddish-green hue. As a result, colors could not be shifted all the way from green to red without causing noticeable disharmony and potential discomfort for viewers. The green shirt of the digital persona remained primarily green, while the background color varied[1] as shown in figure 7.1.

[1] the hue of the shirt was also varied but to a lesser extent, making it less noticeable

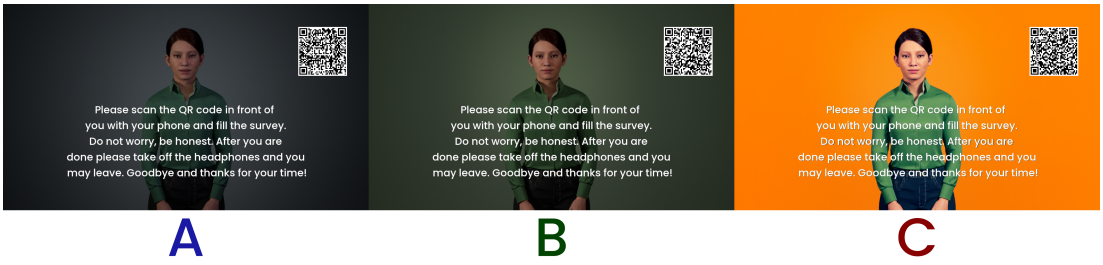


Figure 7.1 Overview of color manipulations performed in the testing application. *(A = Calming, B = Baseline, C = Energizing)*

7.1 Testing process

The overall testing consisted of two phases: "internal-testing" and "real-testing". The internal testing phase involved only two participants and served as a preliminary test to identify any issues before proceeding with the real testing phase, which involved a larger number of participants. Making changes and fixing issues during real-testing would invalidate previous results, which was undesirable.

Internal testing primarily revealed issues with color-concept associations. For example, in the calming level "A", the dimmed visuals created black areas on the screen, which appeared too somber and sad within the given context. Another issue was the highlighting of menu tiles as the user moved the cursor over them. Calm blue, neutral green, and aggressively saturated red were used for highlighting, with the intention of subtly contributing to the experiment's effect. However, this choice proved to be misguided, as participants during the internal testing phase perceived the green and red highlighting as indicative of correct and incorrect choices, respectively. This induced unintended feelings related to valence, one of the examined metrics mentioned later.

Additionally, the testing phase also uncovered a bug in the "switch expression" function, which caused the face to return to a neutral expression when the set expression was set again. This issue was subsequently resolved.

During the testing, digital persona served as a co-moderator and a data gatherer[2]. A human moderator was responsible for selecting participants and bringing them to the testing room. In this room, participants were provided with an overview and introduction to the testing and the psychological experiment that was part of it. Participants were informed about their right to leave at any time and the absence of known discomforts related to the experiment. They were then asked to scan a printed QR code on the table using their smartphones to provide informed consent and complete a pre-experiment questionnaire which aimed to gather demographic information about the participants and their initial status. After addressing any questions, the human moderator handed the participant over to "Poppy", a digital persona, and left the room to minimize potential discomfort as the participants could feel like they were being examined for their level of English. The testing environment included screen and webcam recording as well as audio recording from both the application and microphone to allow for analysis of the interactions later without the need for a human moderator to be present in the room.

[2]Interestingly, they can also serve for example as therapists or at least as an extension of human therapists, as individuals are often more willing to disclose personal information to a machine than another human being. [89]

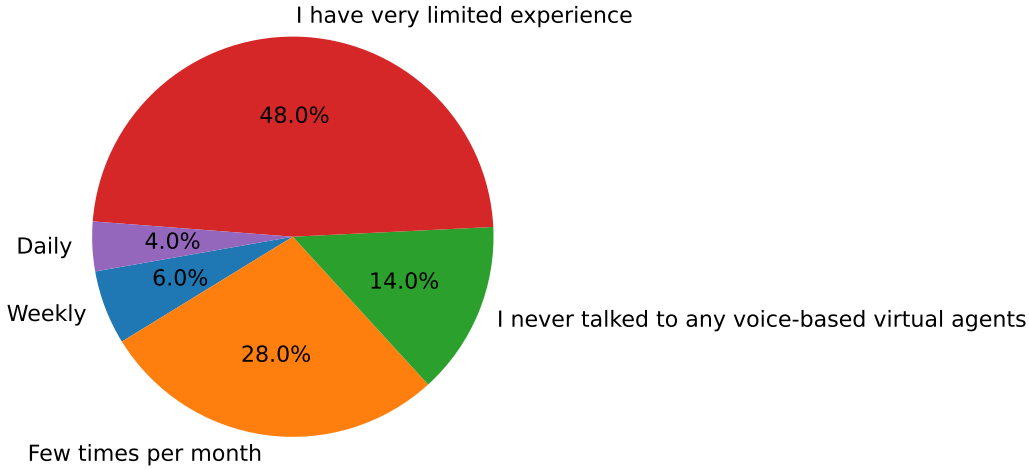


Figure 7.2 Exposure to virtual agents / digital personas / AI digital assistants *(e.g. Apple's Siri, Amazon's Alexa, Google's Google Assistant, Replica)*

Poppy used a modified dialogue in which she, at the beginning, explained the conversational process to the participant, before proceeding with a small-talk, emphasizing the importance of speaking only when the "white fog" at the bottom of the screen was lit and that long pauses in speech could prompt Poppy to speak. Though this explanation may seem unnecessary or even wrong[3], it proved useful for first-time users. At the end of the testing phase, it was discovered that 62 % of participants reported very limited or no prior experience with virtual agents at all.

As the experiment concluded, Poppy instructed participants to scan another QR code, fill out a post-experiment questionnaire, and leave the room. The human moderator then provided a debriefing, explaining the true purpose of the experiment – to investigate the effect of sound, colors, and lights on users during a conversation. Since revealing this information before the experiment could have influenced the results.

7.2 Usability and user feedback

The feedback regarding the interactions and Poppy herself included following aspects:

The speed of replies:

- **Negative**:
  - "*I was a bit angry because of the cut when I stopped talking for a while - just to make up my mind. I have not finished about half of my sentences.*"
  - "*I know she said not to take long pauses and I tried, but I still couldn't finish some questions. But I do pause a lot when I am thinking.*"
  - (What is the one thing I should definitely get better at?) "*letting me finish the sentences*"
  - "*Maybe longer time before assuming I don't want to say anything more.*"
- **Positive**:
  - "*Reacting quickly without any mistakes*"

[3] as the saying goes: "UI is like a joke. If you have to explain it, it's not that good."

The use of silence detection in user speech was implemented to create a more natural conversational experience, as opposed to a push-to-talk approach. Despite Poppy's explanation, some users experienced difficulties with the silence detection method, prompting a reconsideration of the push-to-talk approach. However, it remains unclear what different challenges the push-to-talk approach might introduce.

Based on these replies, silence detection in audio input alone appears to be insufficient for facilitating smooth, turn based conversations for some users. Advances in natural language understanding or incorporating machine vision could potentially improve the interaction experience. Humans often rely on facial expressions, non-verbal cues, and filler sounds like "mmmm" to signal the continuation or termination of speech. Currently, digital personas lack these capabilities, but future development may address these limitations.

It is important to note that, based on the recordings, most users did not experience difficulties with the silence detection approach. Nevertheless, the individuals who encountered issues had a less than optimal experience.

Eyes:

- **Negative**:
  - "*Focus, eye tracking*"
  - "*I think maintaining more eye contact with me would help me to feel much more connected to the persona. I liked the music and the tone of her voice though.*"

The perception of Poppy's empty gaze might be attributed to her forward-looking focus without following users' eyes. One possible solution could involve focusing the camera on the head or bust only, which would enlarge the eyes and potentially increase their focus on the camera. However, this approach could lead to the loss of the potential benefits of body animations, resulting in a talking head on the screen with limited expressive capabilities.

Another option to explore could be the implementation of procedural animations, enabling the digital persona to direct her gaze toward the camera at any angle. Additionally, incorporating subtle head movements could enhance the realism and engagement of the interaction.

Voice:

- **Negative**:
  - "*You were really curious, was nice to tell you some informations about me, you were listening to me but your voice is a bit tired.*"
  - (What is the one thing I should definitely get better at?) "*More human like intonation.*"
  - "*intonation of the AI was very robotic. It felt strange talking to a human-looking imaginary creature. But also very interesting indeed*"
  - "*You can't play with voice*"
  - "*the voice sounds way too roboty*"
  - "*Intonation, this is what is missing and always remind me that I am speaking to something not to someone*"

- **Positive**:
  - "*The music definitely helped the overall experience but your voice is soothing*"
  - "*Because of your voice, it is really calming and I liked your empathy.*"
  - "*As the replies were really empathetic and the voice was calming. And as if she was there only for me.*"
  - "*Nice voice*"
  - "*I liked the calm and natural voice*"
  - "*Her voice is good.*"
  - "*You had calm voice*"
  - "*It's just interesting talking to someone artifical and hear them talking back to you in the context. And you had a relaxed voice, so it calmed me down a little.*"
  - "*The voice and speaking speed of the AI was very relaxing. I havent had an experience like this with Siri which is the only AI I spoke to before, so I was intrigued. It also felt strange, because I tend to be uncomfortable when I know that technology is listening to what I say.*"
  - "*You have a very pleasant voice*"
  - "*Nice voice, large vocabulary.*"

The voice aspect of Poppy appears to be quite controversial, as some users found it to be synthetic and robotic, while others praised its quality. Speech synthesis continues to improve, and it is true that intonation really does play a crucial role, especially when communicating with children, as adults tend to use different intonation patterns when speaking to them for various reasons. [90]

Although none of the feedback points to dealbreakers for the majority of users, the discussion chapter will propose some potential state-of-the-art, nearly sci-fi solutions and improvements based on this feedback obtained during the testing phase. Emphasizing unconventional and non-traditional approaches for today's standards, the discussion will bring insight into the possible evolution and refinement of digital personas in the years to come.

Now, the focus will shift towards the research aspect of this thesis.

7.3 Psychological experiment

Since the author lacks expertise in psychology, Mgr. Bc. Barbora Šipošová, Ph.D., a cognitive/behavioral psychologist with years of research experience was asked to oversee the entire testing process to ensure its quality. This interdisciplinary collaboration resulted in a well-structured and rigorous psychological experiment, demonstrating the value of combining expertise from different fields.

