Design and Implementation of Conversational Humanoid Avatars for Healthcare Applications

Yujing Liu

Department of Electrical & Computer Engineering

McGill University

Montréal, Québec, Canada

A thesis presented for the degree of Masters of Electrical Engineering

Abstract

The integration of conversational avatars in healthcare settings holds significant potential for improving patient engagement and therapeutic outcomes. This thesis examines the design challenges associated with developing effective conversational avatars and explores these challenges through two specific projects: Psychosis and ADiNA project.

The Psychosis project investigates avatar therapy, where patients interact with avatars resembling their hallucinations. This innovative approach provides a controlled environment for therapeutic intervention, aiding patients in confronting and managing their symptoms. The ADiNA project addresses elderly care, employing avatars to assist with daily activities and offer emotional support, thereby alleviating the burden on caregivers and enhancing the quality of life for the elderly.

Both projects underscore the importance of realistic avatar design and the necessity of advanced technologies for natural and emotionally resonant interactions. The thesis details the system architecture, backend control mechanisms, and interface design considerations crucial for creating effective conversational avatars. Pilot studies and user feedback highlight the efficacy of these solutions and suggest areas for further refinement.

This thesis contributes to the digital health field by demonstrating how well-designed conversational humanoid avatars can bridge the gap between technology and patient care, offering innovative solutions to enhance therapeutic experiences and improve overall healthcare delivery.

Abrégé

L'intégration d'avatars conversationnels dans les établissements de soins de santé offre un potentiel important pour améliorer l'engagement des patients et les résultats thérapeutiques. Cette thèse examine les défis de conception associés au développement d'avatars conversationnels efficaces et explore ces défis à travers deux projets spécifiques : Psychose et ADiNA.

Le projet Psychose étudie la thérapie par avatar, où les patients interagissent avec des avatars ressemblant à leurs hallucinations. Cette approche innovante fournit un environnement contrôlé pour l'intervention thérapeutique, aidant les patients à affronter et à gérer leurs symptômes. Le projet ADiNA porte sur les soins aux personnes âgées et utilise des avatars pour les aider dans leurs activités quotidiennes et leur offrir un soutien émotionnel, afin d'alléger le fardeau des soignants et d'améliorer la qualité de vie des personnes âgées.

Les deux projets soulignent l'importance d'une conception réaliste des avatars et la nécessité de technologies avancées pour des interactions naturelles et émotionnelles. La

thèse détaille l'architecture du système, les mécanismes de contrôle de base et les considérations de conception d'interface cruciales pour la création d'avatars conversationnels efficaces. Les études pilotes et les commentaires des utilisateurs soulignent l'efficacité de ces solutions et suggèrent des domaines à affiner.

Cette thèse contribue au domaine de la santé numérique en démontrant comment des avatars humanoïdes conversationnels bien conçus peuvent combler le fossé entre la technologie et les soins aux patients, en offrant des solutions innovantes pour améliorer les expériences thérapeutiques et la prestation globale des soins de santé.

Acknowledgements

I would like to acknowledge the assistance of OpenAI's ChatGPT for polishing the language and enhancing the clarity of this thesis, as well as the DeepL platform for translating the English abstract into French.

I would like to express my deepest gratitude to my supervisor, Prof. Jeremy Cooperstock, for his invaluable guidance and support throughout my project and thesis work. His insightful feedback and suggestions were instrumental in shaping my research.

I am also deeply thankful to my lab members for their cooperation, suggestions, and encouragement. For the psychosis project, I am particularly grateful to Clara Ducher, Hyejin Lee, Vishav Jyoti and Aishwari Talhan for helping me familiarize myself with the project and the lab. I extend my heartfelt thanks to Amanda Essebag and Julia Toledano, with whom I conducted the pilot study, and for their assistance with transportation to the experiment site. For the ADiNA project, I would like to express my sincere appreciation to Mauricio Fontana De Vargas for his guidance and support throughout the entire project, and to Romain Bazin for his collaboration and valuable advice. They have been amazing colleagues to work with.

I am also deeply thankful to Dinae Guo, with whom I spent unforgettable days in Toronto for the demo.

During my master's journey, I have been fortunate to have the support of incredible people in the lab and at McGill. I would like to thank Emmanuel Wilson, Yaxuan Li, and Linnea Kirby for their support and suggestions, which greatly enriched my experience in Montreal. I am also grateful to Xiaoxi Du, Heyang Li, Jano Fu, and Yichen Zou for their companionship and encouragement. A special thanks to my roommate, Yuhe Liu, for providing me with tremendous mental support over the past three years, and to Lingya Wang, who encouraged me throughout this process. I would also like to thank Yinan Wang, Wendy Li, May, and Le Chang for their companionship during our self-study sessions at school. My gratitude also goes to Weiming Ren for his invaluable support and encouragement throughout this journey.

Finally, and most importantly, I am profoundly thankful to my family for their unwavering emotional support and trust. Without them, I could not have completed this unforgettable journey.

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence.
API	application programming interface.
ASR	Automatic Speech Recognition.
CBT	Cognitive-Behavioral Therapy.
DM	Dialogue Management.
HMD	Head-Mounted Display.
NLG	Natural Language Generation.
NLP	natural language processing.
NLU	Natural Language Understanding.
SSML	Synthesis Markup Language.
SV2TTS	Speaker Verification to Text-to-Speech.
TCP	Transmission Control Protocol.
TTS	Text-to-Speech.
UI	User Interface.

Chapter 1

Introduction

Conversational avatars are emerging as a solution to enhance human-computer interaction in many fields. These agents, which utilize natural language processing (NLP) to simulate human conversation, are increasingly employed in various fields such as customer service, education, and healthcare. Their ability to understand and respond to user inputs in a human-like manner makes them valuable tools for enhancing user experience and engagement.

In the healthcare domain, conversational agents hold significant potential to improve patient engagement and provide mental health support. These agents might offer timely information, guidance, and emotional support, thus playing a crucial role in patient care [1–3].

Despite a flurry of interest in this topic, many current healthcare conversational agents

lack the ability to engage users on a deeper emotional level. This shortcoming limits their effectiveness, particularly in scenarios requiring empathy and personalized interaction. To address this gap, there is a growing need for more advanced, humanoid avatars. These avatars, with their human-like appearance and behavior, could simulate realistic interactions, providing a sense of comfort and familiarity that is particularly beneficial in healthcare settings.

This thesis focuses on two healthcare-related use cases and explores the design and implementation of humanoid conversational avatars. It focuses on two main projects: the psychosis project and the ADiNA project, targeting patients with hallucinations and elderly individuals who need care professionals. These projects establish initial requirements and outline a few key challenges for the effective development of humanoid avatars in two health-oriented areas.

Based on the two projects' related domains, this research will explore the technical requirements and design choices, test the effectiveness of the conversational avatar pipeline through pilot studies and demos, and reflect on the challenges and considerations for future implementation in real-world healthcare settings.

Chapter 2

Background

2.1 Conversational Agents

2.1.1 Definition and Development

A conversational agent is defined as a dialogue system that conducts natural language processing (NLP) and responds automatically using human language, simulating human-to-human interaction, and understanding context and meaning [4]. The first conversational agent is widely considered to be ELIZA in the mid-1960s [5]. ELIZA was designed to simulate conversation with a human by using a pattern-matching technique to respond to user inputs. It can have conversations with the general public, and it could also perform psychotherapy simulation: the "DOCTOR" script emulates a Rogerian psychotherapist, responding to user inputs with questions and reflections that encourage

the user to continue the conversation. This shows the potential of conversational agents and inspired later works. Most early-state conversational agents were based on predefined rules and scripts, following a decision tree to respond, and were often used for simple, repetitive tasks [6, 7].

The technique for generating conversational responses has become more advanced with the development of the large language model (LLM). Conversation agents are capable of more complex interactions and can support various areas including customer support, healthcare, e-commerce, education, and entertainment [8–12]. Based on the target user, different conversational agents are used across various platforms with suitable embodiments.

2.1.2 Conversational Agents Visualization Embodiments

The advancement of graphics and display media has significantly contributed to enhancing interaction experiences across various tasks. Conversational agents can be categorized based on their embodiment: they may exist without any visual representation, be visualized as avatars, or have physical embodiments [13]. Each type of embodiment is suited to different kinds of tasks, optimizing the effectiveness of the interaction.

Conversational agents with no visual representation, such as chatbots and voice assistants are commonly used in many commercial areas. For instance, customer service agents like those found in online chat support systems provide quick and efficient responses to user

queries. Specifically, voice assistants such as Siri and Google Assistant offer hands-free interaction for tasks like setting reminders, sending messages, and answering questions [6]. Research indicates that chatbots can handle up to 80% of routine questions, significantly reducing response times and operational costs [14]. Additionally, a survey analyzing the motivational factors behind the use of conversational chatbots found that productivity and entertainment are the top drivers [15]. From an implementation perspective, the lack of visual or physical presence allows for easy integration into various devices and platforms, making them relatively cost-effective for the companies [14, 16, 17]. Some research has also shown that users exhibit higher levels of disclosure when interacting with a chatbot compared to when an human-like avatar face is present [18, 19], suggesting that people perceive less social risk when sharing information with a non-human entity [20, 21].

Conversational avatars provide a visual representation of the agent, often resembling human-like figures. Previous research suggests that the use of human-like avatars has positive effects in fostering long-term human-chatbot relationships [22], increasing the perceived social presence of a chatbot [23], and enhancing customer satisfaction [20, 24]. It is widely studied in domains that require engaging and empathetic interaction. For example, in mental health support, Woebot [25] used facial expressions and body language to create a more relatable and comforting user experience. Similarly, digital nurses such as Molly [26] guide patients through chronic disease management with a human-like presence that can build trust and adherence to medical advice. The visual embodiment of avatars

enhances user engagement and emotional connection [27–30], demonstrating varying effectiveness across different tasks, such as improving job interview training efficacy [2] or reducing symptom severity in healthcare settings [31, 32]. Additionally, Leon et al. [33] found that participants experienced fewer uncanny effects and fewer negative effects when interacting with a simpler text-based chatbot compared to a more complex, animated avatar chatbot. This suggests careful consideration is needed during design to avoid triggering negative responses.

Physical embodiments, like robots, are best suited for tasks requiring physical interaction or presence. In hospital settings, robots like Pepper [34–36] provide guiding assistance to patients and visitors, helping them navigate complex environments. Additionally, in elder care, robots like Paro [37–39] offer companionship and assist with routine tasks, improving the quality of life for seniors. These robots can perform physical tasks and interact with the environment, offering a tangible presence that can be particularly reassuring and helpful in scenarios where human-like physical interaction is beneficial.

In conclusion, the choice of embodiment depends on effectiveness and suitability for various tasks. By leveraging the appropriate embodiment, developers can optimize conversational agents to meet the specific needs of different applications, enhancing overall user experience and task performance.

This thesis focuses on avatar-based visualization and discusses the system design for two healthcare projects, both of which utilize realistic avatars to enhance user engagement

and empathy, fulfilling supportive and interactive roles. However, several broader questions remain beyond the scope of this work, such as the long-term effectiveness of avatar-based interactions on patient outcomes, response generation within dialogue system design, algorithmic improvements for graphical rendering, the scalability of such systems for widespread adoption, and the ethical implications. Further research is needed to explore these larger questions in the related field.

2.2 Conversational Avatars

2.2.1 Effectiveness of Humanoid Avatars in Healthcare

The use of humanoid conversational avatars in healthcare shows potential to improve outcomes across various domains due to their ability to engage patients in human-like interactions. For example, the SimSensei Kiosk [1] has been instrumental in supporting individuals with psychological distress, such as depression and PTSD, by creating engaging interactions where users feel comfortable sharing personal information, leading to improved assessments and more informed healthcare decisions. Similarly, Virtual Reality Job Interview Training (VR-JIT) [2] has shown significant promise in vocational training for adults with autism spectrum disorder, resulting in improved job interview skills, and increased self-confidence. Conversational avatar platforms like “superhero therapy” [3] for chronic pain management have shown the potential to reduce pain intensity.

The consistent positive outcomes across diverse healthcare scenarios underscore the potential of humanoid avatars to become an integral component of modern healthcare delivery. As the field continues to evolve, integrating these avatars into healthcare systems promises to improve patient interactions, ensuring more personalized and effective care.

2.2.2 2D and 3D Avatar Generation

This subsection addresses the method of producing static 2D and 3D avatars, discussing current trends and existing tools. The following Section 2.4 will specifically explore techniques for animating these representations.

2D Avatar Generation

For 2D avatars, the design can encompass various styles and forms, facilitated by numerous tools capable of generating stylized avatars. Neural Style Transfer (NST) [40], utilizes convolution network enables the creation of avatars that combine realistic features with artistic styles. generative adversarial networks (GANs), particularly the StyleGAN [41] architecture have revolutionized image generation by producing high-quality, diverse avatars with customizable stylistic attributes. Diffusion models [42] have also emerged as powerful generative models capable of producing high-quality images. Different mature commercial platforms were also developed, for example, Cartoonify could transform photos into cartoon-style images, Zmoji provides emoji-style avatar creation, and tools like Stable

Diffusion [43] and Midjourney also allow users to generate avatars based on text input.

These methods and architectures have been further refined to handle complex style transformations, such as converting avatars into cartoonish, Pixar-like or claymation styles. There are also advancements in different platforms in enhancing the quality and realism of these transformations, reducing artifacts, improving the efficiency of the algorithms, providing artists and developers with powerful tools to create unique and personalized digital representations, and expanding the creative possibilities in avatar design.

3D Avatar Generation

There are also numerous styles for 3D avatars. A comprehensive review [30] summarizes various types, including but not limited to cartoon, realistic, robotic, stylized, point cloud, and video avatars. The body parts represented can range from full-body to head-only, or head and hands.

Simplistic 3D avatars can be created using 3D asset creation software like MAYA or game engines. For more complex humanoid avatars, tools like Character Creator and Unreal's MetaHuman provide robust features for creating detailed humanoid avatars. Advanced architectures using Neural Radiance Fields (NeRF) [44], as well as products like Avatar SDK and ReadyPlayerMe, are well-developed solutions for generating photorealistic or cartoonish avatars based on input photos.

2.2.3 Advantages of Realistic Avatar

Designers can generate stylized avatars using advanced tools, making the choice of style an open question in conversational agent design. Realistic styles are now a popular choice, as these avatars have been shown to enhance the sense of presence, which is crucial for immersive experiences [27–29]. Specifically, when using immersive VR scenes as display media, it's essential to design a full-body avatar so users can view them from different perspectives. However, in other display media such as 2D screens, it's not necessary to show the full body for realism. Nonetheless, displaying more body parts and body motion may enhance the interaction experience.

