TU WIEN Informatics

Embodied Conversational Agents with Situation Awareness

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Visual Computing

eingereicht von

Ing. Martin Rumpelnik, BSc
Matrikelnummer 1633397

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Peter Kán, Dr.techn.

Wien, 27. September 2023

Martin Rumpelnik	Peter Kán

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Embodied Conversational Agents with Situation Awareness

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Visual Computing

by

Ing. Martin Rumpelnik, BSc
Registration Number 1633397

to the Faculty of Informatics

at the TU Wien

Advisor: Peter Kán, Dr.techn.

Vienna, 27th September, 2023

Martin Rumpelnik	Peter Kán

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Ing. Martin Rumpelnik, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 27. September 2023

Martin Rumpelnik

Acknowledgements

I want to thank Peter Kán for supporting me during every part of this thesis. Thank you for introducing me to this topic and all the useful tips and support to create a higher quality result. I also want to thank my family and friends for always motivating and supporting me throughout all phases of live.

Kurzfassung

Virtuelle Welten bieten unbegrenzte Möglichkeiten für die Erstellung von Lernszenarien in verschiedenen Bereichen. Diese Welten werden oft mit verkörperten Agenten angereichert, um menschliches Verhalten bei verschiedenen Interaktionen zwischen menschlichen Benutzern und virtuellen Agenten zu simulieren. Allerdings verfügen diese Agenten in der Regel nur über begrenztes Wissen und Verhalten und ihre Kommunikationsfähigkeiten sind in der Regel vordefiniert oder sie sind überhaupt nicht in der Lage zu kommunizieren. In dieser Arbeit untersuchen wir die Auswirkungen von verkörperten Agenten mit Konversationsfähigkeiten und Situationsbewusstsein auf die menschliche Wahrnehmung und Leistung in einem Trainingsszenario für Ersthelfer. Wir stellen eine neuartige Lösung vor, um verkörperten Agenten ein Situationsbewusstsein zu ermöglichen, welches ihnen erlaubt, Veränderungen in ihrer Umgebung und ihrem eigenen Zustand zu erfassen und darauf zu reagieren. Die Agenten sind in der Lage, dieses erfasste Wissen durch umfassende Konversationsfähigkeiten zu vermitteln, indem sie eine Kombination aus neuartigen Methoden von NVIDIA für die automatische Spracherkennung und Sprachsynthese und der industrieerprobten Konversations-Artificial Intelligence (AI) Rasa verwenden. Um unsere konversationellen Agenten zu evaluieren, führten wir eine Between-Groups Nutzerstudie mit 24 Teilnehmern in einer Trainingsanwendung in der Unity Spiel-Engine durch und untersuchten die Unterschiede zwischen Agenten mit vollständigen Konversationsfähigkeiten und Agenten mit geskriptetem Audio. Während der Studie haben wir verschiedene quantitative Metriken gemessen, darunter Präsenz, Kopräsenz, Aufgabenleistung, Realismus, Lernerfolg, Informationspräsentation, Agenteninteraktion und Trainingsdauer sowie qualitative Messungen in Form von offenen Fragen. Während unsere quantitativen Ergebnisse keine signifikanten Unterschiede in allen gemessenen Metriken aufzeigten, fanden wir einen signifikanten Unterschied zu Gunsten von Agenten mit vollen Konversationsfähigkeiten in der Metrik Kopräsenz. Darüber hinaus fanden wir signifikante Unterschiede zwischen den Geschlechtern in den Metriken subjektive Aufgabenleistung und Trainingsdauer. Abschließend diskutierten wir das Nutzerfeedback zu unseren konversationsfähigen Agenten und leiteten aus unseren qualitativen Ergebnissen Richtlinien für die zukünftige Entwicklung und Forschung von Trainingsanwendungen mit verkörperten konversationsfähigen Agenten mit Situationsbewusstsein in VR ab.

Abstract

Virtual worlds offer unlimited possibilities for creating educational training scenarios in various domains. These worlds are often enriched with embodied agents to simulate human behavior in various interactions between human users and virtual agents. However, these agents usually only have limited knowledge and behavior and their communication skills are usually predefined or they are not able to communicate at all. In this thesis, we investigate the impact of embodied agents with conversational abilities and situation awareness in a first responder training scenario on human perception and performance. We present a novel solution to enabling situation awareness for embodied agents which allows them to capture and react to changes in their environment and their own state. The agents are capable of conveying this captured knowledge through full conversational capabilities by utilizing a combination of novel methods from NVIDIA for automatic speech recognition and speech synthesis and the industry proven conversational Artificial Intelligence (AI) Rasa. To evaluate our conversational agents, we conducted a between-groups user study with 24 participants in a Virtual Reality (VR) training application in the Unity game engine and investigated the differences between agents with full conversational capabilities and agents with scripted audio. During the study we measured several quantitative metrics including presence, co-presence, task performance, realism, learning outcome, information presentation, agents interaction and training duration as well as qualitative measurements in the form of open questions. While our quantitative results did not indicate significant differences in all measured metrics, we found a significant difference in favor of agents with full conversational capabilities in the metric co-presence. In addition, we discovered significant differences between genders in the metrics subjective task performance and training duration. Finally, we discussed user feedback on our conversational enabled agents and derived guidelines for future research and development of training applications with embodied conversational agents with situation awareness in VR from our qualitative results.

Contents

CHAPTER 1

Introduction

Disaster response training is an important educational exercise for first responders to acquire knowledge and skills needed to be prepared for unexpected situations during an actual disaster. However, traditional trainings, such as classroom courses or real-life simulations may not be effective or cost-efficient enough for today's challenges. Recent advancements in computer and mixed reality technologies opened opportunities to use these tools for training and exercise purposes [ALK+]. Specifically, Virtual Reality (VR) based training can offer an immersive virtual scenario that is usually not present in real-life simulations or classroom-based training. Moreover, VR training does not only offer a unique experience that is hard to simulate in the real world, but is also considerably more cost efficient compared to large-scale real-life exercises [HLB+13]. In real world scenarios, communication with people on site is an important part for first responders to get additional information about the actual situation and is also important to, e.g. properly understand the state of injured people and decide on the best treatment for them. Non-player Character (NPC)s can simulate the behaviour of real people in various virtual situations where it would potentially endanger a real person in a real simulation. These virtual characters paired with a humanoid body, also known as embodied agents, play an important role in VR training applications [CW19], and have the potential to improve realism and increase presence [SBS19]. Moreover, they can provide important social cues, such as eye contact and turning towards the user to appeal more human to a user. This human design can be perceived as appealing by user and make the interaction with the agent feel natural [RHW20]. In addition, conversational capabilities can be immensely beneficial to provide trainees with additional information, needed to solve a given problem of the training. However, virtual agents in VR training applications typically only have a limited set of predefined behaviours and their communication skills are either also predefined and scripted or they are not able to communicate with a user at all. Another important factor which is often missing in VR, is situation awareness. Agents are often not able to capture and react to changes in their environment or their

own state and reliably communicate them to participants of the training. Figure 1.1 shows two training scenarios with virtual agents as bystanders of two incidents. The agents in Figure 1.1a may have experienced the cause of the car crash and may be able to provide important information that is not directly obvious to a trainee after just arriving at the scene. If the agent inside of the car is conscious, he could be able to report about his condition which may influence his rescue and medical treatment. The virtual agents in Figure 1.1b may have knowledge on how the fire spread outside of the fireplace and provide important hints to a trainee on how to extinguish the fire if there were e.g. chemicals involved. Therefore, an investigation on the impact of embodied conversational agents with situation awareness in a VR training scenario is needed to gain more insights for future research and development of training applications in VR.

(a) A trainee in a VR first responder training on the way to help the agent inside of the crashed car with a first aid kit. One agent on the right and one agent behind the car on the left are acting as bystanders in this scenario.

(b) A trainee immersed in VR trying to put out a the spread fire from the fireplace with a fire extinguisher. Two agents, who may be involved in this incident, are watching this scenario.

Figure 1.1: Figure 1.1a and Figure 1.1b show two VR training scenarios where a trainee is immersed in VR and tries to help an agent inside a car and put out a fire. Both scenarios include virtual agents as bystanders, who may have experienced the cause of the incidents.

1.1 Motivation

As part of the Virtual Enhanced Reality for interoperable training of CBRN military and civilian Operators (VERTIgO) project, we want to address the limited communication problem with digital characters by proposing methods for the embodiment of Artificial Intelligence (AI) chatbots into a virtual humanoid body. This thesis therefore investigates and contributes to embodied conversational agents with advanced AI capabilities, including natural language processing, speech synthesis and speech-to-text conversion. Additionally, our developed AI agents use data about their surrounding environment from a defined data source to enable situation awareness. We hypothesize, that the

conversational capabilities paired with AI and situation awareness are beneficial for training scenarios in VR due to higher immersion and therefore increases the learning outcome and self-reported performance of trainees. While previous research already studied the usage of embodied agents in VR [KBH+18, KdMN+20, SBS19, WSR19], they suggest, that more studies are required to learn about the benefits and disadvantages on metrics such as realism, learning outcome and task performance. In addition, more research about the design of such agents will be beneficial for designing and implementing future VR training applications that utilize conversational agents.

1.2 Aim of the Work

The main goal of the thesis is the connection of the Rasa [BFPN17, Ras] conversational AI, see Section 2.4, and the novel NVIDIA Riva [Rivb] speech services, see Section 2.5, to virtual humanoid bodies in 3D. In addition, a suitable model for talking about specific training scenarios was trained for Rasa, which is able to get information from the Unity 3D [Uni] game engine to enable situation awareness. As part of the VERTIgO project, the connection to Riva and Rasa services were implemented in Unity 3D. These components were connected with 3D humanoid bodies in a training scenario to create a humanoid agent with the ability to respond to users in a meaningful and situation aware way. Speaking to a humanoid agent should be fluent by just looking at the agent and start talking without e.g. the need to press a button. Each agent listens and responds to users if the user is nearby, talks and looks at the agent. The virtual agents have fully rigged models as bodies with basic animations. The agents shall also have basic facial expressions such as mouth movements synchronised to their on-the-fly created spoken responses. To enable situation awareness, the Unity application contains a server to provide dynamic scene information to Rasa. Rasa was configured and specifically trained for disaster training scenarios to help with proper intent extraction from the users spoken content. These training scenarios include scenarios where medical aid is required and dangerous substances, such as chemicals leak out and their removal is needed. To evaluate our agents, we created one training scenario as a 3D-VR scene using Unity 3D and we allowed users to move through the virtual world and approach virtual agents by looking and talking. We are expecting that each agent is situation aware and can answer questions such as "What happened?", "Does your leg hurt?", "How many people are inside this house?" or "Do you know how many people are injured?". Compared to traditional NPCs, we studied if our situation aware agents provide a higher level of immersion and a better training outcome. We evaluated our training through a user study to investigate the impact of our conversational agents on human perception.

1.3 Contribution

The main contribution of this thesis are embodied conversational agents with situation awareness for a first responder training scenario. We propose and implement a novel method for enabling full conversational capabilities on virtual agents by utilizing state

of the art Natural Language Processing (NLP) and novel speech services for Automatic Speech Recognition (ASR) and Text to Speech (TTS) in Unity 3D [Uni]. We train a suitable model for NLP to give agents a basic understanding of disaster situations and allow them to answer first responder trainee's questions. Additionally, we propose a novel technique for enabling situation awareness for virtual agents, that is utilized by the NLP service to provide situation dependent answers. To verify our proposed method, we compared two conditions in a between-group user study: Conversation and No-Conversation. The Conversation condition contains agents with full conversational capabilities, i.e. participants can speak naturally with an agent and get answers to their questions. The No-Conversation condition has agents that can only provide information by speaking pre-build text and are not able to answer questions from participants. Both conditions were applied on the exact same training scenario and participants could receive the same information in both conditions. Situation awareness was also enabled for both conditions, so the agents in both conditions reported actual information. Our goal was to investigate the impact and benefit of using embodied agents with conversational capabilities in comparison to agents that can only provide static information. Finally, we performed a qualitative analysis of our study results and we provide guidelines that together with the study results can be useful for future research and development of training applications in VR with conversational embodies agents.

In summary, the main contributions of this thesis are:

1. A novel method for including embodied conversational agents with situation awareness into a first responder training scenario in VR.
2. A user study, showing the impact of conversational agents in a first responder training scenario in VR.
3. Guidelines based on our qualitative analysis of the study results for designing future conversational agents and training applications in VR.

1.4 Outline

Chapter 2 introduces previous research done in the field of virtual training such as disaster response training, rescue tasks and emergency evacuation. Furthermore, we describe how embodied conversational agents were used in the past and how researchers enabled the embodiment of AI conversational agents into virtual humanoid bodies and present previous techniques for enabling situation awareness. In addition, we provide some general background about AI chatbots, NVIDIA Riva speech services and the conversational AI Rasa.

In Chapter 3, we explain the technical details behind our novel method for our embodied conversational agents and how we enable situation awareness. We show how our pipeline works by explaining how we connected NVIDIA speech services and Rasa to Unity and

how the different parts of our pipeline communicate with each other. Finally, we explain the configuration and training data we were using to train the language model for Rasa.

Chapter 4 presents the virtual environment we created as a training scenario for our user study. We describe the environment including walkable buildings as well as the locations of the agents and incidents. We are also explaining how locomotion, object interactions and the interaction with the agents work and what tasks participants are asked to solve for completing the training.

In Chapter 5, we present some general data about our participants and explain how we conducted the user study and what data we collected during the study. We finish the chapter by explaining the design of our study questionnaire and present all of the questions from study questionnaire.

The results of the study are presented in Chapter 6 and we explain how we assessed the data collected during the study. The chapter ends with a discussion of the study results and presents guidelines based on our qualitative analysis for future research and experiments on embodied conversational agents with situation awareness in VR training scenarios.

In Chapter 7, we are reflecting on the limitations of our work and study and propose solutions to overcome these limitations in future research and experiments. Finally, we conclude this thesis with a summary of our work and results.

CHAPTER 2

Background and Related Work

Virtual worlds offer a high potential for interactive experiences in a variety of genres such as entertainment, education and training. There are unlimited ways to design and build a world that users can explore and experience, e.g. navy personal can become familiar with a ship and biology students can learn about anatomy while inside a human body. These worlds that range from fantasy to factual and from past to future are often enriched with intelligent virtual agents to make it more lively and to give users the possibility for interactions and human like conversations. However, these interactions are often very limited and scripted and do not offer users a natural way to carry on a dialogue [Ric01, JRL00]. Striving away from these simple interactions to interactive experiences with face-to-face communications is perhaps the greatest challenge when designing virtual humans. Intelligent virtual agents must not only be concerned with themselves but their surrounding, multiple characters and multiple conversations. They should know when they are talked to and who is talking to them and give dynamic answers dependent on the virtual world where they exist [TR02]. In this chapter, we are exploring previous work and research of training in Virtual Reality (VR) and embodied conversational agents in VR and Augmented Reality (AR) and provide technical background for the reminder of this thesis.

2.1 Training in Virtual Reality

First responders are facing quiet unique challenges when disaster occur and therefore need adequate training to operate safely in dangerous situations. Consequences of critical incidents are high and experiments of highly dangerous situations are usually not carried out for safety reasons and for being too resource intensive, i.e. firefighting training often does not even include a real fire. Traditional classroom settings and low-fidelity exercises are not always sufficient for training unforeseen incidents [LPL22, HZG+20]. With the advancements of modern technologies, training in a virtual world has a high

potential to overcome these issues. Using VR, trainees can be fully immersed in a virtual environment and practice the skills needed without being threatened of their lives. Figure 2.1a shows a trainee getting instructed to an air monitor by a virtual instructor and Figure 2.1b shows a room filled with smoke from a fire rescue scenario. Immersive VR environments give trainees a sense of being physically there by creating a sufficient believe that the environment is real. This sense can be created by using various technologies such as haptic and force feedback and smell and taste replications. Fully immersive environments are possible but very rare as the majority of existing trainings do not incorporate all aspects such as haptic feedback and smell replications to create a realistic experience [NJD19, SRD15].

(a) A virtual instructor of a fire brigade explains to a trainee how an air monitor works [HZG+20].

(b) A room filled with smoke due to a fire in the building [LYXX20].

Figure 2.1: Figure 2.1a shows a virtual instructor explaining a tool to a trainee and Figure 2.1b shows a room filled with smoke from a fire rescue scenario.

Previous research investigated how to utilize various aspects to create a fully immersive training experience for first responders in VR. Mossel *et al.* [MFS+17] created a platform called VROnSite that supports an immersive training for squad leaders of first responder units. They use an entirely untethered Head Mounted Display (HMD) and created two ways of navigation, abstract and natural, to simulate stress and exhaustion which are important factors for decision making. The abstract navigation technique was a simple two-handed gamepad and the natural navigation an omnidirectional treadmill that enabled real walking. During a user study with real fire brigades, they evaluated the difference of their two navigation methods using quantitative and qualitative measures. Their measures included usability of the platform, perception and perceived task loads of participants when assessing two virtual disaster sites with the different navigation techniques. Quantitative and subjective measurement were collected using a 5-point Likert [Lik32] scale and qualitative measures with open questions on a questionnaire

after the experiment. They also observed the physical stress level (sweating and faster breading) of participants by encouraging participants to think aloud during the experiment. Participants reported a high degree of presence and considered the platform highly suitable for training of decision making in complex first responder scenarios, while not favoring any of the two navigation methods. However, the quantitative data revealed the importance of using a suitable navigation technique in this context due to a higher task load when using the treadmill which more closely resembled real-life drills. The qualitative data showed, that participants felt like free navigation, sound and interaction were the most important aspects to train assessment of a disaster situation.

Velz *et al.* [VAG+14] studied the influence of interaction technologies on the learning process with a focus on teaching industrial assemble tasks in VR. They developed a training which can use one of four interaction methods: mouse-based, haptic system and two configurations of motion capture systems where one configuration had 2D hand tracking and the other had 3D hand tracking. In their user study, four groups of participants were training using one of the interactions methods and a fifth group was trained with a video tutorial, which was showing how to perform each step. The day after the training a post-training test was carried out to evaluate the performance of the participants on a real task. The goal of the experiment was, to study the efficiency and effectiveness of each interaction technology for learning a task. In there evaluation they considered both, quantitative measurements such as training time, real task performance, evolution from the virtual task to the real one as well as qualitative data, i.e. user feedback from a questionnaire. The results did not indicate any significant differences in the final performance between the five groups. However, they found a significant difference in the training time, where users trained with the mouse and 3D tracking motion capture system finished significantly faster then the other groups. Using this results, they conclude that motion capture based interactions can be a valid interaction method for training assembly tasks and the perceived collisions of haptic interactions do not necessarily increase the learning transfer from a virtual task to a real task.

Stansfield *et al.* [SSS98] presented a VR system for training medical first responders that focuses on sorting injured people on the battlefield into groups, based on their need for medical treatment, and treat them in order. Users are represented by an avatar and are able to manipulate virtual instruments and carry out medial procedures. Since users were seeing themselves in a virtual avatar, the avatar must be updated at real-time to reflect immediate actions of the users and not cause discomfort. This real-time positional update of body parts was accomplished by using several tracker input modules worn by the users. A dynamic casualty simulation generated the state of various casualties and provided realistic cues of patients conditions, e.g. changing blood pressure and pulse, and let patients respond to the action of the trainee by, e.g. changing the color of the skin. The focus of the training was rapid decisions making and situational assessment in highly stressful situation. They also implemented voice recognition techniques to let users request information such as vitals and give commands to patients that execute certain actions such as evacuation. While evaluating their system, they found a high

level of task complexity and that users want more visual and tactile cues, i.e. perceptual anchors, prior to committing to a decision or commanding action.

In the domain of Chemical, Biological, Radiological, Nuclear (CBRN) hazards, first responders need a high-quality training to avoid fatal errors. Exercises for CBRN trainings are often expensive, require complex management and only reproduce an approximation of a real hazard due to the need to preserve the trainees safety. To cope with this issue, Laberti *et al.* [LLGPM21] developed a VR training platform, that allowed trainees to train alone or in a team. Trainees are able to interact with Non-player Character (NPC)s and NPCs are able to follow or guide the trainee. In a user study they found a high sense of presence and users recognized the high potential of VR training applications. Based on feedback they think, that machine learning techniques would be beneficial to further improve the behaviour of the NPCs and boost believability of the experience. The study also highlighted the need for a different system than controllers to interact with virtual objects.