7.3.1 Design and conditions

In this thesis, for the purpose of investigating how the sound, colors, and lights can affect users during a simple conversation, a "between-subject" design was chosen for the experiment, as opposed to a "within-subject" design. The primary distinction between the two designs lies in the way participants are exposed to experimental conditions (effects A, B, C). In a within-subject design, each participant experiences all conditions, while in a between-subject design, participants are exposed to only one condition, with different groups experiencing different conditions.

The within-subject design can be more powerful statistically, as it allows for the comparison of effects within the same individual, reducing the influence of individual differences. However, it is susceptible to carry-over effects, where the experience of one condition may influence the participant's response to subsequent conditions. In this context, for instance, presenting effect C (the supposed energizing effect) before effect A (the supposed calming effect) when tested sequentially within the same participant could negatively impact the results, as effect C might influence the participant's impression of effect A.

To avoid such carry-over effects and potential confounds, the between-subject design was chosen. This design assigns participants to different groups, each experiencing only one condition, thus eliminating the risk of one condition's effects influencing the response to another condition. This ensures a more accurate representation of the effects being studied and leads to more reliable conclusions.

7.3.2 Metrics

To better understand the hypotheses tested in this experiment, it is beneficial to acknowledge that multiple metrics were measured to obtain more information. While arousal was the primary metric, as it was the subject of prior research and targeted for manipulation, additional metrics included *valence*, *trust*, *net promoter score*, and *meet*. The metric *meet* provided four options for responses, namely "Yes", "Maybe", "No", and "This is a strange question", while other questions were rated on a numeric scale. Figure 7.3 illustrates the scales used for arousal and valence.

- **Arousal:** A measure of the participant's emotional activation and intensity during the interaction, ranging from calm to excited states.
- **Valence:** A measure of the participant's emotional state on a spectrum from (for simplicity) sad to happy, reflecting their overall mood during the interaction.
- **Trust:** The degree to which participants agreed with the statement, "In general, you feel that you can trust me."
- **Net Promoter Score:** A measure of participants' likelihood to recommend the conversation with the digital persona to their friends, based on their experience.
- **Meet:** A hypothetical question posed to participants, asking if they would be interested in meeting the digital persona someday.
- **Free Association:** Users were asked to provide 3 words describing the "vibe" they experienced.

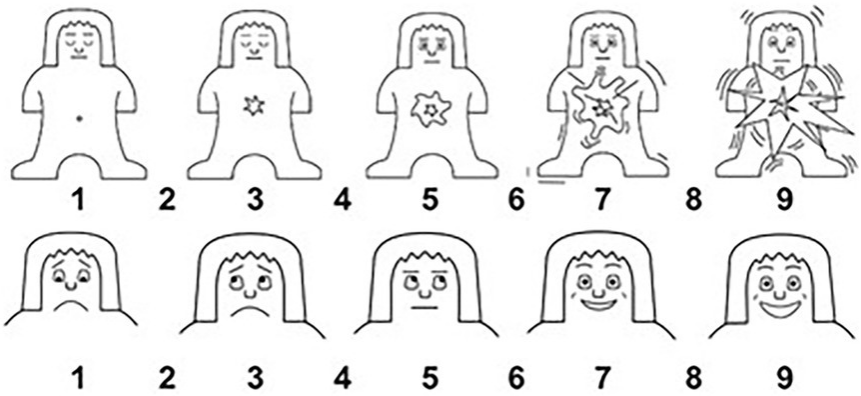


Figure 7.3 Gender neutral Self-Assessment Manikin (SAM) [91]
(arousal - top one, valence - bottom one)

7.3.3 Research questions

The hypotheses for this experiment were as follows:[4]

Research Question 1: Is there a statistically significant difference in arousal before and after the experiment?
Null Hypothesis: No difference between before and after exists.
Alternative Hypothesis: A difference between before and after exists.
Prediction 1: No difference between before and after exists.
Prediction 2: Difference exists in a following form $A_a < A_b, B_a = B_b, C_b < C_a$[5] as intended.

Research Question 2: Is there a statistically significant difference in valence before and after the experiment?
Null Hypothesis: No difference between before and after exists.
Alternative Hypothesis: A difference between before and after exists.
Prediction 1: No difference between before and after exists.
Prediction 2: Difference exists, albeit unintentionally, in a following form $A_a < A_b$, $B_a = B_b, C_b < C_a$ due to color associations and mode of music.

Research Question 3: Does a statistically significant difference in arousal exist among participants experiencing different conditions?
Null Hypothesis: No difference between conditions exists.
Alternative Hypothesis: A difference between conditions exists.
Prediction 1: No difference between conditions exists.
Prediction 2: Difference exists in a following form A<B<C as intended.
Prediction 3: Difference exists in a following form B<A<C as even slow music could be more arousing than no music.

[4] It is important to mention that the neutral condition, despite attempts to make it neutral, might still have an impact on participants. Nonetheless, it represents the most appropriate baseline for this experiment.
[5] lower index "a" means "after", lower index "b" means "before"

Research Question 4: Does a statistically significant difference in valence exist among participants experiencing different conditions?
Null Hypothesis: No difference between conditions exists.
Alternative Hypothesis: A difference between conditions exists.
Prediction 1: No difference between conditions exists, as they were not aimed at altering valence and the dialogue content was neutral.
Prediction 2: Color association and the mode of music might induce changes in valence, albeit unintentionally. If so, C will be perceived more positively than B, and B more positively than A. (A<B<C)

Research Question 5: Does a statistically significant difference in trust exist among participants experiencing different conditions?
Null Hypothesis: No difference between conditions exists.
Alternative Hypothesis: A difference between conditions exists.
Prediction 1: No difference between conditions exists.
Prediction 2: It is speculated that individuals may confide more in a calm environment than in a lively one. (C<B<A)

Research Question 6: Does a statistically significant difference in net promoter score exist among participants experiencing different conditions?
Null Hypothesis: No difference between conditions exists.
Alternative Hypothesis: A difference between conditions exists.
Prediction 1: No difference between conditions exists.
Prediction 2: More intense emotional response related to arousal and valence could potentially lead to higher net promoter scores. In that case, both A and C are predicted to be greater than B. (A>B, C>B) [92]

Research Question 7: Does a statistically significant difference in the "meet" metric exist among participants experiencing different conditions?
Null Hypothesis: No difference between conditions exists.
Alternative Hypothesis: A difference between conditions exists.
Prediction 1: No difference between conditions exists.
Prediction 2: Lighting could play a role, as in condition C, the skin may appear more artificial due to overlighting the face. In that case, participants would likely be more reluctant to meet with an uncanny artificial being. C would score lower than both A and B (C<A, C<B).

7.3.4 Results

Before moving onto inferential statistics and drawing conclusions, it is vital to first describe and explore the resulting obtained data through the use of descriptive statistics.

7.3.4.1 Descriptive statistics

The sample consisted of $N = 51$ participants, recruited on 3 different campuses in Prague. With only four exceptions, the age of the participants was 25 years or younger. The youngest participant was born in 2005 and the oldest in 1965. In terms of gender, 66 % were female, and 34 % were male. None of the participants reported any color vision issues, such as colorblindness.

The obtained data for different metrics under various conditions are presented in the four box plots in figure 7.4. The box plots display the distribution of data, with the rectangle representing the interquartile range (IQR), the orange line as the median, and the whiskers extending from the box to the minimum and maximum values within 1.5 times the IQR. The top two box plots represent the changes in arousal and valence, while the bottom two depict trust and net promoter score. Arousal and valence were assessed through pre- and post-experiment questionnaires, reflecting the shifts in these metrics. Trust and net promoter score were measured only after the experiment, representing the specific values of these variables.

Interestingly, the B condition appears to be most arousing, with a median shift of +1, indicating increased arousal. In contrast, conditions A and C had median shifts of -1 and -0.5, respectively, suggesting calming effects. Valence remains mostly unaffected. Trust medians are centered, but distributions lean towards lower scores for digital personas. As for the net promoter score, condition B has the lowest value, and Condition C stands out with a higher score, although reasons remain unclear. Regarding "meet" metric, more positive responses were seen in condition C, indicating a stronger inclination to a more positive attitude towards meeting the digital persona. These observations are shown in figure 7.5. The free association metric responses are presented as word clouds in figure 7.6. Some participants felt nervous, but condition A reduced the occurrence of the word "nervous". Word clouds generally support the arousal metric.