The realistic design was also shown to improve user experience by providing a more natural and intuitive interaction [28, 29]. A systematic review by Weidner et al. [30] highlights that the use of full-body humanoid avatars in AR and VR offers extensive benefits across various applications. When users can see and control a complete virtual body, they experience higher levels of embodiment and body ownership. Increasing task performance such as in physical activity, communication, and education or training is another area where realistic avatars provide significant advantages [45]. Moreover, full-body avatars contribute to improved social presence and co-presence, which are critical for collaborative and multiplayer applications.

These consistent advantages observed across multiple studies underscore the importance of integrating full-body avatars into AR and VR systems to maximize user engagement and

effectiveness. In the design of formal conversational avatar platforms, opting for a realistic rendering style may be preferable.

2.3 System Architecture

Research on designing conversational avatar pipelines can generally be categorized into two approaches: fully autonomous systems and partially autonomous systems that involve human experts in response generation. Fully autonomous pipelines are often preferred for their efficiency and are considered more advantageous as they do not require human resources. Additionally, research in this area frequently overlaps with studies on dialogue systems, which do not inherently require the presence of an avatar.

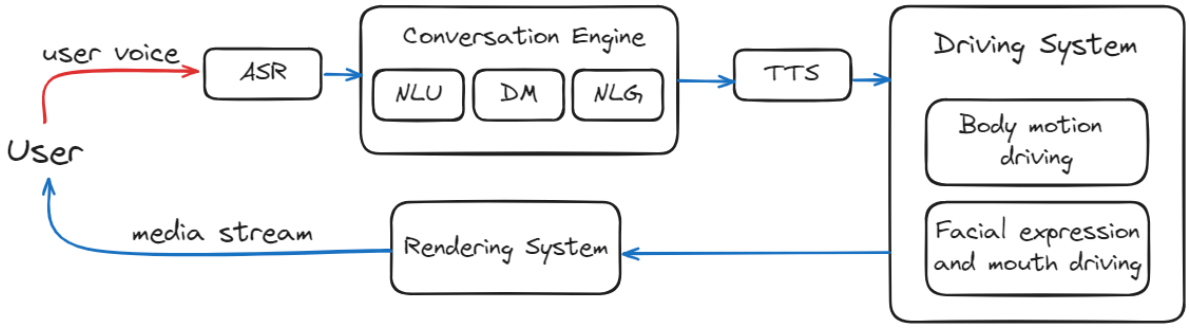


Figure 2.1: An example dialogue system architecture

Figure 2.1 shows an example of a dialogue system architecture design, which follows a structured workflow integrating various components to process and respond to user inputs. While different parts of the system may have different names in various research, the workflow involving ASR, the three stages of the conversational engine, and TTS modules constitutes

a widely recognized pipeline [46–48]. And the key components of the workflow are:

ASR (Automatic Speech Recognition): Converts spoken input from the user into text. This is achieved using advanced speech recognition algorithms and models such as Google Speech-to-Text API and Amazon Transcribe, which have demonstrated high accuracy in transcribing spoken language.

Conversation Engine: The classic architecture [47–49] contains three core components: natural language understanding (NLU), dialogue management (DM), and natural language generation (NLG). NLU focuses on interpreting and understanding the user's input. DM manages the flow of the conversation, maintaining context and determining the appropriate action based on the current state and history of the dialogue. NLG converts the system's response into coherent and contextually appropriate natural language, ensuring the generated text is meaningful and engaging. Together, these components enable seamless and intelligent interactions between humans and machines.

TTS (Text-to-Speech): Converts system-generated text responses back into speech. TTS technologies like Google Text-to-Speech and Amazon Polly enable the system to communicate verbally with the user, ensuring accessibility and a natural interaction flow.

Driving and Rendering System: The driving and rendering system is responsible for the visual and auditory representation of the avatar. This system integrates several components to ensure the avatar's movements and expressions align with the spoken dialogue, creating a cohesive and immersive interaction.

In certain applications, the need for professionally controlled dialogue is paramount. Human expert-controlled feedback systems allow professionals to control the avatar's responses and behaviors, thereby ensuring a high degree of customization and contextual relevance. This is particularly evident in domains such as therapeutic avatars and educational avatars, where the precision and appropriateness of responses are crucial for the efficacy of the intervention. In these cases, the dialogue generation process will be different based on the professional issuing commands to control the avatar, while other parts remain similar.

In the following chapters, this thesis will discuss the design and implementation of two conversational avatar projects. Chapter 3 will cover the Psychosis project, focusing on a workflow involving therapist control, while Chapter 4 will detail the ADiNA project, which centers on an autonomous workflow.

2.4 Motion Design

This section will focus on the non-verbal motion design of the system, including the generation of talking heads, facial expressions, eye movements, and body motions.

2.4.1 2D Avatar Talking Head Animation

The conventional approaches to generating talking-head avatar videos follows a multi-step procedure. Initially, the avatar's response audio is analyzed to identify visemes, which are

the visual correlates of phonemes. A predefined set of phoneme-mouth correspondences is utilized to modify mouth shapes to match the identified visemes.

The synchronization of mouth movements with the audio is then controlled via a scripted animation sequence. For instance, a phoneme such as /p/ might be mapped to a specific mouth shape where the lips are pressed together. However, this method often fails to produce natural transitions between mouth shapes, rendering it more suitable for cartoonish avatars rather than realistic ones.

Recent advancements have favored the use of machine learning methods [50]. Some use pipeline methods, which involve two main steps: mapping low-dimensional driving source data (e.g., audio) to facial parameters and converting these parameters into high-dimensional video output. These methods can be further classified based on the data type used for facial parameters including the landmark-based method, coefficients-based method and 3D vertices-based method.

Others use end-to-end methods to directly generate talking-head videos from audio without intermediate facial parameter steps. For example, Wav2Lip [51] used a generative adversarial networks(GAN) based lip-sync discriminator to ensure accurate lip synchronization. AD-NeRF [52] used DeepSpeech audio features to render dynamic neural radiation fields for speaker-face rendering.

Despite their advantages, these models also have limitations. For example, the original Wav2Lip [51] can only generate fixed-size images with 96×96 resolution, potentially

mismatching the original input source and design specifications of an avatar. Additionally, integrating the generated video clip into an overall system pipeline necessitates a merging step, complicating the process.

Furthermore, head and body movements are not explicitly controllable, as they are learned from training data. This can lead to inconsistent temporal coherence in video, which refers to the consistency of visual elements across consecutive frames, particularly during motion. Poor temporal coherence can result in misalignment between lip movements [51] and motion jitters [53], which adversely affects the video quality and user experience.

2.4.2 3D Avatar Talking Head Animation

In the domain of 3D model-based conversational avatars, transformations can be controlled by manipulating facial 3D information of the models, including mesh vertices, blend shapes, and bone transitions. In the game and movie industries, manipulating mesh vertices, blend shapes, and bone transitions are prevalent techniques. Typically, this involves scripts that control predefined animation clips, triggering changes based on specific rules. For example, these scripts may initiate facial expressions or movements in response to particular audio cues. While this method demands considerable human effort and the design of intricate rules, it offers a high degree of control over the final animation.

Recent advancements have introduced automated techniques that generate 3D

parameters based on audio input without relying on predefined animation clips. Notably, Nvidia's Audio2Face [54] and VOCA (Voice Operated Character Animation) [55] exemplify methods that generate blend shape features directly from audio data. These techniques obviate the need for manually defined animation clips, streamlining the animation process.

Despite their advantages, automated methods present certain limitations. Some techniques can be computationally intensive, potentially affecting real-time performance. Methods often operate independently of the engines that display 3D avatars. Consequently, an additional integration step is required to merge the generated data with the avatar and synchronize it with the spoken audio. Also, the generated animations may not always achieve perfect lip-sync, which can be misleading for users who prioritize accurate synchronization between audio and visual elements.

In conclusion, both manual and automated methods have their respective strengths and challenges, and the choice of method depends on the specific requirements of control, integration, and real-time performance within the application context for different tasks.

2.4.3 Other Non-verbal Animation

Non-verbal animations, such as facial expressions, eye movements, and body motions, play a crucial role in conveying a speaker's intentions and emotional state [56]. It is important to consider these factors in virtual avatar design, as incongruous expressions may exacerbate the uncanny valley effect [57]. The use of non-verbal animations shows potential to enhance the

perception of virtual agents [58, 59]. These behaviors can convey traits such as friendliness, warmth [60], and competence [61], also increasing the sense of presence in virtual reality environments [62].

Facial Expression

Facial expressions are crucial for emphasizing speech content and expressing emotions. They provide non-verbal cues that can significantly enhance the interaction experience. It is important to consider facial expressions in designing conversational avatars [57], as research has shown that a perceived lack of facial expressions increases the "uncanniness" effect. Studies have shown that users perceive conversational agents with well-designed facial expressions as more believable and relatable, which enhances the overall interaction experience [63, 64].

In 2D, advanced models like Styletalk [65] and SadTalker [66] can generate talking heads with facial expressions in video or picture format, helping to make the interaction more engaging. In 3D, models [55, 67] and some advanced platforms such as Nvidia's Audio2Face [54] have the option to generate realistic facial expressions along with lipsync. These models track multiple facial points and adjust the avatar's facial features to match the user's expressions in real time, providing high level of realism. Another direct method involves using scripts in game engines to adjust blendshape coefficients, enabling detailed control over facial animations.

Eyeblink and Gaze

Eyeblinks and gaze are powerful tools for managing conversation flow, providing feedback, and seeking responses. For example, gaze behaviors in animated agents can simulate looking at the user to indicate turn-taking or understanding, which significantly enhances the naturalness of interaction.

In 2D, some recent open source network models [66,68] and commercial platforms [69,70] can achieve these effects and integrate them seamlessly into the animation. In 3D, trained models [55, 67] can output detailed eye movements, which can be further enhanced with game engine plug-ins. Tools like SALSA [71] and Meta's Unity engine plug-in [72] allow for precise control of blendshape coefficients, enabling realistic eye animations and blinks that reflect the avatar's focus and attention.

Body Motion Animation

Effective body motion animation, including hand gestures, head movements, and overall body movements, is essential for enhancing the realism of interactions. For example, Munhall et al. [73] showed that the rhythmic beat of head movements increases speech intelligibility.

In 2D, this can be achieved by utilizing models trained on datasets with body motion data, generating coordinated body movements that match the speech and context of the interaction [74–76]. Many commercial platforms [69, 70] also provide APIs to generate avatar videos. In 3D, some methods generate body control-related parameters. For instance, Nvidia's

Adio2guesture [77] can generate realistic body poses based on audio input, though these methods are often separate from talking face generation and require merging. Game engines offer a more direct approach with animation controllers that trigger different poses at the right time. These animations can be custom-designed or sourced from open platforms like Mixamo, allowing for a wide range of expressive movements.

2.5 Speech Generation Design

This section will focus on the audio component of the pipeline design, including speech generation methods and design details for enhancing overall performance.

2.5.1 Speech Generation Methods

Text-to-speech (TTS) technologies continue to improve over time, with many state-of-the-art techniques sounding remarkably close to the human voice. Many end-to-end neural networks, such as Tacotron [78] and WaveNet [79], directly map text to speech waveform [80]. Models like StyleSpeech and Meta-StyleSpeech [81] further enhance personalization by allowing customization of the speech output to match specific requirements. Additionally, many companies offer cloud services and APIs for TTS, making these advanced capabilities more accessible.

The TTS method typically uses predefined voices for output, which is suitable for tasks that do not require a specific voice. However, in some cases of conversational avatar design,

it is crucial to output the response in a specific type of voice. This requires addressing the challenge of voice synthesis to ensure the generated voice matches the desired characteristics. This field is also being widely researched, with neural models such as HiFi-GAN [82] and VITS [83] enabling personalized and adaptable voice outputs [84,85]. Some models also have the ability to clone voices across different languages and speaker characteristics [86,87].

In the following chapters, Chapter 3 covers the Psychosis project, focusing on a workflow using cloned voices, while Chapter 4 details the ADiNA project using the predefined voice.

2.5.2 Dialogue Response Design Details

Adding more details to the dialogue response can significantly enhance the user experience. For instance, the use of vocal fillers can better control the conversation flow, increase social presence [88], and improve both the comprehension of information presented by an agent and the naturalness of the agent's voice. Additionally, the voice–body continuum is important when selecting a voice to match the avatar's appearance, as mismatches may result in awkwardness. More realistic voices tend to align better with realistic bodies, and social presence is rated higher when the body and voice are matched in terms of pitch [89]. Furthermore, compared to voice alone, combining voice with facial expressions results in higher social agency scores [90]. Incorporating emotion into voices can also greatly enhance interaction quality, as users provide better feedback when the emotional tone of the voice is matched with facial expressions [91].

2.6 Design and Technical Challenge of Implementation

Even though conversational avatars are becoming increasingly popular in healthcare areas and other areas, there are still some open questions and design challenges, both in design and technical aspects.

2.6.1 Design Challenges

Designing avatars requires balancing form realism with behavioral realism. Form realism refers to the extent to which the avatar's shape appears human, while behavioral realism captures the degree to which it behaves as a human would in the physical world [21].

With the advancement of graphics technology, form realism has significantly improved thanks to various design tools and software. As a result, professionals in the field are now increasingly focusing on behavioral realism and the balance between these two factors.

Unsatisfactory design on one of the two aspects or misalignment between these aspects can lead to poor user experiences. For example, Nordnet's AI assistant Amelia [92], a 3D avatar designed with high form realism was initially introduced to help streamline customer onboarding and service to enhance customer interactions. However, despite the realistic appearance and AI capabilities, Amelia's performance did not meet the company's expectations and the response from the customers was OK but not overwhelming. As a result, the company decided to prioritize other areas within its AI strategy and eventually terminated the service.

Another example is IKEA's virtual assistant, Anna [93], a 2D conversational avatar introduced to help customers with IKEA's product inquiries. Users reported that Anna failed to understand and correctly respond to basic queries and failed to handle more complex questions [94]. Consequently, IKEA eventually decided to retire the avatar.

These examples underscore the difficulty and importance of balancing behavioral realism with appropriate form realism to meet user expectations and enhance user experience effectively. According to [95] users tend to have high expectations for avatars, especially those with high form realism. If an avatar looks very human-like, users expect it to behave intelligently and naturally. Failing to meet these expectations can result in negative user experiences and decreased trust in the avatar.

The considerations discussed are important in various fields, including healthcare, which is the focus of this thesis. In healthcare applications, the interface should be intuitive and easy to use for all users, regardless of their technical proficiency. This involves designing user-friendly interfaces with clear instructions, simple navigation, and user-friendly error messages.

The conversational avatar interaction setting should adapt its behavior based on user interaction preferences. For example, some users might prefer interacting with the avatar within a closer physical distance in the VR environment, while others may not. Personalizing avatars to meet the specific needs and preferences of individual users is also challenging, since different users will have varying preferences for appearance, voice, and interaction

style. Therefore, designing adaptive platforms that can accommodate user preferences is important.