Search and Rescue (SAR) skills are important first responders in the firefighting domain and therefore, Doroudian *et al.* [DWW+22] saw the need to improve VR training for SAR tasks. They developed a system with immersive maps that have both, static information about the 3D environment and real-time information collected from the simulated environment. The collected real-time information included dynamic locations of fires and persons to be rescued. In a user study, users were asked to use the dynamic maps about the environment to solve some tasks. For locomotion they experimented with free movement by character controller and teleportation. The free movement approach caused motion sickness for some participants without VR experience, therefore they only used teleportation for the user study. Participants could access the virtual map located on their left hand at any time by raising the hand. The main focus of the study was on the information levels from the virtual maps and the danger degree of the environment that was controlled by the fire simulations. The results confirmed the advantage of using real-time information for training and its effects on changes of locomotion behaviors. In their system, they mainly used visual effects and suggest to also include sound effects for a more realistic training.

Lorenzis *et al.* [LPL22] developed a VR training system for practicing the use of a blower as a firefighting tool. The system aims to assists trainees in learning the procedure and assessing their knowledge afterwards. To reproduce the weight of a real blower and enhance realism, they modified a real tool so that it can be used as an interface to the application. For the modification, the handle was replaced by a Vive Pro controller and a Vive tracker was attached to the body of the blower to track the blower location. They implemented a believable, though not physically accurate simulation of the blowers behavior and the fire so that the blower affects fuel and fires in the scene. Flames could spread across the scenes foliage when not correctly extinguished with the blower. The application started with a guided mode where an NPC illustrated and explained via voice recordings all the necessary information on how to properly use the blower as a firefighting tool. After the user has completed the guided mode and learned how to

correctly use the blower, the evaluation mode was started. In this mode, the trainee was asked to use the blower without any guidance of the NPC. The feedback towards the application was positive and users saw the potential of it being useful for training.

VRescure is a system to help trainees get used to various disaster circumstances. Nguyen *et al.* [NJD19] developed a VR city scenario with an ambulance rescue agent and several rescuees. The rescue agent was automatically searching for the optimal path to save all of the rescuees and the trainee was able to interfere in this rescuing process by placing obstacles or adding more rescuees along the way. Placing more rescuees caused the rescue agent to re-route the initial calculated optimal path. Trainees could practice disaster circumstances through observing the intelligent agent who maps the optimal path and reacts to changes caused by the trainees.

Using VR environments has not only been recognized as an alternative to traditional real-life trainings for first responders but also as an alternative for evacuation drills. Sharma *et al.* [SRD15] proposed an application to create unique ways to train emergencies for university campus safety. Similar to first responder trainings, campus trainings in VR have a considerable cost advantage over large emergency evacuation trainings on a real university campus. A quick evacuation of occupants in a building and the movement of people in threat situations is very critical to save lives. Disorganized evacuation can not only lead to injuries and confusion but also death. In the training application, Sharma *et al.* [SRD15] wanted to gather data on human behaviour and emergency response in an evacuation scenario. They created a virtual campus environment and implemented three ways of crowd behaviour: Rules for computer simulated agents, controls for users to navigate through the VR environment as autonomous agents and direct control through keyboard/joystick along with an immersive VR HMD. They created a multi-user evacuation drill on a virtual campus environment where multiple users could enter as avatars through a HMD and take part in the evacuation drills. The environment also contained computer controlled agents that were programmed to act as obstacles to users. The behavior of the computer controlled agents was either defines through rules or could be controlled through users.

Jin *et al.* [JBG+19] proposed an agent-based virtual interview training system to help college students with high shyness level to improve their interview skill and reduce anxiety before being exposed to a real interview. They developed three virtual agents with different types of personalities and three kind of interview scenarios with a multidimensional evaluation method to meet the most common demands. The interview scenarios included interview trainings for the enterprise, civil servant and college domain. A user study indicated, that the system can help shy college students cope with interview anxiety and improve their interview training performance. During the study the system was evaluating behavior, facial expressions and physiological signals of the participants [JBG+19].

Peretti *et al.* [PSSE21] experimented with a novel training solution for first responders that utilizes a gamification aspect. First responders need to assess and act fast, therefore they put the user in a stressful situation where a timer is running and they need to make decisions quickly. The training scenario had several variables for adjustment to create

a dynamic scenario where the user was presented with multiple choice questions and object manipulation tasks which had to be solved under a time constraint. The questions were created in collaboration with professionals and the object manipulation tasks were executed with a controller or hand tracking technology. In an experiment with users, they found that hand tracking is highly appreciated compared to more commonly used controllers.

As first responders are put in highly stressful situations, they may experience multiple stress levels such as fear, panic and collapse of clear thinking. To organize appropriate support and avoid risk-taking, it is important to stay cognitively under control in these circumstances. Using psychophysiological measurements, Paletta *et al.* [PSR+22] studied levels of stress during training in real and virtual environments in the context of situation reporting under realistically simulated mission conditions. They induced physical stress in real-life by having participants run a 5-minute endurance run on the test site and similar to Mossel *et al.* [MFS+17] used a VR-supported treadmill in VR. To simulate real equipment, participants also wore heavy operational clothing and a 20kg backpack. Cognitive strain was induced by having a operator watch the mission scenario video and informing the participant to prepare a situation report within one minute. When the report was ready, the participant had to report on the scenario within one minute. The scenario was designed to follow a command scheme that relates to observing, considering actions and communicate actions. The evaluation showed, that this was a promising method to measure observation skills and that creating situation reports lead to a high level of cognitive and emotional stress that must not be neglected in trainings.

Haskings *et al.* [HZG+20] studied requirements that first responder trainings have to offer for a high quality training. In a typical classroom training, a trainer presents some situation to a trainee who assesses the situation and responds with the most appropriate action to take. The most appropriate action is usually what the trainee believes is right. While this form of role-playing is valuable, it is mostly based on a verbal story and lacks real visuals or sound that occur during real incidents. Since the situation description is presented verbally, the trainer may give away important cues that the trainee might not have noticed in a real situation. Understanding and noticing important cues in the environment is often called situational awareness and defined by Endsley [End95] as perceiving all the elements in the surrounding environment and knowing the relevant information. Without situational awareness first responders might jump to conclusions based on a bias and endanger themselves or other civilians on site. Another essential skill for first responders is proper communication with either team members or civilians. First responders often find themselves in unique, stressful, difficult and dangerous situations were communication is essential to communicate the problem and find suitable solutions.

2.2 Conversational Agents in Mixed Reality

Conversational agents can be beneficial in VR and AR applications to aid in various training scenarios such as interpersonal skill training, including sales pitching, negotiation

and interviewing, as well as health care trainings. These agents provide a safe opportunity to practice skills needed for real human to human communications by simulating different persona, including guides, mentors, competitors, teammates and patients. Existing virtual worlds are mostly based on military simulations and computer games and their focus is more towards a photo realistic environment than on the human aspects of agents that inhabit the virtual environment. It is important to embed a persona into conversational agents so they can properly engage in a conversation and convey intelligence to provide more than just a sufficient training. Embedding intelligent human behavior into virtual characters can provide new possibilities regarding training and learning opportunities for trainees that currently requires complex real-live exercises, role-playing or classroom contexts. Realistic agents need to have enough realism and intelligence to create the illusion of human-like behavior. The agents need to respond to human users and events around them and they need to be interpretable by users through verbal and nonverbal cues and gestures that people usually use to communicate and understand each other [CW19]. Designing intelligent embodied conversational agents is a complex task where several technologies such as Automatic Speech Recognition (ASR), Text to Speech (TTS), Natural Language Processing (NLP), Artificial Intelligence (AI), deep machine learning, 3D computer graphics and animation have to be combined to create believable agents [CW19]. Multiple factors such as appearance, behaviour and responses of an agent play an important role in how believable and trustworthy an agent appears to a user [SW18, WSR19, KBH+18]. Furthermore, the agent can also improve social richness and social presence [KBH+18] and help reduce the task load in virtual trainings [KdMN+20].

Mission rehearsal exercises are important for Army personnel to gain experience in handling peacekeeping situations. For this purpose, Traum *et al.* [TR02] developed a high-end VR training application with Hollywood storytelling techniques in a VR theatre with immersive, spatialized sound. The virtual training scenario took place in a small village in Bosnia with buildings, vehicles and virtual agents and the user took the role of the lieutenant. The virtual characters had support for speech interactions that were based on a script or dependent on the users actions, variations of that script. Figure 2.2 shows an example of the peacekeeping scenario where multiple conversations are possible. The lieutenant could talk to the sergeant, the mother or to the medic and the medic could e.g. talk to the mother. Furthermore, some agents could also simply listen to conversations near them, e.g. the mother could listen to the conversation between the sergeant and the medic. To handle these various conversation possibilities, the agents had to be smart enough to figure out who is talking to whom and know when they are addressed. In addition, they had to carry nonverbal cues such as looking at the person who is approaching the agent to initiate a conversation.

Bersot *et al.* [BGGN98] found that users often prefer to say what they want rather than "do it" by using traditional input devices such as a mouse. They developed a conversational agent embedded in a virtual world with support for ASR and TTS. Users could navigate through a complex virtual environment by talking to the virtual world and the agent behind the scene would respond or move the user. Due to limitations of

Figure 2.2: Traum *et al.* [TR02] explored embodied conversational agents in an interactive peacekeeping scenario. The image shows a sergeant, a mother and a medic in the foreground and soldiers and bystanders in the background.

the used speech recognition framework, users had to press a button to talk and could not speak completely fluent but had to make a short pause between every word. With this navigation method, users were able to say, e.g. "Go in front of the house" and the agent would move the user towards this location.

The health care domain can be very challenging, not only due to health related issues but also due to violence and aggression from, e.g. relatives and friends of the patient, against health care workers. VR simulations are cost effective ways to complement traditional de-escalation trainings for health care workers worldwide. In the de-escalation trainer developed by Moore *et al.* [MAB+22], the user takes the role of an emergency department nurse and needs to de-escalate a situation where the son of a patient was distressed due to the long waiting time. The son was represented by a conversation enabled agent that was able to interpret voice from the user, generate an answer and give a verbal response through Google TTS. Depending on what the user said to the agent, the agent responds either positively or negatively which furthermore decreases or increases their level of aggression and frustration. The scenario ended with a overview of the performance if the user has either de-escalated the situation or the agents aggression level got too high. Feedback to the application was positive and users liked that it was rather portable and used an untethered HMD.

Job interviews can be very difficult to handle emotionally by some people due to being nervous or anxious. Hartholt *et al.* [HFR+19] tried to tackle this issue by proposing a framework with embodied conversational agents with support for ASR, NLP and TTS. The framework supports room-scale VR as well as seated VR and mobile AR to be accessible for the majority of people. With this framework, users can practice job interview sessions with the agent as interviewer. Before the training starts, the user can choose between a male and a female interviewer. The chosen interviewing agent will

then ask the user common job interview questions. The framework measures eye contact, blink rate and response delay using Magic Leap sensors. These metrics are presented at the end of the interview to the user.

Griol *et al.* [GSMC19] developed enhanced conversational agents with the capability to provide academic information and placed them into social virtual worlds such as Second Life. Their agents were trained on a set of real and simulated dialogues and are able to modify the dialogue strategy by detecting new answers that were not used during the training. The results of their experiment showed, that the agents are able to fully adapt their conversational behavior to the users interaction characteristics.

Sexual violence in colleges is a common problem which existing prevention programs fail to address. This led Schlesener *et al.* [SLB+23] to develop a mobile AR game, that aims to improve current sexual assault bystander intervention training. The training includes a geolocated real campus and can be started from anywhere on the real campus by opening the application on the mobile phone. A communicative agent informs users of a specific harassment scenario while guiding them to the location where it takes place. Through the mobile application users then see digital humans role-play as harassers and victims. After watching the harassment scenario unfold, users are asked by the agent to choose an intervention option. After they have chosen, users watch the consequence of their decision and the agent explains them if their decision was good or why it was not so good. If the decision was not the best option, users are able to choose a different options. The game rewarded users with points to create a gamification aspect, creating a motivation factor of getting all points. The agents used lip sync technology and were fully animated with gestural animations such as waving and head nodding.

Previous research found that a natural appearance of virtual agents is preferred over more simplistic versions [Unc, SWHK15, SWH18]. However, the closer an agent resembles a real human, the more likely a user will notice small details that cause irritations, negative emotions and distrust because it does not meet the users expectations of human features [RHW20].

This effect is called the uncanny valley and was proposed by Mori [MMK12] and marks a region of negative affinity towards an entity with human-like appearance on a graph showing the relation between human likeness of an entity and the perceiver's affinity for it. Mori found that by changing the appearance of robots to more closely resemble a human being also increases our affinity for them until we come to a valley, see Figure 2.3. This phenomenon is not only limited to robots but can also appear with humans that have physical disabilities and wear a prosthetic limb, i.e. an artificial hand. These artificial hands are often indistinguishable from a real one, but when we touch it, we realize that it is actually artificial and loose our sense of affinity for it, i.e. the hand becomes uncanny, hence it is placed near the bottom of the valley in Figure 2.3. Movement further amplifies the peaks and valleys of the graph since our affinity increases towards moving objects, i.e. a turned off industrial robot is just a machine, but when it moves and grabs an item like a human hand, our affinity for it starts to increase.

While the theory of the uncanny valley may suggests, that virtual agents with a more simplistic humanoid design are favored over hyper-realistic humanoid characters, the study from Reinhart *et al.* [RHW20] contrasts this. They found that a carefully designed realistic humanoid agent in AR increases social interaction, social presence and likability compared to a more simplified version.

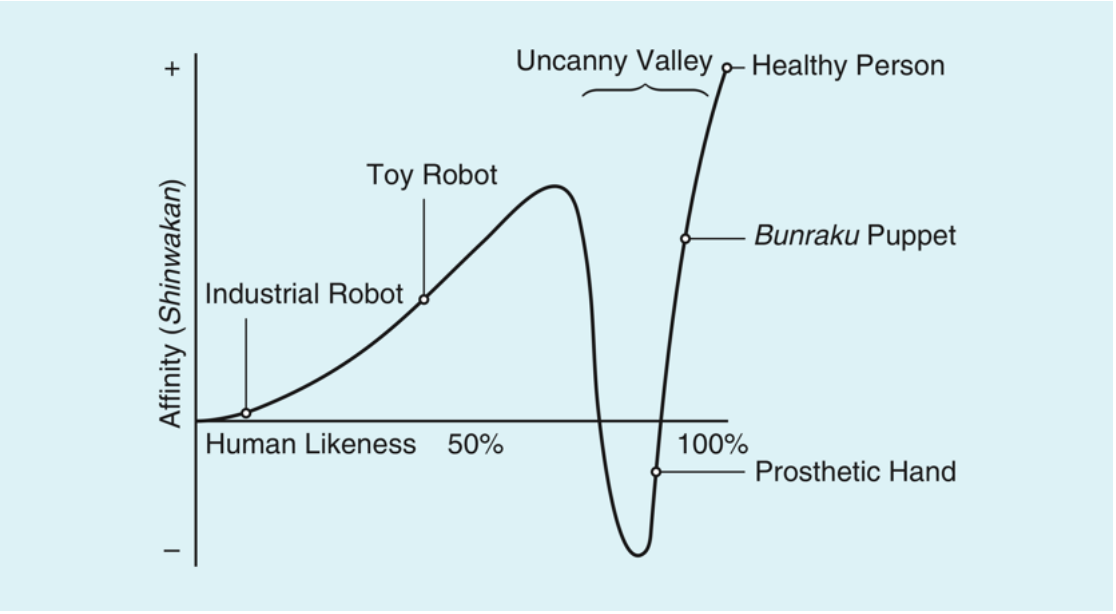


Figure 2.3: The graph is showing the hypothesized relation between the human likeness of an entity and the perceiver's affinity for it. The uncanny valley is the region of negative affinity towards an entity with almost human-like appearance [MMK12].

A lot of people are interacting with intelligent virtual agents in the form voice assistants such as Amazon Alexa and Apples Siri on a day to day basis. However, these agents are only capable of reacting to voice commands through voice feedback and lack any nonverbal cues such as eye contact and body movement which are important for social interactions. Kim *et al.* [KBH+18] tried to overcome this issues by providing natural social behavior and visual embodiment for virtual assistants in AR. The results of their user study indicated that this led to an increase of the users confidence that the agent is able to influence the real world by e.g. walking to a lamp and switch it off. It also led to a higher confidence that the agent will respect the users privacy because the agent left the room during the study when commanded to do so which closely matches the behavior of real human behavior. In addition, they also found a positive effect on social richness and social presence with the agent.

Techasarntikul *et al.* [TRO+19] explored two different styles of locomotion for conversational agents in a virtual guide context. An agent took on the role of a virtual guide, who located interesting locations on a large piece of art. The agent then moved near that location and used hand gestures to point to the location of interest. Since eye

contact is important for a communicative relationship between the agent and the user, the agent always faces the user and makes eye contact with a user. The guide narrative to explain information for a specific part of the image was generated from a commercial TTS software. In a user study, the authors compared locomotion through teleportation and flying. The results indicated that participants preferred flying over teleportation as it was easier to track the position of the agent.

In a mixed reality environment real humans are able to coexist with virtual agents in the same virtually augmented physical space. Schmidt *et al.* [SNS19] researched such boundary crossing agents, which are capable of changing physical properties such as object locations and surface materials in this virtually augmented space. They used robotic actuators to introduce physical interactions from the virtual agents and thermochromic ink, which changes color based on temperature, for changing surface materials. The user study was focused on perceived social and spatial presence for which they developed a golf scenario. Participants had to interact with an agent who was capable of physical manipulations such as hitting the golf ball and writing on a physical paper. The golf ball was not a standard golf ball, but a robotic one that could move along a scripted path, which simulated the interaction between the virtual agent and the golf ball. To enable writing appear on a sheet of paper, a novel device that activated thermochromic ink on a sheet of paper was used. The activation happened through temperature changes of the device based on the golf scores. The possible scores were pre-defined on the paper and mapped to a temperature. They synchronized the temperature change with the agents animations such that it looked like the agent was really writing on the physical paper. Quantitative results did not show any significant difference between boundary crossing agents and virtual agents without these physical capabilities. However, the qualitative results showed, that participants seemed to be more in favor of the physical object manipulations since the boundary crossing agents improve realism and the user experience.

Spatial presence, i.e. the ability to experience a sense of "being there" is an important aspect to create an immersive virtual experience. Khenak *et al.* [KVB20] studied spatial presence and related factors, including affordance, enjoyment, attention allocation and cybersickness, in a within-subject study where users had to complete a navigation task where they had to follow a route and avoid obstacles on the way. Their conditions for the study included a real environment, a remote environment via a telepresence system and a virtual simulation of a real environment. The evaluation was done through a presence questionnaire and they also collected performance measurements regarding task execution and environment recollection. The results did not show a significant difference in spatial presence between the remote and the virtual condition but showed affordance and enjoyment more in favor towards the virtual condition. The remote condition had a higher degree of reality than the virtual condition and the number of collisions was also lower in the remote condition. The authors also found, that the behavior of participants in the remote conditions resembled more closely the behavior in the real environment.

Social and physical presence in VR can be experienced by all kinds of different people

and regardless of age and gender. Felnhofer *et al.* [FKH+14] investigated this in an experiment between a group of older people with an average age of 67 years and a group of younger people with an average age of 25 years. Their scenario included a virtual outside environment, where the participants started, including a virtual coffee shop. Participants had to learn how to navigate outside and afterwards enter the coffee shop. Inside of the coffee shop they had to order a drink from the waiter and interact with a stranger. The results of their experiment did no indicate a significant difference in social and physical presence between the older and the younger groups. However, they found that male participants experienced a higher level of spatial presence than females which supports past findings on gender differences regarding spatial presence.

Research on embodied conversational agents is mostly focused on the agents and ignored the external environment, i.e. research is more focused on the believability aspect of the conversation itself and the non-verbal communication cues such as gestures, gaze and facial expressions. However, believable agents need to be able to also reason about their environment, understand interaction capabilities of other participants, understand their own goals and current state of the environment. Ijaz *et al.* [IBS11] labels this as awareness believability, which can be described by the three components: environment awareness, self-awareness and interaction-awareness. Awareness is an essential part of conversational behaviors. In conversations we are typically aware of where we are (environment awareness), who we are (self-awareness) and generally how the interaction is progressing (interaction-awareness). Similar to this, agents should have up-to-date knowledge of their surroundings and know where and what buildings and objects are around them (environment awareness). They should be able to reflect on their own state in the virtual world, explain reasons for performing certain actions or using certain objects and have awareness of their own goals and plans (self-awareness). Finally, an agent must understand its own opportunities when interacting with other participants in the virtual world and predict possible actions others may perform (interaction-awareness). To implement awareness believability, Ijaz *et al.* [IBS11] proposed two levels of environment annotations: object annotation (annotation of object in the environment with names, type and descriptions) and regulation annotations (annotation of social norms, interaction protocols, roles and other kinds of regulations for interactions). When a user interacts with a virtual agent in the virtual world, the agents are using a communication interface connected to the two layers to generate intelligent responses, see Figure 2.4. When an agent is asked about the environment, the object annotation layer is used to generate object information. Whereas, for questions about the agent's interactions and self-awareness such as goals or plans, the regulation annotation layer is used. This layered approach was primarily used as a generic solution, so some features could be used in a new environment without modifying the core functionality. The annotation layers also allow for dynamic environments, where objects, agents and buildings could be changed, removed or inserted at any time. The approach was evaluated in a user study which showed, that the perception of agents with this awareness approach are considered more as believable.