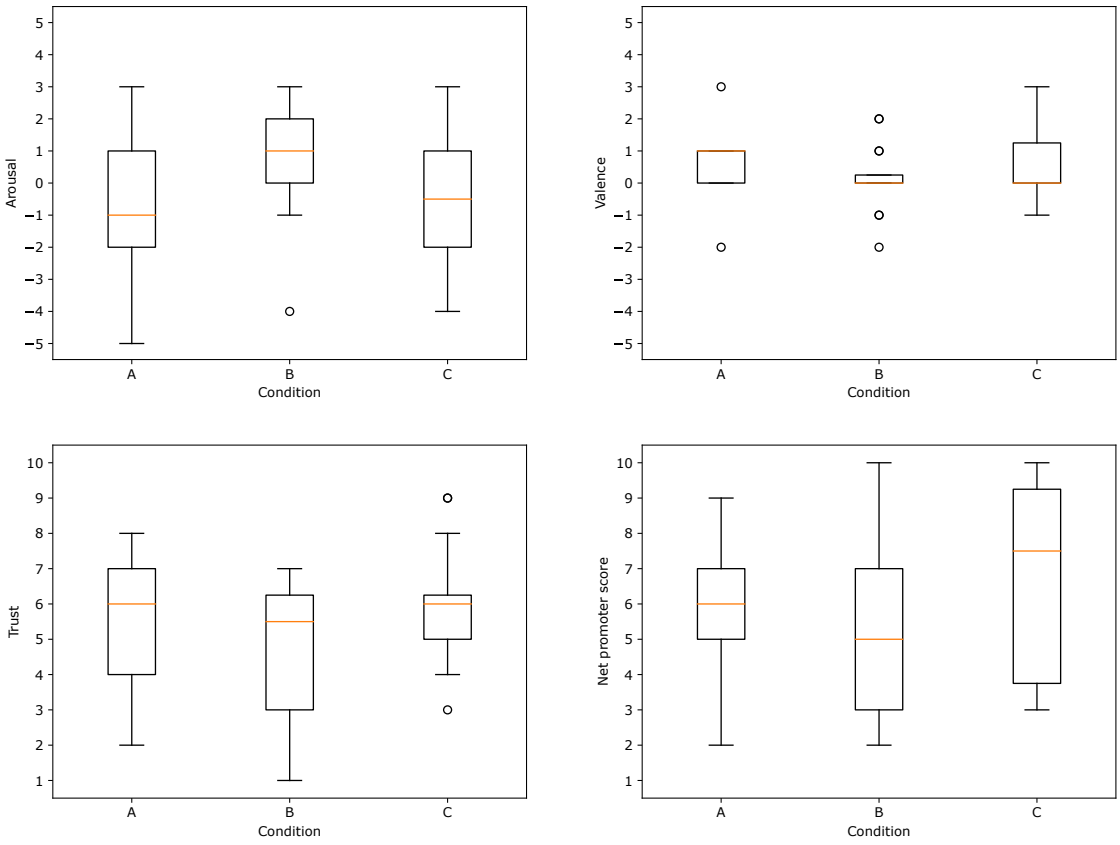


Figure 7.4 Box plots of metrics

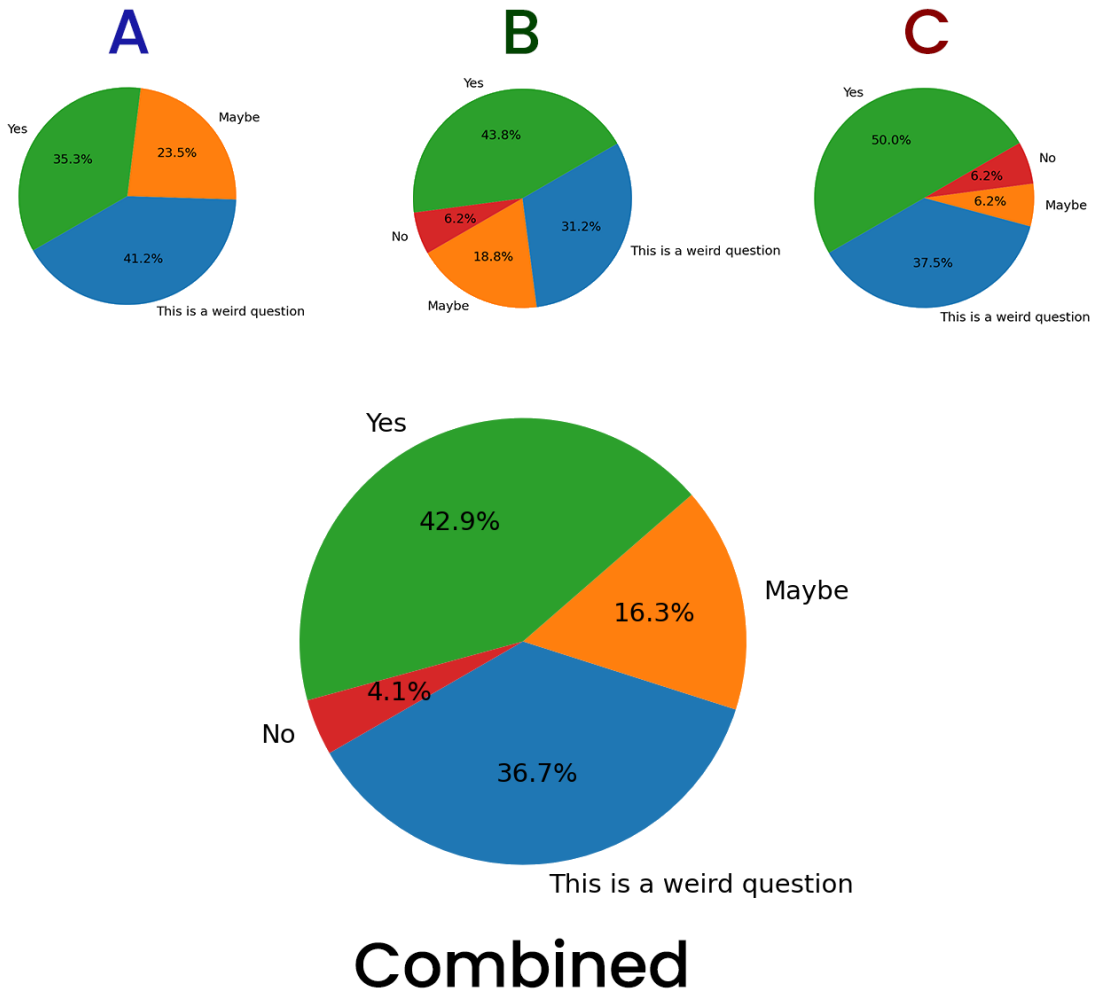


Figure 7.5 Meet metric

Figure 7.6 Word clouds depicting word frequencies obtained from the free association metric

7.3.4.2 Inferential statistics

In examining the inferential statistics, no statistically significant differences were found in any metric, that means the observations made through descriptive statistics could be coincidental. Multiple statistical tests were employed to search for significant differences. The exploration began with a pairwise Mann-Whitney test, accompanied by Bonferroni[6] correction, to examine differences in arousal and valence between conditions before and after the experiment. A t-test was considered, but the Shapiro-Wilk test rejected the normality of the distribution. The Mann-Whitney test yielded results that would support the alternative hypotheses at a 10% confidence level. However, as the standard 5% level is typically employed, the null hypothesis cannot be rejected, and it is concluded that no statistically significant difference exists.

Subsequently, all metrics shown in figure 7.4 were tested for normality using the Shapiro-Wilk test. Trust was the only metric not rejected as coming from a normal distribution. Levene's test was applied to trust to assess equality of variances, and the test did not reject the assumption of equal variances. Thus, one-way ANOVA was used to test for statistically significant differences in trust, followed by Tukey HSD to identify where the differences lie. Non-parametric tests, such as Kruskal-Wallis and Dunn's test, were used analogously for other metrics. The only test that would support the alternative hypothesis at a 10% level of significance was Kruskal-Wallis for arousal. However, as mentioned earlier, the standard 5% level is employed, and the null hypothesis cannot be rejected.

Given the lack of significant differences, power analysis was performed using the Monte Carlo simulation technique of bootstrap resampling. [93] This revealed that the tests on different metrics had low power, with the highest reaching around 50 %. Typically, experiments aim for 80 % or higher power, limiting the chance of a Type II error (failing to detect an effect that exists) to 20 % or less. Figure 7.7 demonstrates that this experiment was essentially a pilot study, with 17 participants per group proving insufficient for a full-fledged experiment. Approximately 100 participants per group would be needed to determine the presence and location of statistically significant differences for all metrics, except for the "meet"[7] metric, with a less than 20% chance of committing a Type II error.

The solid lines in figure 7.7 represent the number of participants needed to achieve a given power for determining statistically significant differences, while the dashed lines represent the number of participants needed to find the particular two or more pairs[8] with significant differences. Since larger effects are more easily detectable, figure 7.7 can also be used to get an idea of comparison of effect sizes. However, for comparing effect sizes, proper tests such as Cohen's D and Cliff's Delta would be employed. Confidence intervals for mean differences could also provide some insight into the impact of conditions on the original numeric scales.

Despite the fact that no definitive conclusions were reached, the discovery of the necessary sample size is valuable in itself, as it enables the calculation of the experiment's cost and the determination of whether the information gained would be worth the investment. This finding leaves the experiment prepared for future execution.

[6] the type of correction chosen does not affect the outcome as the tests fail to reject the null hypothesis even without any correction

[7] string responses were mapped to integers as follows: "Yes" = 1, "Maybe" = 0.5, "No" = 0. "This is a weird question" was not counted as if the participant did not respond

[8] detecting three pairs would necessitate a significantly larger participant pool, which may not be justifiable in terms of time and resources

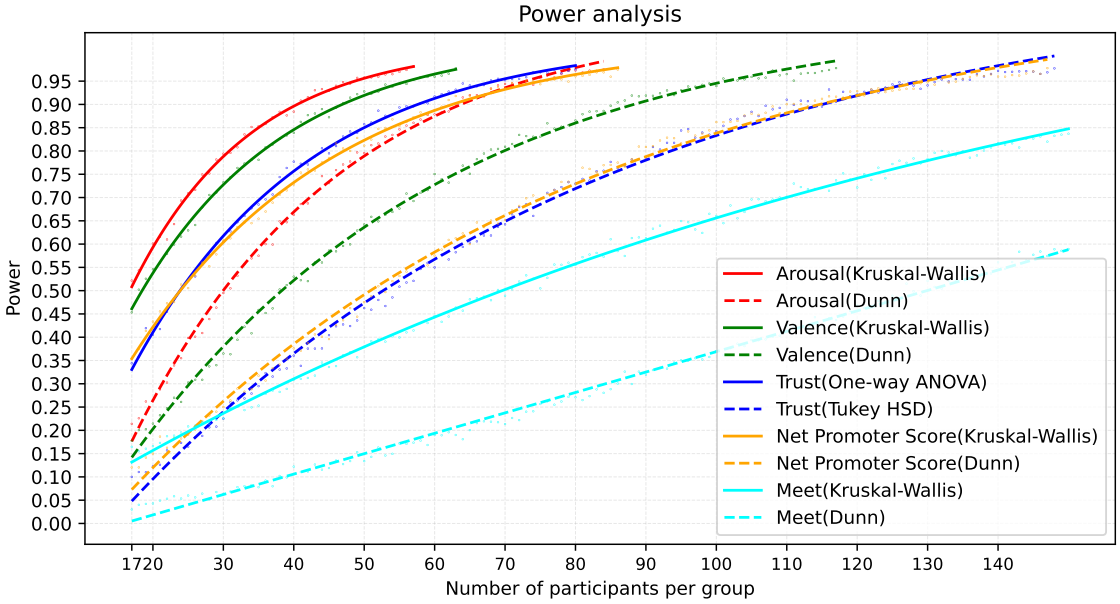


Figure 7.7 Power analysis

Discussion

chap:discussion This section will explore some unconventional and non-traditional solutions and improvements for current standards that could be employed in digital personas, even today. The reason why the proposed improvements are not widely utilized is that they are in their early stages and not yet production-ready, or at least not cost-effective.