2.6.2 Technical Challenges

Integration Complexity

The integration of various components to ensure smooth operation and minimize compatibility issues presents a significant challenge in the design of conversational avatars. The architecture of such systems typically comprises multiple interconnected modules, including natural language processing (NLP), speech recognition, dialogue management, and response generation. Each module must communicate seamlessly with others, necessitating robust data transmission protocols and interoperability standards. The complexity arises from the need to ensure that data flows correctly between modules, maintaining the integrity and consistency of information throughout the system. This involves dealing with different data formats, synchronization issues, and potential bottlenecks that could impede real-time performance.

Time Delay Problem

The issue of time delay, particularly in steps like voice generation, is a critical concern in maintaining smooth and responsive interactions in conversational avatars. A real-time factor (RTF) [52] of less than or equal to 1 is essential to ensure that the system can process inputs

and generate responses in real-time, without perceptible lag. Time-consuming processes within the system, such as complex speech synthesis or extensive NLP computations, must be optimized to meet this requirement. Designers must explore and implement various methods to minimize latency, such as parallel processing, efficient algorithms, and hardware acceleration. The goal is to achieve real-time performance, where the system responds to user inputs instantaneously, thereby enhancing the user experience and ensuring fluid, natural interactions.

Scalability and Robustness

Ensuring the scalability and robustness of conversational avatar systems is crucial for handling varying user loads and maintaining reliability in diverse operational environments. Scalable architectures, such as cloud-based solutions and distributed computing, could dynamically allocate resources based on demand.

In conclusion, the design of conversational avatars entails addressing significant challenges related to integration complexity, time delay, scalability, and robustness. By developing sophisticated integration strategies, optimizing for real-time performance, and ensuring scalability and robustness, designers can create advanced conversational systems that provide seamless, responsive, and reliable user experiences.

Chapter 3

Psychosis Project

Preface

This project is the result of a collaborative effort, currently under submission as a conference paper. The psychosis project focuses on developing a platform for the treatment of patients experiencing auditory or visual hallucinations through avatar therapy. This treatment method helps patients externalize their hallucinations and directly confront them, with the therapist role-playing the hallucination. The project encompasses two stages:

1. Creating a suitable platform for generating the audio and visual representations of hallucinations.

2. Involving patients in interactions with their hallucinations, with the therapist

controlling the flow of these interactions.

In this chapter, the thesis focuses on the second stage of the design and implementation of the interaction interface, discussing the architectural design choices. A pilot study was conducted. While the pilot study employed the full architecture, this thesis focuses on the second stage, analyzing how it contributes to the user's sense of interaction with their hallucinations and its effectiveness in helping them gain a sense of control. The thesis also discusses the limitations of the current platform based on participant feedback. Specifically, the background overview in Section 3.1 is based heavily on the IRB application of the project and the pilot study described in Section 3.4 is derived from the content presented in the conference paper in submission.

Author's Contribution

For the first stage of the hallucination generation platform design Clara Ducher implemented the visual hallucination generation UI, and Hyejin Lee implemented the audio hallucination generation UI used in the pilot study.

In the second stage of interaction interface development, which is the focus of this thesis, Clara Ducher and Vishav Jyoti collaborated on the backend interface implementation, voice cloning implementation, and connection method exploration. Hyejin Lee explored the use of avatar generation and lipsync tools, designing the first version of the Unity frontend. Building

on previous work, Yujing Liu implemented the Unity interfaces for both the therapist and patient sides and finalized the data transmission connection between the backend and the two Unity projects.

For the pilot study, Kaylee Novack wrote the IRB application, which heavily informed the Background section. Amanda Essebag and Kaylee Novack finalized the questionnaire used in the study. Yujing Liu integrated the adjusted pipeline of the backend and the Unity project running on the VR headset. Amanda Essebag, Julia Toledano, and Yujing Liu conducted the study at the Douglas Hospital, with Amanda Essebag and Julia Toledano taking on the roles of therapists, while Yujing Liu provided technical support. Yujing Liu and Amanda Essebag analyzed the collected data.

Professor Jeremy R. Cooperstock supervised the entire project.

3.1 Background

Psychosis describes when a person has lost touch with reality, and often results in a collection of symptoms that affect the mind and cause a loss of grasp on reality, including auditory and visual hallucinations. It is often a symptom of an underlying mental health condition, such as schizophrenia, severe depression, or bipolar disorder [96].

To address the disease, the use of antipsychotic medication often faces challenges, including patient resistance to taking the medicine [97, 98]. A previous one-year randomized trial shows that 20% to 50% of patients do not respond to clozapine, a

standard treatment [99]. Even worse, some patients suffer from treatment-resistant schizophrenia [100] and also experience many side effects [101]. Therefore, different from pharmacological interventions, psychosocial interventions have become extensively endorsed in clinical practice guidelines as part of the treatment [102–104]. The most widely studied evidence-based treatment recommended for psychotic symptoms is cognitive–behavioral therapy (CBT) for psychosis, focusing on helping patients identify and change negative thoughts about the voices they hear through each session [104, 105]. Globally, most studies have found CBT effective in ameliorating psychotic symptoms [106–108]. However, many studies show modest treatment effects compared to other psychotherapies. Research has also shown that up to 50% of patients do not respond to this approach [109].

Given this, researchers are exploring interventions specific to the experience of seeing or hearing hallucination, and a new wave of therapy utilizing visual depiction, known as avatar therapy, is gaining attention [31, 110]. Unlike CBT, avatar therapy focuses on encouraging patients to confront their hallucinations directly. In this treatment, a representation of the patient's hallucination is created during the initial stage. In the following sessions, the therapist role-plays the hallucination to aid and encourage patients in practicing different responses. By being involved in creating the avatar, patients externalize their hallucination, which can give them more courage to confront it. This process also makes them more likely to confront and question the existence of their hallucinations on their own, compared to relying on the therapist's suggestions [111, 112].

Previous experiments conducted in hospital environments have demonstrated the potential for amelioration of symptoms. Dellazizzo et al. [32] presented a single case study using a VR headset display, in which a patient who had been hearing voices for 20 years showed significant improvement and complete amelioration of auditory visual hallucinations (AVH). The same group conducted a one-year single-blind randomized controlled trial [100], comparing the effectiveness of avatar therapy with cognitive-behavioral therapy (CBT). The results indicated that, in the short term, both interventions produced significant improvements in AVH frequency and depressive symptoms. Additionally, avatar therapy demonstrated significant effects on persecutory beliefs and quality of life.

Another team utilized a 2D screen for display, followed by a proof-of-concept study [31], and compared avatar therapy with supportive counseling (SC) delivered over six weekly sessions. Avatar therapy led to greater reductions in auditory hallucinations and related distress at 3 months, with improvements maintained at 6 months, though there were no significant differences between the two methods [113].

The trial by Aali et al. [114] compared avatar therapy with treatment as usual (TAU) and supportive counseling (SC). The study did not demonstrate any positive or promising results for the avatar therapy method. While there were suggestions of effects, the team declared these findings uncertain due to the risk of bias and their unclear clinical significance.

Overall, there are still open questions regarding this treatment, and its generalizability

to patients remains unclear. Overviews indicate that avatar therapy has potential but remains experimental [115, 116]. Most research studies focus on the treatment procedure and analyzing its effectiveness in different domains, with few detailing the design and implementation of the architecture. Some studies introduce the platforms and tools used for generating hallucinations [31, 32] rather than the overall architecture, particularly the interaction capabilities. As display tools continuously evolve, there are increasing possibilities for improving interaction methods. The lack of detailed architectural design introduces difficulties in reproducing these experiments. Therefore, in our psychosis project, one of our goals is to explore and implement a comprehensive pipeline for both patients and therapists, addressing these gaps and enhancing the effectiveness of therapeutic interventions.

3.2 System Design

3.2.1 Design Considerations and System Architecture

The objective of this platform is to facilitate avatar therapy by providing an interface for both therapists and patients. The platform aims to enhance therapeutic outcomes by allowing therapists to control the avatar and simulate the patient's hallucinations during the treatment sessions.

The therapist's interface should be user-friendly and provide functionalities to control

the beginning and end of sessions. Additionally, the system should allow for timely interactions using the therapist's voice, enabling introductions, real-time guidance, or interruptions as needed during the session. Therapists need the ability to pretend to be the patient's hallucination, speaking as if the hallucination is addressing the patient. This requires the functionality of switching between the therapist's voice and the simulated hallucination voice. Additionally, the interface should enable therapists to monitor the patient's status in real-time and switch back to their original voice if necessary.

For the patient side, the platform should ensure that the patient requires minimal interaction with the interface, aside from engaging with the avatar. All controls should be managed by the therapist.

The system is designed with three main components: the patient interface, the therapist interface, and the backend system, as shown in Figure 3.1. The patient interface displays the hallucination avatar, allowing the patient to engage directly with it. The therapist interface is responsible for controlling the session flow and communication content. The backend system, situated on a laboratory server, manages communication, data processing, and voice generation for the entire pipeline.

For the psychosis project, the visual and audio representation of the hallucination was designed in advance using tools developed by other teammates. For the work discussed in this thesis, it begins with a 2D photo representation of the hallucination's face and a voice sample. To build on this foundation and utilize a 3D model, the team employed Avatar

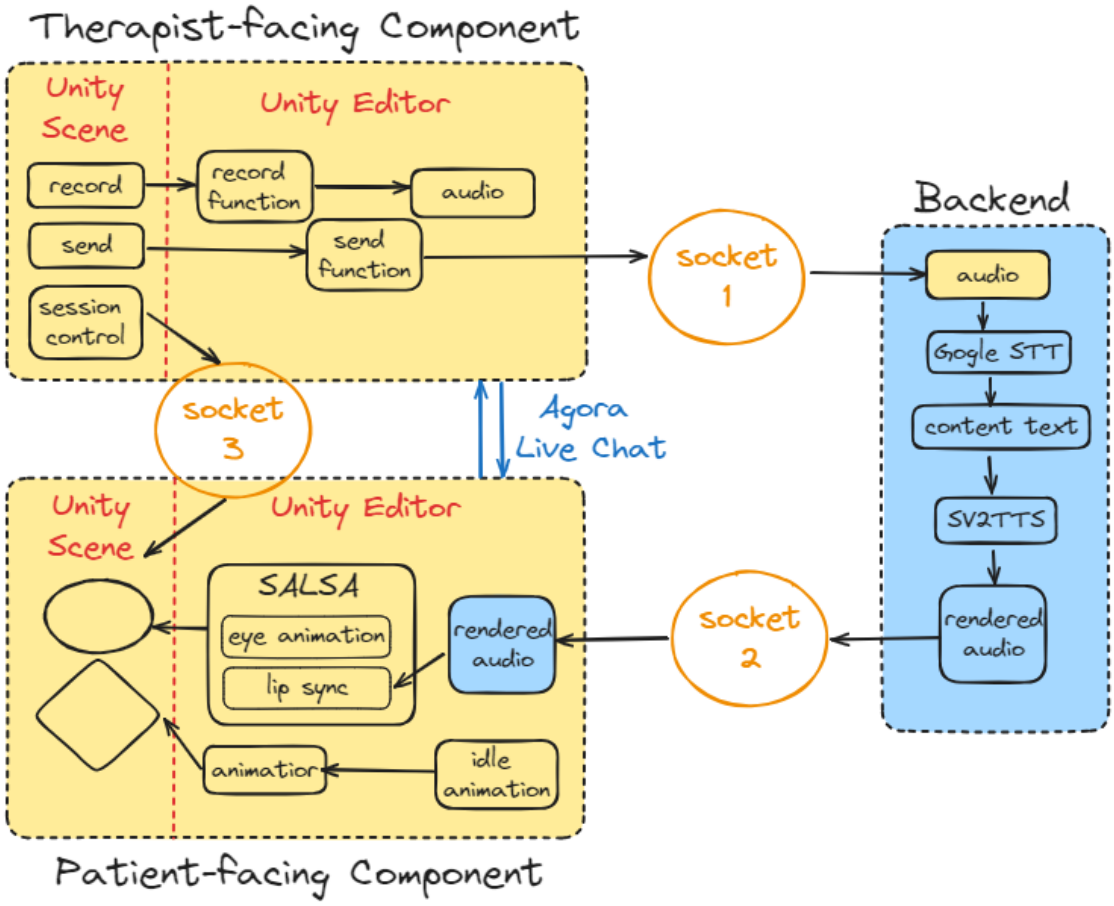


Figure 3.1: Project system architecture

SDK [117], a mature tool for generating photorealistic 3D avatars from 2D images. An example is shown in Figure 3.2 (a). The Unity Editor is selected for its robust support for multiuser development and 3D humanoid model animation, ensuring smooth interaction and control. The audio representation will be used in the response generation part, ensuring that the generated conversation audio spoken by the avatar closely matches the voice of the hallucination.

(a) Example of 2D avatar appearance

(b) Patient-facing interface with the generated 3D avatar from the 2D picture

Figure 3.2: Patient hallucination example and the designed UI

3.2.2 Patient Interface

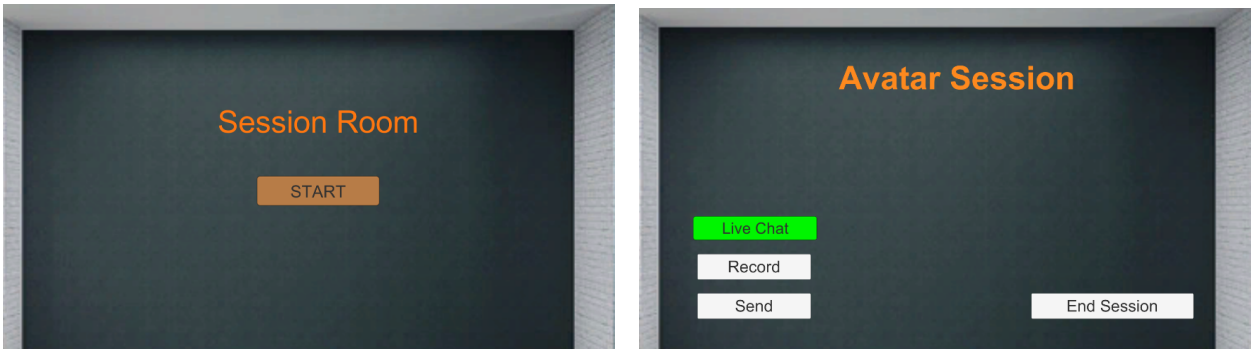
Participants are presented with a Unity project scene containing a humanoid avatar, as shown in in Figure 3.2 (b). They do not need to press any button inside the scene and will focus on interacting with the avatar of their hallucination. When viewed on a PC screen, the scene will exhibit only the upper half body of the humanoid avatar. When the scene is experienced through a VR headset, the entirety of the human form will be visible. The decision choice of display options will be based on the available equipment. In this section, the Figures are from the screenshots on a 2D screen, in the following Pilot Study section, a VR headset is utilized in the experiment.