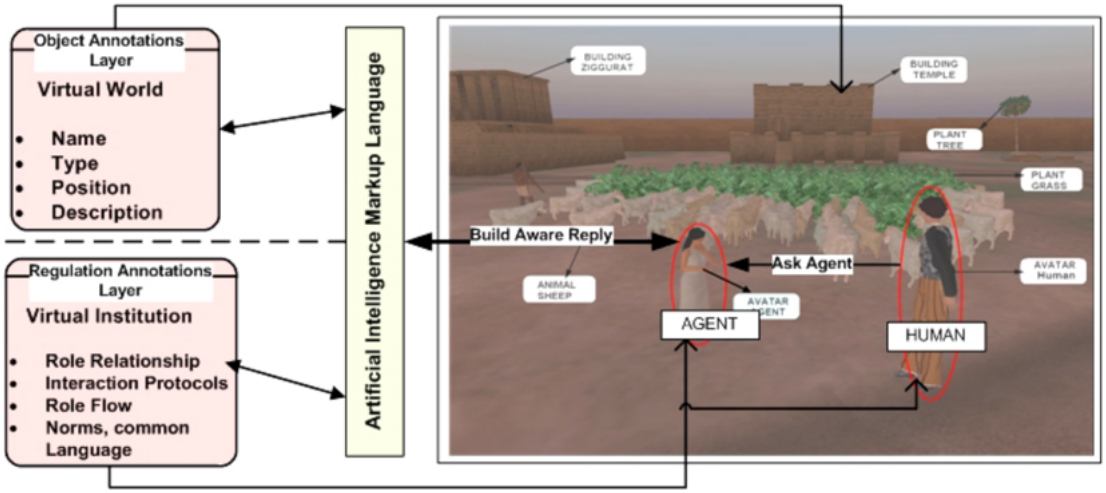


Figure 2.4: Ijaz *et al.* [IBS11] used a layered annotation architecture to achieve believability awareness of virtual agents where objects and regulations are described by various attributes.

Wang *et al.* [WSR19] found that users are more likely to gaze at human-like agents than non-human agents and that users are gazing at an agent who is speaking to them or while they are speaking to the agent. They also found that a user would wait a short time for a response from the agent before looking away. Their tests included agents of different appearance such as non-human, voice only and embodied with voice. The tests were done in AR where users preferred embodied miniature human-like agents to full size embodied human-like agents due to a reduced feeling of uncanniness towards the miniature version.

Additionally to conversational embodiment, Kangsoo *et al.* [KBH+18] found that users perceive an agent more believable when it also has natural social behaviors during interaction. These behaviours create confidence in the agents awareness of events that happen inside the world. Conversations with these agents are also treated more like a real human to human conversation [BC05, KBH+18].

Schrammel *et at.* [SGST07] conducted experiments to research if an agent's gaze can guide the user's attention towards designated locations. While talking, the agent would look at certain locations which are mentioned in speech. This, however, had a negative impact as users seemed to pay less attention to the agents words when eye contact was broken. Since eye contact is often broken in real world human to human conversations, this may also imply that users do not see the agent as human like as expected.

Finding solutions for complicated problems can be more effective in a group. However, collaboration is not an easy task as it requires proper communication between group members to be more effective than finding a solution alone. Kangsoo *et al.* [KdMN+20] investigated the effects of conversational embodied virtual agents on collaborative decision

making. They created a desert survival task with three conditions where participants had to perform the task alone, work with a disembodied voice assistant and work with a conversational embodied agent. The voice assistant and the agent used pre-recorded audio for all the answers they were able to give. Using lip sync technology, the agent also had the lips synchronized to the spoken audio for a more realistic appeal. The results of a within-subject study showed a higher task performance for the conditions with the disembodied assistant and the conversational embodied agent compared to working alone. Furthermore, the agent had a significantly lower reported task load than the disembodied voice assistant. Participants also experienced a higher level of social presence with the embodied agent, supporting previous research findings that embodied agents help to increase immersion. While the pre-recorded audio was fine for the small domain of the authors experiment, intelligent virtual agents would need a learning approach to be able to understand and respond to participants in a more dynamic way.

2.3 Chatbots

Research and experiments have shown, that AI chatbots embodied into virtual humanoid bodies can have a major impact on how virtual agents are perceived by users [SW18, WSR19, KBH+18]. These conversational systems need to process some input from the user and create an appropriate response to be believable. Chatbots are usually constructed by using retrieval-based models or generative models. Bots using a retrieval-based model are able to respond to answers with correct grammar and spelling if they are in the dataset that was used for the training. In contrast, bots using a generative model are able to answer questions outside of the training dataset but these answers may contain spelling or syntax errors. The training data of the bot determines the domain, a chatbot is operating in: closed domain or open domain. Bots operating in a closed domain are only able to answer questions in that specific domain, requiring typically a smaller training dataset. Bots operating in an open domain are able to answer unrestricted questions, therefore a very large amount of training data is needed to support this unrestricted knowledge. There are various ways to build the different types of chatbots but each needs to be able to handle classification and determining and extracting the intent that a user expresses. Smart chatbots are also able to understand acronyms and misspelled words [LLK20].

Serban *et al.* [SSG+17] developed a chatbot, called MILABOT, which allowed interaction in both speech and text form. The chatbot used a deep reinforcement learning approach which is a combination of multiple algorithms and neural networks. Their approach did well in the Amazon Alexa price competition and they hypothesize that the bot could perform better with more training since every component of the chatbot consists of a trainable machine learning model. While they developed their own deep learning models for natural language retrieval and generation, multiple researchers lean towards retraining existing dialogue systems using Rasa, see Section 2.4.

Lam *et al.* [LLK20] created a closed domain chatbot with Rasa and trained its knowledge

domain to the College of Information and Communication Technology of Can Tho University in Vietnam. Their chatbot did quite well for questions belonging to the trained intent but they found, that the answers were sometimes unnatural and the chatbot could not answer question outside of the trained dataset. They used a limited training data set and suggest that more data for training is needed to help the system answer questions outside of the training dataset. Jiao [Jia20] compared Rasa Natural Language Understanding (NLU) with a Neural Network (NN) system using Tensorflow [Ten] and found that Rasa NLU has higher accuracy that the NN model.

Baccinellli *et al.* [BvdBR+22] created a virtual health coach called Perfect Fit which is also build on Rasa. Their chatbot is trained in the health domain and they configured Rasa to get data from a database which stores user information and can be updated with new user information without re-training the model used for data extraction. Similar to Windiatmoko *et al.* [WHR20] with their Facebook chatbot using Rasa and Nenciu [NCD20] with their conversational agent for the Romanian language they show that Rasa has great accuracy when enough data for training is available. Linders *et al.* [LVA+22] used Rasa for dialogue management in their health care domain agent since Rasa is an often used open source conversational AI.

2.4 Rasa

Previous research showed, that Rasa is often a popular choice among researchers from multiple domains [BvdBR+22, LVA+22, Jia20, WHR20, NCD20] to enable a believable conversational system. Rasa provides easy to use open source tools for building conversational systems. It consists of the two modules Rasa Natural Language Understanding (NLU) and the dialogue management part Rasa Core. The NLU module is responsible for understanding what a user wants by analyzing a message from the user. This process usually splits the text into tokens based on some rules and annotates them using a model trained for the domain of the system. Rasa Core manages the state and history of the conversation so that the system always knows, how the conversation progressed. The basic architecture of Rasa can be seen in Figure 2.5. When Rasa receives a message from the user, it is passed to an Interpreter for the extraction of intent, entity and other structured information. The dialogue state is saved in the tracker which is the only stateful component in the system and there is only one tracker per conversation session. The tracker stores slots and a log of events that led to the state and occurred during a conversation. The state of the conversation can be reconstructed by replaying all of the saved events inside the tracker. The current state of the conversation is sent to the policy, which chooses the next action based on that state. The chosen action is logged by the tracker and executed. The executed action performs some user defined tasks and may send back a message to the user [BFPN17].

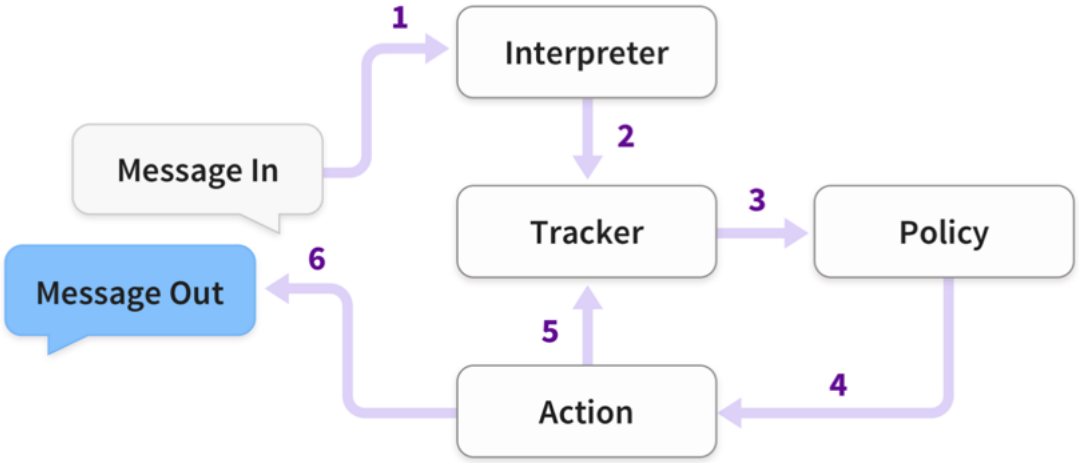


Figure 2.5: 1. The received message from the user is passed to an Interpreter for intent and entity extraction. 2. The state of the conversation is stored inside the tracker. 3. The policy receives the current state and 4. chooses the next action to take. 5. The tracker logs the chosen action. 6. The chosen action is executed and may send back a message to the user.[BFPN17]

2.5 NVIDIA Riva

Different solutions were used in the past to add speech input and/or output to conversation enabled agents. In our research, we are interested in a natural user-agent conversation and therefore see the need for real-time speech recognition and text-to-speech technologies. In 2022, NVIDIA introduced a novel GPU-accelerated Software Development Kit (SDK) called Riva [Rivb] for building multilingual speech applications with real-time performance and support for Automatic Speech Recognition (ASR) and Text to Speech (TTS) and neural machine translation. The SDK can be deployed in clouds, data centers or on embedded devices and streamlines the end-to-end process of developing speech AI services with real-time performance. Riva achieves a latency of under 300 ms to interact with users naturally and provide a human-like interaction. It includes pretrained speech models, trained and evaluated on wide variety of real-world datasets including telecommunications and healthcare vocabulary, and tools to customize and build new models for specific use cases. ASR, also known as speech-to-text, speech recognition or voice recognition is the process of converting a raw audio signal of spoken content into text. In contrast, TTS, also known as speech synthesis, takes some text and converts it to an audio signal, i.e. it generates human-like speech from plain text. Riva services are exposed through an Application Programming Interface (API) operations which are accessible via Remote Procedure Call (RPC) [RPC] endpoints to hide the complexity of the services. The RPC API is exposed to client applications through a API server running inside a Docker [Doc] container. Figure 2.6 shows the Docker encapsulated Riva services on the left and client applications on the right. The services are responsible for processing all the speech

incoming and outgoing data [Riva].

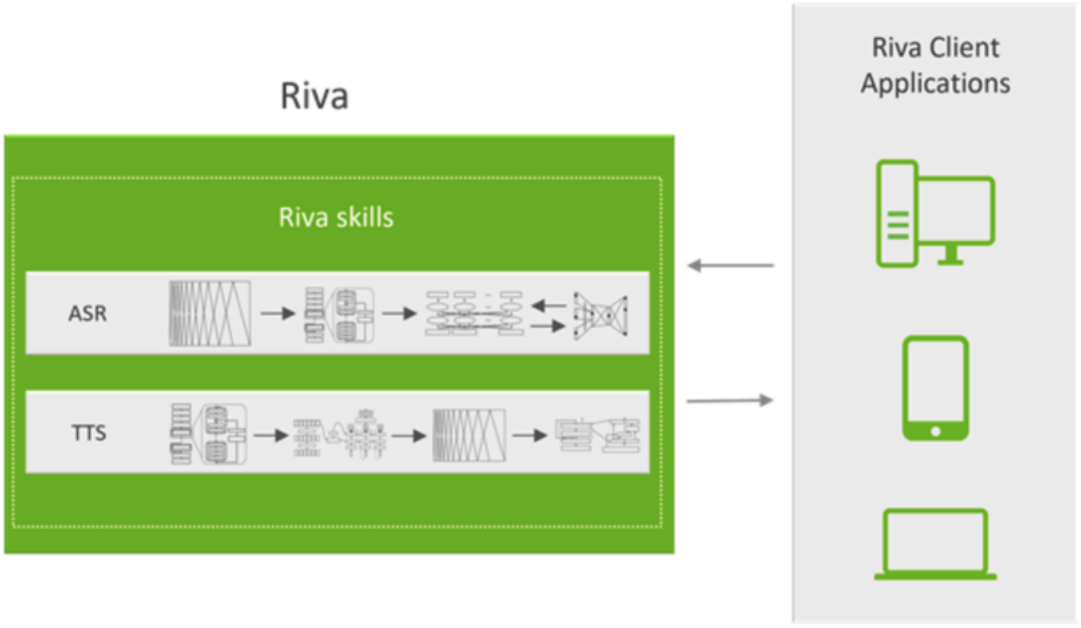


Figure 2.6: Riva services are exposed to client applications through RPC calls and are responsible for processing all of the incoming and outgoing data. ASR converts speech audio signal into text and TTS turns text into a verbal, audio form. [Riva]

CHAPTER 3

Methodology

The main goal of this thesis is to explore agents embodied in a 3D humanoid body in Virtual Reality (VR) with speech capabilities that know about their surrounding and are able to have a natural conversation with a human user. There are several challenges to achieve this goal. First we need to define agents with certain properties and knowledge acquiring capabilities, so that they can observe their own state and their surroundings including other agents, objects, buildings, incidents and human participants. The agents then need to be able to process spoken content coming from the user and respond accordingly. To process and analyze spoken content, we need to create a transcript out of it, which is where we use use the Automatic Speech Recognition (ASR) service from NVIDIA Riva. The created transcript is sent to Rasa for Natural Language Processing (NLP) which uses scene specific information from our Unity application to generate a suitable response for the user. This response is synthesized into speech again with the Text to Speech (TTS) service from NVIDIA Riva and output through the speakers to the user. In this chapter we are first giving an overview of the speech pipeline and how the Unity application, Riva and Rasa communicate with each other. The remainder of this chapter explains how our scene is represented and stored to enable situation awareness followed by how we use Riva and how we configured and trained a suitable model with Rasa.

3.1 Speech Pipeline

The voice of the user is recorded through the default microphone input and audio is played back through the default audio output configured on the computer. For convenience reasons and to make the whole setup easy for the users, we are using the microphone and the speakers of the HTC Vive Pro HTC Head Mounted Display (HMD), therefore the user only needs to wear the HMD and no external headset or microphone. Since the microphone of the HMD is not located directly in front of the mouth, it is by default

more sensitive than other microphones and we found that it is easy to record unwanted noise in the background even when the user is not speaking. This unwanted background noise, combined with the static noise that nearly every microphone has, would also be used for speech recognition if we directly route the incoming voice to Riva. We choose to route the incoming audio through Voicemeeter Banana [Voi] and add a noise gate to suppress the unwanted noise when nobody is speaking. Therefore in the Unity application, we are able to sent any input that passed through the noise gate to Riva and do not need to filter out noise by ourselves. Though Riva would be able to handle and ignore some amount of noise, pre-filtering the microphone input also reduces the number of Application Programming Interface (API) calls to Riva and therefore also reduces Graphics Processing Unit (GPU) load since the GPU is not tasked with analyzing and transcribing noise.

Figure 3.1 shows the pipeline for one agent, e.g. in the real application, the pipeline is executed for every agent that received voice input from the user. The incoming voice input in Figure 3.1 to the Unity block is therefore already the filtered input that passed through the noise gate, which is packed into a bytestream and send via Remote Procedure Call (RPC) [RPC] to the Riva ASR service for processing. Riva will send back a transcript of the received audio data which the Unity application will send to Rasa for language processing. Rasa will analyze the transcript, extract relevant data and determine an intent. Depending on the determined intent, it may need further information from the Unity application and make a *HTTP* request to get it from the webservice on the Unity side. The webservice handles the request and queries a SQLite [SQL] database, that stores the information, and sends the data in *json* format back to Rasa. With the received information, Rasa can create a suitable response and send the response to the Unity application where it is queued for speech synthesis. The response queue is checked every frame and if it contains a transcript for speech synthesis, the transcript is de-queued and send via RPC call to the Riva TTS service. The audio data returned from Riva is queued into another queue for playback. Audio data is de-queued every frame except if there is already audio playing on the audio source of the agent, i.e. the agent is currently speaking. This insures that the agent will respond to multiple quick questions in order and the agent will not say two or more responses at the same time. The audio, i.e. the voice of the agents will be played back through the default speakers of the computer, e.g. in our case the integrated speakers of the HTC Vive Pro.

3.2 Unity

Rescue training scenarios can have multiple different events, incidents, objects and agents with various attributes that take part in the scenario. One challenge is to design the scene in a way to have a structured representation of all the important objects and events, so that our agents can talk about them. We also want to keep scene creation flexible and therefore decided on a component based system similar to the one used in Unity, where objects and agents are composed of different components to give them attributes,

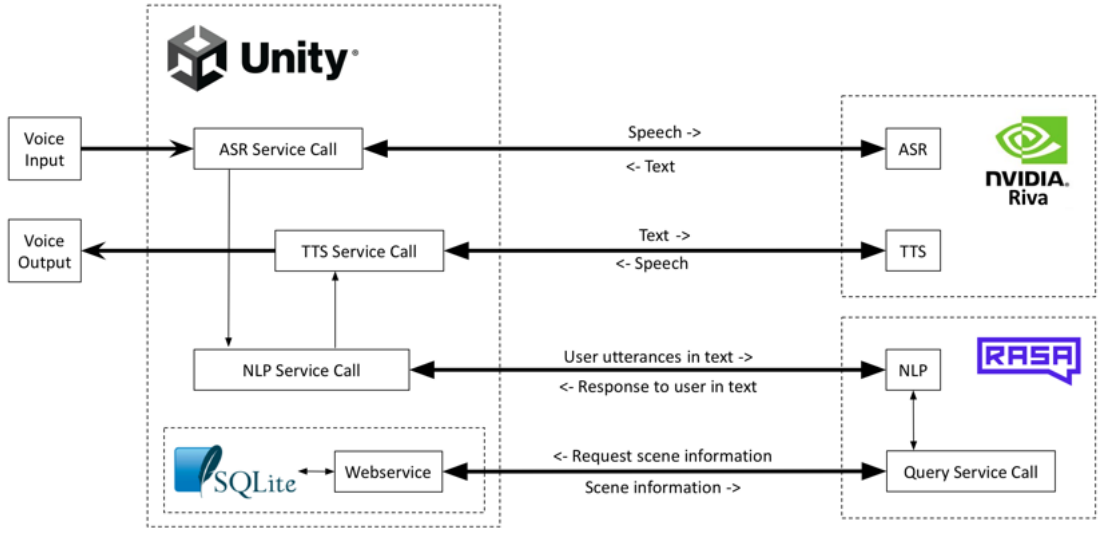


Figure 3.1: Overview of the speech pipeline. The Unity application receives voice input through the microphone and uses the Riva ASR service to convert it into text. The text is send to Rasa for language processing, which will callback the Unity application to receive scene specific information. The received information about the scene is used to create a proper response to the user and send back to the Unity application which will send the response to the Riva TTS service for text to audio conversion. The audio response is played back through an *AudioSource* on the responding agent.

capabilities or personality. The goal of this is to make it easy to compose a variety of objects in the Unity editor to create multiple scenarios without changing any code.