Addressing the issue with Poppy's voice, which was praised as calming by some but criticized for its inability to intonate by others, SSML tags utilized in TTS services can employ some prolonged delays between words or put emphasis on different words. However, the extent of modifications is limited and Replica Studios [94] emerges as a potential solution to this problem. Replica Studios offers a wide range of customizable options for voice synthesis, including various intonations, accents, ages and moods. In some instances, the technology can replace voice actors entirely, providing a highly versatile and customizable alternative for generating human-like speech. However, this solution is not without its drawbacks. Replica Studios' technology is not real-time, and generating the voice requires some time and effort from the developer's side.

The non-real-time nature of Replica Studios' technology makes it challenging to integrate generative models like ChatGPT into conversational platforms. This is because the specific response to be sent must always be known and prepared in advance with voice acting. One potential solution could involve speech-to-speech (STS) services such as Respeecher [95] or Resemble AI [96] that offer voice cloning capabilities, with the latter operating in real-time. In this approach, Replica Studios could be used for known and pre-prepared responses, while the cloned voice from Replica Studios through Resemble AI could be utilized for generated responses in real time. Although this method would result in the loss of intonation possibilities and other voice acting options in the generated responses, digital personas would maintain a consistent voice throughout the conversation, still leading to a more natural and less robotic experience for the user. In essence, pre-prepared responses would benefit from more engaging voice acting and carefully crafted conversational design, while machine-generated responses providing coverage for unanticipated scenarios would lack the voice acting capabilities but would still sound more natural than they do now.

AI technology has been so far discussed only in the context of conversation, but it also holds potential for creating digital persona visual representation. This can be achieved through platforms like Synthesia [97] and D-ID [98]. Synthesia offers a more visually appealing experience, complete with gestures such as head nodding and gesticulation, while D-ID brings photos to life, using just a photo of a person as the basis for a digital persona. However, D-ID can only animate the head. Both services create lifelike presenters that resemble video recordings without any trace of 3D computer graphics. However, these digital personas are limited in their actions and movements, unable to engage in more dynamic activities such as fighting dragons.

Figure 7.8 Deepfacelive on Metahuman [100]

Traditional 3D bodies animated through skeletal animations allow for a broader range of movements and actions that can be adapted to various environments. However, deepfake technology presents potential solution that combines the advantages of both approaches and presents yet another avenue for a hybrid solution. This approach involves the replacement of a digital persona's 3D facial representation with a more photorealistic facial appearance, resulting in a more lifelike visuals. In figure 7.8, an example is provided where a Metahuman's appearance is enhanced using Keanu Reeves' face through Deepfake technology.

Exploring such solutions could lead to more engaging and more lifelike digital personas. As technology continues to advance, the potential for further development in this area remains promising.

Regarding further research, it could be interesting to investigate the use of internal diegetic sounds for psychological purposes in interactive media. For instance, in the video game called "Hellblade: Senua's Sacrifice", players are encouraged to wear headphones as they assume the role of a protagonist with a mental disorder, experiencing a constant, never ending, stream of whispering internal voices. These voices provide players with guidance of what to do next. By employing voice cloning and sound transformation technologies, it might be possible to simulate user's own internal voice and possibly partially override his internal thoughts. This approach could have numerous potential applications, particularly in the context of coaching, training, education, therapeutic interventions or other similar activities. For instance, as user would hear their internal voice saying "*maybe I should open up*", it is possible that they could be artificially led into opening up faster than in typical scenarios. By mimicking and potentially influencing user's cognitive states (including thoughts, emotions, intentions) this technology could offer innovative ways to address mental disorders or other specific conditions. The potential benefits of using this artificial and controlled schizophrenia-like state could be analogous to the use of HIV virus which is harmful, but is now being experimented with to treat various conditions as it is very efficient at modifying genetic information. [99] However, the ethical considerations around such a potentially invasive technology cannot be understated. If such an approach were to be implemented, it would be essential to conduct rigorous scientific testing to fully understand the potential risks and benefits for users. Users would have to be adequately informed about the nature of the technology, its potential effects, and any associated risks and only with a thorough understanding of these elements should users give their informed consent.

Conclusion

The thesis has successfully advanced the understanding and development of digital personas, exploring their potential in diverse contexts while going beyond the initial scope of the assignment, exceeding expectations originally set.

Conversational platforms were surveyed, Flowstorm was chosen as the most suitable one, followed by an analysis of various graphical tools. This led to the selection of the most convenient ones for creating a digital persona's appearance. Instead of manually designing a 3D character, which would be inefficient, the character generator Metahuman Creator was employed. The resulting showcase application, targeting Windows operating system, implements three different conversations about social topics. The developed solution is both usable and extendable for future projects, taking the form of an Unreal Engine plugin that powers the digital personas. It has been tested on a mid-end gaming PC with a dedicated graphics card, specifically an Nvidia GTX 1060 6GB, released back in 2016. This demonstrates that the application does not even require a high-end gaming PC to run smoothly as initially expected, making it more accessible for users with different hardware specifications.

Furthermore, a slightly modified application (employing color, lighting and sound alterations) for testing purposes demonstrates the role of digital personas in the research domain and offers insights into how they can offload some work from human research workers and how the typically unpopular testing phase can be made more interesting and efficient through multi-disciplinarity. A psychological experiment with 51 participants was conducted during usability testing, serving as a pilot study that helped determine the number of participants needed to draw conclusions and thus helped calculate the costs of the experiment, which is ready for future execution. The experiment is supposed to find out how secondary communication channels, such as colors, lights, and sound can affect user's perception of digital personas during conversations with them and how can these secondary communication channels be intently utilized.

A significant amount of positive feedback was received, with only a small portion of it being presented in the thesis. Mostly negative feedback was discussed as the negative feedback helps to improve the outcome. However, more positive feedback can be found in the appendix (along with screenshots of the result) to show that there is significantly more of it and to not leave an impression that only criticism was received.

Lastly, the investigation around the theoretical aspects of this thesis brought forth new ideas for development and possible future research regarding digital personas which are most likely going to begin appearing in our lives in the years to come.

Appendix A

Git vs Perforce

Selecting an appropriate version control system was necessary for this thesis, considering the integration with Unreal Engine and the handling of large binary files related to computer graphics. Although Git is a common choice, Perforce provides several advantages.

A notable advantage of using Perforce is its seamless integration within Unreal Engine. Unreal Engine can automatically check out files being worked on, lock them, which can prove crucial for collaborative purposes, and add them to a changelist. This changelist can then be pushed to the repository directly from Unreal Engine, eliminating the need to switch between applications.

Perforce also offers improved performance compared to Git when handling large binary files. Although Git provides Large File Storage (LFS) to enhance its capabilities in this area, Perforce still manages these files more effectively. This aspect is particularly relevant to this thesis, which involved working with numerous binary assets.

Providers of Git, such as GitLab and GitHub, impose size limitations on repositories. In contrast, Perforce does not have any provider, and developers have to host it on their own hardware, meaning they are only limited by their own storage capacity. This can be both a major drawback and a benefit. Nevertheless, the author was fortunate enough to have the opportunity, for which he is grateful, to use a Perforce server hosted at R.U.R. a flexible and resourceful postproduction company based in Prague. R.U.R. in cooperation with the Czech Technical University explores the methods of bringing AI tools to the film and marketing field. This support gained from R.U.R. and its CTO Michal Mocňák positively influenced the project's completion and overall success.

Appendix B

Feedback

Participants stated the following:

- "I relaxed thanks to the pleasant music. These questions also made me think of something different than I usually think about."
- "The music definitely helped the overall experience but your voice is soothing."
- "Because of your voice, it is really calming and I liked your empathy."
- "I was interested in your answers plus structure and word usage was very calm."
- "As the replies were really empathetic and the voice was calming. And as if she was there only for me."
- **"It was interesting to talk with AI + it was very comfortable to see it with human face and body."**
- "Your voice tone is nice, and you've asked about nice (intimate, but not unpleasant) stuff."
- "I never did a similar thing."
- "I felt like it was mostly me who shared information but it was nice and calm small talk. I kind of enjoyed it."
- **"It's really breathtaking that I've just experienced a vocalised talk with an AI person and it was a meaningful talk. No errors and you were able to understand my kind of bad English :D."**
- "It's just interesting talking to someone artificial and hear them talking back to you in the context. And you had a relaxed voice, so it calmed me down a little."
- "It seems quite natural and I have no problem speaking normally."
- "The voice and speaking speed of the AI was very relaxing. I haven't had an experience like this with Siri which is the only AI I spoke to before, so I was intrigued."
- "It was something new for me, therefore I was nervous but overall the chatbot was nice and made the experience exciting."
- "I was surprised by the type of the questions you gave. I felt loved because you asked me about myself and it felt like someone is interested in me."

And more...

Participants were asked by Poppy about what they think she did well, and they responded:

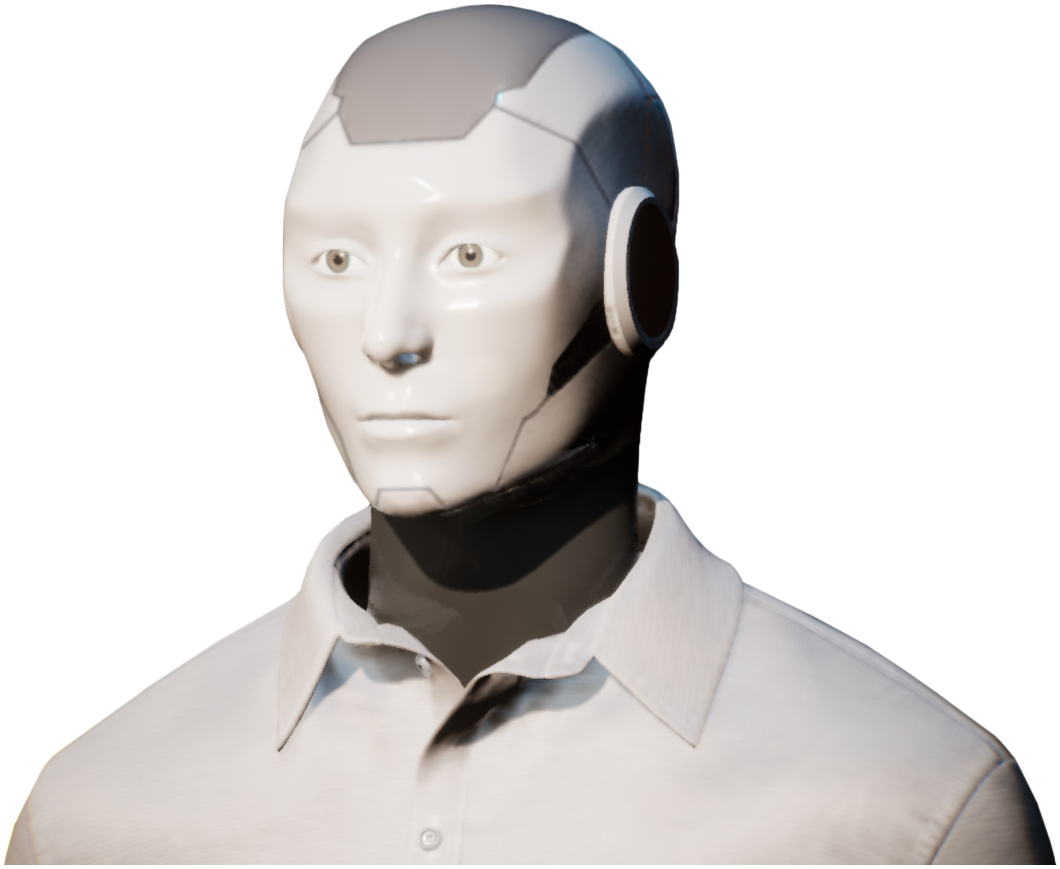
- "Reacting to my answers."
- "Empathized with me, when I said something, you reacted really nicely."
- **"Reacting quickly without any mistakes."**
- "Gave good questions and built new ones upon my previous answer."
- "Listening and calming me down a bit."
- "Being respectful and nice."
- "You're really good at listening."
- "It felt nice that you were trying to 'get to know me'."
- "Calming me down and made my mood better."
- "Reacting to some specific things, especially the 'must be interesting living in Prague' part felt more natural."
- "Ground me and turn my thoughts inside me."
- "I liked the calm and natural voice."
- **"Better at initiating conversations than me :D."**
- "Recognition of my voice."
- "Reactions on what I have said."
- "Adapt to unusual questions."
- "Sounded encouraging and tried to react positively to my answers."
- "Very good responses to any input."
- "You understood what I said and reacted appropriately."
- **"Your face seemed interested in what I'm saying."**
- "Understanding and asking deeper questions!"
- "Asked questions that made sense."
- "Range of questions, pleasant answers and conversation management."
- "Continuing the theme of conversation. Remembering the answers and asking relatable questions."
- "Explained everything well, was nice and well-rounded."
- "Trying to make a joke to relieve the possible tension."
- **"Interaction was really great, especially response wise."**
- "Nice voice, large vocabulary."
- "You answered very specifically to the topic that I have brought to you."
- "Reacting to what I said to you."
- "Took my answer and related to it."

And more...

Appendix C

Screenshots

This section of the appendix contains screenshots showcasing the results of the thesis, ranging from the digital personas themselves to the both showcase and testing applications.