During the session, participants will be confronted with the avatar of their hallucination from the display device. Participants could hear two kinds of audio: the therapist's voice for introduction and direct communication; and hallucinatory voice, which will be spoken by

the avatar in the scene. The hallucinatory voice will be synchronized with lip movements, enhancing the realism of the interaction.

3.2.3 Therapist Interface

A TCP socket establishes a communication link between the therapist and patient Unity projects. Utilizing this socket, the therapist could manipulate the scene experienced by the participants. This includes different avatar scenes for interaction, as well as the scene that signifies the conclusion of the session.



(a) Starting interface **(b)** Therapy-session interface

Figure 3.3: Therapist-facing UI

As shown in Figure 3.3, the therapist initiates the session by clicking the START button, which triggers the patient's Unity project to display the scene with the humanoid avatar. By default, both Unity scenes use the Agora [118] plugin to enable live chat upon switching to the session scene. The Agora plugin provides straightforward tools for implementing a live voice call platform. Upon entering the session scene, both the patient and therapist

projects automatically join the same voice chat channel, with the therapist having an on-screen button to mute their side and communicate using a hallucinatory voice. Initially, the therapist will use the live chat channel to introduce the study to the patient and engage in some light conversation. Before transitioning to the dialogue focused on communicating with the hallucination, the therapist will inform the patient and seek their permission to begin. This procedure is designed based on established practices in previous avatar therapy studies [31, 32, 100, 113], where the therapist and patient sit in separate rooms, making it essential for the therapist to monitor the patient's reactions.

After obtaining the participant's permission, the therapist will deactivate the live chat recording feature on their end by pressing the mute button, which will change the color of the button from green to grey. The therapist will continue to hear and monitor the patient's reactions from the Agora livechat channel. This allows the therapist to intervene or end the session promptly if the patient wishes to pause or if any unforeseen circumstances arise.

To continue the avatar therapy session and role-play the patient's hallucination, the therapist will use the start and stop buttons on the screen to record their voice. Once the recording is complete, the audio clips are transmitted to the backend for processing. In the backend, the recorded audio is processed, as detailed in the following subsection, and a new audio file is generated. This file contains the same content, but the voice is altered to match the patient's hallucination. The patient's interface will then receive the rendered audio, which contains the same content but in the voice of their hallucination. The therapist will

assess the patient's reactions and determine appropriate responses to continue the dialogue. This iterative process continues until either the therapist or the patient decides to end the session.

3.2.4 Backend Design

The backend serves as a central channel in the process, receiving audio inputs from the therapist's Unity project and subsequently delivering the processed audio to the patient's scene, as shown in Figure 3.1. The two Unity projects communicate with the backend over a TCP socket.

Upon receiving the audio data from the therapist side, the backend employs Google's Speech-to-Text API to transcribe the content of the audio. Following this initial step, the synthesizer, vocoder, and two-stage text-to-speech (SV2TTS) [85] framework will be called with the preselected target voice sample and transcribed content as input, rendering the audio content into a voice that emulates the participant's hallucination. The newly transformed audio will then be transmitted back to the Unity project on the patient's side, where it is played back upon arrival. This bidirectional flow of audio will work in a loop as long as the therapist's Unity project continues to provide new audio inputs.

3.3 Animation Control Design

3.3.1 Body Animation

The humanoid avatar after generation will be imported into Unity. Under the Inspector setting, the model could be selected to be rigged as a humanoid avatar and the model's bone structure could be mapped to the humanoid bone structure defined by Unity.

For users experiencing hallucinations, our goal is to present them with an avatar that closely resembles their hallucinations, necessitating a high level of behavioral realism to make the avatars more human-like. Since we lack prior knowledge of how these hallucinations occur in terms of motion, introducing large or mismatched body movements could disrupt the user's perception, leading to a less satisfactory interaction.

To address this and reduce cognitive load, we selected a natural idle animation clip from Mixamo [119], an online platform offering 3D character animations and rigging services, and integrated it into the animator controller. This idle animation depicts the avatar standing still and breathing without significant movement of other body parts. The clip is set to loop with smooth transitions between the end and beginning motions in the Unity Inspector, ensuring the animation runs seamlessly. When the scene starts, the animator controller drives the avatar to continuously perform the idle motion, maintaining a consistent and realistic presence.

3.3.2 Facial Animation

Non-verbal interaction is important, as introduced in the previous chapter. To have some non-verbal communication with users, the project used the plug-in LipSync v2 to achieve eye blinking and movement. The plugin's algorithmic approximation of natural eye movements enables the avatar to simulate various eye behaviors, including blinking and gaze shifting. This process is automatic, with movement controllable through frequency parameters.

The lip-syncing method of SALSA offers a seamless and automated approach to synchronizing an avatar's mouth movements with speech audio. The built-in functions operates on the avatar's blendshape in real-time, ensuring that the avatar's lip movements are synchronized with the audio playback. This capability enhances the naturalness and engagement of interactions, providing a more immersive user experience.

3.4 Pilot Study

The team conducted a pilot study with part of the introduced pipeline at the Douglas Hospital. There are three steps in the study: creating the visual representation of the participant's hallucination, creating the audio representation of the participant's hallucination, and finally creating the scene for interacting with the hallucination. This thesis will focus on the third part of the pilot study, which utilizes the result of the first two parts since these parts are carried out by other team members.

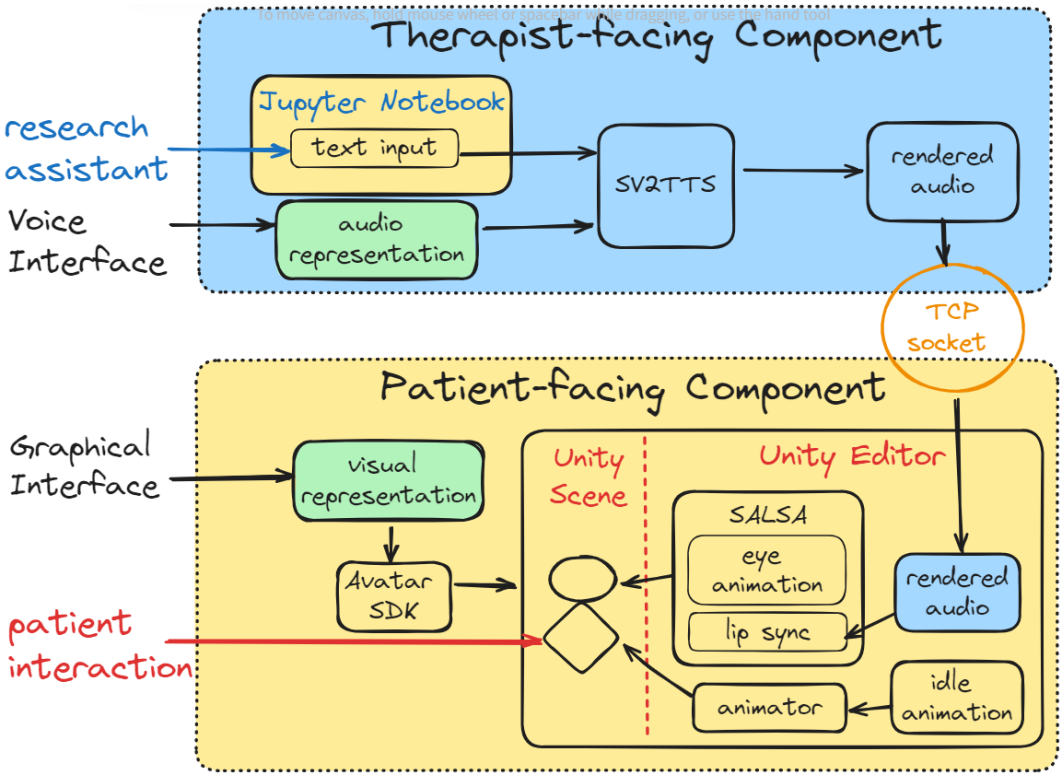


Figure 3.4: System components include a therapist-facing and a patient-facing interface. For our study, the avatar's speech content was entered by a research assistant, but for intended future therapeutic use, this would be controlled (spoken) by a therapist, and transformed into the voice of the avatar through the lower pathway of the "therapist-facing" interface.

3.4.1 System Architecture

Instead of using the therapist side interface and controlling the rendered voice content with the therapist's input audio, the voice content was generated from the therapist's text input. During the experiment, researchers used a Jupyter Notebook connected to the backend lab machine where they could input the target text. With given input of the target voice and

target text, SV2TTS [85] generated the audio in around 4-5 seconds on the lab server, then the generated voice was transferred to the local computer via a TCP connection. Only the patient-side Unity project is used in the study, and an Oculus Quest 2 VR headset was used to display the scene to the patient. The architecture consists of a therapist-facing component (in our pilot study, used by a research assistant) and a patient-facing component, illustrated in Figure 3.4, and described in the later section.

Therapist/Experimenter interaction interface

During the experimental session, the research assistant controlled the content spoken by the avatar using a Jupyter Notebook interface and entering the reply message through a text prompt. The topic of the interaction content was chosen to be about daily routines or hobbies, for example asking about participants's favorite books or movies. For the implementation, the Jupyter notebook and dockerized system components were hosted on a remote laboratory computer with an AMD Ryzen 7 3800X CPU (8 cores), 32 GB DDR4 RAM, and Nvidia Titan Xp 12GB graphics cards.

With the inserted text message from the text prompt, the notebook then calls a synthesizer, vocoder, and Speaker Verification to SV2TTS [85] process. The synthesizer uses the voice embedding (latent parameters) of the target voice, along with the input text to generate a mel-spectrogram. The vocoder then converts the mel-spectrogram into the final audio output, producing speech with the content of the input text. Although newer

speech synthesis techniques are available, SV2TTS offered satisfactory performance in terms of voice generation speed and quality, making it suitable for our project timelines and requirements.

The output audio waveform of the converted text-to-speech is then transferred via TCP to the Unity client, running on a laptop, where it is played back synchronously with the avatar animation for the patient to observe.

Patient Interaction Interface

Although the full-body avatar, representing the patient's hallucination, may be experienced either on a conventional desktop display or within a VR environment, we opted to use the latter for our user study in order to maximize the possibility of immersion in the experience. For the this purpose, we used an Oculus Quest 2 VR headset, in which the avatar is initially positioned a few steps away from the patient.

Same as the introduction in Section 3.3, the avatar is driven by the animator component, linked to an idle animation clip to exhibit lifelike movements, and with lip movement approximately synchronized to the synthesized speech playback. An example of the generated full-body avatar from user selection in VR scene is shown in Figure 3.5.

Communication between the patient and the therapist (playing the part of the hallucination) was managed via TCP. The biggest contributor to latency was the speech synthesis process, resulting in a total end-to-end delay typically in the range of 5-8 s, which

Figure 3.5: 2D facial appearance and resulting 3D avatars generated by patients P147 and P145 to represent their hallucinations.

felt sluggish as far as typical conversational interaction is concerned.

3.4.2 VR Experimental Session

To conduct a preliminary assessment of our system, we recruited seven patients with schzophrenia from the Douglas Mental Health University Institute. The study was approved by the Comité d'éthique de la recherche du CIUSSS de l'Ouest-de-l'Île-de-Montréal, IUSMD-21-18.

Participants had to be at least 18 years of age, and have a primary diagnosis of a schizophrenia spectrum disorder or affective disorder with psychotic symptoms. Exclusion criteria were an inability to provide written consent, currently undergoing another form of

psychological therapy for voices, paranoia with regards to monitoring that would preclude wearing of finger-sensor, acute mental or physical health crisis, diagnosis of organic brain disease, neurological disorder, or substance use disorder, or visual or hearing impairment.

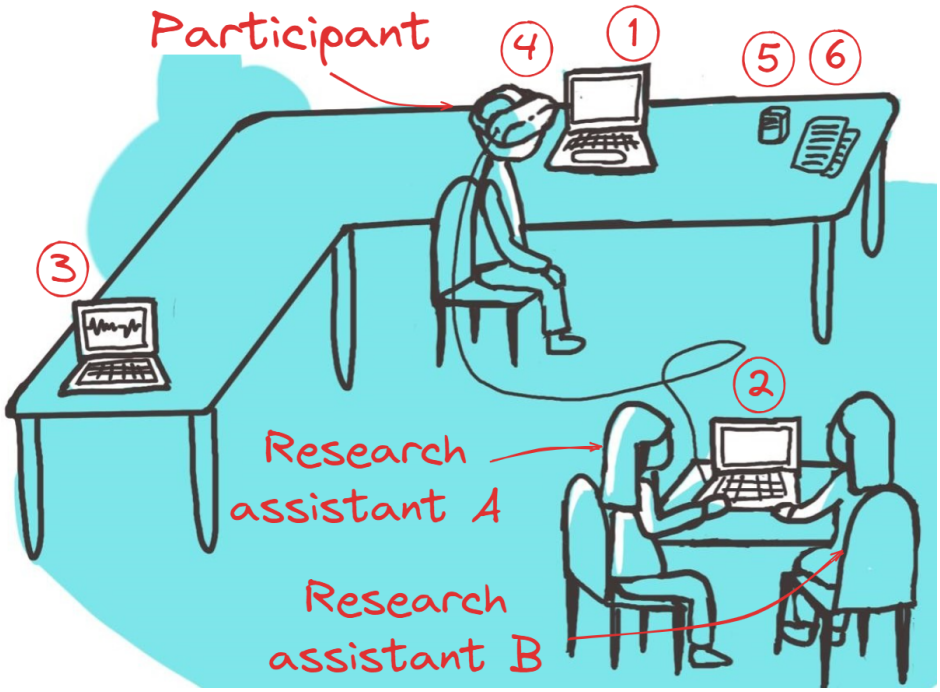


Figure 3.6: Experiment setup for the pilot study with equipment: 1. Laptop displaying graphical and audio representations; 2. Laptop running Unity VR scene and displaying Jupyter Notebook interface; 3. Laptop recording biosignals; 4. Oculus HMD presenting the VR scene; 5. Voice recorder; 6. Questionnaire.

The experimental setup of our avatar therapy system, illustrated in Figure 3.6, provides an interactive environment for participants to create, and subsequently interact with, avatars of their hallucinations.

During the study, the avatar's speech content was transformed into a voice matching

the vocal characteristics as selected by the patient through the previous hallucination externalization process and appears to be spoken by a 3D avatar, generated based on the previous visual representation of their hallucination.

Interaction in the VR Environment

When interacting with the avatar in the VR environment, participants were encouraged to adjust its position to a comfortable location with the assistance of a research team member. They then engaged in a brief interaction, during which the avatar prompted the patient with casual questions such as “What’s your favorite movie?”, “Do you like today’s weather”.