The virtual world is divided into multiple parts using invisible bounding boxes which represent named areas or locations. Hence the enclosing bounding box for a factory building has the name of the factory. With the location name, we also define a description where in the scene the location is located so agents can help users find a location. Just like in the real world, all objects and agents in the scene must be inside a certain location. We use this location division to give agents limited knowledge, so by default, they do not know about the whole scene. By default agents only knows about their enclosing location but we can add additional locations and therefore knowledge about locations in the Unity editor. Defining different locations throughout the scene enables questions such as "What happened in the factory?", "Where is the car?" or "How many people are injured inside the hotel?".

We define incidents as an object that has a name, e.g. fire, a description how it happened and a description how to resolve the incident, e.g. use a fire extinguisher. This definition enables questions such as "How did the fire start?" or "What can I do against the fire?". Incidents can also affect agents, such as when an agent is locked away, e.g. inside a room.

All of the incidents can be resolved by using a tool such as a fire extinguisher or an axe.

We give these tools a name and a description what the tool can be used for. Resolving an incident happens by either touching the incident with the tool, e.g. touching a closed door with an axe, or by pointing a particle effect towards the incident, e.g. pointing the water beam of a fire extinguisher towards the flames.

A simple agent in our scene is defined by the attributes first name, last name, age and gender and can be enhanced by adding more components in the Unity editor. The gender also defines which voice the agent will use for speech synthesis. Since NVIDIA, at the time of writing this thesis, only provides two voices: one male and one female voice, for Riva, our gender selection is limited to these two genders. We enrich the agents visually by giving them a humanoid body from the Rocketbox Avatar Library [GFOP+20, roc] created by Microsoft. The library provides realistic looking, fully rigged avatars optimized for VR. Despite the realistic look of the Rocketbox avatars, the meshes are relatively low-poly, making them perfect for real-time embodiment in VR.

The avatars also come with a couple of animations where we use one of the *breathing idle* animations when no agent-user interaction happens, the turning animations so that the agents face the user when the user gets close and one of the listening animations when the user is talking with the agent. While the animations are rather simple and not very complex, we found these animations gave the agents a more realistic looking appearance and made them more human. An agent starts listening to the users voice as soon as the user gets within a range of about 3 meters and looks at the agent. When this condition is satisfied, the agent automatically starts the listening animation and stops it, as soon as the user leaves the range of about 3 meters. As long as the user is within the range and looked at the agent once, the agent will listen, e.g. it is not required to always look straight at the agent.

The agents gain speech and Artificial Intelligence (AI) capabilities, through a separate component that communicates with Riva and Rasa. Each agent also receives a Unity *AudioSource*, which will play back the responses to the user. We also enable 3D spatial audio on these *AudioSources* so that the user perceives the audio output dependent on the location and head rotation differently. Since the Rocketbox avatars are fully rigged, we integrated Oculus Lipsync [Ocu] for Unity, which analyzes the synthesized audio from the TTS service and creates lip movements that correspond to the particular spoken sound during playback.

Agents can have zero or multiple conditions that represent the agents physical or mental state, e.g. head ache or ankle pain. We define a condition as a description, a cause describing how it happened, if the condition disables movement of the agent and which body part is affected. Conditions are healed with the first aid kit, that can be found in the scene. For ease of use, the user just has to touch the agent with the first aid kit to heal all of the conditions. Our Unity application needs a modern NVIDIA GPU with support for Riva and about $1GB$ of main memory during runtime.

3.2.1 Situation-Awareness

Enabling situation awareness for the agents means we need a way to store and access the state of all the components defined in the scene in real-time. We use a SQLite [SQL] database as it is used by several web browsers and mobile phones, serverless and designed to be light-weight and fast. At application startup, all the data from our defined objects (agents, conditions, locations, etc.) in the Unity scene are collected and written into the database. Every object also receives a unique id, that uniquely identifies it and can be used by Rasa to fetch, e.g. conditions for an agent. Figure 3.2 shows the Entity-Relationship diagram of our database. Storing the data inside a database greatly simplifies accessing scene specific information as, e.g. getting conditions for an agent is a simple *SELECT* SQL-statement instead of a scene or game object traversal. This also provides some flexibility in where the data is stored as the database can be on the same computer as the application or on different computer. Since all our services are running on the same computer for the experiment, we store the database file with the Unity application. The database is exposed to Rasa through a *HTTP* server that handles incoming requests and returns the result of the database query in *json* format.

3.3 NVIDIA Riva

NVIDIA Riva provides [GPU] accelerated services for speech recognition and speech synthesis. We deploy these services as a docker image and interface with them through [RPC] [RPC] calls. The configuration parameters are unchanged to the ones from the *riva_quickstart* package from Riva release *2.8.1*, as we found they provided sufficient performance for our case. Hence, for speech recognition we are using the *Conformer-CTC* [GQC+20] model with voice activity detection algorithms and for speech synthesis, we use the out-of-the-box *FastPitch* [RRT+19] English language female and male voice models. The riva service needs about $12GB$ of [GPU] memory during runtime. The average response time of the agents is around $32ms$. This time was measured from the start of the audio recording until the response audio playback is started.

3.3.1 Automatic Speech Recognition (ASR)

There are two modes for sending audio to the Riva [ASR] service, offline and streaming. In offline mode, the full audio signal is first captured from the microphone and the send to Riva for processing. Depending on the length of the full audio clip, this could introduce high latency, as the audio processing only begins after the full audio signal has been captured and received by Riva. In streaming mode, small audio chunks are send to Riva and processing starts as soon as the chunks arrive on the Riva side. We are using streaming mode and send audio chunks every frame after encoding the captured segment into a bytestream to achieve low-latency. Riva expects uncompressed *16-bit* signed little endian samples as a bytestream, e.g. we need to convert the float samples which we get from Unity to a bytestream. Riva will respond with a partial transcript of the received audio signal as soon as an intermediate transcript is available. The response from Riva

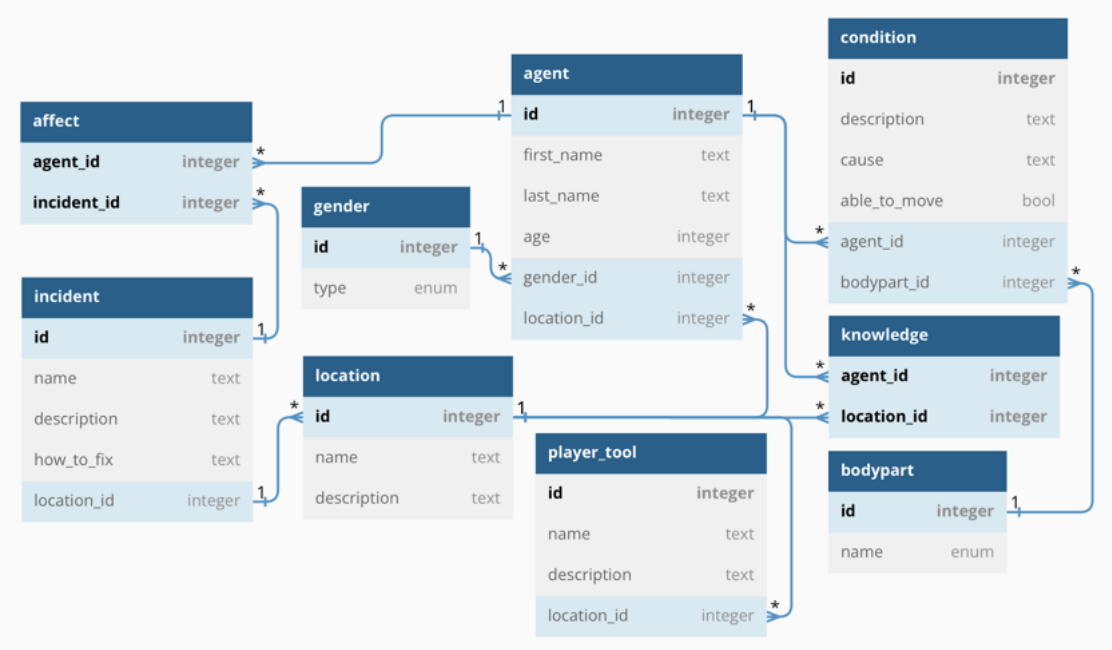


Figure 3.2: Entity-Relationship diagram for saving scene specific data in a structured way. Agents have a gender, can have a condition, knowledge about certain locations and be affected by incidents. Conditions are related to a bodypart and have a description, a cause and an indicator if the agent is able to move the bodypart with the condition. Incidents have a name, a description of what happened, a description on how to resolve the incident and must be in a location. Player tools have a name, a description for what the tool can be used and must be in a location. Locations have a name and a description of where the location is located.

contains the transcript and a flag, indicating if the transcript is final or not. As long as we are not receiving the final transcript, the whole transcript could change due to the model deciding to change words after receiving more audio signals from the spoken sentence. Since we need the full sentence to make intent recognition easier for Rasa, we are not using the partial transcripts and wait for the final one. This unfortunately adds a small amount of latency but avoids getting wrong answers from Rasa and conveying them to the user.

3.3.2 Text to Speech (TTS)

We send the full transcript of the response to the Riva TTS server deployed as a docker container. Similar to ASR, the audio data can either be received for the whole transcript once or in small chunks as it becomes available. While the streaming method would introduce less latency, we choose to wait for the whole transcript to be converted as latency was not an issue and it avoids having the agents potentially speak chopped of words or sentences. The TTS service returns the audio samples as a bytestream which

we need to convert back to float samples for usage in Unity's audio source. Similar to ASR, the returned audio data has a sample rate of *16Khz* and a bit depth of *16bit*, e.g. we need to pack *2 bytes* from the bytestream into one float sample until every byte of the bytestream was used. The resulting array of float samples can then be used as an input to the audio source of an agent. The audio source on each agent is also configured to support spatial audio, so that the user perceives the talking agents with changing volume on both ears, dependent on distance and rotation to the agent.

3.4 Rasa

Rasa provides tools for dialogue management, a Natural Language Understanding (NLU) part that detects intents and extracts entities from user utterances and a action server that provides actions that the NLU part can use to handle detected intents. We deploy all parts inside a single docker [Doc] container with open ports for the communication between Rasa and the Unity application. Our application will submit user utterances to Rasa as a *HTTP POST* request and Rasa will fetch scene specific information as a *HTTP GET* request. In this section, we give insights into how our training data, stories and actions are defined and how the pipeline for our model is configured. Our Rasa configuration needs about 1.1*GB* of main memory during runtime.

3.4.1 Training Data

The training data for the NLU pipeline consists of example utterances that a user would say to the agents. These example utterances are categorized by intents and also contain example entities that the model should extract from these utterances. Intents relate to which task a user wants the agent to perform and entities are pieces of information that are needed to accomplish this task. For example, if the user input to the model is the question "What happened in the factory?", we want the model to detect it as the intent *what_happened* with the location entity set to *factory*. We also include synonyms for certain entities in the training utterances so that the model learns to map multiple words to one common representation, e.g. we want "you", "your" and "yourself" to be always mapped to "self" for the subject entity because they are referring to the same meaning and that is we want to know something about the agent we are having a conversation with. Defining these synonym mappings makes it easier to know which data is requested for which entity. Our training data set contains 492 example sentences such as "What happened here?", "Where is the fire?" and "How can I help you?" for 27 intents with 7 entities that can be extracted.

3.4.2 Actions

After receiving a user utterance, the model will predict an action that will be executed. Actions can be simple responses which sends text to the user or custom actions that can run any python code. We are using simple responses as an answer to, e.g. "Hello", and use custom actions any time, scene specific information is requested such as conditions of

agents or incidents at locations. Requesting scene specific information is done through a *HTTP Get* request with parameters that define the type of the request to the webservice of the Unity application. The webservice will parse the sent parameters, query the database and return the request data in *json* format. Each custom action is then responsible for parsing the json data and building a response out it. For building the response we have defined 33 template response messages such as "Great, thank you for healing my bodypart", "I have description at my bodypart" and "My name is firstname lastName". Our custom actions replace the identifier inside the curly braces with the data received from the Unity application. We defined multiple template messages for each intent to have some variety in the answers. Rasa will randomly select one of the template messages during runtime so agents may, e.g. respond to the utterance "I have extinguished the fire" with "Great, thank you", "Good job" or "Thank you". In total, we have 17 custom actions that request and build responses for the agents.

3.4.3 Stories

Stories are example conversation representations between a user and an agent that are needed for training the dialogue management part of the model so it is able to either follow an example story or generalize to unseen conversation paths. Stories are a list of steps where user inputs are expressed as intents and the responses are expressed as actions. If an entity is extracted for an intent, it should also be listed in the story because the model learns to predict the next action based on the combination of the intent and the entities. In the stories we also need to make sure that we have converging paths dependent on entities that are set. For example, every story path where a subject entity is extracted during intent recognition we need to determine if the agent actually knows who is meant. Consider: "What happened to you?", *you* is recognized as a subject and mapped to *self*. Every agent can answer questions about itself, therefore the agent will not have to ask who is meant by *you*. Now consider "What happened to him?": If this is the first question a user asks an agent, the agent has no idea who the user is talking about, so a story which asked for the agent the user is talking about must exist. If there was already some conversation between the user and the agent and they already talked about another agent, therefore it is defined who is meant by *him*, another story that does not ask "Who are you talking about?" must exist. We defined 45 different stories with various conversation paths.

3.4.4 Pipeline

When Rasa tries to identify intent and entities in a user utterance, the data flows through a pipeline which is performing a sequence of operations. The basic steps of a pipeline are tokenization, featurization and training or inference. Tokenization extracts a list of words (tokens) from the utterance, e.g. "What happened here" is extracted to the tokens "What", "happened" and "here". Featurization transforms words into meaningful numbers that can be used by a training algorithm. Training is where the algorithm learns from the derived user utterance and inference is when the trained model is used to

make predictions for any user utterance. When a trained model is used, a user utterance follows the same path as during training, e.g. utterances are tokenized, features extracted and used to predict intent and entities. Figure 3.3 shows our pipeline configuration for Rasa which will now be described.

The *WhitespaceTokenizer* separates the words by white spaces and the *CountVectorsFeaturizer* is configured to recognize combinations of letters in a word so even if there are a few miss spellings in the text after ASR, the model can still predict the correct intent. The first *CountVectorFeaturizer* creates a bag-of-words representation of user utterances, intents and responses. The second looks at sub-word sequences of characters.

The *DIETClassifier* receives sparse features from the *CountVectorsFeaturizers*, where sparse means that all values in a vector are 0 except for one which is 1. During training, boundaries between groups of data points are learned so every utterance in the training data is represented in a vector space. The position is calculated from the features and the algorithm must learn where the separation between all groups of utterances (intents) is. The model should also be able to generalize, meaning it should apply what it learned from the training data to unseen user utterances. We choose 200 epochs since our training dataset is not that big and using more increased the time to train and created overfitting where it did perform worse on unseen data. *use_masked_transformer_layers* helps the classifier to obtain additional domain knowledge by adding context to embeddings which helps to differentiate subtle nuances between intents. *constrain_similarities* applies a sigmoid cross entropy loss over all similar terms which helps in better generalization of the model to user utterances.

During training, the *DIETClassifier* knows from the provided training data which tokens of the training utterances are entities. At inference time, it will go through all tokens of an utterance and evaluate if a token belongs to an entity. If there are two or more neighboring tokens that belong to the same entity, the token sequence is tagged as the entity. To evaluate if a token belongs to an entity, the algorithm looks at the features of the token being evaluated and the tokens before and after the current evaluated token: e.g. "What happened to your head" where head is a *bodypart* entity. If the current evaluated token is *head*, the token before is *your* and there is no token after. The model learns that every token following *your* has a likelihood of being a *bodypart* entity.

LexicalSyntacticFeaturizer creates features for entity extraction by moving a sliding window over every token of the user utterance. We use default configuration which defines features for the token before, the current token and the token after. These features are used to check if, i.e. a token is at the beginning or end of a sentence, if the token is lower or upper case or if the token is a title or contains just digits.

The *EntitySynonymMapper* will map values of detected entities to predefined synonyms if the mapping is defined in the training data, e.g. we have defined to map the subject entities *you*, *your*, *yourself* to *self* so that our custom actions only need to check for the word *self* in the subject entity when they want to know if they are processing information on the asked agent or someone else.

ResponseSelector follows a similar architecture as the *DIETClassifier* and is used to directly predict a response from a set of candidate responses to handle single-turn interactions better.

FallbackClassifier classifies an utterance with the intent *nlu_fallback* if the NLU was not able to classify an intent with a confidence greater or equal to the defined threshold of the classifier. We also define the *ambiguity_threshold* so that the *FallbackClassifier* will also predict *nlu_fallback* if the confidence score difference between the two highest ranked intents is smaller that the threshold of 0.1.

```
pipeline:
- name: WhitespaceTokenizer
- name: RegexFeaturizer
- name: LexicalSyntacticFeaturizer
- name: CountVectorsFeaturizer
- name: CountVectorsFeaturizer
  analyzer: char_wb
  min_ngram: 2
  max_ngram: 4
- name: DIETClassifier
  epochs: 200
  constrain_similarities: true
  use_masked_language_model: true
  number_of_transformer_layers: 4
- name: EntitySynonymMapper
- name: ResponseSelector
  epochs: 200
  constrain_similarities: true
- name: FallbackClassifier
  threshold: 0.5
  ambiguity_threshold: 0.1
```

Figure 3.3: Rasa pipeline

3.4.5 Policies

At each step in a conversation between the user and an agent, the conversational model will utilize policies to decide which actions will be chosen next. Figure 3.4 shows our policy configuration. Every turn, each defined policy will predict the next action with a certain confidence level. Policies have a default priority, in case two or more policies predict an action with equal confidence. In our case, *RulePolicy* has a higher priority than *MemoizationPolicy* which has a higher priority than *TEDPolicy*.

The *MemoizationPolicy* is able to remember stories from the training data and performs checks if a conversation matches one of them. If a story matches, it predicts the next action from the matching story. E.g. The user has asked agent A for the state of another agent and agent A asked who the user is talking about. If the user answers with the

name of an agent, the policy is able to predict the action that gets the state for that agent.

Transformer Embedding Dialogue Policy (*TEDPolicy*) predicts actions and recognizes entities. We configure the epochs to be 200 where one *epoch* is equal to one forward and on backward pass of all the training examples. We found, that more epochs did not improve performance and only made training slower and a lower number of epochs resulted in worse performance. With the parameter *max_history* we control the number of dialogue history used by the model to decide which action is chosen next. We limited the history to 5 as we do not have very long dialogues with each agent where the history is important for enough. A higher history count also did not improve action prediction since we defined short stories for every possible dialogue we could think of. Similar to the *DIETClassifier*, *constrain_similarities* helps with model generalization to real world user utterances.

RulePolicy handles parts of a conversation that follow a fixed pattern and makes predictions based on rules in the training data. We only define one rule (Figure 3.4) which is for the fallback if a user says unexpected messages, leading to unknown conversation parts. *TEDPolicy* is optimized to handle unknown paths but predicts the actions with low NLU confidence. If this confidence is lower than the *core_fallback_threshold*, a fallback action is predicted. We found that a threshold of 0.3 was suitable for our model. The fallback action is defined to return a default answer such as "I did not understand that, could you please rephrase?" to inform the user that an agent did not understand the utterance and prompting the user to try again with a rephrased sentence. Similar to the template messages in Subsection 3.4.2, we defined multiple of these messages to have some variety in the response from the agents.

```
policies:
- name: MemoizationPolicy
- name: TEDPolicy
  max_history: 5
  epochs: 200
  constrain_similarities: true
- name: RulePolicy
  constrain_similarities: true
  core_fallback_threshold: 0.3
  core_fallback_action_name: action_default_fallback
```

Figure 3.4: Rasa policies

3.4.6 Limitations

The training data consists mainly of sentences from the authors knowledge and research on how first responders would respond and ask questions which is a limited representation of what real first responders would ask. Since there is also a near unlimited variety in how different users would ask questions, the model will have some difficulty to predict the

right intent for some users and ask the user to rephrase their sentence. The responses to utterances are either pre-made or template sentences where Rasa just fills in the received information from Unity. We tried to give the agent a small variety of pre-made sentences to every utterance, e.g. the bot will randomly select between "Hey", "Hi", and "Hello" to an utterance of "Hello" or "Hi". We expect this to feel more natural when not every agent is responding with the exact same message. An example of a template response is, e.g. "I have description at my bodypart", where Rasa will fill in the variables in the curly braces with the information received from Unity. We tried to formulate the description in Unity so that it fits every template response that we have, but we do recognize that this is not the optimal solution for bigger scenarios. As a future work, we think connecting Rasa to some kind of Natural Language Generation (NLG) service to create a natural sounding answer with the given information would enhance immersion.