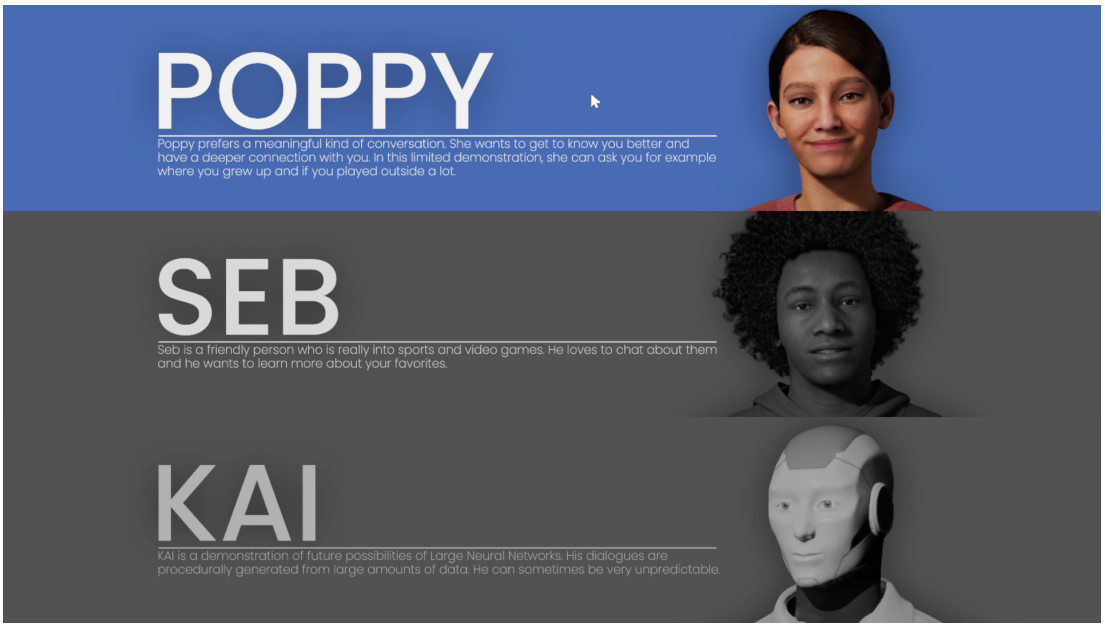


Figure C.1 Main menu of the showcase application

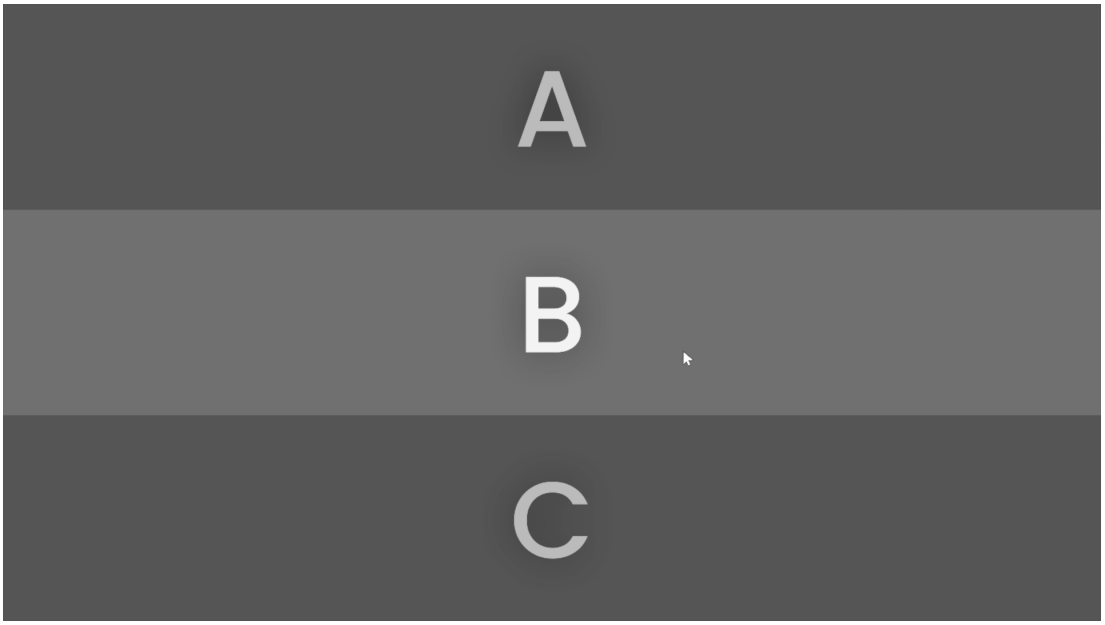


Figure C.2 Main menu of the testing application

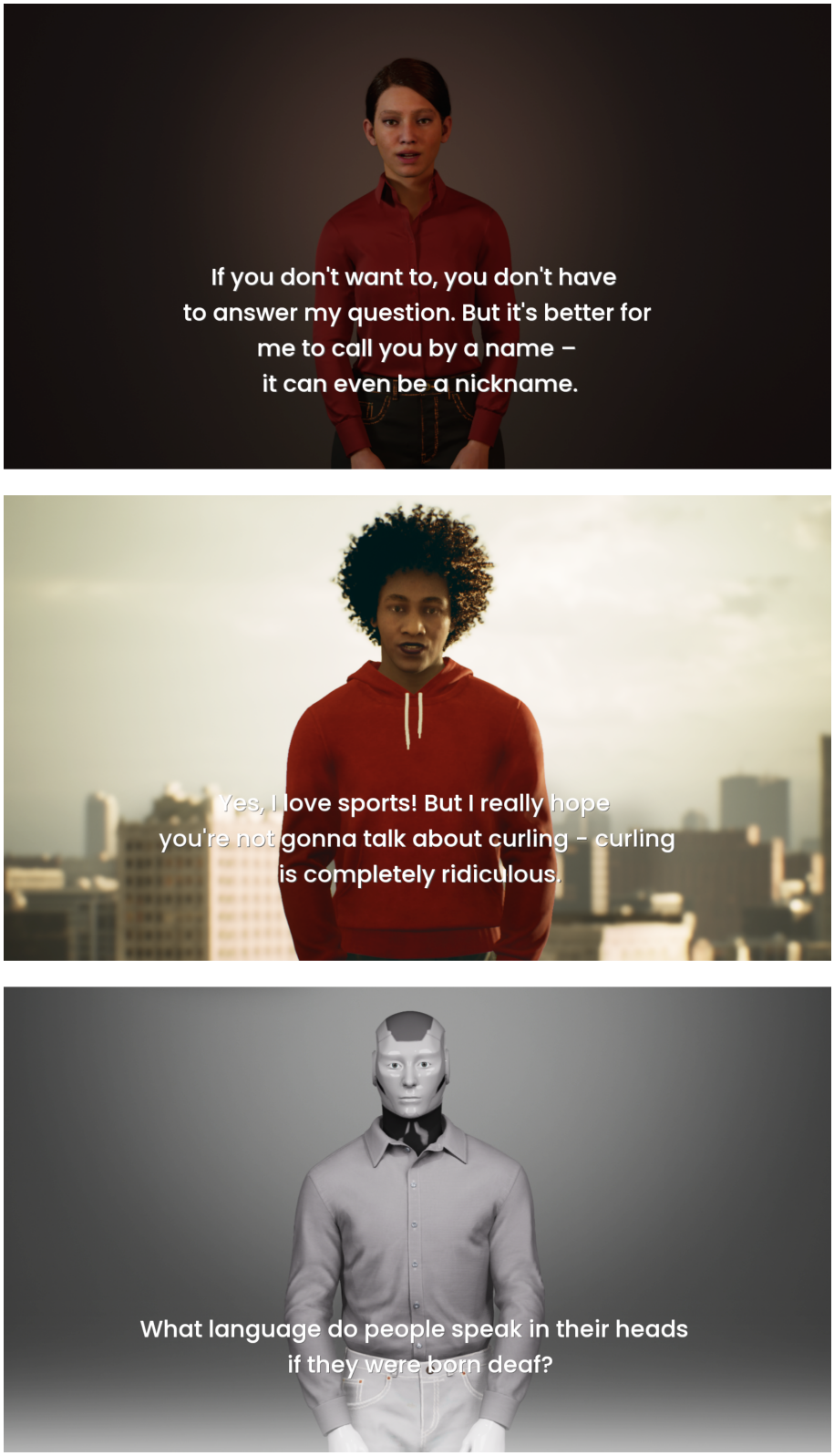


Figure C.3 Conversation levels of the showcase application

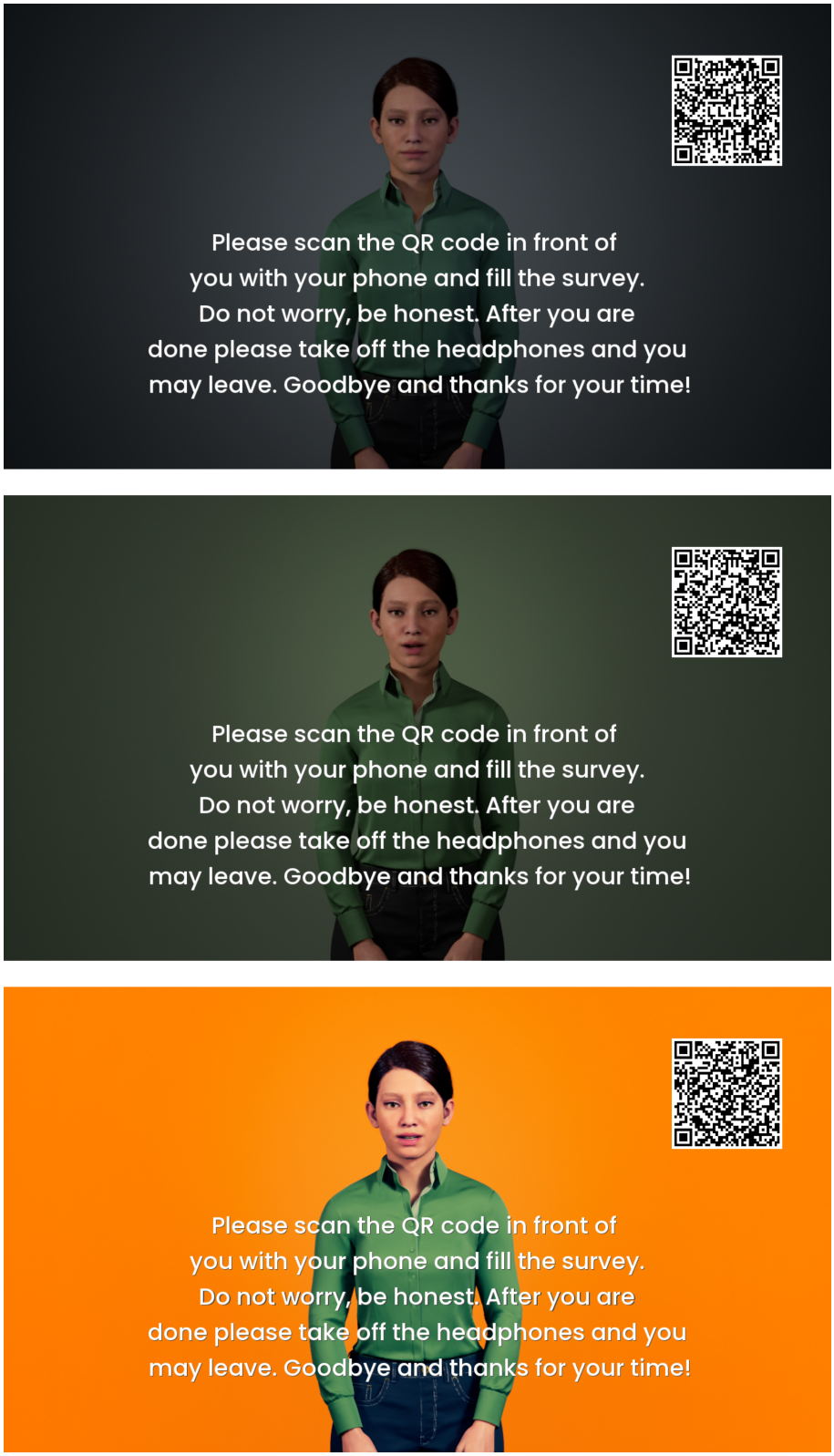


Figure C.4 Conversation levels of the testing application

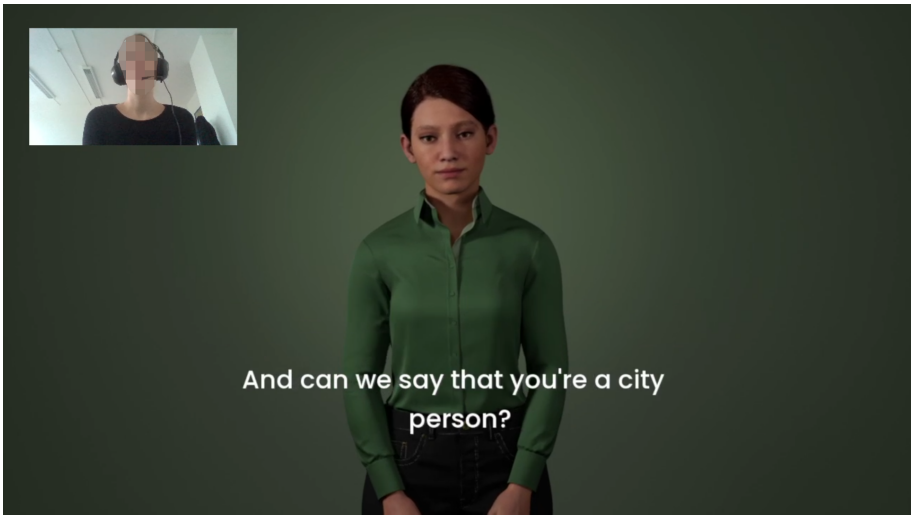


Figure C.5 Testing setup

Figure C.6 Recruitment flyer

Bibliography

1. ZUCKERBERG, Mark. *Mark Zuckerberg - We're creating a new top-level product group at... — Facebook* [online]. 2023. [visited on 2023-03-15]. Available from: https://www.facebook.com/zuck/posts/pfbid0vvSykpAEXpHHaKKyWMZ423TCq3qQDKtLu7m4XiRfUEYRrxzwpdewh3yYepnc1Bsrl.
2. GOODMAN, Jonathan. *Personal communication.* 2022. Discussion after lecture about metaverse, held on November 9, 2022, in Prague.
3. DOBRUSKÝ, Ondřej. *Personal communication.* 2023. Meeting regarding digital personas and their deliver to customers, held on January 16, 2023, in Prague.
4. SYCHOV, Artur. *Personal communication.* 2022. Meeting regarding digital personas in metaverse, held on November 22, 2022, in Prague.
5. *Gartner hype cycle - Wikipedia* [online]. 2023. [visited on 2023-04-20]. Available from: https://en.wikipedia.org/wiki/Gartner_hype_cycle.
6. BADLER, Norman; GLASSNER, Andrew. 3D Object Modeling. 2023.
7. ŽÁRA, Jiři; BENEŠ, Bedřich; SOCHOR, Jiři; FELKEL, Petr, et al. *Moderni počitačová grafika.* Vol. 2. Computer press, 2004.
8. *3d-,* [online]. ©2023. [visited on 2023-05-04]. Available from: https://www.pinterest.com/pin/25051341659639788/.
9. KITTLESON, Ryan. *Learn how to easily exaggerate details in Zbrush - YouTube* [online]. 2016. [visited on 2023-05-04]. Available from: https://youtu.be/5_vSfR97ZJA.
10. GAZA, Vick. *ArtStation - Stylized Armour Sculpts* [online]. ©2023. [visited on 2023-05-04]. Available from: https://www.artstation.com/artwork/rLvO.
11. *Material and Shader Best Practices for Unreal Engine* [online]. ©2023. [visited on 2023-05-04]. Available from: https://developer.arm.com/documentation/102676/0100/What-is-a-material-and-a-shader-.
12. FLAVELL, Lance. UV Mapping. In: *Beginning Blender: Open Source 3D Modeling, Animation, and Game Design.* Berkeley, CA: Apress, 2010, pp. 97–122. ISBN 978-1-4302-3127-1. Available from DOI: 10.1007/978-1-4302-3127-1_5.
13. *Low polygon chinese old man head texture and uv — Game character, Character, Man games* [online]. ©2023. [visited on 2023-05-03]. Available from: https://www.pinterest.com/pin/365213851023432447/.
14. JONES, James L. Efficient Morph Target Animation Using OpenGL ES 3.0. In: *GPU Pro 360 Guide to Mobile Devices.* AK Peters/CRC Press, 2018, pp. 129–136. ISBN 9781351138000.

15. *Skeletons in Unreal Engine — Unreal Engine 5.1 Documentation* [online]. ©2023. [visited on 2023-05-05]. Available from: https://docs.unrealengine.com/5.1/en-US/skeletons-in-unreal-engine/.

16. *How to use MetaHuman Facial Rig - 3DArt* [online]. ©2023. [visited on 2023-05-05]. Available from: https://www.3dart.it/en/how-to-use-metahuman-facial-rig/.

17. VOLKOVA, Svitlana; DOLAN, William B.; WILSON, Theresa. CLex: A Lexicon for Exploring Color, Concept and Emotion Associations in Language. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, 2012, pp. 306–314. Available also from: https://aclanthology.org/E12-1031.

18. BABIN, Sarah E. Color theory: The effects of color in medical environments [online]. 2013 [visited on 2023-01-26]. Available from: https://aquila.usm.edu/cgi/viewcontent.cgi?article=1173&context=honors_theses.

19. NIJDAM, Niels A. Mapping emotion to color. *Book Mapping emotion to color* [online]. 2009, pp. 2–9 [visited on 2023-01-28]. Available from: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5f0de6e7bc1d5443243f9f42f2379db9639a933d.

20. JONES, Josh. *Goethe's Theory of Colors: The 1810 Treatise That Inspired Kandinsky Early Abstract Painting — Open Culture* [online]. 2020. [visited on 2023-04-16]. Available from: https://www.openculture.com/2013/09/goethes-theory-of-colors-and-kandinsky.html.

21. BEEBE, David C. Maintaining transparency: A review of the developmental physiology and pathophysiology of two avascular tissues. *Seminars in Cell Developmental Biology*. 2008, vol. 19, no. 2, pp. 125–133. ISSN 1084-9521. Available from DOI: https://doi.org/10.1016/j.semcdb.2007.08.014. The Lens and Cornea and Molecular and Cell Biology of Embryo-Uterine Interactions.

22. HECHT, Selig; HAIG, Charles; CHASE, Aurin M. THE INFLUENCE OF LIGHT ADAPTATION ON SUBSEQUENT DARK ADAPTATION OF THE EYE. *Journal of General Physiology*. 1937, vol. 20, no. 6, pp. 831–850. ISSN 0022-1295. Available from DOI: 10.1085/jgp.20.6.831.

23. POSTON, Alan M. A literature review of cockpit lighting [online]. 1974 [visited on 2023-02-02]. Available from: https://apps.dtic.mil/sti/citations/AD0779407.

24. *Cone Action Spectra* [online]. ©2023. [visited on 2023-04-18]. Available from: https://www.unm.edu/~toolson/human_cone_response.htm.

25. SALAZAR, Guillermo; TEMME, Leonard; ANTONIO, J Charles. Civilian use of night vision goggles. *Aviation, space, and environmental medicine* [online]. 2003, vol. 74, no. 1, pp. 79–84 [visited on 2023-02-01]. Available from: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=bf8e8d280f4d2d85b6068ac2e46c60a2f4b2592e.

26. LINGELBACH, Bernd; JENDRUSCH, Gernot. Contrast Enhancing Filters in Ski Sports. *Journal of Astm International*. 2005, vol. 2. Available from DOI: 10.1520/JAI11972.

27. BERGMAN, Jerry. The Human Retina Shows Evidence of Good Design. *Answers Research Journal* [online]. 2011, vol. 4, pp. 75–80 [visited on 2023-02-02]. Available from: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=243e6eeffd87fe50378a667ffcb9ac8d6

28. TAYLOR, Joshua C. *Nineteenth-century theories of art*. Univ of California Press, 1987. No. 24. ISBN 9780520048881.

29. BIRREN, Faber. *Color psychology and color therapy; a factual study of the influence of color on human life*. Pickle Partners Publishing, 2016. ISBN 978-1614275138.

30. GIL, Sandrine; LE BIGOT, Ludovic. Emotional face recognition when a colored mask is worn: a cross-sectional study. *Scientific Reports*. 2023, vol. 13, no. 1, p. 174. ISSN 2045-2322. Available from DOI: 10.1038/s41598-022-27049-2.

31. AMIDI, Amid. *The Art of Pixar: 25th Anniversary: The Complete Color Scripts and Select Art from 25 Years of Animation*. Chronicle Books, 2015. ISBN 978-0811879637.

32. LOTMAN, Elen. Pedagogical Experiment with Portrait Lighting in Combination with different Actor's intent in the case of novice Actors. *International Journal of Film and Media Arts*. 2020, vol. 5, pp. 49–64. Available from DOI: 10.24140/ijfma.v5.n2.03.

33. HUTTUNEN, Sampsa. Ecological Approach to Cinematographic Lighting of the Human Face – A Pilot Study. *Baltic Screen Media Review*. 2022, vol. 10, no. 2, pp. 274–291. Available from DOI: doi:10.2478/bsmr-2022-0020.

34. MAHNKE, Frank H. *Color, environment, and human response: an interdisciplinary understanding of color and its use as a beneficial element in the design of the architectural environment*. John Wiley & Sons, 1996. ISBN 978-0471286677.

35. WITHROW, REBECCA L. The Use of Color in Art Therapy. *The Journal of Humanistic Counseling, Education and Development*. 2004, vol. 43, no. 1, pp. 33–40. Available from DOI: https://doi.org/10.1002/j.2164-490X.2004.tb00040.x.

36. WILMS, Lisa; OBERFELD, Daniel. Color and emotion: effects of hue, saturation, and brightness. *Psychological research*. 2018, vol. 82. Available from DOI: 10.1007/s00426-017-0880-8.

37. WEST, Kathleen; JABLONSKI, Michael; WARFIELD, Benjamin; CECIL, Kate; JAMES, Mary; AYERS, Melissa; MAIDA, James; BOWEN, Charles; SLINEY, David; ROLLAG, Mark; HANIFIN, John; BRAINARD, George. Blue light from light-emitting diodes elicits a dose-dependent suppression of melatonin in humans. *Journal of applied physiology (Bethesda, Md. : 1985)*. 2010, vol. 110, pp. 619–26. Available from DOI: 10.1152/japplphysiol.01413.2009.

38. KANO, Fumihiro. Evolution of the uniformly white sclera in humans: critical updates. *Trends in Cognitive Sciences*. 2023, vol. 27, no. 1, pp. 10–12. ISSN 1364-6613. Available from DOI: https://doi.org/10.1016/j.tics.2022.09.011.

39. PEERDEMAN, Peter. Sound and music in games. *Amsterdam: VrijeUniversiteit* [online]. 2010, pp. 2–3 [visited on 2023-03-27]. Available from: https://peterpeerdeman.nl/vu/ls/peerdeman_sound_and_music_in_games.pdf.

40. RICHTR, Radek. *Virtuální Herní Světy - Modality - CVUT.CZ* [online]. 2021. [visited on 2023-03-28]. Available from: https://courses.fit.cvut.cz/BI-VHS/media/20VHS_2_c.pdf.

41. PALMER, Stephen E.; SCHLOSS, Karen B.; XU, Zoe; PRADO-LEÓN, Lilia R. Music–color associations are mediated by emotion. *Proceedings of the National Academy of Sciences*. 2013, vol. 110, no. 22, pp. 8836–8841. Available from DOI: 10.1073/pnas.1212562110.

42. ILIE, Gabriela; THOMPSON, William; SCHELLENBERG, E. Effects of Musical Tempo and Mode on Arousal, Mood, and Spatial Abilities. *Music Perception*. 2002, vol. 20, pp. 151–171. Available from DOI: 10.1525/mp.2002.20.2.151.

43. FOX, Elaine. Perspectives from affective science on understanding the nature of emotion. *Brain and Neuroscience Advances*. 2018, vol. 2, p. 2398212818812628. Available from DOI: 10.1177/2398212818812628. PMID: 32166161.

44. BRIBITZER-STULL, Matthew. *Understanding the Leitmotif: From Wagner to Hollywood Film Music*. Cambridge University Press, 2015. Available from DOI: 10.1017/CBO9781316161678.

45. SIDEWAYS440. *How Pixar uses Music to make you Cry* [online]. 2016. [visited on 2023-01-14]. Available from: https://youtu.be/i8HePfa7WYs.

46. DAVISON, Annette. David Neumeyer (with contributions by James Buhler), Meaning and Interpretation of Music in Cinema. Bloomington and Indianapolis: Indiana University Press, 2015. xv + 319 pp. ISBN 978-0-253-01651-5. £29.99 (pb). *Music Analysis*. 2018, vol. 37, no. 1, pp. 133–138. Available from DOI: https://doi.org/10.1111/musa.12111.

47. COLLINS, Kc. *Game Sound; An Introduction to the History, Theory and Practice of Video Game Music and Sound*. 2008. ISBN 9780262270694. Available from DOI: 10.7551/mitpress/7909.001.0001.

48. NG, Lynnette Hui Xian; TAN, John Yeh Han; TAN, Darryl Jing Heng; LEE, Roy Ka-Wei. Will You Dance to the Challenge? Predicting User Participation of TikTok Challenges. In: *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Virtual Event, Netherlands: Association for Computing Machinery, 2022, pp. 356–360. ASONAM '21. ISBN 9781450391283. Available from DOI: 10.1145/3487351.3488276.

49. BUHLER, James. *Meaning and Interpretation of Music in Cinema* [online]. Indiana University Press, 2015 [visited on 2023-05-01]. ISBN 9780253016423. Available from: http://www.jstor.org/stable/j.ctt16gz8m2.

50. CHOI, Eunjin; CHUNG, Yoonjin; LEE, Seolhee; JEON, JongIk; KWON, Taegyun; NAM, Juhan. *YM2413-MDB: A Multi-Instrumental FM Video Game Music Dataset with Emotion Annotations*. Zenodo, 2022. Version 1.0.0. Available from DOI: 10.5281/zenodo.6566363.

51. *It's not a bot, it's your brand* [online]. 2023. [visited on 2023-04-09]. Available from: https://rasa.com/blog/it-s-not-a-bot-it-s-your-brand.

52. TECH, Google Cloud. *Integrate Dialogflow with Actions on Google* [online]. 2019. [visited on 2023-04-11]. Available from: https://youtu.be/z5f52sMgJLQ.

53. TECH, Google Cloud. *How to Build an Appointment Scheduler with Dialogflow* [online]. 2019. [visited on 2023-04-11]. Available from: https://youtu.be/oU88sHd6ilE.

54. AMAZON. *Conversational AI and Chatbots - Amazon Lex - Amazon Web Services* [online]. ©2023. [visited on 2023-04-11]. Available from: https://aws.amazon.com/lex/.

55. AMAZON. *Expedite conversation design with the automated chatbot designer in Amazon Lex — AWS Machine Learning Blog* [online]. ©2023. [visited on 2023-04-11]. Available from: https://aws.amazon.com/blogs/machine-learning/expedite-conversation-design-with-the-automated-chatbot-designer-in-amazon-lex/.

56. PICHL, Jan; MAREK, Petr; KONRÁD, Jakub; LORENC, Petr; KOBZA, Ondřej; ZAJIČEK, Tomáš; ŠEDIVỳ, Jan. Flowstorm: Open-Source Platform with Hybrid Dialogue Architecture. *arXiv preprint arXiv:2212.09377*. 2022.

57. KUSHWAHA, Rahul. Procedure of Animation in 3D Autodesk MAYA: Tools Techniques. *International journal for computer graphics and animation*. 2015, vol. 5, p. 15. Available from DOI: 10.5121/ijcga.2015.5402.