3.4.3 User Evaluation of the Avatar

Participants evaluated the voice and face creation interfaces through a questionnaire using the questionnaire of Table 3.1, and their interaction with the avatar in the VR environment using the questionnaire of Table 3.2. Both questionnaires used a 5-point Likert scale.

Participants rated the avatar in the 3D environment with neutral to moderately positive scores with regard to similarity of appearance and voice. Contrary to our expectations, the VR did not enhance their perceived realness, with feedback being neutral to negative; in fact, scores for perceived realism in the VR environment were actually lower than the perceived realism of 2D graphical appearance and isolated voice output. In addition, participants reported a lower sense of control as compared to those reported in Q5 where individual face

Question	Face UI		Voice UI	
	$\overline{x}$	σ	$\overline{x}$	σ
1. How did you find the process of creating your avatar's {face|voice} on the computer?	3.3	1.6	3.3	1.2
2. How similar is the avatar's {face|voice} to the one you were trying to make?	3.7	1.0	3.0	0.8
3. Recall your creation of the {face|voice} of your avatar. How real did the avatar feel compared to a previous experience you had with the voice of your hallucination?	3.3	1.0	2.7	1.2
4. How often did you think that your avatar felt real compared to a previous experience you had with the voice?	3.0	0.6	2.7	1.0
5. How much control did you feel you had over your avatar's {appearance|voice}?	3.3	1.5	3.1	1.6

Table 3.1: Questionnaires and results of evaluating patient experience with the avatar appearance and voice creation interfaces.

(mean = 3.3, SD = 1.5) and voice user interfaces (mean = 3.1, SD = 1.6). Interestingly, the average response to Q9 was 2.3 (SD = 1.0), indicating a more significant influence from the graphical aspects of the avatar during interaction. Overall, despite being immersed in the VR environment and interacting with the 3D hallucination avatar created from their earlier inputs, patients did not perceive an increased sense of realism or control over the process.

Responses to questions about perceived control (Q11) show a neutral to high correlation with questions concerning overall similarity (Q8, corr=0.67) and frequency of perceived realism (Q10, corr=0.67). These correlations are higher than those with similarity in face (Q6, corr=0.23) and voice (Q7, corr=0.57) interfaces individually, indicating participants who perceive the overall avatar as similar and realistic are likely to feel they have better control over the entire creation process.

Question	$\overline{x}$	σ
6. Looking at the 3D avatar you created in the virtual reality headset, how similar did it look compared to the face you were trying to make?	3.1	1.3
7. Hearing the 3D avatar you created in the virtual reality headset, how similar did it sound compared to the voice you were trying to make?	3.0	1.3
8. How similar to your normal experience of hearing the voice you were trying to make was your experience with the avatar?	2.4	1.3
9. During your experience with the avatar, did you think of it more as an image you saw or as the voice you were trying to re-create talking to you, on a scale of 1 (image) to 5 (voice)?	2.3	1.0
10. During the time of your experience, how often did you think to yourself that you were actually hearing the voice you were trying to make?	2.4	1.3
11. How much control did you feel over both the face and voice of the avatar you made?	2.6	1.8

Table 3.2: Questionnaire and results of evaluating patient experience during VR interaction with their created avatars.

Despite this, our pilot study offered potential insight into the therapeutic benefit of such interventions, even for those who did not consider their created avatar to be an accurate representation of their hallucination. For one such participant, P150 the inaccuracy of the representation was due to the fact that in her prior experience, the hallucination never interacted with her, and on the contrary, actually ignored her. This was, obviously, unlike the scenario we created with an interactive avatar. Nevertheless, the patient expressed the following: "I've always been a bit curious about you, actually"; "I was trying to interact, but she thought that I couldn't see her... It puts it into the everyday, into the real world type of thing." In this case, she was addressing me which was rather weird... It puts it into the everyday, into the real world type of thing." Another patient, P148, had similar feedback

regarding their experience with the avatar. She found the VR interaction interesting since this was her first experience with the technology. However, she did not find the avatar to have high realism, noting "She didn't know I could see her... I was trying to interact by eye contact... she thought that I couldn't see her... she was addressing me, which was rather weird... "

Patients generally responded positively to all the three sessions of avatar therapy, giving neutral to positive feedback on the VR session. This shows the potential of using the designed platform to enhance the immersion and therapeutic effect. However, the overall neutral rating to the VR session also highlights the potential for improving all three parts of the system to enhance the perception of realism for therapeutic benefit.

Feedback from participants generally focused on the misalignment of behavioral realism. This occurs because the redesigned avatar motion could differ from their existing hallucinations, making them feel uneasy and realize they are not seeing the right person. The mismatch between the visual and audio hallucination creation from the previous stage also exacerbates this issue.

For future work, it is worth considering adding another layer of customization to the hallucination behavior in the VR component. This could include options for having a conversation or not, as well as non-verbal behaviors such as eye interaction. Allowing therapists to choose these options could result in a higher and more frequent perceived realism of the hallucination.

Additionally, improving the lip-syncing method, which is not currently fully viseme-based, and speeding up the TTS process using more advanced tools or APIs could enhance the overall system performance. Furthermore, since not all patients experience humanoid hallucinations, incorporating different styles of avatar generation methods should be considered for future implementation.

Chapter 4

ADiNA Project

Preface

This project is the result of a collaborative effort that has been demoed at AGE-WELL AgeTech Innovation Week Conference in 2023. The project focuses on developing a digital nurse to assist caregivers and elderly individuals in long-term care facilities. It comprises multiple applications running on local machines and servers. This thesis will focus on the visual and speech interaction components of the overall architecture, which depend on the proper workflow of the other parts of the project.

Author's Contribution

On the development side of this project, Mauricio Fontana De Vargas implemented the state machine-based application control, managing the flow of different stages of the whole project, and designed the personal selection interface. Diane Hsueh and Charlene Yin conducted interviews and collected information that guided the nurse personalization model design. Carrie Dai was involved in the early design phase of the study. Romain Bazin implemented the ASR (Automatic Speech Recognition) module and the response emotion detection module. Yujing Liu implemented the TTS (Text-to-Speech) module and the avatar interaction Unity frontend. Prof Jeremy R. Cooperstock and Prof Karyn Moffatt supervised the project.

4.1 Background

The objective of the ADiNA project is to provide the elderly with a virtual nurse conversational avatar, aimed at offering companionship and mitigating the burden on caregivers.

Research related to the digital nurse for elderly people has been an attractive topic for many years, since the global population is aging rapidly. For example, According to the World Health Organization (WHO), the proportion of the world's population over 60 years will nearly double from 12% in 2015 to 22% by 2050 [120], leading to a higher prevalence of

age-related health issues and an increased need for long-term care. [121]

However, traditional healthcare systems are not efficient enough to handle this surge in demand. The shortage of qualified nursing staff is exacerbated by factors such as the aging nursing workforce [122], high turnover rates [123] [124], and insufficient training programs [125]. Also, healthcare costs continue to rise globally, driven by many factors such as increased demand for services, advancements in medical technology [126], and the growing prevalence of chronic diseases [127] [128] [129]. The one-on-one encounters between nurses and patients can be limited by time constraints, especially in busy care centers. Moreover, the current system is unable to ensure round-the-clock care for elderly patients [130]. This issue was exacerbated during the pandemic when maintaining social distance becomes crucial and accessing caregivers becomes more challenging [131].

Nursing avatars can fill this gap by providing continuous monitoring, personalized care, and timely interventions, ensuring that elderly individuals receive the necessary attention and support without overwhelming the healthcare infrastructure [132–134]. These avatars can alleviate some of the pressures caused by the nursing shortage by handling routine tasks, thereby allowing human nurses to focus on more complex and personalized care activities [132, 135]. By engaging with patients around the clock, nursing avatars can solve the issue of limited one-on-one interactions, answering questions, providing information, and offering emotional support. Traditional health education materials, such as pamphlets and posters, are often static and may not effectively engage or inform patients. Nursing avatars, on

the other hand, can deliver dynamic, interactive content tailored to individual needs and preferences, making health information more accessible and engaging [136]. This approach not only enhances the quality of care but also helps reduce burnout and turnover among nursing staff [137].

There has been a lot of research and development in this area. Previous works like Molly [138, 139] developed by Sensely drew much attention among users. Molly provides a realistic humanoid digital nurse platform for patients and can be used for elderly people groups. Inviting elderly people to interact with the avatar, it can perform different tasks including symptom checking and triage, managing medications, providing remote monitoring, delivering personalized health education and counseling and assisting with appointment scheduling and follow-ups. These capabilities could enhance patient engagement, increase accessibility, reduce healthcare costs, and improve health outcomes.

Another example is Hippocratic AI's healthcare agent, focusing on the safety-focused large language model(LLM) [140], they provide a similar digital conversational nursing platform and also offer different ages and races of avatars for users to choose from. Many other digital nurse examples are being developed in academia to provide a better platform with suitable human-computer interaction [139, 141, 142].

Research has demonstrated the positive impact of digital nurse avatars in many aspects. For instance, McCue et al. [143] found that subjects who interacted with a digital nurse avatar reported 100% compliance with stress reduction suggestions and 89% compliance

with healthy eating recommendations given by the avatar. Additionally, 44% of the participants indicated a preference for interacting with an avatar over a clinician. Similarly, Bickmore et al. [144] discovered that an avatar nurse incorporating social content in its scripted conversations led to improved patient experiences. This aligns with broader findings [145] that digital avatars can provide crucial psychosocial support for older adults in hospital settings. Furthermore, a study by Gingele et al. [26] on integrating nurse-look-alike avatars into telemedicine applications received positive feedback from heart failure patients, highlighting the potential of digital nurse avatars to enhance patient engagement and adherence to health recommendations.

However, there have also been concerns that nursing avatars cannot fulfill the socio-emotional aspects of nursing care that are essential to the profession, particularly the deeper sense of care that includes compassion and altruism [132]. Additionally, Sestino et al.'s study [146] shows that higher-level human-like interactions positively influence individuals' intention to use such healthcare services, indicating that enhancing the form and behavior realism of digital nurse avatars could improve patient outcomes and satisfaction.

As the effectiveness of digital nurses continues to be enhanced with advancements in display technology and dialogue systems, our goal is to address these concerns by utilizing advanced tools to design a digital nurse platform with emotional feedback, making interactions more attractive and engaging.

Our platform is designed for both caregivers and older adults. For caregivers, the goal is

to assist them with daily routines by enabling them to input care goals for each elderly person through conversation, eliminating the need for form-filling and reducing their workload. For elderly patients, the digital nurse will initiate conversations based on information provided by their caregivers, inform them of their daily routines, and engage in free dialogue based on their individual preferences. This thesis will focus on the visual and speech interaction part of the whole architecture.

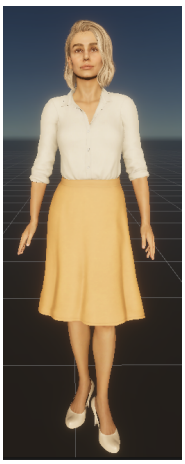
4.2 System Overview

4.2.1 Design Considerations and System Architecture

In the project design, elderly individuals face the Unity interface and interact with the avatar. To create a smooth and engaging interaction, we aim to avoid the need for additional devices for input or control, allowing users to focus solely on conversation. Consequently, the interface is designed to contain only a photo-realistic avatar, as shown in Figure 4.1a. The avatar is created using Character Creator 4 [147], a 3D photorealistic character creation software with advanced tools that enable detailed customization and creation of digital personas, ensuring the form realism of the avatar. The selected avatar is depicted in Figure 4.1b.

As discussed in Section 2.6, it is vital to ensure that the avatar exhibits a high level of behavioral realism, as users tend to have higher expectations for a photorealistic avatar.

(a) Unity interaction interface



(b) Idle animation

Figure 4.1: ADiNA interaction interface and full-body appearance

Failure to meet these expectations may result in a poor user experience.

In this project, to achieve a high level of behavioral realism, we focus on several key aspects of the avatar's performance: realistic body motion, accurate lip-syncing, natural eye movements, and facial expressions that correspond to the speech content. We also prioritize the naturalness of dialogue responses, which includes high-quality audio, fast response times, the use of filler sentences, and emotionally expressive audio. These elements are controlled by various components of the system. Our overarching goal is to enhance the quality of communication modality, response type, and social content, thereby creating a more engaging and lifelike interaction with the avatar.

The system contains three primary components: the backend, situated on a laboratory machine, responsible for speech detection and audio generation-related API calls, a Python project tasked with managing the conversational state, and the Unity project, responsible

for facilitating appropriate interactions. This thesis will focus on the Unity project and the backend components that directly interact with the Unity framework. The related architecture is shown in Figure 4.2.

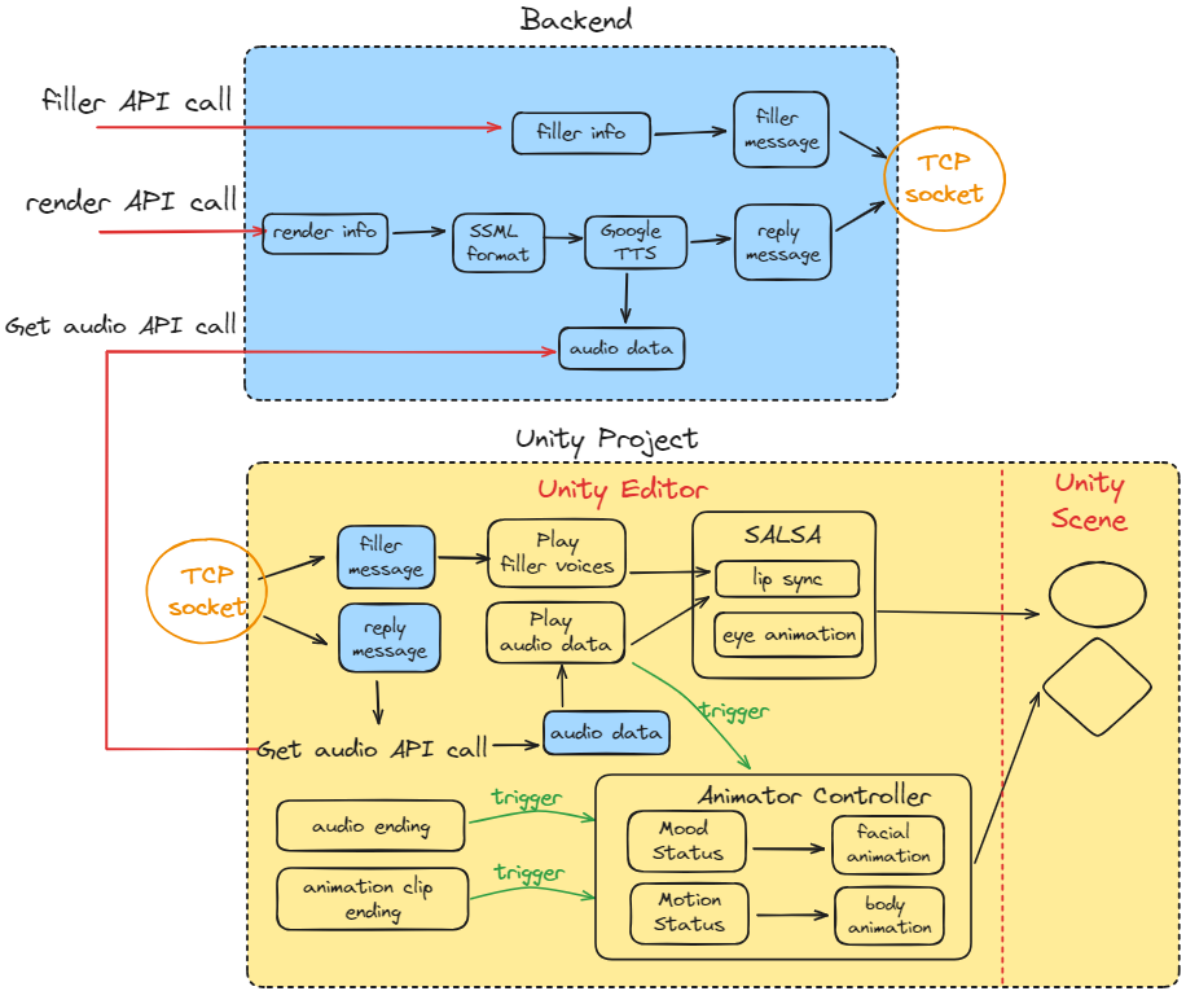


Figure 4.2: System architecture design

4.2.2 Backend Control Design

The portion of the backend architecture discussed in this thesis, as shown in Figure 4.2, efficiently manages the timing and generation of audio responses. The higher-level backend system controls the initiation of its filler and sentence rendering functions by invoking the corresponding API calls. The backend components then manage the generation of audio responses and the latest status information is sent to the Unity project for playing the audio content, whether it is a response sentence or a filler.