CHAPTER 4

Training Scenario

The purpose of the study is to explore training and task solving guidance in Virtual Reality (VR) through virtual agents with different interaction methods. Throughout the study the user will wear a Head Mounted Display (HMD) and traverse a virtual world using teleportation with controllers and solve several tasks. We want to compare our speech enabled agent interaction method with a static spoken text interaction, where the agent just tells the user what is happening in the environment. In this chapter we describe the virtual environment that a user will traverse during the study, the mechanics, interactions and the tasks to solve. We are also explaining how the two different agent interaction modes work.

4.1 Virtual Environment

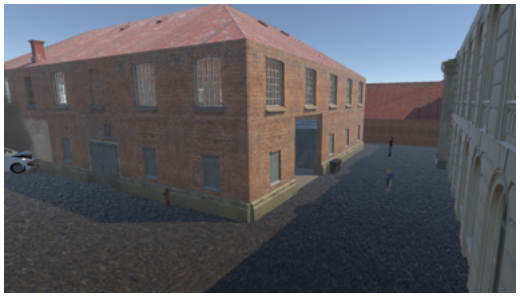
The virtual environment was created by the authors in the Unity game engine, using assets from the Unity Asset Store and the Quixel megascans [Qui] library. Our requirement for the scene was, that the user cannot immediately see all the agents and incidents, but has some space to navigate and explore using the agents as a travel guide, e.g. since the agents are aware of their location and surroundings, they know the answer to questions such as "Where is the first aid kit?", "What can I do about the fire?" and "How can I help you?". Figure 4.1 shows an overview of our created environment. Agents are marked with a person symbol, which is purple if they are injured or green if they are healthy. Incidents are marked with a red symbol. There is a locked car door, chemicals, fire and a locked room door. The yellow keys indicate locked doors to an empty room. The blue symbols mark the tools first aid kit, tongs, sand bag, fire extinguisher and axe, which are needed to resolve the incidents and heal the agents. The user is represented by the orange glasses at the top and starts at the beginning of a one-way street, surrounded by large buildings. To prevent the user from leaving the environment, the street is blocked by an ambulance at the top. The building on the left side, where the roof is visible in the

top-down view, is not accessible to the user. The big building in the middle represents a factory and is accessible through an open entrance at the bottom where the two agents are located. The size of the virtual environment is $62x54$ meters with a walkable area of approximately $1688m^2$.

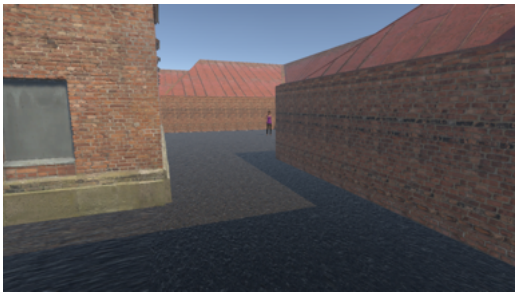
Figure 4.1: An overview of the virtual environment with symbols marking starting position, tools, incidents and agents. The user starts at the position of the orange glasses symbol at the top. Agents are marked with a person symbol, which is purple if they are injured or green if they are healthy. Incidents are marked with a red symbol. There is a locked car door, chemicals, fire and a locked room door. The yellow keys indicate locked doors to an empty room. The blue symbols mark the tools first aid kit, tongs, sand bag, fire extinguisher and axe, which are needed to resolve the incidents and heal the agents.

Figure 4.2 shows different views from the street. Figure 4.2a shows the view from the top left corner with the user starting location and the ambulance and the first aid kit on the left and the street down to the factory entrance. On the street are two agents and a

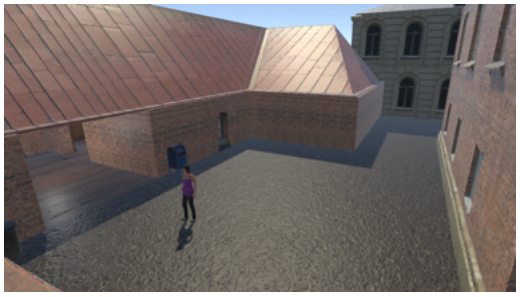
(a) View of the street down to the factory. On the left is an ambulance and on the right are two agents and a car cash.



(b) View of the factory entrance with the car crash to the left and the street to the hotel on the right.



(c) View of the street leading to the hotel on the right. To the left is the factory building.



(d) View of the area in front of the hotel with the entrance on the left and an agent.

Figure 4.2: Figure 4.2a shows two agents, a car crash and the street to the factory entrance. Figure 4.2b shows the car crash on the left and the factory building with the entrance in the middle. Figure 4.2c shows the street going along to the hotel. Figure 4.2d shows the area in front of the hotel, reachable through the street on the right.

smoking car that crashed into the factory. Inside of the car is an injured agent and near the car door are the tongs to cut open the car doors.

Figure 4.2b shows the view from the bottom left corner towards the factory in Figure 4.1. On the left is the car and near the right side are two agents with the factory entrance. The user needs to walk around a few corners to see what is actually happening inside the factory. The factory contains one injured agent and chemical waste from the barrels that fell due to the car crash.

Figure 4.3 shows a view from a corner inside the factory with the agent, the chemicals and the sand bag to resolve the chemical issue. Figure 4.2c shows the view from the last corner of the street leading towards a small area with an injured agent. Figure 4.2d shows the small area with the agent from the opposite side. On the left side of Figure 4.2d is the entrance to the second accessible building, a hotel. Figure 4.4 shows view from inside of the hotel. Walking from the entrance into the hotel, there is a room on the left which contains a fire hazard and the fire extinguisher to extinguish the flames, see Figure 4.4a. On the right side of Figure 4.4a is the axe to destroy wooden doors and a entrance into a

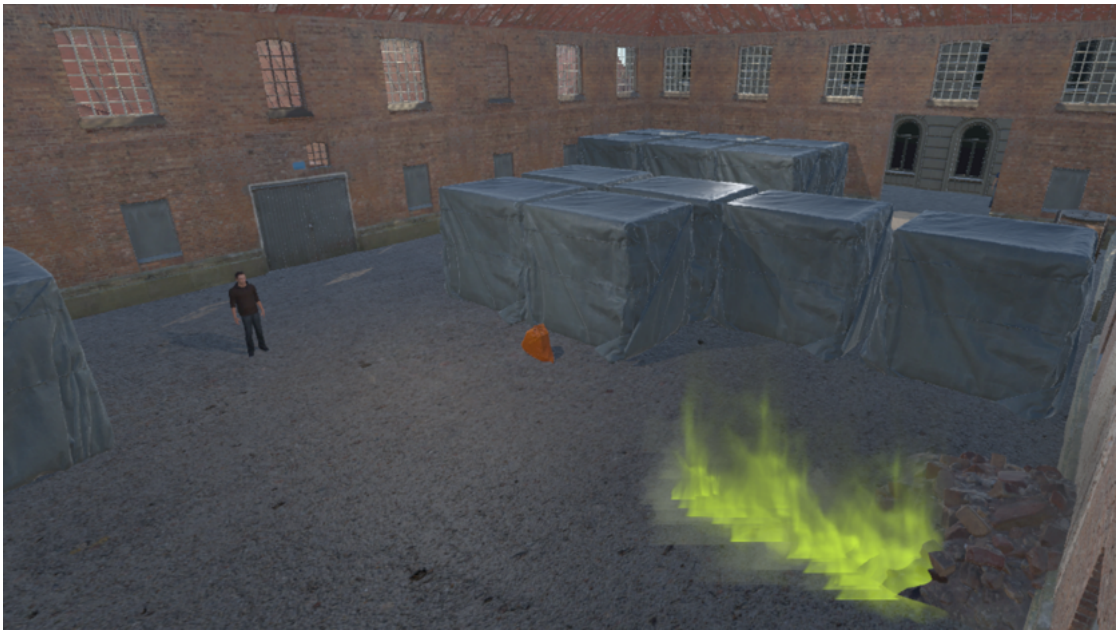


Figure 4.3: View inside the factory, showing an agent, the sand bag and chemical waste on the ground.

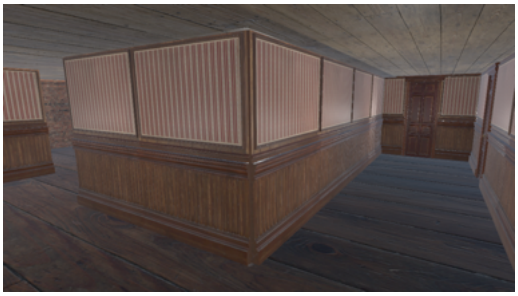
small maze with 5 closed doors, where the user has to find the room where an injured agent is locked away. Figure 4.4b shows the beginning of the maze, Figure 4.4c, the corridor leading to the locked door with the agent behind and Figure 4.4d an empty room. We choose to put the tools needed to resolve an incident near the actual incident, so users do not need to search for them. Consider a real rescue scenario, where first responders probably have the tools needed with them.

4.2 Locomotion

Locomotion is an important part of a VR application and involves multiple considerations such as realistic movement and motion sickness. Since we have access to the Cyberith Virtualizer [vir], a device where the user is strapped into a machine that allows movement in all directions by sliding the feet over sensors in the ground plate of the device, we considered using it for traversing our scene. Due to real feet movement it has a more realistic walking feel than, e.g. teleportation and users are not as prone to motion sickness. However, after having a few test runs with people and getting early feedback, we found that people who are not familiar with the device have difficulties using it. This is due to the need of leaning into the direction they want to go which quickly becomes exhausting. We want to measure the time it takes a user to complete the rescue scenario and compare these times between the two groups with different agent interaction methods. With the Virtualizer we are measuring how familiar the user is with the Virtualizer instead of how long it takes to complete the scenario. Therefore, we decided to use the classic

(a) View from the hotel entrance on an axe and the community room with a fire extinguisher and a fire incident.

(b) View from the entrance of the hotel to the corridor inside the hotel leading to the rooms.



(c) View from the corridor inside of the hotel. The exit is to the left.

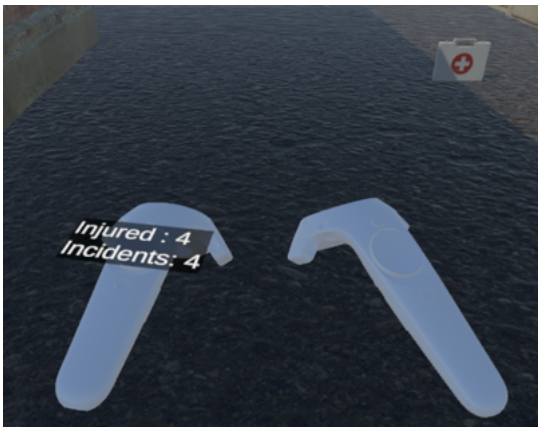
(d) View from a room inside of the hotel with a table and a chair towards the corridor.

Figure 4.4: Various views from the inside of the hotel. Figure 4.4a shows a community room with a fire incident. Figure 4.4b and Figure 4.4c show a corridor leading to the rooms of the hotel and Figure 4.4d shows one of the rooms.

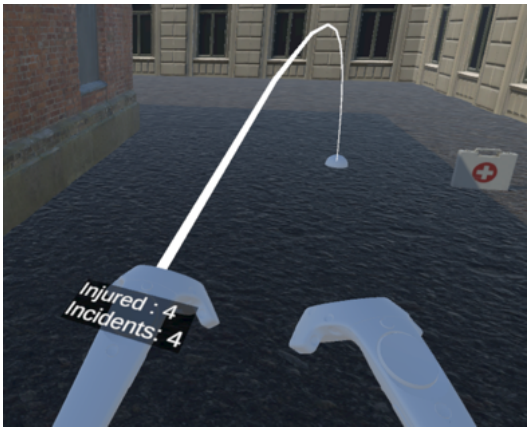
teleportation method, since it is easy to learn and also not as prone to motion sickness. The teleportation is initiated by pressing down the *Trigger Button* on the back side of the left hand controller. While pressing the button, a beam shoots out of the controller in an arc and ends in a ball, see Figure 4.5b. Dependent on the area or surface, the ball on the end of the arc is hitting, the beam will either be red or white. A red beam means the desired area is not accessible, therefore teleportation to that area is not allowed. A white beam means, teleportation to that area is possible by releasing the trigger button. In our virtual environment, movement is only possible on flat surfaces on the ground, hence no teleportation onto objects is possible.

4.3 Object Interaction

The virtual environment contains several objects that the user can interact with by moving the right hand controller near them. To pick them up, the user must press and hold one or both of the *Side Buttons* of the controller. 5 of these objects are tools necessary to complete all of the tasks. The tools are the following:



(a) The left hand controller shows the number of injured people and incidents left to solve at every time.



(b) Pressing the trigger button of the controller on the left hand side controller triggers a beam which shows the teleportation destination as a ball.

Figure 4.5: Figure 4.5a shows the left hand and right hand controllers of the user. The left hand controller shows a task counter and is used for teleportation. Figure 4.5b shows a ray out of the left hand controller, indication the teleportation destination.

1. First-aid kit
2. Sand bag
3. Axe
4. Tongs
5. Fire extinguisher

There are two ways to resolve an issue with a tool, by touching the issue and by pointing a special effect from the tool onto the issue, see Figure 4.6. When picking up the *fire extinguisher* or the *sand bag*, a particle effect will start and the corresponding issue can be solved by pointing the particle effect onto the issue, i.e. point the water jet from the fire extinguisher onto a fire, see Figure 4.6b. The particle effect will also disappear when the user let go of the tool. When picking up the *axe*, *tongs* or *first-aid kit*, the objects interactable with the tool will get an orange outline. To solve issues with these tools, the outlined objects must be touched with the tool until the outline disappears. Figure 4.6a shows an injured agent with an orange outline while the user is holding the first aid kit. The agent is outlined as long as the agent is injured or the user is holding the first aid kit.

(a) When the user holds the first aid kit, injured agents get a orange outline, indicating that they are healable by the first aid kit.

(b) The fire extinguisher shots out a hose of water as soon as it is picked up by the user. To extinguish a fire, the hose must be pointed at the fire.

Figure 4.6: Figure 4.6a shows a injured agent, indicated by the orange outline and the first aid kit in the users hand. Figure 4.6b shows the fire extinguisher shooting out a hose of water to extinguish a fire.

(a) When the user holds the tongs, the car doors get a orange outline, indicating that they are destroyable by the tongs.

(b) When the user holds the axe, the interactable doors get a orange outline, indicating that they are destroyable by the axe.

Figure 4.7: Figure 4.7a and Figure 4.7b show a orange outline on a door and the car doors when the tongs or the axe is in the users hand, indicating that the user can interact with the door and car doors.

Figure 4.8: Message shown when all of the incidents have been resolved and all of the agents are healed.

4.4 Tasks

Users are instructed to solve 8 tasks to successfully complete the rescue scenario. 4 of these tasks will be to heal some condition on injured agents and 4 will be incidents. The first incident visible to the user after the start will be a car crash, where an agent is trapped inside the car. The user will need to use the *tongs* to cut open the front left car door to help the agent out of the car. The agent inside the car also has some injuries from the crash, making him one of the agents where the *first aid kit* is required for healing. From the street it is also possible to see that the car is stuck inside the wall of the factory. While not visible from the outside, behind this wall is another incident to solve. The car impact, damaged barrels containing a chemical substance and the user needs to use the *sandbag* to pour sand over the chemicals to resolve the issue. Inside the factory is also another injured agent, that tried to resolve the chemical incident but got injured in doing so. The agent at the end of the one-way street is injured due to a fire inside the hotel, another incident to solve. To extinguish the fire, the user must use the *fire extinguisher*. Besides the fire, the hotel contains another incident, where an agent is locked away because an earthquake moved the door angles out of position. To get to the agent, the user must use the *axe* to destroy the door. The locked away agent is also injured and must be healed with the *first aid kit*. The controller in the left hand of the user will display an up to date number of tasks left to solve at all time. In case one gets solved, the number of tasks showing is also reduced. The task display can be seen in Figure 4.5a, showing injured agents and incidents as separate numbers. After solving every task, a text explaining that all tasks are completed (Figure 4.8) will appear in front of the users vision. This text marks the end of the VR-Experience and means, the user can put down the controllers and remove the HMD.

4.5 Agent Interaction

The default interaction with the agents is through a speech based conversation between the user and the agents. To interact with an agent, the user has to walk towards the agent. When the user gets in the range of about 3 meters to the agent, the agent will

(a) One of the male agents talking.

(b) One of the female agents talking.

Figure 4.9: Figure 4.9a shows one of the male agents and Figure 4.9b shows one of the female agents talking to the user. The avatars are from the Rocketbox [GFOP+20] library.

perform a turn animation to face the user. If the user is looking towards the agent withing this range, the agent will start to listen to the microphone of the user, i.e. a user-agent conversation is started. Figure 4.9 shows two agents facing and talking to the user. At this point, the user is free to look anywhere else and can have a natural conversation with the agent. The conversation is stopped if the user moves away, outside of the 3 meter range or looks at another agent who is also within 3 meters to the user. If the user looks at another agent within range, the conversation between the other agent and the user is started. During an agent-user conversation, the user is able to ask the agents any question that concerns the scenario, such as *What is your name?*, *What happened here?* or *How many people are injured?*. Figure 4.9a and Figure 4.9b show a male and a female agent facing the user and talking. We compare this to the second interaction mode, with the same interaction range and interaction switching as just mentioned but where the agent will not listen to the users microphone but just start to talk. In this mode the agent will tell the user everything that the agent has knowledge about. This will essentially be the same information that can also be obtained by asking the agent questions in the other mode but more limited as the user cannot only ask for specific details. After the agent finishes to tell the user all the information, the agent will repeat everything again as long as the user stays within range. If a task is solved, the agent will also stop mentioning the incident or injured people concerning that specific task.

CHAPTER 5

User Study

The user study was conducted using between-groups design in the Virtual Reality (VR)-Lab at TU Wien. In the room, we prepared a freely movable space of about *3x3 meters* and used the HTC Vive Pro as a Head Mounted Display (HMD) (Figure 5.1) for immersing our participants in VR. For convenience and comfort reasons, we used the integrated microphone and speakers of the HTC Vive Pro as the default audio and communication channel. The HMD was connected to a computer which was running our Unity application, as well as NVIDIA Riva and Rasa containers. Table 5.1 lists the whole hardware setup, including a more detailed specification of the computer that was used to run our Unity application, NVIDIA Riva and Rasa. Figure 5.2 shows the author of the thesis wearing the HMD and using the setup in the room we prepared for the user study.

Hardware	Type
CPU	Intel Core i9-11900K 3.50GHz
GPU	NVIDIA RTX 3090 12GB VRAM
RAM	32GB
HMD	HTC Vive Pro + Controller
Microphone	HTC Vive Pro
Speakers	HTC Vive Pro

Table 5.1: The table shows the hardware we were using for the study.

In the study we wanted to investigate how embodied conversational agents with situation awareness are perceived by trainees during a first responder training in VR. We choose a between-group design with two groups of the same size and have each group experience one of the two following conditions:

1) **Conversation**: The embodied agents in the training scenario are capable of having a speech conversation with the participants. This allows participants to obtain

Figure 5.1: The HTC Vive Pro with controllers used during the user study.

information regarding the environment, incidents and conditions of agents by asking questions in natural language. Obtaining this information was intended to help the participant in discovering and solving all of the open tasks.

2) **No-Conversation**: In this condition, the agents also had speech capabilities but were not able to have a natural conversation with the participants. When participants approaches an agent, e.g. walk within a certain range, the agents reveal all their available information by speech. This was a monologue from an agent towards the participant, e.g. the agent would not listen to questions or utterances of a participant.

Both of the conditions were experienced in the same virtual environment, e.g. the placement of the agents, tools incidents and buildings was as described in Chapter 4 for both conditions, allowing for a proper comparison. Compared to within-group design, having the exact same scenario avoids a potentially unfair comparison by creating two different scenarios where one of them could be easier to solve than the other. We decided to have each participant experience only one of the two conditions to avoid the learning factor where participants learn too much about the scenario and tasks, hence making them complete the second condition faster. To compare both conditions, we were measuring various metrics, see Section 5.3, to determine how our agents are perceived by participants. Our main hypotheses in the user study was:

H1: The Conversation condition will achieve overall better scores in the measured metrics compared to the No-Conversation condition.

5.1 Participants

The invitation to the study stated, that the purpose of the study is to explore first responder training and task solving guidance in Virtual Reality VR through virtual

(a) The author of the thesis from the side immersed VR.

(b) The author the thesis from the front immersed in VR.