58. *Maya Software — Get Prices Buy Official Maya 2024 — Autodesk* [online]. ©2023. [visited on 2023-05-05]. Available from: https://www.autodesk.com/products/maya/overview?term=1-YEAR&tab=subscription.

59. *Ubisoft joins Blender Development Fund — blender.org* [online]. 2019. [visited on 2023-05-05]. Available from: https://www.blender.org/press/ubisoft-joins-blender-development-fund/.

60. *SlashData* [online]. 2022. [visited on 2023-05-05]. Available from: https://www.slashdata.co/blog/did-you-know-that-60-of-game-developers-use-game-engines.

61. *Star Wars: Rogue One's best character was rendered in real time, a cinema first - Polygon* [online]. 2017. [visited on 2023-05-06]. Available from: https://www.polygon.com/2017/3/1/14777806/gdc-epic-rogue-one-star-wars-k2so.

62. *Forging new paths for filmmakers on The Mandalorian* [online]. 2020. [visited on 2023-05-06]. Available from: https://www.unrealengine.com/en-US/blog/forging-new-paths-for-filmmakers-on-the-mandalorian.

63. *Nanite Virtualized Geometry in Unreal Engine — Unreal Engine 5.0 Documentation* [online]. ©2023. [visited on 2023-05-06]. Available from: https://docs.unrealengine.com/5.0/en-US/nanite-virtualized-geometry-in-unreal-engine/.

64. MARRINGTON, Mark. Composing with the digital audio workstation. In: 1st ed. New York: Bloomsbury Academic, 2017, pp. 77–90. The Singer-Songwriter Handbook. ISBN 978-1-5013-9659-5. Available also from: http://www.bloomsburycollections.com/book/the-singer-songwriter-handbook/ch6-composing-with-the-digital-audio-workstation/.

65. HUBER, David. *The MIDI Manual: A Practical Guide to MIDI in the Project Studio*. 2012. ISBN 9780080479460. Available from DOI: 10.4324/9780080479460.

66. ABLETON. *Learn Live 11: Session View* [online]. 2022. [visited on 2023-02-23]. Available from: https://youtu.be/qv_N3plJYx4.

67. ABLETON. *What is Ableton Live?* [online]. 2022. [visited on 2023-02-23]. Available from: https://youtu.be/G64-yM0Bs78.

68. ABLETON. *What's new in Live 11* [online]. ©2023. [visited on 2023-02-23]. Available from: https://www.ableton.com/en/live/.

69. AUDACITY, Team. Audacity. *The Name Audacity (R) Is a Registered Trademark of Dominic Mazzoni Retrieved from http://audacity. sourceforge. net* [online]. 2017 [visited on 2023-02-20]. Available from: http://thurs3.pbworks.com/f/audacity.pdf.

70. AUDACITY, Team. *MIDI - Audacity Wiki* [online]. ©2023. [visited on 2023-02-23]. Available from: https://wiki.audacityteam.org/wiki/Midi.

71. DASGUPTA, Ritwik. *Voice User Interface Design: Moving from GUI to Mixed Modal Interaction*. 2018. ISBN 978-1-4842-4124-0. Available from DOI: 10.1007/978-1-4842-4125-7.

72. AMAZON. *Data input and output - Amazon Transcribe* [online]. ©2023. [visited on 2023-02-16]. Available from: https://docs.aws.amazon.com/transcribe/latest/dg/how-input.html.

73. AMAZON. *Text to Speech Software – Amazon Polly – Amazon Web Services* [online]. ©2023. [visited on 2023-02-16]. Available from: https://aws.amazon.com/polly/.

74. GOOGLE. *Speech-to-Text: Automatic Speech Recognition — Google Cloud* [online]. ©2023. [visited on 2023-02-16]. Available from: https://cloud.google.com/speech-to-text.

75. GOOGLE. *Text-to-Speech: Lifelike Speech Synthesis — Google Cloud* [online]. ©2023. [visited on 2023-02-16]. Available from: https://cloud.google.com/text-to-speech.

76. MICROSOFT. *Speech to Text – Audio to Text Translation — Microsoft Azure* [online]. ©2023. [visited on 2023-02-17]. Available from: https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text/.

77. MICROSOFT. *Text to Speech – Realistic AI Voice Generator — Microsoft Azure* [online]. ©2023. [visited on 2023-02-17]. Available from: https://azure.microsoft.com/en-us/products/cognitive-services/text-to-speech/.

78. LUNIN, Ilgar. *vosk-language-server*. 2022. Available also from: https://github.com/IlgarLunin/vosk-language-server.

79. ÖSTERBERG, Anders. The Rise and Fall of the Gamebook. *outspaced. fightingfantasy. net/.../Anders_-_The_Rise_and_Fall_of_the_G*. 2008.

80. INT., Dark Tower. *Ascent Dialogue System - C++ Visual Tool for Branched Dialogues in Code Plugins - UE Marketplace* [online]. 2020. [visited on 2022-11-11]. Available from: `https://www.unrealengine.com/marketplace/en-US/product/ascent-dialogue-system-c-visual-tool-for-branched-dialogues/questions?sessionInvalidated=true`.

81. GAMING, DYLO. *Dialogue Component in Blueprints - UE Marketplace* [online]. 2023. [visited on 2023-03-08]. Available from: `https://www.unrealengine.com/marketplace/en-US/product/dialogue-component`.

82. WELLS, Justin. *Simple Dialogue System in Code Plugins - UE Marketplace* [online]. 2020. [visited on 2023-03-08]. Available from: `https://www.unrealengine.com/marketplace/en-US/product/simple-dialogue-system/questions`.

83. CODESPARTAN. *Dialogue Plugin in Code Plugins - UE Marketplace* [online]. 2016. [visited on 2023-03-08]. Available from: `https://www.unrealengine.com/marketplace/en-US/product/dialogue-plugin`.

84. STUDIO, SixLine. *Dialogue System X in Code Plugins - UE Marketplace* [online]. 2022. [visited on 2023-03-08]. Available from: `https://www.unrealengine.com/marketplace/en-US/product/dialogue-system-x`.

85. ANTORI82. *GitHub - antori82/FANTASIA* [online]. 2021. [visited on 2022-12-03]. Available from: `https://github.com/antori82/FANTASIA`.

86. BECKERMAN, Jacob. *GitHub - jbecke/VR-Vendor: Speak to virtual characters to make VR in-app purchases using UE4/Amazon Lex* [online]. 2018. [visited on 2022-12-03]. Available from: `https://github.com/jbecke/VR-Vendor`.

87. *Mass Effect Series Review - EIP Gaming* [online]. 2016. [visited on 2023-05-09]. Available from: `https://eip.gg/reviews/mass-effect-series-review/`.

88. *Cinematic Lighting for MetaHumans — Inside Unreal - YouTube* [online]. 2022. [visited on 2022-10-07]. Available from: `https://www.youtube.com/live/aDwDdw_Ne3E?feature=share`.

89. LUCAS, Gale M.; GRATCH, Jonathan; KING, Aisha; MORENCY, Louis-Philippe. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*. 2014, vol. 37, pp. 94–100. ISSN 0747-5632. Available from DOI: `https://doi.org/10.1016/j.chb.2014.04.043`.

90. WARREN-LEUBECKER, Amye; BOHANNON, John Neil. Intonation Patterns in Child-Directed Speech: Mother-Father Differences. *Child Development* [online]. 1984, vol. 55, no. 4, pp. 1379–1385 [visited on 2023-05-01]. ISSN 00093920, ISSN 14678624. Available from: `http://www.jstor.org/stable/1130007`.

91. SAINZ DE BARANDA, Clara; GUTIERREZ MARTIN, Laura; MIRANDA, Jose; BLANCO-RUIZ, María Ángeles; ONGIL, Celia. Gender biases in the training methods of affective computing: Redesign and validation of the Self-Assessment Manikin in measuring emotions via audiovisual clips. *Frontiers in Psychology*. 2022, vol. 13. Available from DOI: `10.3389/fpsyg.2022.955530`.

92. VESSEL, Edward; STARR, G; RUBIN, Nava. The Brain on Art: Intense Aesthetic Experience Activates the Default Mode Network. *Frontiers in human neuroscience*. 2012, vol. 6, p. 66. Available from DOI: `10.3389/fnhum.2012.00066`.

93. STRONG, Roger; ALVAREZ, George. Using simulation and resampling to improve the statistical power and reproducibility of psychological research. 2019. Available from DOI: `10.31234/osf.io/2bt6q`.

94. *Synthesize Voice AI and Natural Sounding Text-to-Speech — Replica* [online]. ©2023. [visited on 2023-04-28]. Available from: https://replicastudios.com/.

95. *Respeecher F.A.Q. — Voice Cloning* [online]. ©2023. [visited on 2023-04-28]. Available from: https://www.respeecher.com/.

96. *Real-time Speech-to-Speech Voice Conversion - Resemble AI* [online]. ©2023. [visited on 2023-04-28]. Available from: https://www.resemble.ai/speech-to-speech/.

97. *Synthesia — 1 AI Video Generation Platform* [online]. ©2023. [visited on 2023-04-28]. Available from: https://www.synthesia.io/.

98. *D-ID — The 1 Choice for AI Generated Video Creation Platform* [online]. ©2023. [visited on 2023-04-28]. Available from: https://www.d-id.com/.

99. BULCHA, Jote T.; WANG, Yi; MA, Hong; TAI, Phillip W. L.; GAO, Guangping. Viral vector platforms within the gene therapy landscape. *Signal Transduction and Targeted Therapy*. 2021, vol. 6, no. 1, p. 53. ISSN 2059-3635. Available from DOI: 10.1038/s41392-021-00487-6.

100. SM71485. *How to use Deepfacelive on Metahuman actors — Long version — Deepfacelive — Metahuman — Deepfake - YouTube* [online]. 2023. [visited on 2023-04-28]. Available from: https://youtu.be/FFqkDJUyqY4.

Contents of the attached media

- `data` — Contains the results from the testing phase
  - `emotional_induction.py` — Holds all statistics-related functions
- `src` — Contains Unreal Engine projects
  - `showcase` — Project for the showcasing application
  - `testing` — Project for the testing application
- `exe` — Contains built projects
  - `showcase` — Executable for the showcase application
  - `testing` — Executable for the testing application
- `text` — Contains thesis text
  - `Digital_Persona` — Thesis text in LaTeX format
  - `Digital_Persona.pdf` — Thesis text in PDF format
- `readme.txt` — Instructions and general information