The backend components are hosted on a remote laboratory computer equipped with an AMD Ryzen 7 3800X CPU (8 cores) and an Nvidia GeForce RTX 3080 GPU. These components run in a dockerized environment, which includes a Flask application managed by Gunicorn to handle incoming API calls as part of the response process control. Additionally, the system operates a TCP socket to manage data transmission between the backend machine and the Unity user interface of the whole process.

Audio Generation And Control

Upon receiving a filler POST request, the Flask application determines the language type from the request data. Subsequently, a message indicating the 'filler' status, along with the language type, is packaged and prepared for sending to the Unity client. This ensures that the appropriate filler audio is rendered in the correct language.

The audio rendering POST request API call contains the necessary information for audio

generation, including the text to be rendered, language type, and mood information. The corresponding TTS function is invoked to generate the audio response.

For the TTS functionality, the team prioritized speed and opted to use Google TTS to achieve a natural conversational pace. After evaluating the overall speed of the ASR and response generation stages, these preliminary steps already consumed a significant amount of processing time. Deploying an open source TTS module on the remote lab machine would further increase response duration and disrupt the overall flow. Therefore, the team leveraged the fast generation speed of Google TTS, which provides faster synthesis output than we are able to achieve when running on a local machine.

To generate response audio with emotion, the content is first converted to the speech synthesis markup language (SSML) format. SSML is a powerful tool used to control various aspects of speech synthesis, such as pronunciation, volume, pitch, and rate. The ADiNA project aims to convey three types of emotions: neutral, joyful, and worried. Although Google TTS does not natively support emotional rendering, previous research [148, 149] has demonstrated that direct manipulation of pitch can effectively convey emotions to the audience. Utilizing SSML, the synthesized speech conveys the desired emotional tone by adjusting the pitch: increasing it for a joyful tone and decreasing it for a worried tone.

The SSML content is then sent to the Google TTS API for rendering. The voice chosen for synthesis is that of a mature female, aligning with the appearance of ADiNA. By integrating emotional context into the speech synthesis process, the system enhances the naturalness

and engagement of the conversational agent, thereby improving user interactions.

After receiving the audio data, it is saved on the remote lab machine. Subsequently, a text message containing the 'replies' information and the mood information is packaged and prepared for sending to the Unity client. This systematic approach ensures that the audio response is not only fast but also emotionally appropriate, thereby enhancing the overall conversational experience.

TCP Socket Connection

A TCP socket connection is established between the backend and the Unity project to ensure smooth communication and state transition controls, with the remote laboratory machine acting as the server and the Unity project as the client.

The backend is responsible for transmitting packaged data to the Unity side, controlling two primary states: the "reply state" and the "filler state". During the reply state, the backend sends the previously packaged message to Unity, prompting it to perform an API call to retrieve the audio data and play it out directly, ensuring that the digital nurse continues the conversation fluidly. Conversely, in the filler state, the backend sends the previously packed filler message, triggering the Unity project to play a filler sentence, and maintaining the conversational flow when a meaningful response is not immediately available.

This communication protocol ensures synchronized interaction between the backend and the Unity-based digital nurse, facilitating a coherent conversational flow with a real-time

factor (RTF) of approximately 0.2, as measured previously. This setup allows the Unity project to terminate and re-establish connections as needed, supporting the conduction of various experimental sessions at different times.

The backend's role in managing data transmission and controlling the interaction stages is critical for the digital nurse's smooth and responsive performance.

4.2.3 Unity Interface Interaction Design

During the interaction, the elderly or their caregivers will face a 2D monitor screen displaying a Unity scene that shows the upper part of ADiNA's body, as depicted in Figure 4.1a. Users can initiate the conversation, and the avatar on display will respond and interact with them, with functionalities powered by the backend dialogue system architecture.

Conversational Interaction

Depending on whether a 'reply' or 'filler' message is received from the TCP socket and parsed by the control script, the Unity scene will trigger the corresponding audio playback function. In the case of a 'filler' message, the scene will randomly play a pre-generated filler sentence. Six filler sentences, such as "hummm", "I see", and "Je vois", have been created using the same TTS voice, with three in English and three in French. The filler function then triggers the appropriate filler sentence based on the language information contained in the TCP socket message.

For a 'reply' message, Unity will play the rendered response audio. When audio playback is triggered, non-verbal animations, including lip motion and body motion, will also be played accordingly, providing a more immersive and interactive communication experience.

Non-Verbal Interaction

Similar to the previous psychosis project, the Unity engine integrates the SALSA plug-in to facilitate both lip synchronization and eye-blinking animations, providing natural interactions during conversations, and enhancing the expressiveness of characters.

To increase the level of behavioral realism, a two-layered animator controller for facial and body animation has been designed, complemented by custom scripting to orchestrate the animations. The Animator Controller allows users to arrange and maintain a set of animation clips and associate animation transitions. Users can manage the transitions between different clips using the state machine or script control to trigger switches between different animations based on their designed logic.

(a) neutral **(b)** joyful **(c)** worried

Figure 4.3: Different facial expressions of ADiNA

In the facial animation layer, animations correspond to the emotion of the response sentence, with three types: neutral, joyful, and worried, as shown in Figure 4.3. Three distinct animation clips with blend shape transformation information have been crafted. Each clip is tailored to convey the specified emotion, enhancing the avatar's ability to communicate its emotional state in line with the tone of the spoken sentence. When the avatar is idle or speaking with neutral emotion, the expression will randomly switch between neutral and joyful to convey a sense of warmth. For speech with joyful or worried emotions, the corresponding animation clip will be triggered. Each animation clip has a set duration, and before it ends, or if the talking or idle state changes (such as audio starting or stopping), the Animator Controller will smoothly transition to the next clip based on the latest state information.

(a) **(b)**

Figure 4.4: ADiNA talking animation example

In the body animation layer, two types of body movements are implemented in the

Unity project: 'Standing Idle' as shown in Figure 4.1b and 'Talking' as shown in Figure 4.4, each with three different animation clips. These clips are activated in a stochastic manner, governed by the underlying logic within the control script. This random triggering of animations within the same category ensures a more natural and varied representation of the avatar's movements. Similar to facial animation, when the animation clip ends or if the talking or idle state changes, the Animator Controller will smoothly transition to the next animation clip based on the latest state information.

4.3 Demo at AGE-WELL Conference

The research team attended the AGE-WELL conference [150] in 2023 and conducted a demo session with the conference attendees. Most of the attendees have a research background or relevant knowledge in elder care, and some are older adults themselves. While they are not the target users of our system, their feedback is valuable for future design improvements. This demonstration is described in the main body of the thesis since it provides some degree of feedback relevant to the evaluation of the overall system architecture as described in the previous sections.

4.3.1 Demo Procedure

During the demo session, participants were introduced to the research goals of ADiNA and were invited to interact with the digital nurse avatar. They faced a monitor displaying

ADiNA's interface and communicated with the digital nurse through a microphone. The team provided scripts containing marked-up demographic information of elderly people in a care center, including names, disease backgrounds, treatment plans, and hobbies, as handy guidance. Participants were also encouraged to talk freely with ADiNA. They could choose to perform either as caregivers, inputting a patient's information through conversation, or as patients, initiating a conversation with the avatar. In the former case, ADiNA engaged naturally and followed a designed process for collecting elderly people's demographic data through communication. In the latter case, the digital nurse selected topics based on the previously inputted information and informed the participants about their daily routines.

Approximately ten people participated in the demo session at the conference. All participants successfully performed the task, either as a caregiver or as a patient. After the interaction, they are encouraged to give some feedback and suggestions based on their interaction.

4.3.2 Feedback Results and Discussion

Participants provided generally positive feedback, expressing interest in the realistic appearance and varied motions of the avatar. They were particularly impressed by the avatar's ability to respond naturally and select topics that catered to elderly people's interests when looking at the provided script. This indicates that our combination of body motion, facial expression, and speech content performs well together, achieving a

satisfactory level of alignment in behavior realism and form realism.

The team also received valuable advice regarding the details of behavior realism. Consistent with feedback from the previous Psychosis project, some participants recommended incorporating eye contact to make the avatar more interactive and human-like. Others felt that the current lip-sync method was not accurate enough; one participant, an elderly person herself, noted that elderly users with hearing difficulties rely on lip motion as auxiliary information to audio. A misalignment between what they hear and what they see could result in a strange and disconcerting experience, making them more aware that they are talking to a machine. This feedback is crucial, as an imbalance in behavior and form realism can lead to questions about the intelligence of the communication avatar.

Future work on the project should focus on enhancing interaction by adding participant identification to enable eye contact and read participants' facial expressions to provide suitable responses. Better lip-sync solutions, particularly those based on visemes, should be integrated without increasing the overall response time. Currently, this decision involves a trade-off, as testing has shown that other non-Unity plug-in methods have longer processing times when run locally. NVIDIA's Audio2Face [54] method could be a future direction, but its API is still under development at the time of this demo.

Additional suggestions included improving the response time of the ASR module in noisy environments and controlling the length of the generated response to avoid overwhelming

users with information. While this thesis will not delve into these topics in detail, as they are not the main focus, they are important considerations for future enhancements.

Chapter 5

Discussion

This thesis explored the design and implementation of conversational humanoid avatars within the healthcare domain, specifically through the psychosis project and the ADiNA project. The pilot study and demo for these two projects tested their functionality and gathered feedback on both user experience and design limitations. Participants expressed interest in the potential of conversational avatars to address real-world problems and its effectiveness in handling typical tasks. In the psychosis project, users appreciated the ability to externalize their hallucinations and engage in brief communication with the avatar, while in the ADiNA project, participants saw value in using digital nurses to reduce healthcare professionals workload and provide companionship. However, the feedback also pointed out the design limitations in each project. In general, participants felt that the experience could be improved with more interactive features to enhance behavioral realism, as well as more

personalizing options to give users more control in designing their experience. These aspects should be considered in future research.

For the psychosis project, the design architecture of avatar therapy process is being studied. Aiming to help patients to externalize and confront their hallucinations in a controlled virtual environment. The pilot study indicated potential benefits in reducing the severity of hallucinations and improving patients' ability to manage their symptoms. However, the feedback also highlighted the need for high behavioral realism in avatar interactions matching with participants' hallucinations, as discrepancies in visual and auditory cues could disrupt the therapeutic experience.

For the ADiNA Project, a virtual nurse avatar for elder care is developed to address the increasing demand for care services and the limitations of traditional healthcare systems. The demo at the AGE-WELL conference received some positive feedback for its potential to assist care professionals and help reduce their workload. Feedback also emphasized the importance of the avatar's emotional expressiveness and increased detail of interaction. While the avatar successfully engaged users in the designated tasks, suggestions were made to enhance user engagement and realism by improving lip-sync accuracy, and adapting the avatar's interactions to users' emotional states. This feedback also highlights the importance of incorporating more advanced emotion detection and response capabilities to improve the overall user experience. Additionally, for future development, there is also potential to improve the system's scalability with multiple users at the same time.

As already demonstrated by many previous research, the two projects showed that humanoid avatars can effectively engage users by simulating natural human interactions. which is beneficial in healthcare settings where empathy and personalized communication are crucial. Both projects underscored the importance of achieving high behavioral realism in digital avatars, highlighting the need for continuous advancements in animation technologies and real-time processing capabilities. The user experience also highlighted that, for different user groups, the focus of similar designs can vary. For example, in the psychosis project, the patient's typical interaction with their hallucinations influenced their preference for eye contact with the avatar. In the ADiNA project, one participant emphasized the importance of lip-sync accuracy, noting that many elderly individuals with hearing difficulties rely on reading lip movements to understand speech.

The findings and insights from the two projects can provide some inspiration for future innovations that can further enhance the effectiveness of conversational avatars in healthcare settings. For future research, it is vital to adapt to the rapid advancements in animation tools and real-time processing technologies. It is also essential to enhance animation techniques, incorporate more interactive features to improve behavioral realism and ensure system robustness. Additionally, developing adaptive systems that personalize avatar interactions based on individual user preferences and needs will significantly enhance user experience and therapeutic outcomes.

By addressing the identified challenges and leveraging user feedback, future developments

can create more realistic, responsive, and scalable avatar systems that significantly improve patient engagement and healthcare outcomes.

Bibliography

[1] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, et al., "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), pp. 1061–1068, International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[2] M. J. Smith, E. Ginger, K. Wright, M. A. Wright, J. L. Taylor, L. B. Humm, D. Olsen, M. D. Bell, and M. F. Fleming, "Virtual reality job interview training in adults with autism spectrum disorder," Journal of Autism and Developmental Disorders, vol. 44, no. 10, pp. 2450–2463, 2014.

[3] D. S. Harvie, J. Kelly, J. Kluver, M. Deen, E. Spitzer, and M. W. Coppieters, "A randomized controlled pilot study examining immediate effects of embodying a virtual reality superhero in people with chronic low back pain," Disability and Rehabilitation: Assistive Technology, vol. 19, no. 3, pp. 851–858, 2024.

[4] A. C. Griffin, Z. Xing, S. Khairat, Y. Wang, S. Bailey, J. Arguello, and A. E. Chung, "Conversational agents for chronic disease self-management: a systematic review," in AMIA Annual Symposium Proceedings, vol. 2020, p. 504, American Medical Informatics Association, 2020.

[5] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," Commun. ACM, vol. 9, p. 36–45, jan 1966.

[6] K. Ramesh, S. Ravishankaran, A. Joshi, and K. Chandrasekaran, "A survey of design techniques for conversational agents," in International conference on information, communication and computing technology, pp. 336–350, Springer, 2017.

[7] M. Zemčík, "A brief history of chatbots," DEStech Transactions on Computer Science and Engineering, vol. 10, 2019.

[8] M. M. Mariani, N. Hashemi, and J. Wirtz, "Artificial intelligence empowered conversational agents: A systematic literature review and research agenda," Journal of Business Research, vol. 161, p. 113838, 2023.

[9] L. Athota, V. K. Shukla, N. Pandey, and A. Rana, "Chatbot for healthcare system using artificial intelligence," in 2020 8th International conference on reliability, infocom technologies and optimization (trends and future directions)(ICRITO), pp. 619–622, IEEE, 2020.

[10] Nuance Communications, "Hm extends nuance virtual assistant and live chat to google's business messages," 2020.

[11] L. Qiu and I. Benbasat, "Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems," Journal of management information systems, vol. 25, no. 4, pp. 145–182, 2009.

[12] A. K. Goel and L. Polepeddi, "Jill watson: A virtual teaching assistant for online education," in Learning engineering for online education, pp. 120–143, Routledge, 2018.

[13] P. Hsu, M. Aurisicchio, and P. Angeloudis, "Sciencedirect centeris-international conference on enterprise information systems/projman-investigating schedule deviation in construction projects through root cause analysis," in CENTERIS-International Conference on ENTERprise Information Systems/ProjMAN-International Conference on Project MANagement/HCist-International Conference on Health and Social Care Information Systems and Technologies., 00, 2017.

[14] C. Rzepka, B. Berger, and T. Hess, "Voice assistant vs. chatbot–examining the fit between conversational agents' interaction modalities and information search tasks," Information Systems Frontiers, vol. 24, no. 3, pp. 839–856, 2022.

[15] P. B. Brandtzaeg and A. Følstad, "Why people use chatbots," in Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings 4, pp. 377–392, Springer, 2017.

[16] M. Adam, M. Wessel, and A. Benlian, "Ai-based chatbots in customer service and their effects on user compliance," Electronic Markets, vol. 31, no. 2, pp. 427–445, 2021.

[17] C. Bălan, "Chatbots and voice assistants: digital transformers of the company–customer interface—a systematic review of the business research literature," Journal of Theoretical and Applied Electronic Commerce Research, vol. 18, no. 2, pp. 995–1019, 2023.

[18] L. H. Lind, M. F. Schober, F. G. Conrad, and H. Reichert, "Why do survey respondents disclose more when computers ask the questions?," Public opinion quarterly, vol. 77, no. 4, pp. 888–935, 2013.

[19] M. D. Pickard and C. A. Roster, "Using computer automated systems to conduct personal interviews: Does the mere presence of a human face inhibit disclosure?," Computers in Human Behavior, vol. 105, p. 106197, 2020.

[20] A. Stock, S. Schlögl, and A. Groth, "Tell me, what are you most afraid of? exploring the effects of agent representation on information disclosure in human-chatbot interaction," in International Conference on Human-Computer Interaction, pp. 179–191, Springer, 2023.

[21] J. N. Bailenson, N. Yee, D. Merget, and R. Schroeder, "The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure,

emotion recognition, and copresence in dyadic interaction," Presence: Teleoperators and Virtual Environments, vol. 15, no. 4, pp. 359–372, 2006.

[22] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," ACM Transactions on Computer-Human Interaction (TOCHI), vol. 12, no. 2, pp. 293–327, 2005.

[23] T. Araujo, "Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions," Computers in human behavior, vol. 85, pp. 183–189, 2018.

[24] T. Verhagen, J. Van Nes, F. Feldberg, and W. Van Dolen, "Virtual customer service agents: Using social presence and personalization to shape online service encounters," Journal of Computer-Mediated Communication, vol. 19, no. 3, pp. 529–545, 2014.

[25] J. J. Prochaska, E. A. Vogel, A. Chieng, M. Kendra, M. Baiocchi, S. Pajarito, and A. Robinson, "A therapeutic relational agent for reducing problematic substance use (woebot): development and usability study," Journal of medical Internet research, vol. 23, no. 3, p. e24850, 2021.

[26] A. J. Gingele, H. Amin, A. Vaassen, I. Schnur, C. Pearl, H.-P. Brunner-La Rocca, and J. Boyne, "Integrating avatar technology into a telemedicine application in heart failure patients: A pilot study," Wiener klinische Wochenschrift, vol. 135, no. 23, pp. 680–684, 2023.

[27] M. Butz, D. Hepperle, and M. Wölfel, "Influence of visual appearance of agents on presence, attractiveness, and agency in virtual reality," in ArtsIT, Interactivity and Game Creation (M. Wölfel, J. Bernhardt, and S. Thiel, eds.), 2022.

[28] C. Krogmeier and C. Mousas, "Eye fixations and electrodermal activity during low-budget virtual reality embodiment," Comput. Animat. Virtual Worlds, vol. 31, no. 4-5, 2020.

[29] R. Kondo, M. Sugimoto, K. Minamizawa, T. Hoshi, M. Inami, and M. Kitazaki, "Illusory body ownership of an invisible body interpolated between virtual hands and feet via visual-motor synchronicity," Scientific Reports, vol. 8, no. 1, pp. 1–8, 2018.

[30] F. Weidner, G. Boettcher, S. A. Arboleda, C. Diao, L. Sinani, C. Kunert, C. Gerhardt, W. Broll, and A. Raake, "A systematic review on the visualization of avatars and agents in ar & vr displayed using head-mounted displays," IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 5, pp. 2596–2606, 2023.

[31] J. Leff, G. Williams, M. A. Huckvale, M. Arbuthnot, and A. P. Leff, "Computer-assisted therapy for medication-resistant auditory hallucinations: proof-of-concept study," The British Journal of Psychiatry, vol. 202, no. 6, pp. 428–433, 2013.

[32] L. Dellazizzo, S. Potvin, K. Phraxayavong, P. Lalonde, and A. Dumais, "Avatar therapy for persistent auditory verbal hallucinations in an ultra-resistant schizophrenia patient: a case report," Frontiers in Psychiatry, vol. 9, p. 131, 2018.

[33] L. Ciechanowski, A. Przegalinska, M. Magnuski, and P. Gloor, "In the shades of the uncanny valley: An experimental study of human–chatbot interaction," Future Generation Computer Systems, vol. 92, pp. 539–548, 2019.

[34] A. K. Pandey, R. Gelin, and A. Robot, "Pepper: The first machine of its kind," IEEE Robotics & Automation Magazine, vol. 25, no. 3, pp. 40–48, 2018.

[35] M. Sato, Y. Yasuhara, K. Osaka, H. Ito, M. J. S. Dino, I. L. Ong, Y. Zhao, and T. Tanioka, "Rehabilitation care with pepper humanoid robot: A qualitative case study of older patients with schizophrenia and/or dementia in japan," Enfermeria clinica, vol. 30, pp. 32–36, 2020.

[36] K. Blindheim, M. Solberg, I. A. Hameed, and R. E. Alnes, "Promoting activity in long-term care facilities with the social robot pepper: a pilot study," Informatics for Health and Social Care, vol. 48, no. 2, pp. 181–195, 2023.

[37] L. Hung, C. Liu, E. Woldum, A. Au-Yeung, A. Berndt, C. Wallsworth, N. Horne, M. Gregorio, J. Mann, and H. Chaudhury, "The benefits of and barriers to using a social robot paro in care settings: a scoping review," BMC geriatrics, vol. 19, pp. 1–10, 2019.

[38] G. W. Lane, D. Noronha, A. Rivera, K. Craig, C. Yee, B. Mills, and E. Villanueva, "Effectiveness of a social robot,"paro," in a va long-term care setting.," Psychological services, vol. 13, no. 3, p. 292, 2016.

[39] R. Bemelmans, G. J. Gelderblom, P. Jonker, and L. de Witte, "Effectiveness of robot paro in intramural psychogeriatric care: a multicenter quasi-experimental study," Journal of the American Medical Directors Association, vol. 16, no. 11, pp. 946–950, 2015.

[40] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2414–2423, 2016.

[41] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2287–2296, 2021.

[42] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Advances in neural information processing systems, vol. 33, pp. 6840–6851, 2020.

[43] B. Cheung, "Stable diffusion training for embeddings." https://bennycheung.github.io/stable-diffusion-training-for-embeddings, 2023. Accessed: 2024-07-13.

[44] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," Communications of the ACM, vol. 65, no. 1, pp. 99–106, 2021.

[45] H. T. W. S. B. Gao, J. Lee and H. Kim, "The effects of avatar visibility on behavioral response with or without mirror-visual feedback in virtual environments," in VR Workshops, IEEE, 2020.

[46] T. Shen, J. Zuo, F. Shi, J. Zhang, L. Jiang, M. Chen, Z. Zhang, W. Zhang, X. He, and T. Mei, "Vida-man: visual dialog with digital humans," in Proceedings of the 29th ACM International Conference on Multimedia, pp. 2789–2791, 2021.

[47] M. Allouch, A. Azaria, and R. Azoulay, "Conversational agents: Goals, technologies, vision and challenges," Sensors, vol. 21, no. 24, p. 8448, 2021.

[48] E. Merdivan, D. Singh, S. Hanke, and A. Holzinger, "Dialogue systems for intelligent human computer interactions," Electronic Notes in Theoretical Computer Science, vol. 343, pp. 57–71, 2019.

[49] S. M. Huq, R. Maskeliūnas, and R. Damaševičius, "Dialogue agents for artificial intelligence-based conversational systems for cognitively disabled: A systematic review," Disability and Rehabilitation: Assistive Technology, vol. 19, no. 3, pp. 1059–1078, 2024.

[50] R. Zhen, W. Song, Q. He, J. Cao, L. Shi, and J. Luo, "Human-computer interaction system: A survey of talking-head generation," Electronics, vol. 12, no. 1, p. 218, 2023.

[51] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in Proceedings of the 28th ACM international conference on multimedia, pp. 484–492, 2020.

[52] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 5784–5794, 2021.

[53] J. Ling, X. Tan, L. Chen, R. Li, Y. Zhang, S. Zhao, and L. Song, "Stableface: Analyzing and improving motion stability for talking face generation," IEEE Journal of Selected Topics in Signal Processing, 2023.

[54] NVIDIA Corporation, "Nvidia audio2face." https://www.nvidia.com/en-us/ai-data-science/audio2face/, 2023. Accessed: 2024-07-06.

[55] I. Habibie, W. Xu, D. Mehta, L. Liu, H.-P. Seidel, G. Pons-Moll, M. Elgharib, and C. Theobalt, "Learning speech-driven 3d conversational gestures from video," in Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, pp. 101–108, 2021.

[56] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, and M. Neff, "A comprehensive review of data-driven co-speech gesture generation," in Computer Graphics Forum, vol. 42, pp. 569–596, Wiley Online Library, 2023.

[57] A. Tinwell, M. Grimshaw, D. A. Nabi, and A. Williams, "Facial expression of emotion and perception of the uncanny valley in virtual characters," Computers in Human behavior, vol. 27, no. 2, pp. 741–749, 2011.

[58] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," IEEE transactions on audio, speech, and language processing, vol. 15, no. 3, pp. 1075–1086, 2007.

[59] S. Mariooryad and C. Busso, "Generating human-like behaviors using joint, speech-driven models for conversational agents," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 8, pp. 2329–2340, 2012.

[60] T. Randhavane, A. Bera, K. Kapsaskis, K. Gray, and D. Manocha, "Fva: Modeling perceived friendliness of virtual agents using movement characteristics," IEEE transactions on visualization and computer graphics, vol. 25, no. 11, pp. 3135–3145, 2019.

[61] T.-H. D. Nguyen, E. Carstensdottir, N. Ngo, M. S. El-Nasr, M. Gray, D. Isaacowitz, and D. Desteno, "Modeling warmth and competence in virtual characters," in Intelligent Virtual Agents: 15th International Conference, IVA 2015, Delft, The Netherlands, August 26-28, 2015, Proceedings 15, pp. 167–180, Springer, 2015.

[62] T. Randhavane, A. Bera, K. Kapsaskis, R. Sheth, K. Gray, and D. Manocha, "Eva: Generating emotional behavior of virtual agents using expressive features of gait and gaze," in ACM symposium on applied perception 2019, pp. 1–10, 2019.

[63] S. Hyniewska, R. Niewiadomski, M. Mancini, C. Pelachaud, and T. P. LTCI, "Expression of affects in embodied conversational agents," Blueprint for Affective Computing, pp. 213–221, 2010.

[64] S. Gobron, J. Ahn, D. Thalmann, M. Skowron, and A. Kappas, "Impact study of nonverbal facial cues on spontaneous chatting with virtual humans," JVRB-Journal of Virtual Reality and Broadcasting, vol. 10, no. 6, 2013.

[65] Y. Ma, S. Wang, Z. Hu, C. Fan, T. Lv, Y. Ding, Z. Deng, and X. Yu, "Styletalk: One-shot talking head generation with controllable speaking styles," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1896–1904, 2023.

[66] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8652–8661, 2023.

[67] Y. Pan, S. Tan, S. Cheng, Q. Lin, Z. Zeng, and K. Mitchell, "Expressive talking avatars," IEEE Transactions on Visualization and Computer Graphics, 2024.

[68] L. Tian, Q. Wang, B. Zhang, and L. Bo, "Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions," arXiv preprint arXiv:2402.17485, 2024.

[69] Synthesia Ltd, "Synthesia: Ai video generator platform." https://www.synthesia.io, 2024. https://www.synthesia.io.

[70] HeyGen AI, "Heygen: Ai video generator platform." https://www.heygen.com, 2024. https://www.heygen.com.

[71] Crazy Minnow Studio, SALSA LipSync Suite, 2023. Accessed: 2024-07-09.

[72] Oculus Developer, Oculus OVRLipSync for Unity, 2023. Accessed: 2024-07-09.

[73] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," Psychological science, vol. 15, no. 2, pp. 133–137, 2004.

[74] L. Hu, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8153–8163, 2024.

[75] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman, "Dreampose: Fashion image-to-video synthesis via stable diffusion," in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 22623–22633, IEEE, 2023.

[76] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "Animating arbitrary objects via deep motion transfer," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2377–2386, 2019.

[77] NVIDIA, "Audio2gesture," 2024. Accessed: 2024-07-07.

[78] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," arXiv preprint arXiv:1703.10135, 2017.

[79] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, et al., "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, vol. 12, 2016.

[80] N. Kireev and E. Ilyushin, "Review of existing text-to-speech algorithms," International Journal of Open Information Technologies, vol. 8, no. 7, pp. 84–90, 2020.

[81] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in International Conference on Machine Learning, pp. 7748–7759, PMLR, 2021.

[82] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," Advances in neural information processing systems, vol. 33, pp. 17022–17033, 2020.

[83] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in International Conference on Machine Learning, pp. 5530–5540, PMLR, 2021.

[84] F. Lux, J. Koch, and N. T. Vu, "Exact prosody cloning in zero-shot multispeaker text-to-speech," in 2022 IEEE Spoken Language Technology Workshop (SLT), pp. 962–969, IEEE, 2023.

[85] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," 2019.

[86] B. Chen, C. Du, and K. Yu, "Neural fusion for voice cloning," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1993–2001, 2022.

[87] X. Zhou, H. Che, X. Wang, and L. Xie, "A novel cross-lingual voice cloning approach with a few text-free samples," arXiv preprint arXiv:1910.13276, 2019.

[88] H. Goble and C. Edwards, "A robot that communicates with vocal fillers has ... uhhh ... greater social presence," Communication Research Reports, vol. 35, no. 3, pp. 256–260, 2018.

[89] N. Lubold, E. Walker, and H. Pon-Barry, "Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion," in 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 255–262, 2016.

[90] A. S. Ghazali, J. Ham, P. Markopoulos, and E. Barakova, "Investigating the effect of social cues on social agency judgement," in 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 586–587, 2019.

[91] S. Yilmazyildiz, W. Verhelst, and H. Sahli, "Gibberish speech as a tool for the study of affective expressiveness for robotic agents," Multimedia Tools Applications, vol. 74, pp. 9959–9982, Nov. 2015.

[92] Finextra, "Nordnet fires ai assistant amelia." https://www.finextra.com/newsarticle/32371/nordnet-fires-ai-assistant-amelia, 2018. Accessed: 2024-07-01.

[93] Chatbots.org, "Anna virtual assistant." https://chatbots.org/virtual_assistant/anna3/. Accessed: 2024-07-01.

[94] D. M. Scott, "Anna from ikea." https://www.davidmeermanscott.com/blog/2008/08/anna-from-ikea.html, 2008. Accessed: 2024-07-01.

[95] F. Miao, I. V. Kozlenkova, H. Wang, T. Xie, and R. W. Palmatier, "An emerging theory of avatar marketing," Journal of Marketing, vol. 86, no. 1, pp. 67–90, 2022.

[96] M. N. Today, "Psychosis vs. schizophrenia: What's the difference?," 2024. Accessed: 2024-07-03.

[97] S. M. Essock, W. A. Hargreaves, N. H. Covell, and J. Goethe, "Clozapine's effectiveness for patients in state hospitals: results from a randomized trial.," Psychopharmacology Bulletin, vol. 32, no. 4, pp. 683–697, 1996.

[98] J. A. Lieberman, "Pathophysiologic mechanisms in the pathogenesis and clinical course of schizophrenia," Journal of Clinical Psychiatry, vol. 60, no. 12, pp. 9–12, 1999.

[99] D. M. Taylor, T. R. Barnes, and A. H. Young, The Maudsley prescribing guidelines in psychiatry. John Wiley & Sons, 2021.

[100] L. Dellazizzo, S. Potvin, K. Phraxayavong, and A. Dumais, "One-year randomized trial comparing virtual reality-assisted therapy to cognitive–behavioral therapy for patients with treatment-resistant schizophrenia," npj Schizophrenia, vol. 7, no. 1, p. 9, 2021.

[101] D. Siskind, L. McCartney, R. Goldschlager, and S. Kisely, "Clozapine v. first-and second-generation antipsychotics in treatment-refractory schizophrenia: systematic review and meta-analysis," The British Journal of Psychiatry, vol. 209, no. 5, pp. 385–392, 2016.

[102] J. Kreyenbuhl, R. W. Buchanan, F. B. Dickerson, and L. B. Dixon, "The schizophrenia patient outcomes research team (port): updated treatment recommendations 2009," Schizophrenia bulletin, vol. 36, no. 1, pp. 94–103, 2010.

[103] E. Kuipers, A. Yesufu-Udechuku, C. Taylor, and T. Kendall, "Management of psychosis and schizophrenia in adults: summary of updated nice guidance," bmj, vol. 348, 2014.

[104] O. P. du Sert, S. Potvin, O. Lipp, L. Dellazizzo, M. Laurelli, R. Breton, P. Lalonde, K. Phraxayavong, K. O'Connor, J.-F. Pelletier, T. Boukhalfi, P. Renaud, and A. Dumais, "Virtual reality therapy for refractory auditory verbal hallucinations in schizophrenia: A pilot clinical trial," Schizophrenia research, vol. 197, p. 176—181, July 2018.

[105] P. McGorry, "Royal australian and new zealand college of psychiatrists clinical practice guidelines for the treatment of schizophrenia and related disorders.," Australian & New Zealand Journal of Psychiatry, vol. 39, 2005.

[106] M. Van der Gaag, L. R. Valmaggia, and F. Smit, "The effects of individually tailored formulation-based cognitive behavioural therapy in auditory hallucinations and delusions: a meta-analysis," Schizophrenia research, vol. 156, no. 1, pp. 30–37, 2014.

[107] S. Jauhar, P. McKenna, J. Radua, E. Fung, R. Salvador, and K. Laws, "Cognitive–behavioural therapy for the symptoms of schizophrenia: systematic review and meta-

analysis with examination of potential bias," The British Journal of Psychiatry, vol. 204, no. 1, pp. 20–29, 2014.

[108] S. V. Kapadia, "Adapting avatar therapy: Using available digital technology for people living with auditory verbal hallucinations in low-and middle-income countries," Indian Journal of Psychological Medicine, vol. 44, no. 4, pp. 405–408, 2022.

[109] N. Thomas, S. Rossell, J. Farhall, F. Shawyer, and D. Castle, "Cognitive behavioural therapy for auditory hallucinations: effectiveness and predictors of outcome in a specialist clinic," Behavioural and Cognitive Psychotherapy, vol. 39, no. 2, pp. 129–138, 2011.

[110] D. Freeman, "Studying and treating schizophrenia using virtual reality: a new paradigm," Schizophrenia bulletin, vol. 34, no. 4, pp. 605–610, 2008.

[111] J. Leff, G. Williams, M. Huckvale, M. Arbuthnot, and A. P. Leff, "Avatar therapy for persecutory auditory hallucinations: What is it and how does it work?," Psychosis, vol. 6, no. 2, pp. 166–176, 2014.

[112] T. Ward, T. Craig, and M. Rus-Calafell, "Avatar therapy for refractory auditory hallucinations," Brief interventions for psychosis, 2016.

[113] T. K. J. Craig, M. Rus-Calafell, T. Ward, et al., "Avatar therapy for auditory verbal hallucinations in people with psychosis: a single-blind, randomised controlled trial," The Lancet Psychiatry, vol. 5, no. 1, pp. 31–40, 2018.

[114] G. Aali, T. Kariotis, and F. Shokraneh, "Avatar therapy for people with schizophrenia or related disorders," Cochrane Database of Systematic Reviews, vol. 2020, no. 5, p. CD011898, 2020.

[115] E. Bisso, M. S. Signorelli, M. Milazzo, M. Maglia, R. Polosa, E. Aguglia, and P. Caponnetto, "Immersive virtual reality applications in schizophrenia spectrum therapy: A systematic review," International Journal of Environmental Research and Public Health, vol. 17, no. 17, 2020.

[116] M. Rus-Calafell, P. Garety, E. Sason, T. J. K. Craig, and L. R. Valmaggia, "Virtual reality in the assessment and treatment of psychosis: a systematic review of its utility, acceptability and effectiveness," Psychological Medicine, vol. 48, no. 3, p. 362–391, 2018.

[117] Avatar SDK, "Avatar sdk: 3d avatar creation tool." https://avatarsdk.com/, 2024. Accessed: 2024-08-12.

[118] Agora.io, "Agora unity plugin." https://www.agora.io/en/developer/unity-sdk/, 2024. Accessed: 2024-08-12.

[119] "Mixamo: 3d character animation." https://www.mixamo.com. Accessed: 2024-08-13.

[120] World Health Organization, "Ageing and health." https://www.who.int/news-room/fact-sheets/detail/ageing-and-health, 2022. Accessed: 2024-06-25.

[121] United Nations Department of Economic and Social Affairs, "World social report 2023: Leaving no one behind in an ageing world." https://www.un.org/development/desa/dspd/world-social-report/2023-2/, 2023. Accessed: 2024-06-25.

[122] A. S. of Registered Nurses, "How the aging population is affecting the nursing shortage," Journal of Advanced Practice Nursing, 2023. Accessed: 2024-06-25.

[123] B. Nursing, "Sustaining the nursing workforce - exploring enabling and motivating factors for the retention of returning nurses: a qualitative descriptive design," BMC Nursing, 2022. Accessed: 2024-06-25.

[124] UNAC/UHCP, "The dangerous impact of the national nursing shortage," 2022. Accessed: 2024-06-25.

[125] American Association of Colleges of Nursing, "Nursing shortage fact sheet," 2023. Accessed: 2024-06-25.

[126] Kaiser Family Foundation, "Snapshots: How changes in medical technology affect health care costs," 2023. Accessed: 2024-06-25.

[127] L. Zhou, S. Ampon-Wireko, H. Asante Antwi, X. Xu, M. Salman, M. O. Antwi, and T. M. N. Afua, "An empirical study on the determinants of health care expenses in emerging economies," BMC Health Services Research, vol. 20, pp. 1–16, 2020.

[128] U.S. Bureau of Labor Statistics, "What is driving increases in healthcare spending? observations from bls disease-based price indexes," 2023.

[129] P. Maresova, E. Javanmardi, S. Barakovic, J. Barakovic Husic, S. Tomsone, O. Krejcar, and K. Kuca, "Consequences of chronic diseases and other limitations associated with old age–a scoping review," BMC public health, vol. 19, pp. 1–17, 2019.

[130] E. Busca, A. Savatteri, T. L. Calafato, B. Mazzoleni, M. Barisone, and A. Dal Molin, "Barriers and facilitators to the implementation of nurse's role in primary care settings: an integrative review," BMC nursing, vol. 20, pp. 1–12, 2021.

[131] S. H. Sharkiya, "Quality communication can improve patient-centred health outcomes among older patients: a rapid review," BMC Health Services Research, vol. 23, no. 1, p. 886, 2023.

[132] M. B. Abbott and M. Peggy Shaw, "Virtual nursing avatars: Nurse roles and evolving concepts of care," Online Journal of Issues in Nursing, vol. 21, no. 3, p. 1C, 2016.

[133] S. Laacke, R. Mueller, G. Schomerus, and S. Salloch, "Artificial intelligence, social media and depression. a new concept of health-related digital autonomy," The American Journal of Bioethics, vol. 21, no. 7, pp. 4–20, 2021.

[134] B. Cloyd and J. Thompson, "Virtual care nursing:: the wave of the future," Nurse Leader, vol. 18, no. 2, pp. 147–150, 2020.

[135] T. R. Roche, S. Said, J. Braun, E. J. Maas, C. Machado, B. Grande, M. Kolbe, D. R. Spahn, C. B. Nöthiger, and D. W. Tscholl, "Avatar-based patient monitoring in critical anaesthesia events: A randomised high-fidelity simulation study," British journal of anaesthesia, vol. 126, no. 5, pp. 1046–1054, 2021.

[136] M. Miller and R. Jensen, "Avatars in nursing: An integrative review," Nurse Educator, vol. 39, no. 1, pp. 38–41, 2014.

[137] N. Bott, S. Wexler, L. Drury, C. Pollak, V. Wang, K. Scher, and S. Narducci, "A protocol-driven, bedside digital conversational agent to support nurse teams and mitigate risks of hospitalization in older adults: case control pre-post study," Journal of medical Internet research, vol. 21, no. 10, p. e13440, 2019.

[138] Sensely, Inc., "Sensely," 2024.

[139] A. Khadija, F. F. Zahra, and A. Naceur, "Ai-powered health chatbots: toward a general architecture," Procedia Computer Science, vol. 191, pp. 355–360, 2021.

[140] S. Mukherjee, P. Gamble, M. S. Ausin, N. Kant, K. Aggarwal, N. Manjunath, D. Datta, Z. Liu, J. Ding, S. Busacca, et al., "Polaris: A safety-focused llm constellation architecture for healthcare," arXiv preprint arXiv:2403.13313, 2024.

[141] American Nurses Association, "Virtual nursing avatars," 2016.

[142] T. Krick, K. Huter, K. Seibert, D. Domhoff, and K. Wolf-Ostermann, "Measuring the effectiveness of digital nursing technologies: development of a comprehensive digital nursing technology outcome framework based on a scoping review," BMC health services research, vol. 20, pp. 1–17, 2020.

[143] K. Mccue, A. Shamekhi, D. Crooks, T. Bickmore, K. Gergen-Barnett, G. Johnson, et al., "A feasibility study to introduce an embodied conversational agent (eca) on a tablet computer into a group medical visit," American Public Health Association (APHA), Chicago, IL. APHA, 2015.

[144] T. W. Bickmore, L. M. Pfeifer, and B. W. Jack, "Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents," in Proceedings of the SIGCHI conference on human factors in computing systems, pp. 1265–1274, 2009.

[145] V. Wang, S. Wexler, L. Drury, B. Wang, et al., "A protocol-driven, digital conversational agent at the hospital bedside to support nurse teams and to mitigate delirium and falls risk," Iproceedings, vol. 4, no. 2, p. e11883, 2018.

[146] A. Sestino and A. D'Angelo, "My doctor is an avatar! the effect of anthropomorphism and emotional receptivity on individuals' intention to use digital-based healthcare services," Technological Forecasting and Social Change, vol. 191, p. 122505, 2023.

[147] Reallusion, "Character creator," 2024. Accessed: 2024-07-07.

[148] R. G. Kamiloğlu, A. H. Fischer, and D. A. Sauter, "Good vibrations: A review of vocal expressions of positive emotions," Psychonomic bulletin & review, vol. 27, pp. 237–265, 2020.

[149] C. Quam and D. Swingley, "Development in children's interpretation of pitch cues to emotions," Child development, vol. 83, no. 1, pp. 236–250, 2012.

[150] AGE-WELL NCE, "Age-well annual conference 2024." https://agewell-nce.ca/conference, 2024. Accessed: 2024-07-04.