Figure 5.2: Figure 5.2a and Figure 5.2b show the author of the thesis using the setup that was used during the user study.

agents. We did not inform participants, that there are two conditions as that could have an impact on the answers during the study. In total, 24 people freely volunteered for the study, knowing that they could quit the experiment at any time. The participants included 16 males and 8 females with an average age of 32.5 in the range of 24 to 65. Most of them had a technical university background or work at a university as a researcher or post-doc. Their VR experience ranged from "Never" to "Every Day", see Figure 5.3. We divided the participants into two groups, each containing 8 males and 4 females. One group was assigned to experience the Conversation condition and one was assigned to experience the No-Conversation condition.

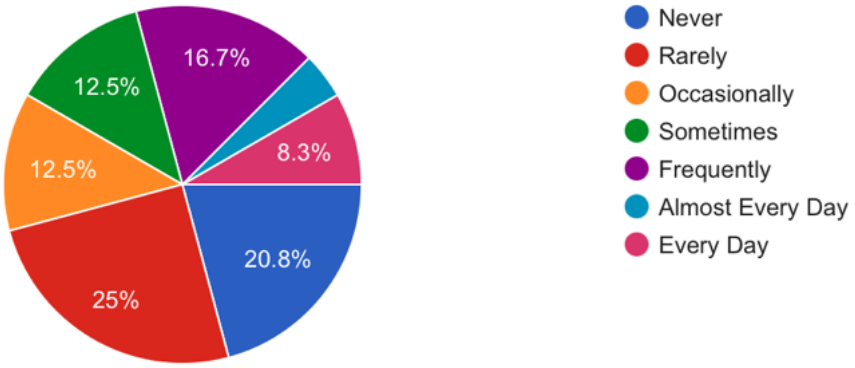


Figure 5.3: The participants VR experience ranged from Never to Every Day.

5.2 Procedure

In this section, we describe the procedure of the study for each participant. Before and after every user, we opened the windows of the room for a couple of minutes to fill the room with fresh air and cool it down to a comfortable temperature since it can get quite warm during the procedure with the computer running and the effort made in VR. In addition, this was to minimize health related risks due to bad air quality in the room. During the procedure the windows and door were closed to make it as quiet as possible for the user to not get distracted by outside noises. We also cleaned the HMD and the controllers with antibacterial tissues after every user. While the participants were using the VR setup, we took care that they do not bump into a something or fall over the cable of the HMD. The procedure was split into the following steps:

1. Study information
2. Consent form
3. Simulator sickness questionnaire
4. Introduction scenario
5. Main training scenario
6. Simulator sickness questionnaire
7. Study questionnaire

First, we have the participants read an information paper which explains what the study is about, what they will do, how the controls and the agent interaction work. We prepared one information sheet for each condition, where the only difference was the description of the agent interaction. After reading through the information sheet, we have the participant read and sign a consent form, informing them of health related issues with VR, data collection and stating that they can quit at any time. Before we let them put on the HMD we had them fill out a demographics questionnaire, asking for their age, gender, job and how often they have used VR. After that, participants filled out a Simulator Sickness Questionnaire (SSQ) that was from Kennedy *et al.* [KLBL93].

The first part of the VR experience was a small introduction scenario where a participant could get familiar with the mechanics used in the main scenario, teleportation, and with all of the tools that are necessary to complete the main scenario. The introduction scenario was just a small area containing the 5 tools, 1 agent with speech capabilities (either with speech input or just spoken text, depending on the condition), 1 locked room and a 1 fire. This was especially useful for participants, who have never or rarely used VR previously so they could get used to grabbing virtual objects and teleportation. During the introduction we let users ask any question regarding movement, interaction and what they can do.

When the participant indicated to be familiar with the mechanics we started the main rescue scenario. During this part, questions regarding what to do and how to solve certain tasks were not allowed as it was part of the study to get that information from the agents. We allowed questions regarding movement and object interaction but no one was asking such questions due to learning everything in the introduction scenario. Participants had unlimited time to complete the rescue scenario but were allowed to stop anytime for any reason. If a participant decided to abort the training early, i.e. before every task was completed, we counted the try as not completed. None of the participants decided to stop early and everyone was able to complete all the tasks and finish the training. The average time to complete all the tasks was 7 minutes and 3 seconds. The shortest completion time was 3 minutes and 47 seconds and the longest 14 minutes and 54 seconds.

After completing every task, users were informed through there HMD, that they can put down the controllers and remove the HMD, see Figure 4.8. The participant was asked to fill out the same SSQ as before the exposure to VR again before filling out the main study questionnaire.

5.3 Study Questionnaire

We designed our study questionnaire to investigate user perception of embodied conversational agents with situation awareness in a first responder training in VR. Our main interest was in the impact on the sense of presence, co-presence, perceived realism, learning outcome, subjective task performance, training duration and the quality how relevant information was presented. We measured the responses to the perception-oriented metrics using the post-experiment questionnaire in Table 5.2. All metrics, except *open questions*, were measured using a 7-point Likert scale [Lik32] ranging from 1 - "Strongly Disagree" to 7 - "Strongly Agree". The statements about presence were inspired by previous research from Vorderer *et al.* [VWG+04] and Witmer *et al.* [WS98]. Open questions were designed to give participants the freedom to express their own opinions and suggestions, hence they were answered in plain text. The training duration was measured automatically by our application from the start of the training until all of the tasks are solved.

In addition to the questions from Table 5.2, the questionnaire from the Conversation condition had three more questions regarding the conversational part of the agents, see Table 5.3. These questions were answered using the same 7-point Likert scale as the perception-oriented metrics. As our contribution are conversational agents with situation awareness, we were interested how participants felt towards talking to our agents.

Metric	Statement/Question
Presence	I could concentrate on the assigned tasks rather than focusing on the mechanisms used to perform the task. I felt like I was part of the virtual environment. I felt that I was physically present in the virtual environment. I felt that I actually took part in the rescue scenario. Even now, I could still find my way around in the virtual environment. I didn't really pay attention to the existence of errors or inconsistencies in the virtual environment.
Subjective task performance	I think my task solving performance was good.
Learning outcome	This training scenario helped me to learn about handling rescue situations.
Agents interaction	The agents gave me the feeling, that I could interact with them. The interaction with the agents was pleasant.
Information presentation	The agents were helpful and the information provided by them was useful. The information was presented in an understandable way.
Realism	The agents seemed realistic.
Co-presence	The agents gave me the impression, that someone else was in the scene.
Open questions	What did you like/dislike about the training? What is your opinion about the agents? How would you improve the training? What would have helped you improving your performance?

Table 5.2: Responses of participants to given metrics were measured after exposure to our VR training using this questionnaire. Open questions were answered in plain text and all the statements were answered using a 7-point Likert scale ranging from 1 - "Strongly Disagree" to 7 - "Strongly Agree". The presence statements were inspired by previous research from Vorderer *et al.* [VWG+04] and Witmer *et al.* [WS98].

Metric	Statement/Question
Conversational agents related questions	I would like to talk to the agents again. The communication with the agents felt natural. Talking to an agent felt like I was talking to a real person

Table 5.3: Additional questions on the questionnaire for the Conversation condition created by us to get feedback regarding our conversational agents. The questions were answered using a 7-point Likert scale ranging from 1 - "Strongly Disagree" to 7 - "Strongly Agree".

CHAPTER 6

Evaluation and Results

In this chapter, we present the results and evaluation methods of our conducted user study at TU Wien. All participants successfully completed the first responder training in Virtual Reality (VR) and were able to solve all of the eight tasks. First, we discuss the results between the two conditions (1) Conversation and (2) No-Conversation and the results between genders by investigating the statistical significance of differences between the compared groups using Mann-Whitney U test. Then we analyze the simulator sickness questionnaire between the pre- and post-experiment measurements using Wilcoxon signed-rank test and present our qualitative analysis of the open questions from the main study questionnaire (Table 5.2). Furthermore, we present the results between genders to the speech related questions we added to the questionnaire of the Conversation condition (Table 5.3). Since the additional questions were answered on a 7-point Likert scale, we also investigate statistical significance of differences between genders for these questions. We finish the chapter with a discussion of the user study results and provide guidelines based on our qualitative analysis for future training applications with embodied conversational agents with situation awareness in VR.

6.1 Comparison between Conditions

The main study questionnaire in Table 5.2 contains multiple questions that belong to the same metric. To measure metrics with multiple questions, we took the average over all the answers to the corresponding questions. The answers of the participants to the main study questionnaire can be seen in the box plots in Figure 6.1. The x-Axis shows our metrics with two box plots for the two conditions and the y-Axis shows the measured values using the Likert scale. The box plots show the data values of the mean value marked by a x symbol inside the Interquartile Range (IQR), which shows were 50% of the data points lie. The minimum and maximum values of the data points are at the end of the whiskers. For the Conversation condition, the data values for lower or first

quartile (Q1) and upper or second quartile (Q2) and the median are placed on the left side, while for the No-Conversation condition they are placed to the right side. Outliers are shown as circles. In Figure 6.1 we can observe that, the IQR is relatively small for

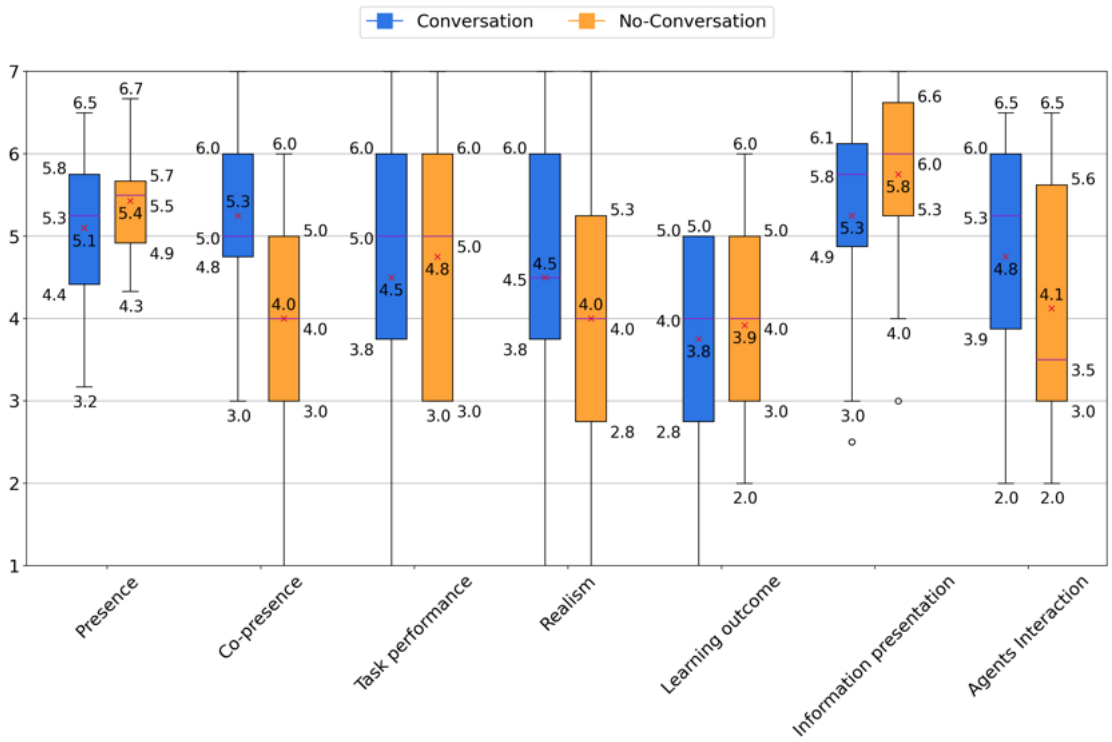


Figure 6.1: The results of subjective responses of participants to our study questionnaire. The x-Axis shows our metrics with two box plots for the two conditions and the y-Axis shows the values of the Likert scale. The mean value is marked by a x symbol and the median by a bar inside of the IQR. The values on the end of the whiskers show the minimum and maximum values of each plot and the numbers on the side of a box plot show the Q1, Q2 and median values. The box plots for the Conversation condition have its numbers on the left and the box plots for the No-Conversation condition on the right.

the metrics presence and information presentation for both conditions and co-presence for the Conversation condition. This means that most of the data points have similar values compared to the metrics task performance, realism, learning outcome and agents interaction for both conditions and co-presence for the No-Conversation condition where the range over which the values are spread is larger. We can also observe whiskers reaching over the whole value range of the 7-point Likert scale for the metric realism on both conditions and task performance on the Conversation condition, indicating a high standard deviation and variance for these metrics. We hypothesize the high variation for the realism metric results from different expectations that our participants had when starting the training. As can be seen in Chapter 5, the VR experience of our participants

was also spread between "Never" and "Almost Every Day", which could explain different expectations leading to this high variation. The median of the metric co-presence for the Conversation and the median of the metric agents interaction for the No-Conversation condition is relatively close to Q1, indicating a right- or positively-skewed distribution of the data points. The median for the metric presence for the No-Conversation and the median for the metric information presentation for the Conversation condition is relatively close to Q2, indicating a left- or negatively-skewed distribution of the data points. We can observe a balanced distribution for the No-Conversation condition in the metrics co-presence, realism and learning outcome. All of the other metrics are either slightly right- or left-skewed. If the mean value is greater than the median, most data points are small and few are very large compared to the smaller values. We can observe this for the No-Conversation condition in the metric agents interaction, indicating that most of the participants rated agents interaction lower. In contrast, if the median is greater than the mean value, most data points are larger and few are very small compared to the large values. This can be seen for the Conversation conditions in the metrics task performance, information presentation and agents interaction, indicating that these metrics were rated higher by participants. For the metric Co-presence, we observe the median line of the Conversation condition to be equal to Q2 of the the No-Conversation condition, indicating that there is likely to be a difference between both conditions.

We used Mann-Whitney U test to calculate significance differences between our two compared groups (two conditions) which can be seen in Table 6.1. To interpret significance, a α value of 0.05 is typically used, meaning there is a 5% likelihood of accepting a false-positive result. However, previous research [Ric89, Mil12, CM00] showed, that tests for statistical differences are often biased and wrongly reject a true null hypothesis, i.e. declaring a difference statistically significant when it actually is not. This likelihood of discovering a false-positive is called Type I error and rises the more comparisons are made during the calculation of significance differences with dependent variables. A common way to reduce the risk of Type I error, is to use a Bonferroni adjusted significance threshold by dividing α by the number of dependent variables, i.e. the number of comparisons made. In contrast, overzealous use of Bonferroni adjustments lead to an increase of Type II errors, i.e. accepting the null hypotheses even though it is actually false. Cabin *et al.* [CM00] found, there is no clear indication of when to use and when to not use Bonferroni correction, thus leaving the choice of considering Bonferroni correction and determining dependent groups to authors. Although our hypothesis was, that the Conversation condition will achieve higher values across all metrics, we consider all of our metrics as independent variables for the test, i.e. only one comparison per metric is performed: Conversation vs. No-Conversation. Hence, we do not see the need to Bonferroni adjust our α and therefore use the significance threshold $\alpha = 0.05$ which means we accept of having a Type I error rate of 5%, i.e. In 5% of the time we will accept a false-positive result.

Using this alpha we see co-presence as statistically significant, supporting what we observed in the box plot where the median of the Conversation condition was equal to Q2

of the No-Conversation condition. Hence participants of the Conversation group rated the co-presence of embodied conversational agents more favorable than participants of the No-Conversation group ($p - value = 0.035$). In addition to co-presence, the Conversation condition achieved higher scores in the metrics realism and agents interaction. In all the other metrics, including presence, subjective task performance, learning outcome and information presentation the No-Conversation condition achieved slightly higher scores. However, none of these differences were statistically significant.

The average completion time of the training was 7 minutes and 39 seconds with a Standard Deviation (SD) of 3 minutes and 28 seconds for the Conversation condition and 6 minutes and 28 seconds with a SD of 2 minutes and 10 seconds for the No-Conversation condition. This difference was not statistically significant as can be seen in Table 6.1.

Metric	U-value	p-value
Presence	60.5	0.52
Co-presence	36	**0.035**
Task performance	68.5	0.87
Realism	60	0.5
Learning outcome	70	0.94
Information presentation	56.5	0.38
Agents interaction	55	0.34
Duration	61	0.55

Table 6.1: Significance of differences between our two compared conditions, calculated by Mann-Whitney U test. Using an α value of 0.05 for interpreting the significance, co-presence can be interpreted as statistically significant, hence participants of the Conversation group rated co-presence more favorable than participants of the No-Conversation group ($p - value = 0.035$).

6.2 Comparison between Genders

Additionally to our main hypothesis, we were interested in differences across genders. The results for our metrics grouped by gender can be seen in Figure 6.2. Similar to the previous section, the x-Axis shows our metrics with two box plots for the two genders and the y-Axis shows the measured values using the Likert scale. We can observe a relatively small IQR and therefore a small amount of data spread for the metrics presence, task performance and information presentation, though information presentation has some outliers. The median is close to Q1 for females in the metrics task performance and realism, meaning the data distribution is positively-skewed. We can see a highly negatively-skewed distribution for females in the metrics co-presence where the median is even equal to Q2. For males, we can see data points with a balanced distribution in the metrics co-presence, realism, learning outcome and agents interaction. Females have a balanced distribution in the metrics learning outcome and agents interaction. The metric realism of the female group has a mean value greater than the median, indicating

that most data points are small, i.e. females rated realism rather lower. The metrics co-presence and information presentation for males and the metrics presence, co-presence and information presentation for females have a median greater value compared to the mean value, i.e. participants rated these metrics rather higher. As previously mentioned, the median for females in the metric co-presence is even equal to Q2. We can observe that for males, the median line for the metric task performance is significantly higher than Q2 of the same metric for females, i.e. there is a high likely-hood, that there is a significant difference for this metric between the two genders. Like in the previous section,

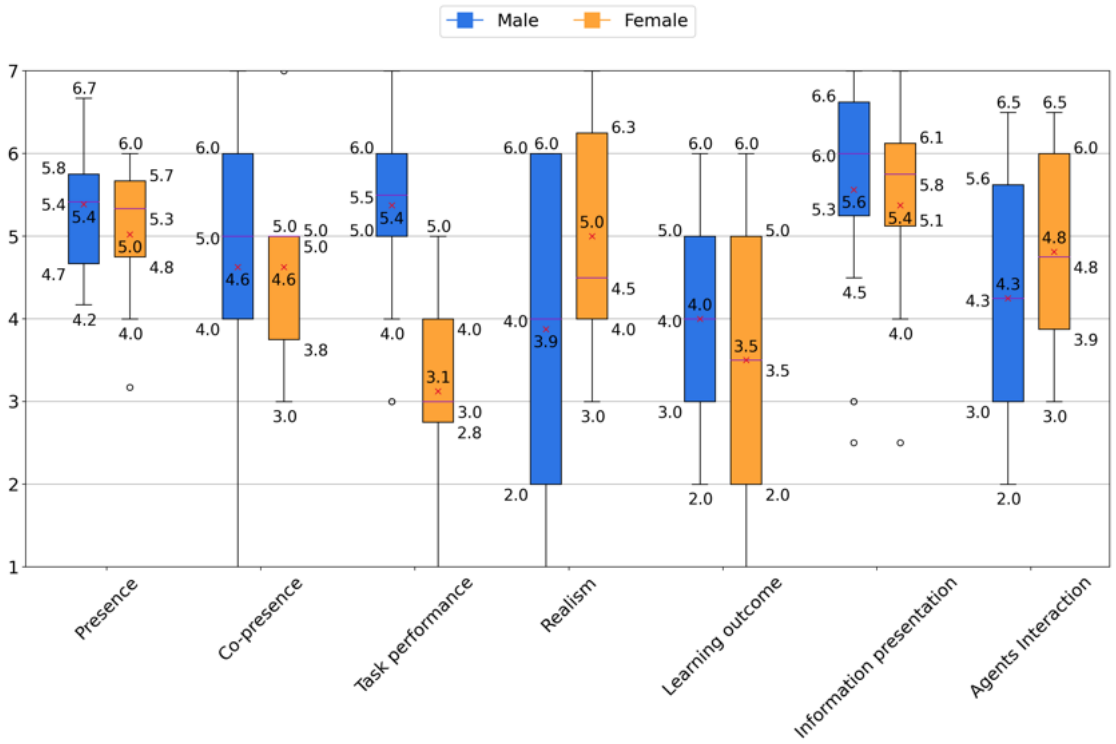


Figure 6.2: The results of subjective responses of participants to our study questionnaire between genders. The x-Axis shows our metrics with two box plots representing the two genders and the y-Axis shows the values of the Likert scale. The mean value is marked by a x symbol and the median by a bar inside of the IQR. The values on the end of the whiskers show the minimum and maximum values of each plot and the numbers on the side of a box plot show the Q1, Q2 and median values. The box plots for males have its numbers on the left and the box plots for females on the right.

we used Mann-Whitney U test to calculate the significance values that can be seen in Table 6.2. For a similar reason as in the last section, we used a significance threshold of $\alpha = 0.05$ since we are only having one comparison for each metric: male vs. female. Using this α we can see, that males rated subjective task performance significantly higher than females ($p - value = 0.001$), supporting what we saw in the box plot. Males also finished the training in a significantly lower duration than females ($p - value = 0.02$).

The other metrics did not show statistically significant differences.

The completion time was on average 6 minutes and 28 seconds with SD of 3 minutes and 6 seconds for males and 8 minutes and 15 seconds, with a SD of 2 minutes and 18 seconds for females. This difference was statistically significant, i.e. males were significantly faster in finishing the training than females ($p - value = 0.02$).

Metric	U-value	p-value
Presence	55.5	0.62
Co-presence	60.5	0.084
Task performance	13.5	**0.001**
Realism	42.5	0.19
Learning outcome	53.5	0.52
Information presentation	54.5	0.58
Agents interaction	47.5	0.32
Duration	27	**0.02**

Table 6.2: Significance differences between genders, calculated by Mann-Whitney U test. Using an α value of 0.05 for interpreting the significance, we see that males rated task performance significantly higher than women ($p - value = 0.084$). In addition, males were significantly faster in finishing the training ($p - value = 0.02$).

6.3 Simulator Sickness Questionnaire

We analyzed simulator sickness by using a Simulator Sickness Questionnaire (SSQ) from Kennedy *et al.* [KLBL93]. Each participant filled out the SSQ twice, once before and once after the exposure to our VR application. Our SSQ included all of the 16 items proposed by Kennedy *et al.* and we accumulated the items into the four factors proposed by the authors: Disorientation (D), Oculomotor (O), Nausea (N) and Total Severity (TS). A weight of 1 is assigned to each of the symptom variables. The total weighted score [1] for Disorientation was calculated by summing the weights for difficulty focusing, nausea, fullness of head, blurred vision, dizzy (eyes open), dizzy (eyes closed) and vertigo. The total weighted score [2] for Disorientation was calculated by summing the weights for general discomfort, fatigue, headache, eyestrain, difficulty focusing, difficulty concentrating, blurred vision. The total weighted score [3] for Disorientation was calculated by summing the weights for general discomfort, increased salivation, sweating, nausea, difficulty concentrating, stomach awareness, burping. The final scores D, O, N, TS are then obtained by using the formulas given in Equation 6.1.

$$\begin{aligned} D &= [1] * 13.92 \\ O &= [2] * 7.58 \\ N &= [3] * 9.54 \\ TS &= ([1] + [2] + [3]) * 3.74 \end{aligned} \tag{6.1}$$

The results of the accumulated scores can be seen in Figure 6.3 and Table 6.3. We can observe that there is not a lot of difference between the two exposure points. The median is zero or near zero for almost all of the factors and outliers after VR exposure are either the same or lower than before VR exposure. None of the box plots show signs of a significant difference. With the SSQ, we have the same measurement at two different time

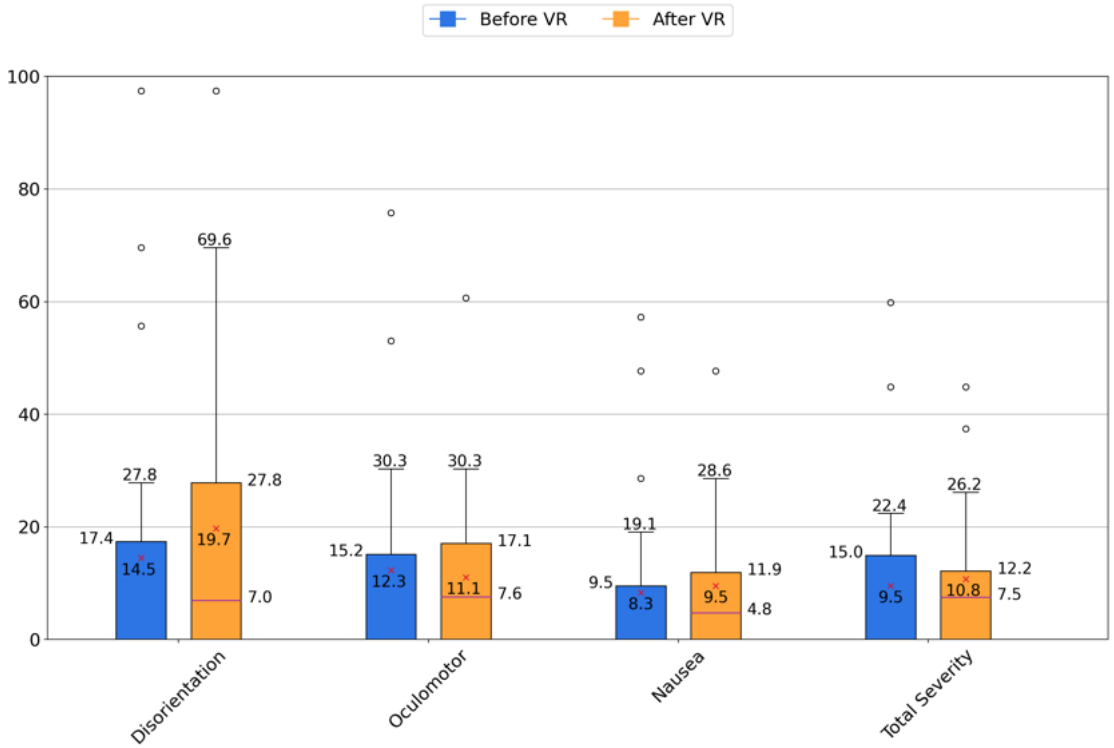


Figure 6.3: The results of the responses to the SSQ questionnaire from both conditions. The x-Axis shows the metrics according to Kennedy *et al.* and the y-Axis a scale between 0 and 100. The mean value is marked by a x symbol and the median by a bar inside of the IQR. The values on the end of the whiskers show the maximum values of each plot and the numbers on the side of a box plot show the Q2 and median values. Outliers are represented by black circles. The box plots for "Before VR" have its numbers on the left and the box plots for "After VR" on the right. We can see that the difference between the two observations is not significantly different in any of the four metrics.

points, hence, we used Wilcoxon signed-rank test to investigate statistical significance of the difference in SSQ scores between the two VR exposure points. The results of the test can be seen in Table 6.3. Using an α of 0.05 for interpreting the significance, we can see that none of the differences were statistically significant.

Metric	Before VR	After VR	Z	p
Disorientation	14.5	19.72	-1.35	0.18
Oculomotor	12.32	11.05	-0.05	0.96
Nausea	8.35	9.54	-0.68	0.5
Total Severity	9.51	10.75	-1.29	0.2

Table 6.3: The SSQ scores according to Kennedy *et al.* before and after exposure to VR with the Z and p values of the Wilcoxon signed-rank test. Using an α of 0.05 for interpreting the significance, we can see that none of the differences are statistically significant.

6.4 Qualitative Analysis

Our study questionnaire from Table 5.2 contained three open questions, which were answered by each participant in plain text. The intent behind these open questions was to give participants some freedom in expressing their feelings about the training scenario and the agents as well as give us valuable feedback to see further enhancements for future first responder training scenarios in VR. We focused the first of these questions on the training itself and asked participants what they liked and disliked about the training. The second question asked about the opinion a participant had towards our agents, giving us specifically opinions on both agents with communication capabilities and agents with predefined voice output only. The third question was designed to let participants state what they would change to improve the training and what they think would have improved their performance. For our qualitative analysis, we summarized all of the answers with similar meaning and extracted categories from these answers to get a list of positive and negative judgements from the participants. The results of this analysis can be seen in Table 6.4. Some of the categories are either very similar or the same for positive and negative effects, revealing the different expectations our participants had towards a first responder training with agents in VR when participating in the training. Comments with a positive effect highlight the benefits of agents and the favourably aspects of our training scenario and comments with negative effect underline the need for improvements in these categories when creating future first responder training scenarios with embodied agents.

6.5 Speech Related Analysis

In Addition to the difference between the two conditions, we were also interested in how participants of the Conversation condition felt towards talking to our agents. We therefore included three additional questions in the questionnaire, see Table 5.3. All three of these questions were also answered on a 7-point Likert scale. The results of these questions can be seen in Figure 6.4. The data distribution is negatively-skewed for males for the first question and normally distributed for females for second question. For males, the median to the first question is smaller than the mean value, indicating that

Effect	Category	Description
Positive	Realism	The agents and the environment were seen as realistic by most of the participants.
	Gamification	Participants liked the gamification aspect and saw it as an advantage.
	VR experience	The scenario on training and the user experience was enjoyable.
	Agent interaction	The speech interaction with agents was seen as useful and a more realistic approach to interact with virtual humans.
	Tools	Participants liked the usage of tools to solve incidents in the virtual environment.
	Helpfulness	Agents were seen as helpful due to providing information about the environment and situation related information needed to solve the tasks.
	Animation	Turn animations to the participant when getting close to an agent and lip animations while talking were rated positively.
Negative	Realism	The agents were perceived as unnatural, unresponsive and with insufficient conversation skills by some participants.
	Agent interaction	Limited and repeated answers to the same question as well as responses to every statement of the participant was seen as unnatural in a conversation.
	Locomotion	Teleportation felt unnatural and caused a loss of orientation.
	Tool Interaction	The interaction between tool and object felt unnatural due to the simple "touch to solve" mechanic.
	Animation	Agents were seen as stiff and unresponsive because they did not move around and could not react according to their situation or express there pain through body language.
	Tools	Participants did not like that they could only carry one tool with them.
	Emotions	Agents showed no emotions and injured agents did not express their feelings and pain with through a different tonal pitch in their voice.

Table 6.4: The results of our qualitative analysis to the three open questions of the main study questionnaire from Table 5.2. We summarized all of the answers with similar meaning and extracted categories to get a list of positive and negative judgements. Positive judgements highlight the benefits of agents and the favourable aspects of out training and negative judgements underline the need for improvement in these categories.

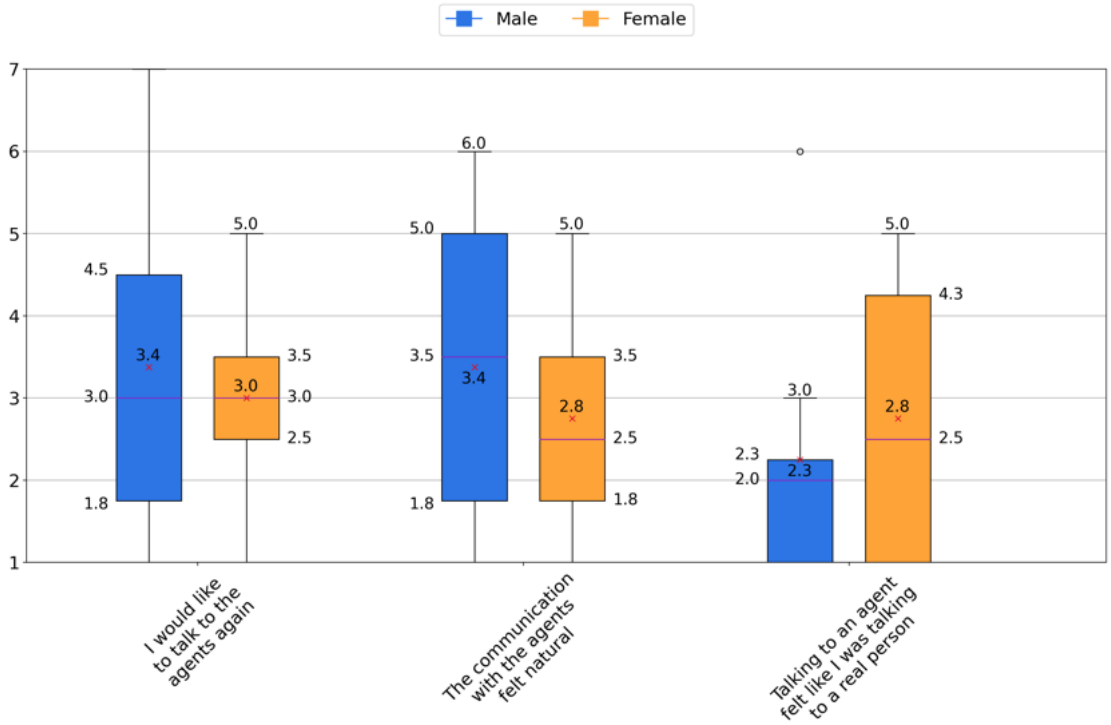


Figure 6.4: The results of subjective responses of participants to our study questionnaire between genders. The x-Axis shows our metrics with two box plots representing the two genders and the y-Axis shows the values of the Likert scale. The mean value is marked by a x symbol and the median by a bar inside of the IQR. The values on the end of the whiskers show the minimum and maximum values of each plot and the numbers on the side of a box plot show the Q1, Q2 and median values. The box plots for males have its numbers on the left and the box plots for females on the right.

the answers were rather lower than higher. We cannot observe a statistically significant difference between genders in the box plots. We also calculated significance differences between genders using Mann-Whitney U test, which can be seen in Table 6.5. Using an α of 0.05 without Bonferroni adjustment for interpreting the significance, as there is only one comparison: male vs. female, we can see that none of the differences were statistically significant.

6.6 Discussion

The main hypothesis, we defined in Chapter 5 was that, the Conversation condition achieves higher scores than the No-Conversation condition across all metrics. We can see in Table 6.1 that only co-presence was statistically significant, i.e. participants of the Conversation group rated co-presence higher than participants of the No-Conversation group. Therefore, our hypothesis was only partially supported by the results of our study.

Question	U-value	p-value
I would like to talk to the agents again	15	0.93
The communication with the agents felt natural	13.5	0.76
Talking to an agent felt like I was talking to a real person	15	0.95

Table 6.5: Significance differences between male and female for speech related questions from the questionnaire in Table 5.3 for the Conversation condition. The values were calculated using Mann-Whitney U test. Using an α value of 0.05 for interpreting the significance, we can see that none of the differences can be interpreted as statistically significant.

Additionally, some of the other metrics such as realism and agents interaction showed trends towards supporting our hypothesis. We hypothesize, that these metrics would have been more significant with a richer conversational model and more participants in the study. In contrast, some of the metrics achieved lower scores in the Conversation condition and were more in favor of the No-Conversation condition, namely general presence, task performance, learning outcome and information presentation. We can observe that in our experiment, the trends of general presence and co-presence do not correlate. Our participants might have felt less present due to the limited conversation model and behaviour of the agents, but perceived a higher co-presence due to the conversation enabled embodied agents. Similar to the metrics realism and agents interaction, we hypothesize that we could steer the trend of general presence towards supporting our hypothesis by improving the conversational model and the static behaviour of our agents. Interestingly the Conversation condition achieved slightly lower scores in the metric information presentation even though both conditions used speech to convey information to the participants. We hypothesize, that this is due to participants of the Conversation conditions receiving the exact same answer again after asking the same question. Participants might have only asked the question again because they did not clearly understand what the agent meant during the first time and expected a different and maybe more clear answer. We can also see, that the training duration was on average longer for the Conversation condition. This does not necessarily mean, that participants of the Conversation condition were slower in performing the task of the training but spend more time talking to the agents since they could have a real conversation with the agents compared to the No-Conversation condition where the interaction between participant and agent was more like a monologue of the agent.
In addition to the differences between the Conversation and No-Conversation conditions, we were also interested in differences between genders across the same metrics. In Table 6.2, we can see that females rated subjective task performance significantly lower than male participants. Furthermore, the training duration of female participants was significantly longer compared to male participants. The significance in these two metrics might be related and females rated their task performance significantly lower due to thinking that it took too much time to complete the training. These findings complement previous research about gender influence on various metrics such as the

sense of presence [FKH+14] and task performance [NG22]. While females rated Realism on average a bit higher than males, none of the other metrics were statistically significant or showed favorable trends towards significance.
We investigated simulator sickness in our VR training application by using an existing Simulator Sickness Questionnaire (SSQ) [KLBL93] before and after the VR experiment and observed 5.22 as the highest increase in simulator sickness. This value can be categorized as minimal increase [BWK20]. Furthermore, we analyzed statistical differences between the pre-experiment and post-experiment answers and did not found any significant differences in SSQ scores. I.e. Our VR training scenario did not have a negative impact on the participants.
We added three additional questions to the main study questionnaire of the Conversation condition to receive additional feedback from participants to our conversational agents. Since only one of the two conditions contained these questions, we only investigated differences between genders. Figure 6.4 indicates that on average, participants of both genders rated all three of the statements rather low. Interestingly, males were more in favor to talking to the agent again and that the communication felt natural but did not thought that talking to an agent felt like talking to a real person. Based on our qualitative analysis in Table 6.4, we hypothesize that this is mainly due to the agents stiffness, limited answers and lack of emotions. In Table 6.5 we can see, that none of differences between genders were statistically significant, indicating that neither males nor females are more in favor of talking to our agents again, felt that the conversation was natural and similar to talking with a real person.

6.7 Guidelines for VR trainings with embodied conversational agents

In Table 6.4 we summarized similar answers to the open questions on our main study questionnaire into categories to get a list of positive and negative judgements. Based on this qualitative analysis, we provide the following guidelines for future research and development of VR training applications which include embodied conversational agents with situation awareness:

Realism is important: While participants rated the realism of the training scenario positively, they also highlighted the need for some improvements in this aspect. Realism does not only concern visual realism but also emotions, decisions, conversations and situation dependent animations. A believable and more realistic training simulation through agents that show emotions through their voice and body language is desired by participants. E.g. It is expected, that agents do not stay in place and speak calmly when they are involved in a car accident or other disasters. Agents are expected to express their emotions, e.g. show panic or happiness, through tonal voice changes and moving their body accordingly.

Richness of conversation is required: The conversation ability of the agents was seen as helpful, but participants were sensitive to noticing conversational mistakes and

unexpected conversational behaviours. They disliked, that an agent responded to the same question with the same answer every time and wanted to hear some variety in the answers. Participants also disliked when an agent responded to every statement that they said out loud. Agents should know when they are addressed and realize when a participant is just saying thoughts out loud. Therefore, the richness of the conversational model as well as smart decision making of when an agent is actually addressed and should speak is desired.

Situation awareness is beneficial: The agents were seen as helpful due to providing information about the environment and situation related information needed to solve the tasks. While our agents responded to the same question with the same answer, which was seen as unnatural, they always provided situation or location related information that helped participants in solving all of the tasks and complete the training. The provided information was updated as the participants progressed through the training scenario, i.e. agents knew when something changed around them. We see a great benefit in further developing situation awareness to provide even more detailed information.

Autonomy of agents is expected: Agents were expected to be autonomous, active and behave believable according to events in their environment. While participants liked that our agents turn towards them when they get close, they expected more animations and locomotion to reflect their current situation. E.g. An agent with a broken ankle is not expected to stand straight but to sit or lie on the floor and maybe holding their hands around their injury.

Natural user locomotion and object interaction is desired: While our teleportation locomotion system did not lead to simulator sickness on any of the participants and allowed for the training to be completed in a smaller room than the actual walkable area in VR, participants would have preferred a real walking locomotion. Therefore, mapping the virtual walkable area onto a real area of equal size is beneficial for future experiments. According to some participants, real walking and turning would have also avoided some disorientation in the virtual environment caused by teleportation. In addition, participants disliked the simplicity of grabbing objects by pressing a button and solving incidents by touching. A more natural way, i.e. grabbing objects with bare hands and having a real interaction between tool and incident to solve it, is desired. Participants also wanted the capability of carrying multiple items at once, e.g. carrying the first aid kit and the axe, like it is possible in the real world.

Gamification is advantageous: The gamification aspect of our training scenario was seen as advantageous for training in VR. The sense of achieving something and progressing could be a motivational aspect in future training applications or serious games for training purposes. This may not only increase the motivation but also the learning outcome. Previous work by Palmas *et al.* [PLPK19] and Ulmer *et al.* [UBC+22] already showed the positive impact of gamification in VR trainings.

CHAPTER 7

Conclusion and Future Work

In this chapter, we are first reflecting on the limitations of our conversational agents and user study. Based on our discoveries during the user study and our qualitative analysis of user feedback, we propose solutions to overcome these limitations in future research and experiments. Finally, we conclude this thesis with a summary of our work and discoveries during our user study.

7.1 Limitations and Future Work

While our study participants considered conversational agents as helpful, there is room for improvements in future experiments. Participants criticised that the agents responded with a limited set of predefined answers with injected entities and they also disliked, that the answer to the same question is always the same. The model used in our training was only trained on the authors knowledge about some first responder trainings. Therefore, we see a high potential in improving the richness of the conversation by using larger language models or Natural Language Generation (NLG) to generate more natural answers with a greater variety of words for conveying the requested information. Dependent on the gender, the agents used one of the two default language models provided by NVIDIA Riva [Rivb] with a static voice tone. To have more realism when listening to an agent, different voice tones dependent on the current emotions of the agent could be used. Additionally, intelligent decision-making of when to respond or not is necessary to avoid unwanted responses from the agent to i.e. thoughts that a trainee is saying out loud.

Our agents had mostly static spatial behaviour with on-place body animations. Participants liked that the agents turned towards them on approach and moved their lips according to the spoken content but a more realistic scenario would require advanced animations and spatial movement capabilities. In future training applications agents could move around, follow directions if asked by the first responder, e.g. to walk to a save place. When asked what happened to them, agents could point to the injured bodypart,

e.g. to a sprained ankle or wound. Additionally, agents could also point to locations when asked where something is.

The results of our user study showed interesting trends in the metrics Realism and Agents interaction, but these results were not statistically significant. Our study had a limited number of participants and more participants may be needed in future experiments to increase the statistical power of the experiment. Most of our participants had a technical or academic background which may have set certain expectations towards our agents. A few participants tried asking our agents answers outside of the first responder domain because they expected some answer to every question similar to chatbots such as ChatGPT [Cha]. Since our main goal of the user study was to explore the benefit of embodied conversational agents with situation awareness in a first responder scenario in VR, we did not evaluate the learning outcome on defined pedagogic objectives. Therefore, in addition to more participants with different backgrounds, we think future studies with a well defined pedagogic scenario and experts in rescuing people such as real first responders are needed to validate the learning outcome of VR trainings compared to real trainings.

7.2 Conclusion

In this thesis, we presented a methodology for first responder training in VR using embodied conversational agents. We implemented and demonstrated a novel solution to enabling situation awareness for embodied agents with AI and speech capabilities in VR. In addition, we conducted a user study to investigate the impact of conversational abilities of embodied agents by comparing the two conditions Conversation and No-Conversation. During the user study, we investigated various metrics such as presence, co-presence, task performance, realism, learning outcome, information presentation, agents interaction and training duration. The only difference between these conditions was that the Conversation condition had conversation enabled agents and the No-Conversation had monologue-only agents. Our results suggest that co-presence is rated significantly higher for conversational agent interaction compared to monologue-only interaction. In addition, we discovered a significant difference in subjectively reported task performance and duration between genders. After the experiment, we offered participants from the No-Conversation group to also try out the Conversation condition if they are interested. Since the training is the same for both conditions, we did not record and asses any data but only asked them how they feel about the conversational agents compared to the monologue-only agents. Most of these participants had a positive reaction and preferred the conversational part as it felt more realistic and natural. Based on our qualitative analysis on the answers from participants to open questions, we provided guidelines for future research and development of training applications in VR with embodied conversational agents with situation awareness.

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence. ix, xi, 2–4, 13, 20–22, 28

API Application Programming Interface. 22, 26

AR Augmented Reality. 7, 12, 14–16, 19

ASR Automatic Speech Recognition. 4, 13, 14, 22, 23, 25–27, 29–31, 33

CBRN Chemical, Biological, Radiological, Nuclear. 2, 10, 76

D Disorientation. 60

GPU Graphics Processing Unit. 22, 26, 28, 29

HMD Head Mounted Display. 8, 11, 14, 25, 37, 44, 47, 50, 51, 71

IQR Interquartile Range. 55, 56, 58, 59, 61, 64

N Nausea. 60

NLG Natural Language Generation. 36, 69

NLP Natural Language Processing. 4, 13, 14, 25

NLU Natural Language Understanding. 21, 31, 34, 35

NN Neural Network. 21

NPC Non-player Character. 1, 3, 10, 11

O Oculomotor. 60

RPC Remote Procedure Call. 22, 23, 26, 29

SAR Search and Rescue. 10

SD Standard Deviation. 58, 60

SDK Software Development Kit. 22

SSQ Simulator Sickness Questionnaire. 50, 51, 60–62, 66, 71, 73

TS Total Severity. 60

TTS Text to Speech. 4, 13, 14, 17, 22, 23, 25–28, 30

VERTIgO Virtual Enhanced Reality for interoperable training of CBRN military and civilian Operators. 2, 3

VR Virtual Reality. ix, xi, 1–5, 7–14, 17, 25, 28, 37, 40, 44, 47–52, 55, 56, 60–63, 66, 67, 70, 71

Bibliography

[ALK+] Sultan A. Alharthi, Nick LaLone, Ahmed S. Khalaf, Ruth Torres, Lennart Nacke, Igor Dolgov, and Zachary O. Toups. Practical insights into the design of future disaster response training simulations. *Proceedings of the ... International ISCRAM Conference*.

[BC05] Timothy Bickmore and Justine Cassell. *Social Dialogue with Embodied Conversational Agents*. 01 2005.

[BFPN17] Tom Bocklisch, Joe Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *ArXiv*, abs/1712.05181, 2017.

[BGGN98] O. Bersot, P. O. Guedj, C. Godéreaux, and P. Nugues. A conversational agent to help navigation and collaboration in virtual worlds. *Virtual Real.*, 3(1):71–82, mar 1998.

[BvdBR+22] Walter Baccinelli, Sven van der Burg, Robin Richardson, Djura Smits, Cunliang Geng, Lars Ridder, Bouke Scheltinga, Nele Albers, Willem-Paul Brinkman, Eline Meijer, and Jasper Reenalda. Reusable virtual coach for smoking cessation and physical activity coaching. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, IVA '22, New York, NY, USA, 2022. Association for Computing Machinery.

[BWK20] Pauline Bimberg, Tim Weissker, and Alexander Kulik. On the usage of the simulator sickness questionnaire for virtual reality research. pages 464–467, 03 2020.

[Cha] OpenAI ChatGPT. `https://openai.com/chatgpt`. Accessed: 2023-09-27.

[CM00] Robert Cabin and Randall Mitchell. To bonferroni or not to bonferroni: When and how are the questions. *Bulletin of the Ecological Society of America*, 81:246–248, 01 2000.

[CW19] Girija Chetty and Matthew White. Embodied conversational agents and interactive virtual humans for training simulators. pages 73–77, 08 2019.

[Doc] Docker. https://www.docker.com. Accessed: 2023-09-27.

[DWW+22] Shahin Doroudian, Zekun Wu, Weichao Wang, Alexia Galati, and Aidong Lu. A study of real-time information on user behaviors during search and rescue (sar) training of firefighters. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 387–394, 2022.

[End95] Mica Endsley. Measurement of situation awareness in dynamic systems. *Human Factors*, 37:65–, 03 1995.

[FKH+14] Anna Felnhofer, Oswald D. Kothgassner, Nathalie Hauk, Leon Beutl, Helmut Hlavacs, and Ilse Kryspin-Exner. Physical and social presence in collaborative virtual environments: Exploring age and gender differences with respect to empathy. *Computers in Human Behavior*, 31:272–279, 2014.

[GFOP+20] Mar Gonzalez-Franco, Eyal Ofek, Ye Pan, Angus Antley, Anthony Steed, Bernhard Spanlang, Antonella Maselli, Domna Banakou, Nuria Pelechano, Sergio Orts-Escolano, Veronica Orvalho, Laura Trutoiu, Markus Wojcik, Maria V. Sanchez-Vives, Jeremy Bailenson, Mel Slater, and Jaron Lanier. The rocketbox library and the utility of freely available rigged avatars. *Frontiers in Virtual Reality*, 1, 2020.

[GQC+20] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition, 2020.

[GSMC19] David Griol, Araceli Sanchis, José Manuel Molina, and Zoraida Callejas. Developing enhanced conversational agents for social virtual worlds. *Neurocomputing*, 354:27–40, 2019. Recent Advancements in Hybrid Artificial Intelligence Systems.

[HFR+19] Arno Hartholt, Edward Fast, Adam Reilly, Wendy R. Whitcup, Matt Liewer, and Sharon Mozgai. Ubiquitous virtual humans: A multi-platform framework for embodied ai agents in xr. *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 308–3084, 2019.

[HLB+13] Edbert B. Hsu, Yang Li, Jamil D. Bayram, David Levinson, Samuel Yang, and Colleen Monahan. State of virtual reality based disaster preparedness and response training. *PLoS Currents*, 5, 2013.

[HTC] HTC Vive Pro. https://www.vive.com/us/product/vive-pro/. Accessed: 2023-09-27.

[HZG+20] Jason Haskins, Bolin Zhu, Scott Gainer, Will Huse, Suraj Eadara, Blake Boyd, Charles Laird, JJ Farantatos, and Jason Jerald. Exploring vr training for first responders. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 57–62, 2020.

[IBS11] Kiran Ijaz, Anton Bogdanovych, and Simeon Simoff. Enhancing the believability of embodied conversational agents through environment-, self- and interaction-awareness. In *Proceedings of the Thirty-Fourth Australasian Computer Science Conference - Volume 113*, ACSC '11, page 107–116, AUS, 2011. Australian Computer Society, Inc.

[JBG+19] Xinpei Jin, Yulong Bian, Wenxiu Geng, Yeqing Chen, Ke Chu, Hao Hu, Juan Liu, Yuliang Shi, and Chenglei Yang. Developing an agent-based virtual interview training system for college students with high shyness level. *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 998–999, 2019.

[Jia20] An Quan Jiao. An intelligent chatbot system based on entity extraction using rasa nlu and neural network. *Journal of Physics: Conference Series*, 1487, 2020.

[JRL00] W. Johnson, J. Rickel, and J. Lester. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11:47–, 01 2000.

[KBH+18] Kangsoo Kim, Luke Boelling, Steffen Haesler, Jeremy Bailenson, Gerd Bruder, and Greg F. Welch. Does a digital assistant need a body? the influence of visual embodiment and social behavior on the perception of intelligent virtual agents in ar. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 105–114, 2018.

[KdMN+20] Kangsoo Kim, Celso M. de Melo, Nahal Norouzi, Gerd Bruder, and Gregory F. Welch. Reducing task load with an embodied intelligent virtual assistant for improved performance in collaborative decision making. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 529–538, 2020.

[KLBL93] Robert S. Kennedy, Norman E. Lane, Kevin S. Berbaum, and Michael G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology*, 3(3):203–220, 1993.

[KVB20] Nawel Khenak, Jeanne Vézien, and Patrick Bourdot. Spatial presence, performance, and behavior between real, remote, and virtual immersive environments. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3467–3478, 2020.

[Lik32] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[LLGPM21] Fabrizio Lamberti, Federico De Lorenzis, F. Gabriele Pratticò, and Massimo Migliorini. An immersive virtual reality platform for training cbrn operators. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 133–137, 2021.

[LLK20] Khang Nhut Lam, Nam Nhat Le, and Jugal Kalita. Building a chatbot on a closed domain using rasa. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, NLPIR 2020, page 144–148, New York, NY, USA, 2020. Association for Computing Machinery.

[LPL22] Federico De Lorenzis, F. Gabriele Pratticò, and Fabrizio Lamberti. Work-in-progress—blower vr: A virtual reality experience to support the training of forest firefighters. In *2022 8th International Conference of the Immersive Learning Research Network (iLRN)*, pages 1–3, 2022.

[LVA+22] Guido M. Linders, Julija Vaitonytundefined, Maryam Alimardani, Kiril O. Mitev, and Max M. Louwerse. A realistic, multimodal virtual agent for the healthcare domain. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, IVA '22, New York, NY, USA, 2022. Association for Computing Machinery.

[LYXX20] Xinzheng Lu, Zhebiao Yang, Zhen Xu, and Chen Xiong. Scenario simulation of indoor post-earthquake fire rescue based on building information model and virtual reality. *Advances in Engineering Software*, 143:102792, 2020.

[MAB+22] Nathan Moore, Naseem Ahmadpour, Martin Brown, Philip Poronnik, and Jennifer Davids. Designing virtual reality–based conversational agents to train clinicians in verbal de-escalation skills: Exploratory usability study. *JMIR Serious Games*, 10, 2022.

[MFS+17] Annette Mossel, Mario Froeschl, Christian Schönauer, Andreas Peer, Johannes Göllner, and Hannes Kaufmann. Vronsite: Towards immersive training of first responder squad leaders in untethered virtual reality. In *Proceedings of IEEE Virtual Reality*, pages 357–358, 2017. poster presentation: IEEE Virtual Reality, Los Angeles, CA, USA; 2017-03-18 – 2017-03-22.

[Mil12] R.G.J. Miller. *Simultaneous Statistical Inference*. Springer Series in Statistics. Springer New York, 2012.

[MMK12] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012.

[NCD20] Bianca Nenciu, Dragos-Georgian Corlatescu, and Mihai Dascalu. Rasa conversational agent in romanian for predefined microworlds. pages 87–94, 01 2020.

[NG22] Federica Nenna and Luciano Gamberini. The influence of gaming experience, gender and other individual factors on robot teleoperations in vr. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '22, page 945–949. IEEE Press, 2022.

[NJD19] Vinh T. Nguyen, Kwanghee Jung, and Tommy Dang. Vrescuer: A virtual reality application for disaster response training. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pages 199–1993, 2019.

[Ocu] Oculus Lipsync for Unity Development. https://developer.oculus.com/documentation/unity/audio-ovrlipsync-unity/. Accessed: 2023-09-27.

[PLPK19] Fabrizio Palmas, David Labode, David Alexander Plecher, and Gudrun Johanna Klinker. Comparison of a gamified and non-gamified virtual reality training assembly task. *2019 11th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pages 1–8, 2019.

[PSR+22] Lucas Paletta, Michael Schneeberger, Lilian Reim, Wolfgang Kallus, Andreas Peer, Christian Schönauer, Martin Pszeida, Amir Dini, Stefan Ladstätter, Anna Weber, Richard Feischl, and Georg Aumayr. Work-in-progress—digital human factors measurements in first responder virtual reality-based skill training. In *2022 8th International Conference of the Immersive Learning Research Network (iLRN)*, pages 1–3, 2022.

[PSSE21] Oceane Peretti, Yannis Spyridis, Achilleas Sesis, and Georgios Efstathopoulos. Gamified first responder training solution in virtual reality. In *2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 295–301, 2021.

[Qui] Quixel Megascans. https://quixel.com/megascans. Accessed: 2023-09-27.

[Ras] Rasa Open source conversational AI. https://rasa.com/. Accessed: 2023-09-27.

[RHW20] Jens Reinhardt, Luca Hillen, and Katrin Wolf. Embedding conversational agents into ar: Invisible or with a realistic human body? In *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '20, page 299–310, New York, NY, USA, 2020. Association for Computing Machinery.

[Ric89] William R. Rice. ANALYZING TABLES OF STATISTICAL TESTS. *Evolution*, 43(1):223–225, 01 1989.

[Ric01] Jeff Rickel. Intelligent virtual agents for education and training: Opportunities and challenges. In Angélica de Antonio, Ruth Aylett, and Daniel Ballin, editors, *Intelligent Virtual Agents*, pages 15–22, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[Riva] Introducing NVIDIA Riva: A GPU-Accelerated SDK for Developing Speech AI Applications. https://developer.nvidia.com/blog/introducing-riva-a-gpu-accelerated-sdk-for-developing-speech-ai- Accessed: 2023-09-27.

[Rivb] NVIDIA Riva Speech AI SDK. https://developer.nvidia.com/riva. Accessed: 2023-09-27.

[roc] Microsoft Rocketbox Avatar Library. https://github.com/microsoft/Microsoft-Rocketbox. Accessed: 2023-09-27.

[RPC] gRPC. https://grpc.io. Accessed: 2023-09-27.

[RRT+19] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. *FastSpeech: Fast, Robust and Controllable Text to Speech.* Curran Associates Inc., Red Hook, NY, USA, 2019.

[SBS19] Susanne Schmidt, Gerd Bruder, and Frank Steinicke. Effects of virtual agent and object representation on experiencing exhibited artifacts. *Comput. Graph.*, 83(C):1–10, oct 2019.

[SGST07] Johann Schrammel, Arjan Geven, Reinhard Sefelin, and Manfred Tscheligi. "look!": Using the gaze direction of embodied agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, page 1187–1190, New York, NY, USA, 2007. Association for Computing Machinery.

[SLB+23] Elizabeth A Schlesener, Caitlin Marie Lancaster, Catherine Barwulor, Chandni Murmu, and Kelsea Schulenberg. Titleix: Step up & step in! a mobile augmented reality game featuring interactive embodied conversational agents for sexual assault bystander intervention training on us college campuses. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA, 2023. Association for Computing Machinery.

[SNS19] Susanne Schmidt, Oscar Javier Ariza Nunez, and Frank Steinicke. Blended agents: Manipulation of physical objects within mixed reality environments and beyond. In *Symposium on Spatial User Interaction*, SUI '19, New York, NY, USA, 2019. Association for Computing Machinery.

[SQL] SQLite: self-contained database engine. https://www.sqlite.org/index.html. Accessed: 2023-09-27.

[SRD15] Sharad Sharma, Shanmukha Pranay Rajeev, and Phillip Devearux. An immersive collaborative virtual environment of a university campus for performing virtual campus evacuation drills and tours for campus safety. In *2015 International Conference on Collaboration Technologies and Systems (CTS)*, pages 84–89, 2015.

[SSG+17] Iulian Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Ke, Sai Mudumba, Alexandre Brebisson, Jose Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Y. Bengio. A deep reinforcement learning chatbot. 09 2017.

[SSS98] S. Stansfield, D. Shawver, and A. Sobel. Medisim: a prototype vr system for training medical first responders. In *Proceedings. IEEE 1998 Virtual Reality Annual International Symposium (Cat. No.98CB36180)*, pages 198–205, 1998.

[SW18] Keng Siau and Weiyu Wang. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, 31:47–53, 03 2018.

[SWH18] Valentin Schwind, Katrin Wolf, and Niels Henze. Avoiding the uncanny valley in virtual character design. *Interactions*, 25(5):45–49, aug 2018.

[SWHK15] Valentin Schwind, Katrin Wolf, Niels Henze, and Oliver Korn. Determining the characteristics of preferred virtual faces using an avatar generator. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '15, page 221–230, New York, NY, USA, 2015. Association for Computing Machinery.

[Ten] Tensoflow. https://www.tensorflow.org/. Accessed: 2023-09-27.

[TR02] David Traum and Jeff Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 2*, AAMAS '02, page 766–773, New York, NY, USA, 2002. Association for Computing Machinery.

[TRO+19] Nattaon Techasarntikul, Photchara Ratsamee, Jason Orlosky, Tomohiro Mashita, Yuki Uranishi, Kiyoshi Kiyokawa, and Haruo Takemura. Evaluation of embodied agent positioning and moving interfaces for an ar virtual guide. 09 2019.

[UBC+22] Jessica Ulmer, Sebastian Braun, Chi-Tsun Cheng, Steve Dowey, and Jörg Wollert. Gamification of virtual reality assembly training: Effects of a combined point and level system on motivation and training results. *International Journal of Human-Computer Studies*, 165:102854, 2022.

[Unc] Monsters of Photorealism. https://www.wired.com/2005/12/monsters-of-photorealism/. Accessed: 2023-09-27.

[Uni] Unity Real-Time Development Platform. https://unity.com/. Accessed: 2023-09-27.

[VAG+14] Yaiza Vélaz, Jorge Rodríguez Arce, Teresa Gutiérrez, Alberto Lozano-Rodero, and Angel Suescun Cruces. The influence of interaction technology on the learning of assembly tasks using virtual reality. *J. Comput. Inf. Sci. Eng.*, 14, 2014.

[vir] Cyberith Virtualizer. https://www.cyberith.com. Accessed: 2023-09-27.

[Voi] Voicemeeter Banana. https://vb-audio.com/Voicemeeter/banana.htm. Accessed: 2023-09-27.

[VWG+04] Peter Vorderer, Werner Wirth, Feliz Gouveia, Frank Biocca, Timo Saari, Lutz Jäncke, Saskia Böcking, Holger Schramm, Andre Gysbers, Tilo Hartmann, Christoph Klimmt, Jari Laarni, Niklas Ravaja, Ana Sacau, Thomas Baumgartner, and Petra Jäncke. Mec spatial presence questionnaire (mec-spq): Short documentation and instructions for application. *Report to the European Community, Project Presence: MEC (IST-2001-37661)*, 06 2004.

[WHR20] Yurio Windiatmoko, Ahmad Fathan Hidayatullah, and Ridho Rahmadi. Developing fb chatbot based on deep learning using rasa framework for university enquiries. *ArXiv*, abs/2009.12341, 2020.

[WS98] Bob G. Witmer and Michael J. Singer. Measuring presence in virtual environments: A presence questionnaire. *Presence*, 7:225–240, 1998.

[WSR19] Isaac Wang, Jesse Smith, and Jaime Ruiz. Exploring virtual agents for augmented reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery.