

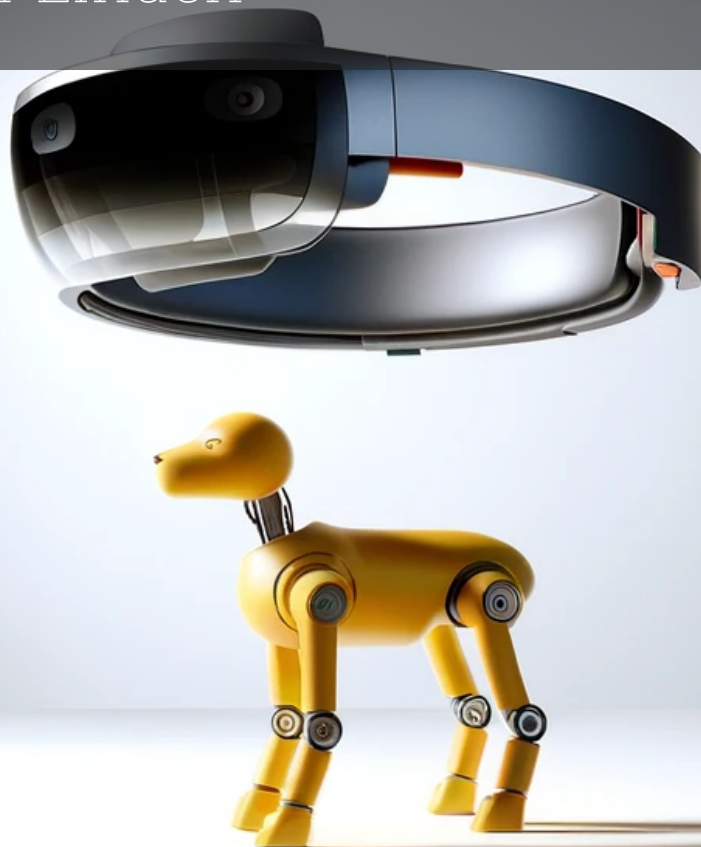
Natural User Interface in Augmented Reality to Control Spot

A Large Scale User Study on Speech and Gesture
Control of Robots With The Microsoft HoloLens

MSc. Thesis

Jesse van der Linden

Delft University of Technology



Natural User Interface in Augmented Reality to Control Spot

A Large Scale User Study on Speech and
Gesture Control of Robots With The Microsoft
HoloLens

by

Jesse van der Linden

Student Name	Student Number
Jesse van der Linden	5419921

Supervisor: Dr.ir. Yke Bauke Eisma
Committee Members: Dr.ir. Yke Bauke Eisma, Dr. Michaël Wiertlewski & Dr. Dimitra Dodou
Project Duration: January 2023 - February 2024
Faculty: Faculty of Mechanical Engineering, Delft

Cover: Dall-E-3, iteratively generated image
Style: TU Delft Report Style, with modifications by Daan Zwaneveld

Preface

This master thesis is written as a conclusion to a one-year-long research in the lab of Yke Bauke Eisma, within the Department of Cognitive Robotics at the Mechanical Engineering Faculty of the Delft University of Technology. Within the lab, I worked on natural ways of controlling a robot, mainly Spot from Boston Dynamics, initially with the mode of gestures, which extended later to other modalities like speech and gaze.

The master thesis aimed to control Spot with the aforementioned modalities, with the Augmented Reality device the Microsoft HoloLens 2 as an integral sensory input suite, as well as visual and audio feedback. Initially, a lot of time was put into studying deep learning pose estimation models and the taxonomy of gestures. After this, I developed four technical implementations to control Spot. Later in the experiment we used two of these implementations and tested them on a large user group, to conclude the characteristics of using speech and gestures to control robots.

At this time I would like to express my gratitude towards Yke Bauke Eisma for his continuous support and supervision of the entire thesis process, for keeping me on track to make this thesis up to the scientific standards, and for motivating me throughout. Next all the people in the experiment team who helped make this possible, I could not have done it without them, Renchi Zhang, Dimitra Dodou, Joost de Winter, and again Yke Bauke Eisma. Finally, to all my family and friends for their unconditional support and for keeping me motivated throughout the entire year.

*Jesse van der Linden
Delft, February 2024*

Contents

Preface	i
1 Introduction	1
1.1 Research Proposal	2
1.2 Hypothesis	2
2 Methodology	4
2.1 Experimental Setup	4
2.2 System Infrastructure	4
2.3 Data Collection	5
2.4 Research Definitions	6
2.4.1 Independent Variables	6
2.4.2 Dependent Variables	6
2.4.3 Eye Gaze Data	7
2.5 Participant pool	8
2.6 Speech and Gesture Recognition	8
2.6.1 Speech	8
2.6.2 Gesture	9
2.7 (Non-) Visual Interfaces and Feedback	10
3 Results	12
3.1 Objective Performance	12
3.1.1 Trial Completion Time	12
3.1.2 Average Commands	12
3.1.3 Command Time evaluation	13
3.1.4 Learning Rate of Conditions	13
3.2 Intra-Experiment User Experience	13
3.2.1 Responsiveness	14
3.2.2 Intuitiveness of Motion	14
3.2.3 Motion Sickness	14
3.3 Gaze analysis	15
3.3.1 Hit Object Analysis	15
3.3.2 Object Attention During Trial	15
3.4 Participant's Positioning	17
3.5 Qualitative Feedback Analysis	17
3.5.1 (Least-) favorite conditions	17
3.5.2 Interview preference reasons	18
4 Discussion	19
4.1 Objective Performance	19
4.2 Learning rate	19
4.3 Subjective Performance	20
4.4 Gaze analysis	20
4.5 Conclusion	21
4.6 Future Work	21
References	22
A Appendix A: Statistical Significance	26
A.1 Chi-Square Test	26

B	Appendix B: Human-Robot Interaction	28
B.1	Heatmap for walking conditions	28
B.2	Orientation Preferences	28
B.3	Comfortable with the robot	28
C	Appendix C: Most mentioned why's for (least) favorite conditions	30
C.1	Speech Walking	30
C.1.1	Why Favorite	30
C.1.2	Why Least Favorite	30
C.2	Speech Stationary	31
C.2.1	Why Favorite	31
C.2.2	Why Least Favorite	31
C.3	Gesture Walking	32
C.3.1	Why Favorite	32
C.3.2	Why Least Favorite	32
C.4	Gesture Stationary	33
C.4.1	Why Favorite	33
C.4.2	Why Least Favorite	33
D	Appendix D : Work done	35

Abstract

The increasing presence of robots calls for a more seamless and information-rich communication method between humans and robots. This paper explores how natural user interface (NUI) modalities, particularly speech and gesture controls, can be used through augmented reality (AR) to operate robots. The increasing presence of robots calls for proper evaluation methods of how to use AR for operating mobile robots.

The study uses the Microsoft HoloLens and the robot, named Spot, from Boston Dynamics as primary technologies. The research consists of a user study consisting of 218 participants, one of the largest participant pools for this field to date. The experiment consists of walking the robot over a trajectory with discrete steps, with the perspective of following the robot or standing on a predetermined stationary point. To support the control of the robot, visual information and feedback are included in the HoloLens.

Speech control showed the best time performance of the experiment, regardless of the perspective condition. Conversely, errors made during the trials were the majority for the speech condition, due to the waiting time of the speech recognition that caused participants to repeat the commands. The walking condition gave participants the impression that control commands were more intuitively mapped to the robot's motion. Overall, the participants preferred the speech control method while walking with the robot, and the least preferred method was using gestures in a stationary perspective.

Even though the speech was the preferred control method and perspective-taking was preferred by participants, this was only for the experiment and task discussed in this paper. Both control methods have different characteristics that make them favorable to be used for specific tasks. Speech and gestures can be used for different tasks when operating a robot with Augmented Reality glasses; preference will depend on the task at hand and the control method design.

Introduction

Robots are becoming more present in the world every year, with the estimated number of robots in operation tripling between 2010 and 2020, and the majority of these robots are robotics manipulators from companies like ABB, KUKA, Fanuc, Kawasaki, and Yaskawa [55]. The market for mobile robotics has also been increasing at a high rate with the total market expected to quadruple between 2018 and 2026[34]. This increasing presence of robots calls for more seamless and effortless ways to interact with robots, as more information needs to be transferred between humans and robots.

Traditional interaction between robots and humans often relies on the robot's physical or visual feedback capabilities, such as movements [11], gaze outputs [25], gestural motion[8], physical transformation [17], small displays [14] or visual feedback through lights [5]. These modalities cause key limitations, such as the pose and form factor of robots that cannot easily be modified on demand. Visual feedback systems like lights and displays are more flexible and are always limited to the robot design or other technical restrictions. For example, a small screen that is relatively far away from the human, such that text becomes unreadable [53].

To support the increasing presence of robots and address the physical limitations of giving feedback, Augmented Reality (AR) has proven to be a useful method and a new way to enhance human-robot interaction (HRI) and robotic interfaces [53]. With the help of AR devices it is possible to display user interfaces, widgets, spatial visualizations, and embedded visual effects [53]. On these interfaces, we can display internal and external robot information, show plans or activities, or virtual objects [53]. The purpose of this visual feedback can be to facilitate programming, have real-time control, improve safety, communicate robot intent, and increase the expressiveness of robots [53].

AR devices are not only useful for displaying information of the robot to users but also allow users to send information to robots, frequently in the form of Natural User Interfaces (NUI) [23]. The reason for this is that AR glasses are integral sensor suites, being able to capture communication methods like speech, hands, and gaze. One well-known AR device is the Microsoft HoloLens 2, this device allows for capturing all modalities at the same time [19]. Methods for humans to interact with a robot through AR are interactable objects in the virtual space [36], speech commands [52, 20, 37], gaze tracking [39, 37], and gestures [4, 7]. All methods can be applied in various ways, but are still in the early stages of their development cycle [53].

There have been several studies on controlling Robots with the help of AR devices. In one of the studies, they gave high-level and descriptive instructions to drones to execute specific tasks [20]. In another study with drones, a gesture-based control interface and a radial visual display initiate the execution of predetermined tasks [7]. In the paper of Park et al. [37], they developed a hands-free interaction method using multi-modal inputs, such as eye gazing, head gestures, and speech. Together with the help of deep learning-based object detection, they could complete a pick-and-place task with a UR3 Robot. All were technical developments of how to control robots with a NUI through AR glasses, but these papers did not include testing the user interface on a prospective user base.

In the literature, some papers discuss user studies to control robots with natural language in immersive environments. In one study, 5 different gesture-based interfaces were tested to teach robots new tasks and behaviors from demonstration. A user study of 35 participants evaluated the performance and experience of the data collection interfaces [22]. The study from T.A.B. de Boer et al. [6], used teleoperation of a robotics arm from different perspectives, such as direct view, mixed reality, and 2D video feed. 24 participants took part to again measure the user performance and experience. Sangyoon Lee et al. [27], developed a haptic-feedback control of a mobile robot within a virtual environment with obstacles from a start location to a target location. In a user study with 20 participants, they found that the number of collisions and the distance between obstacles and the robot decreased due to the haptic feedback. In a task of space teleoperation using Virtual Reality and a wide range of NUI control methods like gestures and speech. An experiment of 50 people showed that using these NUI control methods performed better and was easier to learn compared to conventional control methods [29].

Ever since the first research paper by R. Shepard and J. Metzler on mental rotations [48], there have been many subsequent studies, and spatial ability (mental rotations) became a widely studied topic in the literature [38]. In recent papers, training with augmented reality has proven to improve spatial ability compared to conventional methods [12, 28, 30]. In 2007, a study was published on the influence of perspective taking and mental rotation abilities in space teleoperation. Mental rotation ability showed a correlation with completion time, while perspective-taking ability was negatively correlated [31]. Perspective taking and mental rotations within augmented reality are relevant subjects of study for (tele-)operating robots, for the training spatial ability, and for indicating results.

1.1. Research Proposal

The user studies conducted mostly tested different gesture-based methods, only one paper made a comparison between different modalities, i.e. voice and gesture-based methods. The robots used in the mentioned user studies with VR or AR glasses are all robotic arm manipulators. Control of mobile robot user studies exist [27], but not with the use of any AR devices. Further, the size of user studies did not exceed 50 participants, giving less certainty on the statistical significance of the given studies. Larger sample sizes (more participants) allow small numerical differences to be statistically significant differences [24].

To address this research gap we will perform a large-scale user study that evaluates two different input modalities, namely speech and gestures. In a simple task with three control commands, the participant will make Spot (Boston Dynamics), a mobile robot, follow a laid-out path. The input of these modalities will be done with an established AR platform, the Microsoft HoloLens 2. The HoloLens allows experimenting with visual interfaces and feedback in the operating space. But also extensive data collection like eye gaze data, video and audio recording, and operator position and heading direction.

With this research, we hope to obtain better insight into human preference and performance between speech-based and gesture-based control. Additionally, fully understand both modalities by observing the intrinsic characteristics that make operators prefer or dislike a control method. This could be characteristics like effort, responsiveness, and intuitiveness. The objective measures will consist of metrics like time performance and the number of wrong commands given.

Because of the relevance of spatial ability and mental rotations within the literature, two experimental conditions are included in the experiment. In one condition the operator stands still in one location and in the other condition the operator walks with the robot which means participants can take an advantageous perspective at all times. This allows for discovering if the control methods are influenced when difficult mental rotations need to be made. Within the theme of the research we formulated the following research question:

What are the objective and subjective performances of Natural User Interface control methods, namely speech and gestures, captured by an Augmented Reality device, while using different perspectives to control a mobile robot?

1.2. Hypothesis

In the experimental design, the aim is to make both control methods have the same performance, by making them equivalent in the time it takes to give a command. From this standpoint, we expect that there will be an even performance when it comes to time to complete a trial.

Subjectively, the preference between control conditions is hard to predict as more factors contribute to this than we can account for. When comparing gesture and speech interaction to control an in-vehicle infotainment system it has been proven that both modalities are comparable in perceived usability, mental workload, and mental response [2].

The perspective of being stationary in a predetermined location is predicted to cause more errors made by participants, compared to the walking condition. This is with the knowledge that participants can orient themselves in the same direction as the robot frame, reducing the mental rotation ability needed for determining between left and right, making it less prone to errors.

Spatial ability and education results in engineering fields are connected [38, 12, 28, 30], therefore we can argue that our participant pool, that consists of engineering students, has a good spatial ability. This can be an indicator that the frequency of mistakes will not be much higher than in stationary condition, compared to the walking condition. Subjectively participants will likely review walking easier, but

objectively there should be little difference in performance between perspective conditions.

It has also been proved that people with higher mental rotation ability, were able to complete a robotic manipulation task in VR quicker[31]. Participants who had higher perspective-taking ability inversely correlated with time to complete a task [31]. This could indicate that participants who have a quick time performance have a weaker preference for a favorable perspective and respond quicker by making mental rotations instead of taking a good perspective.

The perspective change is expected to have more impact on gestures as a control method. The visual mapping of hand commands to the robot's motion can be confusing as you would have to point in the opposite direction of the rotation. However, it has been found that making 'assisting' gestures during a spatial visualization task does enhance performance, improving the internal computations of transformation [9]. These 'assisting' gestures are also allowed to be done during the speech control method, without influencing the robot control. Therefore, speech is to be hypothesized to have less influence from the perspective variable on objective performance.

Methodology

2.1. Experimental Setup

Because we want to test mental rotation capability while standing still and walking with the robot, we designed a uniform trajectory depicted in figure 2.2. The trajectory is optimized to have all orientations at least twice and a maximum of three times, making it a balanced and symmetric trajectory. Participants will also face the robot in a mirror situation at least twice, since the third one will not require them to make a rotation.

The experiment area was selected to be at least 5 by 5 meters, to ensure 1 meter padding on all sides of the trajectory for participants to walk around. In the back corner, we put a desktop PC that is needed to establish communication between Spot and the HoloLens; this is described in detail in section 2.2. Next to it is a charger and additional battery for Spot and chargers for the two HoloLenses.



Figure 2.1: Full image of experiment ground, with the experimenter desk on the left to start and monitor the experiment, and the full trajectory with the experimenter explaining the experiment to a participant who is standing in the starting position.

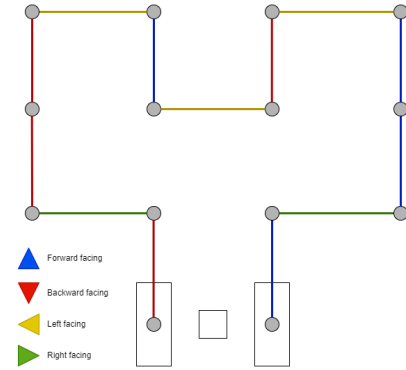


Figure 2.2: Trajectory for the experiment, consisting of 13 forward movements and 10 rotations.

2.2. System Infrastructure

The equipment used to create the experimental setup mainly consists of 3 devices: Spot, the HoloLens, and a computer. All code is collected and publicly available on [GitHub](#).

Programming the robot is done with the help of the Spot-SDK [51], with a Python package the robot can be controlled for getting the lease, sending movement commands, reading out the robot state, and asking for pictures with the image client. The robot's movements are programmed to be done within the odometry frame of Spot. Discrete movements of 1-meter and 90-degree rotations are used for the experiment.

The control input is completely captured with the Microsoft HoloLens 2. HoloLens allows one to track eyes, and hands, and listen to speech input [19]. The augmented reality display also allows giving operators feedback or instructions while controlling the robot.

The development platform for the HoloLens is Unity. The Mixed Reality Toolkit (MRTK) is the plugin to develop for VR/AR devices. The package ROS-TCP-Connector [40] can send ROS messages from a Unity project to any other device running ROS. This device must install the ROS package ROS-TCP-Endpoint and run it inside the node, to receive and forward messages from the ROS-TCP-Connector. In our experiment setup, a desktop PC is the connecting device between the HoloLens and Spot.

To optimize the communication speed between the robot, PC, and HoloLens we installed a high-speed router to provide the network. The PC was connected through an ethernet cable to the router.

Spot and The HoloLens were connected through the 2.4 GHz Wi-Fi connection. The router was placed in an unobstructed location in the experiment ground within a short distance of all devices.

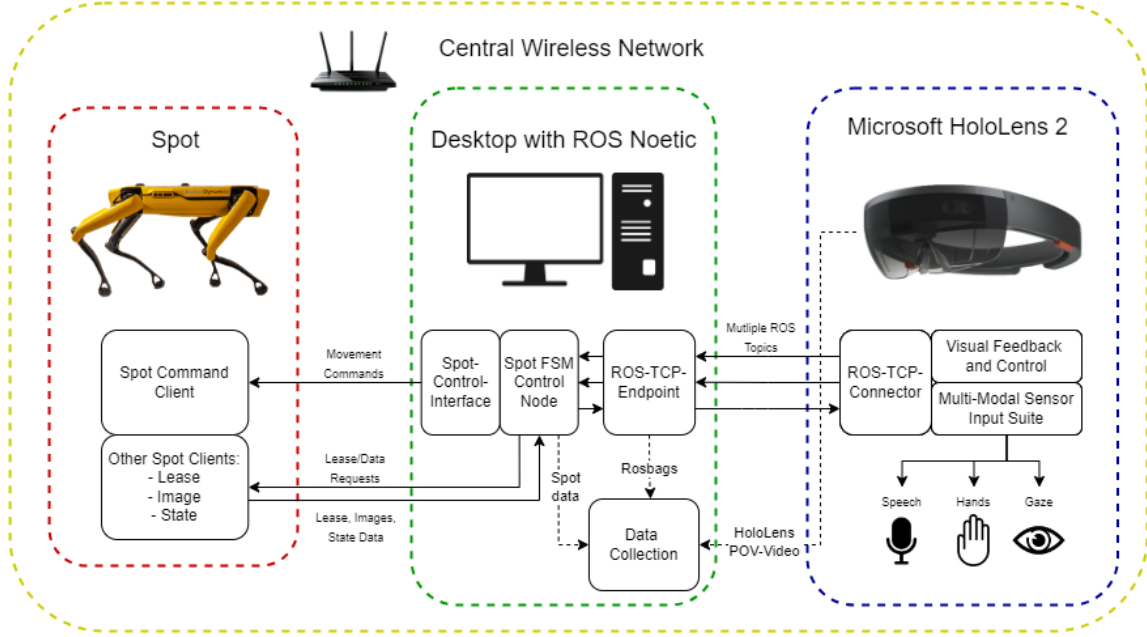


Figure 2.3: The System Infrastructure

2.3. Data Collection

Data is collected from both Spot and the HoloLens, with most of the data being capture from the HoloLens. The data is logged with rosbags, which are storage bags that subscribe to one or more ROS topics, in which data is stored and serialized with a UNIX timestamp [41]. After Rosbags are captured, they are easily exported to CSV files, with the correct columns assigned to the CSV files.

The standard operating frequency of the HoloLens is 60 Hz but during the experiments, the frequency of the HoloLens dropped from 60 Hz to 30 Hz. This happened due to the reduction in processing power, which was required for the video recording of the HoloLens. This also caused the performance of gesture control to drop because it increased the time to give a command, the gesture recognition system is explained in section 2.6. Table 2.1 shows the overview with all data that is collected during the experiment.

What	Device	Frequency [Hz]	Data type	Rostopic
Front camera	Spot	10	video	n.a.
Back camera	Spot	10	video	n.a.
Odometry Frame	Spot	20	3D Pose	/spot_odom
Vision Frame	Spot	20	3D Pose	/spot_odom
Control Commands	HoloLens	n.a.	String	/chatter
Video recording	HoloLens	30	video	n.a.
Hand Keypoints	HoloLens	30	3D Position	/hand_pose
Gaze Origin and Direction	HoloLens	30	3D Pose	/data_collection
Gaze Screen Position	HoloLens	30	2D Position	/data_collection
HoloLens Frame	HoloLens	30	3D Pose	/data_collection
Hit Objects	HoloLens	30	String	/gaze_hit_object

Table 2.1: All data captured during the experiment from Spot and the HoloLens

2.4. Research Definitions

In research, variables are characteristics that take on different values. Independent variables are manipulated and dependent variables are measured to test cause-and-effect relationships [21].

- **Independent variables** are the predetermined experimental conditions and are expected to influence the dependent variables [10].
- **Dependent variables** are the effect and depend on changes in the independent variables [10].

2.4.1. Independent Variables

As previously mentioned the experimental conditions are based on the control methods and the perspective. There are two control methods allow participants to control the robot, and the control methods both have the same three control commands that participants can give to the robot: walk forward, rotate left, and rotate right. To make a fair comparison between the control methods, the effort and time it takes to give them should be aligned, this is further discussed in section 2.6.

- **Speech:** The participant will say the command that will be recognized by the HoloLens.
- **Gestures:** The participant will hold their hand in the view of the HoloLens cameras and the gesture will be recognized.

Perspectives are added to test the mental rotation capabilities of both control methods.

- **Walking:** The participant will be instructed that they need to follow the robot in the orientation that makes giving commands easier.
- **Stationary:** The participants are instructed to stand still on the cross as seen in figure 2.1.

When combining all independent variables you get a total of four possible combinations. In the table below is the summary of what was discussed, the condition names will be used throughout the results section.

Condition Name	Control Method	Perspective
SW	Speech	Walking
SS	Speech	Stationary
GW	Gesture	Walking
GS	Gesture	Stationary

Table 2.2: Experiment conditions overview

Post-Experiment Questionnaire

From the participants, more data was obtained with a post-experiment questionnaire that can later be used as dependent variables. For example, we can measure the difference in time performance in the gesture conditions between Right and Left-handed participants.

- Age
- Gender
- Right-, Left- or Mixed-Handedness
- Visual Aid
- AR Experience

2.4.2. Dependent Variables

In this section, we discuss the dependent variables we expect to obtain from the experiment. We split these into two categories, objective and subjective metrics. Objective metrics are data that are not dependent on the experience of the users, measured from the experiment. Subjective metrics are metrics that are obtained from participants through evaluations.

Objective measures

The performance methods are:

- **Time performance:** The time it took participants to complete the course. We measure this from the first control command given to the last control command given. This gives the fairest results since we start and end the scripts at inconsistent timings during the experiment. The first control command is always done when the participant is actually ready.
- **Number of control commands:** The optimal amount of control commands is 23 and it cannot be done with less. When participants give more than 23 commands, they must have made an error somewhere in the trajectory. This gives a good indication of how many errors an experimental condition gives.
- **Errors:** During the experiment the experimenter monitors the error made by the participants. For this there are two categories for the type errors, duplicate and wrong command errors. Duplicate errors are when the participant gives a command twice without intention. Wrong commands are commands that did not match the trajectory, for example giving the command to rotate left, when a right rotation was required.
- **Timing control commands:** For every control command given we attach a timestamp, with this data we can study how much time participants take between a control command.

Subjective measures

Next to the collection of raw data from the devices, subjects are also asked for subjective evaluations during and after the experiment to measure the users experience of the experimental conditions.

Intra-Experiment Questionnaire

First we want to establish a subjective score of the experimental conditions from the participants. Between experimental trials the participants were asked three questions to evaluate the condition.

- ***The robot properly picked up my control commands.***
 - *Score between strongly disagree (1) and strongly agree (5).*
 - Gives feedback on the speed and recognition of the control method
- ***The mapping of my commands to the robot's motion was intuitive.***
 - *A score between strongly disagree (1) and strongly agree (5)*
 - Gives feedback on the ease of making mental rotations for both the control method and perspective
- ***How do you feel at the moment?***
 - *Based on the MSSS (Motion Sickness Severity Scale)*
 - Standard question when for experiments where people can get nauseous. We gave the opportunity for a break if the score was 4 or higher.

Post-Experiment Interview

After all trials have ended we record an interview with the participants, by recording video with the HoloLens. In this interview, we ask the questions of what conditions were their favorite and their least favorite and elaborate on their reasons behind this decision. This has the goal of getting a qualitative insight into the control methods and perspectives and record new unique and creative takes from participants.

2.4.3. Eye Gaze Data

HoloLens allows for eye tracking at a relatively low frequency of 30 Hz, compared to high-end eye trackers with a frequency of 2000 Hz [13]. This means we cannot analyze eye tracking data with great detail, but does tell us what participants are looking at during the experiment. Within the Mixed Reality Toolkit (MRTK) there is built-in tracking that tracks what digital objects the participants are looking at. Using this we can track if the participants are looking at the following objects:

- Spatial Object Mesh

- Robot
- Hands
- Trajectory
- Instruction Panel
- Voice and Gesture Buttons
- Questionnaire
- Background, miscellaneous objects grouped together

Between experimental conditions, we can establish what the distribution of attention is of the defined objects. With this we can distinguish what objects are more interesting to look at for the experimental conditions. We can tell if participants are exerting more effort when they are more focused on the task or less when they are looking at more arbitrary objects.

2.5. Participant pool

The participant pool is a large pool of 218 people, all of them students or faculty members at Delft University of Technology. The students are doing the course Human-Robot Interaction, and come from Master's programs such as Robotics, Bio-Mechanical Design, Bio-Medical Engineering, and more. The age range of participants is between 21 and 30. There are 194 participants that are right-handed, 18 were left-handed, and 6 were mixed-handed. Out of all participants, 160 were male, 55 were female, and 3 preferred not to respond. 121 participants never wear any visual aids, 49 wore glasses during the experiment, 36 wore contact lenses, and 12 usually wear glasses or contact lenses but not during the experiment. The previous experience with augmented reality of participants was a total of 63 participants making up 23,9% of the group, 155 participants had no prior experience with augmented reality.

The experiment received approval from the Delft Human Research Ethics Committee, approval no. 3502, with each participant providing written informed consent before the commencement of the experiment.

2.6. Speech and Gesture Recognition

Key components for the NUI are speech and gesture recognition. In this section, we describe the inner workings and low-level timings of these recognition systems. Both recognition systems can detect 3 commands that are needed for the experiment: Walk forward, Rotate Right, and Rotate Left.

2.6.1. Speech

Speech recognition is built into the Mixed Reality Toolkit that is used to program the HoloLens. You can add any command by typing the phrase in the speech commands section in Unity. The microphone on the HoloLens is robust and resistant to outside noise and interference [50].

Based on a detailed analysis of a pilot trial for the experiment it takes a participant around 1 second to say the speech command. The recognition of the command takes 1.3 seconds. Making the total time it takes to recognize one speech command is around 2.3 seconds.

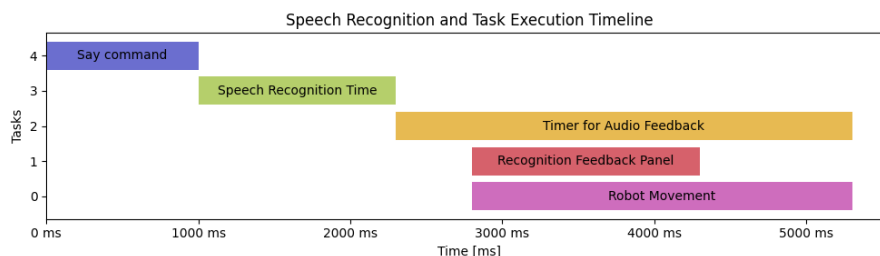


Figure 2.4: Timeline of giving a speech command

2.6.2. Gesture

Custom gesture recognition is not a standard feature that can be implemented in the MRTK. To achieve this we created our dataset of keypoints detected by the MRTK and trained a classification model. The dataset consists of 150 examples of every gesture, recorded with the HoloLens at standard operating frequency from only one person. The collected examples consist of the hand keypoint skeleton, which is available in the MRTK. This skeleton consists of 25 keypoints with 3D coordinates of the location of the joints within the HoloLens frame [16].

Preprocessing is done before the keypoints are put into the classification model. The first data augmentation that is done is MinMaxScaler [43] from the scikit-learn package. This scales all values between a given minimum and upper bound, the standard is between zero and one. The following data preprocessing is the StandardScaler [42], this standardizes features by removing the mean and scaling to unit variance.

The model trained is a Support Vector Machine (SVM). SVMs are good at classifying a low number of classes, ideally binary classification [1]. Therefore it is a good candidate for classifying the 3 necessary commands. With our training data and preprocessing, the SVM achieved a 99% accuracy and has proven to be robust in pilot experiments, demos, and all 218 participants in the experiment.

HoloLens publishes the frequency of the hand keypoints to the Python ROS node at 60 Hz. This allows us to recognize a hand command at 60 Hz. The frequency of the HoloLens is far too high for a human to be able to send discrete steps to a robot, without making errors.

Therefore we introduce a sliding window recognition system. The window considers all recognition from a predetermined window size. Every time an instance is scrutinized to a confidence threshold of 83%. If 70% of all classifications are from a single class, we send the discrete step to the robot. The values are determined by the subjective experience of the system to minimize false positive classifications.

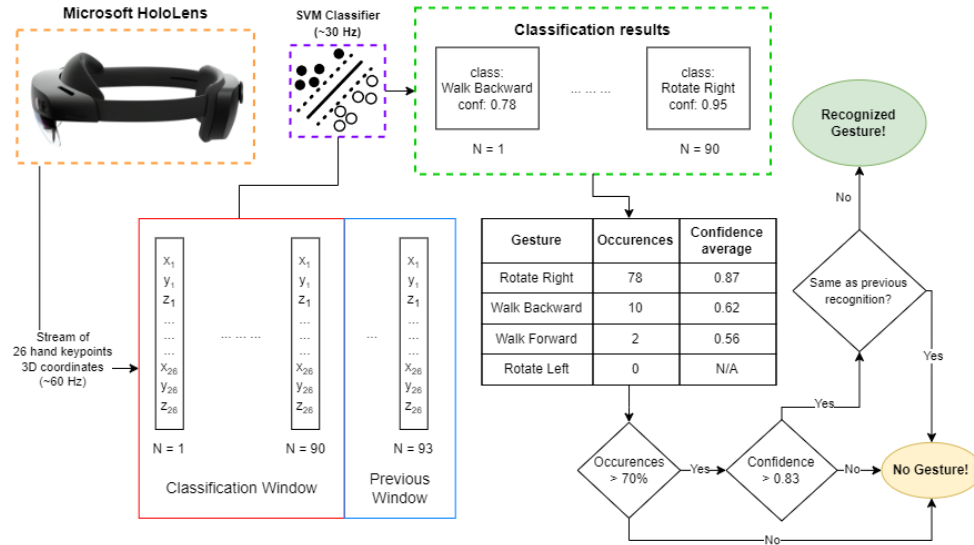


Figure 2.5: Workflow of how the gesture recognition system works

As discussed in section 2.4.2, we want the gesture recognition timing to be similar to the speech recognition timing. Based on this we sized our sliding window to a size of 90, which makes the optimal recognition time based on the publishing frequency, 1.5 seconds. The time for the recognition feedback panel to come up after successful recognition is around 0.5 seconds. Therefore the operator will hold the gesture at a minimum, of 2 seconds.

With Speech recognition to be measured at 2.3 seconds, both control methods are not exactly equivalent in time. But gesture recognition is made quicker as holding a gesture for 2 seconds is more effortful than saying a speech command for 1 second. The increase in recognition speed is given because of the higher perceived effort required to make a gesture command.

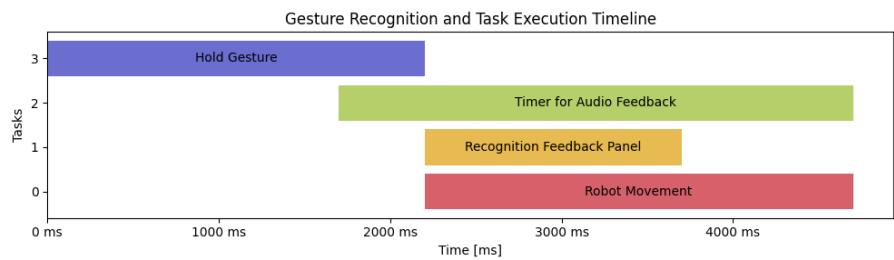


Figure 2.6: Timeline of giving a gesture command

2.7. (Non-) Visual Interfaces and Feedback

Next to the design of the control systems, visual interfaces and feedback are implemented to aid in the operation of the systems.

There are two visual interfaces available to the user. The first one is an instruction panel that indicates the different control commands, with the given movement of the robot. The experimenters will only give the instruction once to the participant. Therefore the instruction panel should be a reminder when the movement commands have been forgotten.




Speech panel	Gesture panel
<div>"Walk Forward" ↑</div>	<div> ↑</div>
<div>"Rotate Left" ↶</div>	<div> ↶</div>
<div>"Rotate Right" ↷</div>	<div> ↷</div>

Table 2.3: Instruction panels seen during the specific experimental condition

After every trial, we also answer the intra-experiment questions with the help of a screen within the HoloLens visual view. This gives the comfort of keeping the HoloLens on instead of putting on and taking off the HoloLens between every trial to answer the questionnaire.

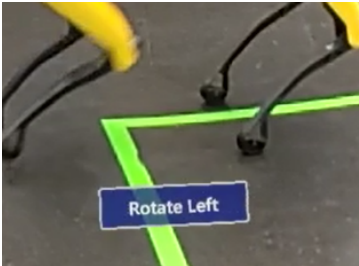


Figure 2.7: Speech and Gesture Recognition Feedback panel

Post-Trial Questionnaire: *Speak Out Your Answers by Three Numbers*

The robot properly picked up my control commands.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1	2	3	4	5

The mapping of my commands to the robot's motion was intuitive.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1	2	3	4	5

How do you feel at the moment?

No problems	Uneasiness (no typical symptoms)	Dizziness, warmth, headache, stomach awareness, sweating, ...				Nausea				
0	1	Vague	Slight	Fairly	Severe	Slight	Fairly	Severe	Retching	Vomiting
		2	3	4	5	6	7	8	9	10

Figure 2.8: Intra-Experiment Questionnaire

Visual feedback is to let operators know their speech or gesture command is recognized. This is done with a simple rectangular feedback panel that pops up on the bottom of the screen. The panel should contain the text of the given control command.

Another form of non-visual feedback is a sound that the HoloLens makes once the robot finishes a discrete movement step. Participants are instructed to wait for this signal before giving the next control command.

Results

In the result sections, we discuss visualizations forthcoming of all data collected from the 218 participants. The aim is to understand the performance and user experience of the participants in the experiment. We are also trying to understand some of the behaviors that can be observed from the captured data, like eye tracking and human-robot interaction. In table 2.2, we provide the condition abbreviations that are used in the graphs.

3.1. Objective Performance

The time performance of a condition is evaluated based on when the first control command is given and when the last command is given. This gives every participant an even playing field in time performance, as the timer is not always started at the same time for all participants.

3.1.1. Trial Completion Time

Figure 3.1 shows the distribution of time performances for all experimental conditions. An analysis of variance (ANOVA) on these scores yielded significant variation among conditions, $F((3, 218) = 7.18, p < .001$. It can be seen that speech walking ($M = 133.3, SD = 24.1$) and speech stationary ($M = 134.0, SD = 26.7$) performed the best with an insignificant difference between them according to a post hoc Tukey test between conditions, ($F(1, 218) = 0.68, p = 0.993$). The slowest performing conditions were gesture walking ($M = 148.1, SD = 27.8$) and gesture stationary ($M = 146.2, SD = 22.0$), which also showed insignificant difference ($F(1, 218) = 1.861, p = 0.87$).

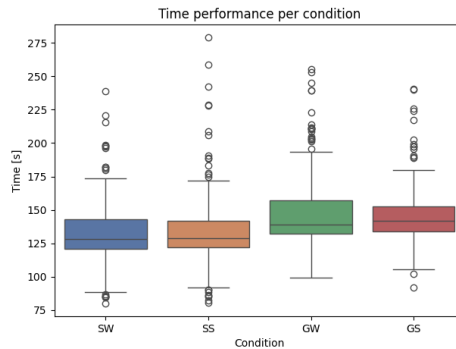


Figure 3.1: Time performance of the conditions

3.1.2. Average Commands

The optimal and minimal number of commands needed to complete the course is 23. Figure 3.2 illustrates the average number of commands for the experimental conditions. The ANOVA test revealed statistically significant differences across all conditions, $F(3, 218) = 4.00, p = .0077$. The speech conditions, both walking ($M = 24.2, SD = 2.27$) and stationary ($M = 24.3, SD = 2.62$) exhibited higher average commands, indicating a greater propensity for errors. Conversely, in the gesture conditions, walking ($M = 23.7, SD = 1.92$) and stationary ($M = 23.8, SD = 1.75$), demonstrated an average number of commands that were closer to the optimum.

To support the analysis of errors, the experimenters recorded the types of errors made under each experimental condition. Table 3.1 presents the count of errors, revealing statistically significant differences across conditions, using the Chi-square test, $\chi^2(3, 218) = 34.4, p < .001$. This proves that the speech control method is more prone to duplicate commands and the gesture stationary condition frequently issues wrong commands.

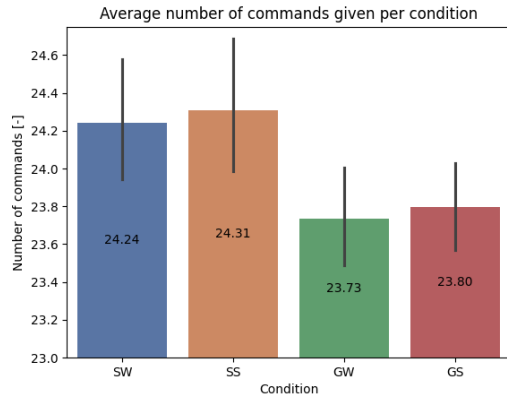
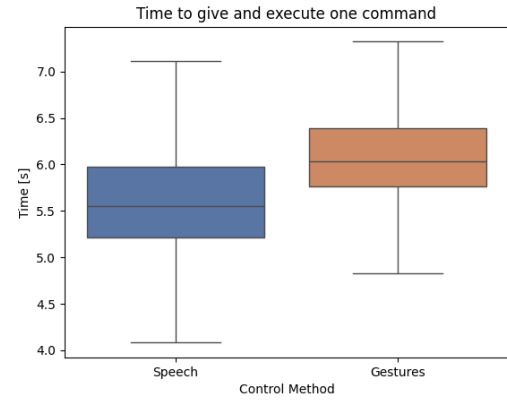
Error types	SW	SS	GW	GS	Total
Wrong command	30	32	25	48	141
Duplicate command	43	34	15	5	106
Total Condition	73	66	40	53	247

Table 3.1: Contingency table of the count of errors per error type and condition

3.1.3. Command Time evaluation

To support the converse correlation between time performance and the average number of commands, figure 3.3 depicts the distribution of the time it took to complete a single command, excluding outliers. Outliers were excluded due to the large sample size of commands ($N = 9857$) that caused excessive outliers, masking the meaningful message of the plot without telling anything significant.

A T-test compares the single command time performance of speech versus gesture control, it reveals a statistically significant out-performance of speech ($M = 5.75$, $SD = 2.06$) compared to gestures ($M = 6.55$, $SD = 3.40$), $t(9857) = -20.1$, $p < .001$.

**Figure 3.2:** The average number of commands given for a condition**Figure 3.3:** The distribution of the time it took to complete a single control command

3.1.4. Learning Rate of Conditions

The comparative analysis of the time to complete the first and second trials of a control method is displayed in figure 3.4. This does not account for the perspective, only for the chronological order of trials done. For the control method of speech, significant improvement is shown between the first ($M = 136.1$, $SD = 25.6$) and the second ($M = 131.1$, $SD = 25.1$) trial, $t(218) = 2.00$, $p = 0.047$. Similarly, the control method of gestures shows a significant improvement in time performance from the first trial ($M = 151.1$, $SD = 25.6$) to the second trial ($M = 143.6$, $SD = 24.2$), $t(218) = 3.11$, $p = 0.002$.

Figure 3.5 shows the improvement in the time performance across all trials in chronological order, irrespective of the experimental condition. An ANOVA test conducted on the trial outcomes showed statistical significance, $F(3, 218) = 4.97$, $p = 0.002$. A detailed comparison for the first ($M = 145.9$, $SD = 27.1$), second ($M = 140.6$, $SD = 26.2$), third ($M = 138.2$, $SD = 23.6$), and fourth ($M = 137.0$, $SD = 26.5$) trial, using a post hoc Tukey test, statistical significance can exclusively be found between the first and third ($F(1, 218) = 7.76$, $p = 0.011$) and first and fourth ($F(1, 218) = 8.91$, $p = 0.002$) condition. Comparison of any other pairs of trials did not yield significant differences.

3.2. Intra-Experiment User Experience

In this section, we discuss the subjective feedback that was given in the intra-experiment questionnaire.

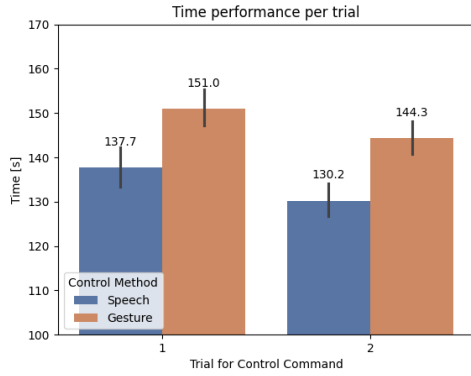


Figure 3.4: Learning rate for both control methods

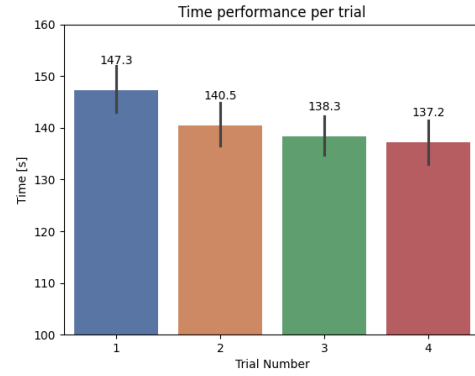


Figure 3.5: Learning rate over all trials

3.2.1. Responsiveness

The scores when asked the question: 'the robot properly picked up my control command', are seen in figure 3.6. For this, a scale where 1 is 'strongly disagree' and 5 is 'strongly agree' is used. The results of an ANOVA test show statistically significant differences between all conditions, $f(3, 218) = 23.9$, $p < .001$. When comparing all condition pairs with a Tukey HSD test insignificance was found within the control methods. The conditions speech walking ($M = 4.62$, $SD = 0.61$) and stationary ($M = 4.58$, $SD = 0.67$) showed very high correlation, $F(1, 218) = 0.037$, $p = 0.944$. Also the comparison of the pair of gesture walking ($M = 4.15$, $SD = 0.77$) and stationary ($M = 4.29$, $SD = 0.68$) were deemed to be statistically insignificant, $F(1, 218) = 0.138$, $p = 0.156$.

3.2.2. Intuitiveness of Motion

The graph in figure 3.7 shows the same scale of evaluation as the responsiveness plot for the question: 'The mapping of my command to the robot's motion was intuitive'. An analysis of variance (ANOVA) across all conditions proved significant differences, $F(3, 218) = 27.4$, $p < .001$. When conducting a post hoc Tukey test a statistical significance was found between all pairs except one. The pair of speech stationary ($M = 4.24$, $SD = 0.74$) and gesture walking ($M = 4.37$, $SD = 0.70$) did not yield significant differences to judge the intuitiveness of the command mapping between this pair, $F(1, 218) = 0.124$, $p = 0.285$. Suggesting that these conditions have the same level of intuitiveness.

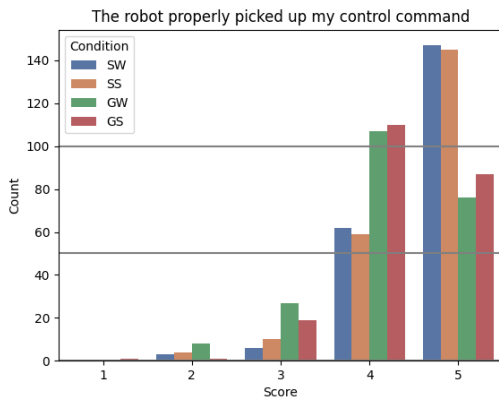


Figure 3.6: The robot properly picked up my control command.

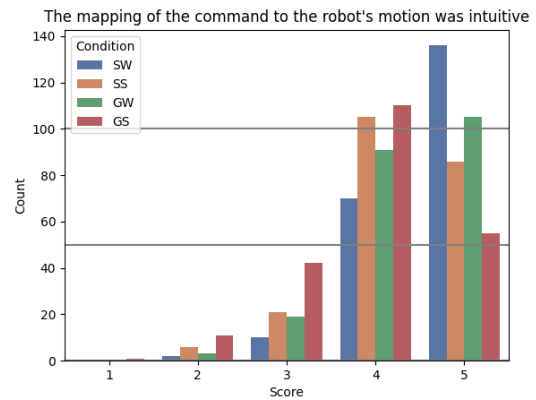


Figure 3.7: The mapping of my command to the robot's motion was intuitive.

3.2.3. Motion Sickness

The experienced motion sickness of participants according to every condition can be seen in figure 3.8. A statistical analysis of variance proved there to be no statistical significance between the control

methods for motion sickness, $F(2, 218) = 0.269$, $p = 0.848$.

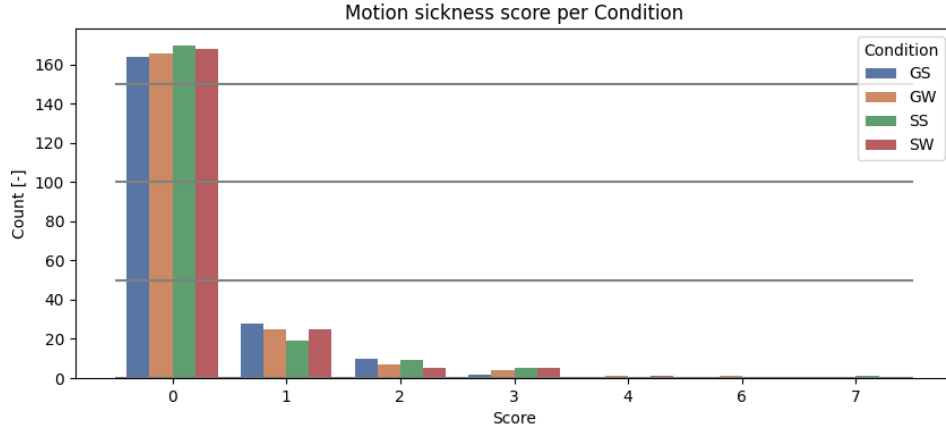


Figure 3.8: Distribution of motion sickness

3.3. Gaze analysis

As real-world objects like the robot, hands, trajectory, and other background objects were not present in the digital space of the HoloLens, they could not be tracked with the *gaze_hit_object* rostopic. To obtain this information, gaze data needs to be plotted on the HoloLens video recording and an object detection or segmentation model needs to be used to classify what part of the Spatial Object Mesh the participants are looking at. This is currently out of the scope of this thesis.

3.3.1. Hit Object Analysis

In figure 3.9 the average percentage that participants are looking at a certain object category is presented. To find out if there are any statistically significant differences between the gaze percentage and experiment condition, every object category is analyzed with a one-way ANOVA test, the results are displayed in table 3.2.

Object	SW Mean / Std.	SS Mean / Std.	GW Mean / Std.	GS Mean / Std.	F	p
Spatial Object Mesh	0.778 / 0.152	0.651 / 0.171	0.839 / 0.108	0.787 / 0.139	65.769	0.0
Questionnaire	0.048 / 0.062	0.04 / 0.044	0.047 / 0.059	0.042 / 0.057	0.461	0.709
Voice Switch	0.032 / 0.022	0.049 / 0.033	0.017 / 0.02	0.022 / 0.024	66.266	0.0
Instruction Panel	0.136 / 0.131	0.223 / 0.152	0.065 / 0.064	0.099 / 0.103	72.335	0.0
Gesture Switch	0.007 / 0.009	0.018 / 0.021	0.024 / 0.017	0.033 / 0.025	59.635	0.0
Background	0.028 / 0.045	0.044 / 0.062	0.032 / 0.061	0.043 / 0.074	3.406	0.017

Table 3.2: Statistical overview of gaze metrics per object. The mean and standard deviation of the percentage looked at the object during a trial per condition. With the ANOVA test for all conditions per object for percentage looked at during trial

For all object categories, there is a statistical significance between the experimental conditions, except for the questionnaire object category.

For all conditions the majority of the attention went to the Spatial Object Mesh, better described as the task at hand, including but not limited to the robot, the hands, and the path. Most noticeable is that for the stationary speech ($M = 0.651$, $SD = 0.171$) condition, the gaze percentage of the spatial object mesh is significantly lower than for all other conditions. The difference can be found in the instruction panel taking a higher percentage of attention of 22.3%. The gesture walking ($M = 0.839$, $SD = 0.108$) condition takes the highest percentage. The speech walking ($M = 0.778$, $SD = 0.152$) and the gesture stationary ($M = 0.787$, $SD = 0.139$) conditions score the same for attention for the Spatial Object Mesh, not showing any statistical significance in attention allocation, $F(1, 218) = 0.010$, $p = 0.900$.

3.3.2. Object Attention During Trial

In figures 3.10 to 3.13, the same gaze data as used in the previous section is categorized over time steps of 10 seconds in the first 3 minutes of the experiment trial.

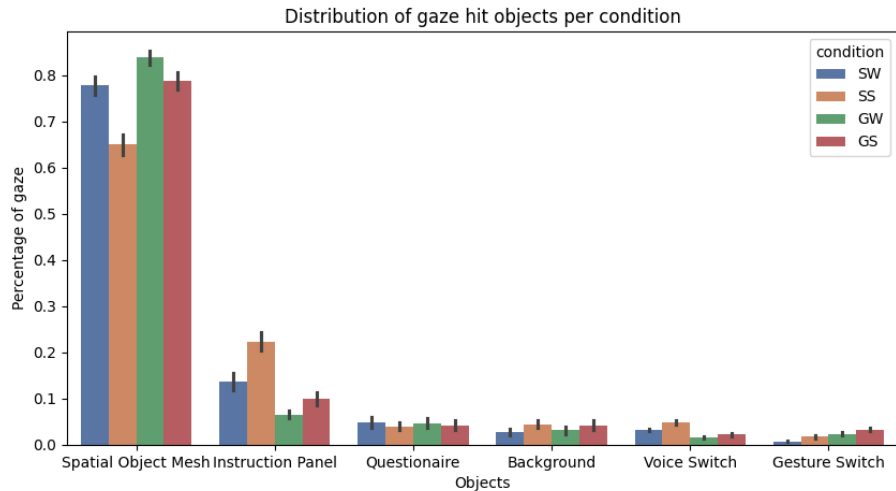


Figure 3.9: Distribution of Gaze per Condition

For all conditions, in the first 10-30 seconds the participants are interacting mostly with the elements inside the digital space of the HoloLens. This happens because of the recording of data starting before the participants start the trial. This also explains why participants are looking at the intra-experiment questionnaire, which will remain visible until one of the conditions is turned on by the participant.

For the walking conditions, you can see a consistent distribution of attention during the entire trial with little changes, except for the intra-experiment questionnaire taking a larger share of attention at the end of the trial. The stationary conditions show an increase in time spent looking at the instruction panel halfway through the trial, with the speech control method having a larger amplification.

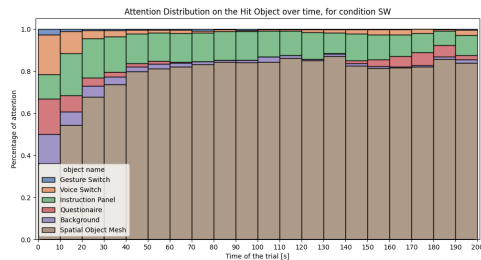


Figure 3.10: Distribution of Gaze Hit Object over time, SW

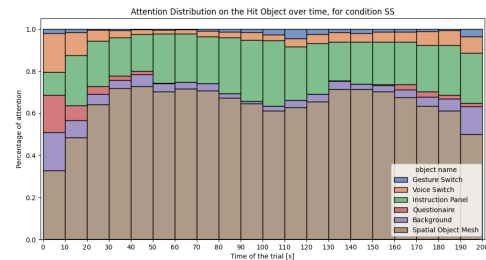


Figure 3.11: Distribution of Gaze Hit Object over time, SS

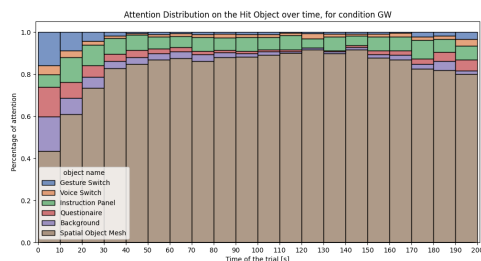


Figure 3.12: Distribution of Gaze Hit Object over time, GW

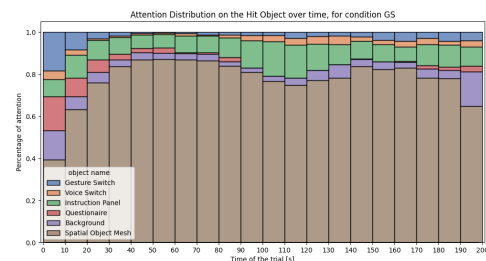


Figure 3.13: Distribution of Gaze Hit Object over time, GS

3.4. Participant's Positioning

To study the positioning of participants during the experiment we tracked the position and orientation of both the HoloLens (participant) and Spot (robot). As an example we made videos for all trials of the experiment with the position and orientation, one example can be seen in [this video](#).

With the data, we created a few visualizations in an attempt to conclude the positioning.

- Heat maps of the position of the participant for the walking conditions (SW & GW).
- The orientation offset of the participant when giving a control command, compared to the robot's orientation for the walking conditions.
- The average distance of a participant to the robot for all conditions/.

The results are found in appendix B but the results were not significant enough to show or draw any major conclusions.

3.5. Qualitative Feedback Analysis

For the HoloLens video recording of the post-experiment interview, the following processing pipeline was implemented to extract the results:

1. The interviews are cut out of the HoloLens recording video
2. The videos are converted to an audio-only format
3. The audio files are transcribed with Whisper large-v3
4. Every individual transcription is evaluated with GPT4 to get the following answers from the interview:
 - Favorite condition
 - Least favorite condition
 - Reason favorite condition
 - Reason least favorite condition

3.5.1. (Least-) favorite conditions

In the last figure 3.14 the votes are shown for the most preferred and least preferred experimental condition of all participants. The results are also shown in the contingency table A.1, which is analyzed using the Chi-square test. The test showed results to be statistically significant, $\chi^2(3, N = 218) = 125.83, p < .001$. Proving that participants had a significant preference for the speech walking condition and the least preference for the gesture stationary condition.

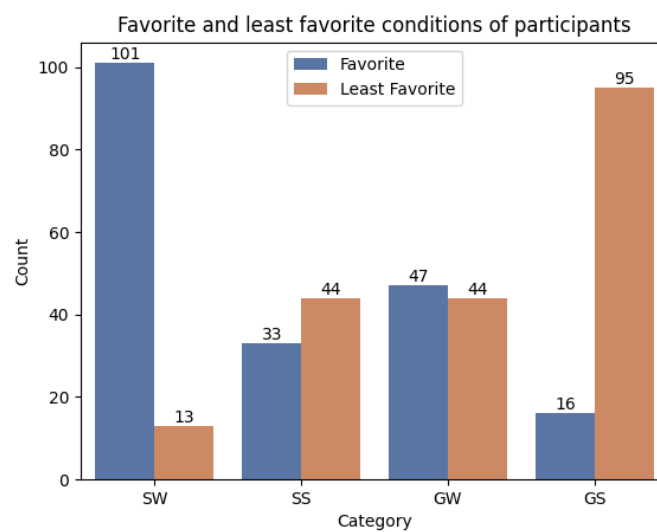


Figure 3.14: Votes for favorite and least favorite control condition

3.5.2. Interview preference reasons

The second part of the interview was for participants to give their reasoning behind the choice of the chosen (least) favorite conditions. This gave substantial insight into the control methods, which can be used to make potential improvements in the future. For all reasons, we created a summary of the most mentioned reason in appendix [C](#), with the reasons ordered from most to least mentioned.

Discussion

In this thesis, we explored the comparison of using different control methods and perspectives to control a robot with an Augmented Reality Device. In the discussion, the findings are discussed to answer the research question proposed in the introduction.

What are the objective and subjective performances of Natural User Interface control methods, namely speech and gestures, captured by an Augmented Reality device, while using different perspectives to control a mobile robot?

4.1. Objective Performance

In the first section, we discuss the factual performance of the four conditions: speech while walking and stationary, gesture while walking and stationary. Both conditions with speech as a control method have a similar time performance, and the conditions with gestures as a control method also have a similar time performance. The optimal number of commands is 23, speech commands had a high average number of commands compared to gestures. This means errors do not contribute to the lower time performance of gestures. Looking at the time distribution to complete a single control command, gestures have a significantly higher time to complete a command. A big reason is because of the drop in frequency as explained in section 2.3.

In the study by A. Matrin-Barrio et al. [29], the teleoperation of a hyper-redundant manipulator with conventional controllers, gestures, speech, or a combination of both showed that the highest efficiency was achieved with the gesture or speech-gesture combination control method. In an experiment similar to ours [18], a drone was controlled with gestures, speech, and a conventional controller. The time it took to complete the course was significantly quicker for the controller. However, the success rate in completing the course with the NUI was the same as the controller. In this paper, no clear differences were presented between speech and gesture control [18].

When comparing our results to the existing literature, there is no clear pattern to which control methods it the quickest, the most secure, or most efficient for operating robots. It is obvious that the performance of a control method is highly dependent on the given task and the design of that control method.

For our task, with the discrete commands for movements of a mobile robot, the chosen design, and unforeseen performance loss in gesture control, speech became the quicker control method. When looking at other tasks like the operation of a robotic arm, gestures can be the better method because they allow to mimic the hand position as the robot's end effector position.

4.2. Learning rate

Both control methods showed significant improvement between their first and second trial. Within all trials statistically significant was only present between the first and the third and the first and the fourth trial. No significant improvements were seen between adjacent trials, only over the whole experiment. We can conclude that both control methods are easy-to-learn systems for controlling robots. We can attribute this to the simplistic design of the control commands.

Control methods based on natural language have previously been shown to perform better despite having less previous experience, this makes it possible to complete an objective efficiently requiring less previous experience [29]. A gesture interface designed to instruct a service robot to complete a grasping task had a high ease of use, with only 3% of all experiments failing [54].

Both our study and the literature showed that natural user interfaces are easy to learn and allow for the completion of the experiment's task. The work of A. Matrin-Barrio et al. [29], actually compared it to the conventional control method. While the work of S. Waldherr et al. [54] and our work only used NUIs in the experiment. This means we cannot determine if the control methods were easier to learn compared to the conventional controller of Spot.

A substantial limitation is that the experimental task is particularly simple, which means it is unknown what would happen when the task and environment increased in complexity. To get more insight, the control methods can be compared with the learning rate of a conventional control method. This can strengthen the claim that NUI control methods are easier to learn than conventional control methods.

4.3. Subjective Performance

The favorite condition according to the participants was the speech walking condition, not only in the interview but also in the intra-experiment questionnaire on responsiveness and intuitiveness. The least favorite condition is gestures while stationary, it scored lowest on intuitiveness but for responsiveness, it scores lowest together with its gesture counterpart. The speech conditions also scored the same for responsiveness, for both perspectives. For ratings of intuitiveness, the control method was again the important factor as speech scored the highest in this regard. However, the perspective influenced the intuitiveness scores notably, the walking condition was highly preferred when comparing this with the stationary condition within the control methods.

For the task of hyper-redundant teleoperation of a robotic arm in VR, various control methods were used: controller, master-slave, local gestures, remote gestures, voice commands, and combined voice and gesture methods. The participants showed the largest preference for the combined gesture and voice method with also lowest levels of frustration. Voice commands showed high levels of frustration, caused by the nature of the task that only allows voice to make a rotation in a single joint [29]. The majority of participants showed a preference for an interaction tool based on natural language over a conventional control method [29].

If you compare our results to the literature, it can be seen that the subjective performance highly depends on the task. For our experiment the preference by participants was for speech, while in another study the preference was for gestures [29]. This difference is completely determined by the task and the control method design.

The design of the gesture recognition system is the main contributing reason for the subjective result, especially for responsiveness. This means that when designing a natural user interface, one should consider what task it aims to complete and design the control methods based on this task. When conducting an experiment, pilot experiments need to be done extensively to see that the system operates in the way that was intended.

4.4. Gaze analysis

The gaze analysis concluded that most of the attention goes to the task at hand, which includes the hand, the robot, and the experiment ground. When comparing the differences between the experimental conditions, in the gesture walking condition (84%) most attention goes to the task and the least for the speech standing condition (65%). The plots that categorize the gaze object category over time show that in the standing conditions, halfway through the experiment less time goes to the task, and more to the instruction panel. It can be argued that standing in the same place gives a better overview, requiring less attention for the task and more for other elements like the UI.

In the literature, metrics like gaze entropy, or fixation duration are commonly used for analysis of eye tracking data [12]. Pupil size seems to be an important metric for monitoring workload during a task [56, 15, 35], but cannot be measured by the HoloLens.

Our current data looks at how much a participant looks at a specific target, expressed in percentages. This is a valid metric to get a lower-level understanding of what are important elements in the workspace, but the metric does not allow to discuss differences in cognitive workload between the conditions.

The available data that consists of x and y locations on the screen, does allow us to calculate the gaze entropy, which is a common metric for determining workload based on eye gaze data [56]. Further analysis can be done with the existing experiment but does not fall within this paper.

4.5. Conclusion

In this thesis, the aim was to answer the following research question. For this, a user study with 218 participants was completed while collecting objective and subjective metrics.

What are the objective and subjective performances of Natural User Interface control methods, namely speech and gestures, captured by an Augmented Reality device, while using different perspectives to control a mobile robot?

Based on the discussion we can conclude that speech was a better method to control a mobile robot for the task of making discrete steps. Speech outperformed gestures based on time while conversely making more errors. Subjectively speech was also preferred by participants scoring higher in responsiveness and intuition. The walking perspective allows for perspective-taking, making participants prefer this over standing on a stationary point.

The higher performance of speech can mainly be attributed to the design of the control system, unforeseen time performance loss in the gesture method reduced the equality of the control. Overall both control methods were received to work well with an easy learning curve.

Both methods have different characteristics that make it a good modality for robot control. Based on the task you want to complete you should consider whether speech or gestures will be a good way to complete the task. For a full range of robotics control through Augmented Reality device you would have to use both modalities, while also including other ones like gaze and digital object interaction. With this information, it can be concluded that a multi-modality system would be the end goal when creating an AR-based robotic control interface.

4.6. Future Work

Both control methods are exciting and quickly developing new ways to interact with robots, it is an exciting field with a lot of potential. With the continuing progress of new AR/MR/VR glasses like Apple Vision Pro [3], Meta Quest 3 [32], and Microsoft HoloLens [33], new capabilities will emerge that will be usable for robotics in the near future.

For this specific thesis, the data collected by the experiment has not been fully analyzed, which means that there are more potential findings in the conducted experiment. For example, it can be shown how gestures and commands are performed between right-handed and left-handed people. Gaze data needs to be calibrated and plotted on the HoloLens video to study the gaze attention in greater detail. Or more experiments can be done with shorter gesture detection windows to increase the speed and create a closer experience for the control methods. There are a plethora of new research questions that can be derived from this research.

In the end, a holistic implementation is envisioned, where speech is used to convey information to the robot. Gaze tracking to show the intentions of the operator or point to locations where the robot must execute a task. Gestures can be used for a more direct type of control like manipulative control of the robot arm of Spot, deictic pointing to specific locations or objects, and even some semaphoric gestures to communicate discrete information [26].

References

- [1] “A comprehensive survey on support vector machine classification: Applications, challenges and trends”. In: *Neurocomputing* 408 (2020), pp. 189–215. ISSN: 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2019.10.118>. URL: <https://www.sciencedirect.com/science/article/pii/S0925231220307153>.
- [2] Leonardo Angelini et al. “Comparing Gesture, Speech and Touch Interaction Modalities for In-Vehicle Infotainment Systems”. In: *Actes de la 28ième conférence francophone sur l’Interaction Homme-Machine*. Actes de la 28ième conférence francophone sur l’Interaction Homme-Machine. Fribourg, Switzerland, Oct. 2016, pp. 188–196. DOI: [10.1145/3004107.3004118](https://hal.science/hal-01384007). URL: <https://hal.science/hal-01384007>.
- [3] *Apple Vision Pro*. <https://www.apple.com/apple-vision-pro/>. Accessed: 2024-01-31.
- [4] Stephanie Arevalo Arboleda et al. “Assisting Manipulation and Grasping in Robot Teleoperation with Augmented Reality Visual Cues”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. <conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: [10.1145/3411764.3445398](https://doi.org/10.1145/3411764.3445398). URL: <https://doi.org/10.1145/3411764.3445398>.
- [5] Kim Baraka, Ana Paiva, and Manuela Veloso. “Expressive Lights for Revealing Mobile Service Robot State”. In: *Robot 2015: Second Iberian Robotics Conference*. Ed. by Luís Paulo Reis et al. Cham: Springer International Publishing, 2016, pp. 107–119.
- [6] Thomas A. B. de Boer, Joost C. F. de Winter, and Yke Bauke Eisma. “Augmented reality-based telepresence in a robotic manipulation task: An experimental evaluation”. In: *IET Collaborative Intelligent Manufacturing* 5.4 (2023), e12085. DOI: <https://doi.org/10.1049/cim2.12085>. eprint: <https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/cim2.12085>. URL: <https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cim2.12085>.
- [7] Jessica R. Cauchard et al. “Drone.io: A Gestural and Visual Interface for Human-Drone Interaction”. In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2019, pp. 153–162. DOI: [10.1109/HRI.2019.8673011](https://doi.org/10.1109/HRI.2019.8673011).
- [8] Vijay Chidambaram, Yueh-Hsuan Chiang, and Bilge Mutlu. “Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues”. In: *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*. HRI ’12. Boston, Massachusetts, USA: Association for Computing Machinery, 2012, pp. 293–300. ISBN: 9781450310635. DOI: [10.1145/2157689.2157798](https://doi.org/10.1145/2157689.2157798). URL: <https://doi.org/10.1145/2157689.2157798>.
- [9] Mingyuan Chu and Sotaro Kita. “The nature of gestures’ beneficial role in spatial problem solving.” In: *Journal of experimental psychology. General*, 2011. URL: <https://pubmed.ncbi.nlm.nih.gov/21299319/>.
- [10] *Common Terms and Equations: Dependent and Independent Variables*. <https://www.nlm.nih.gov/oet/ed/stats/02-200.html>. Accessed: 2024-01-31.
- [11] Anca D. Dragan, Kenton C.T. Lee, and Siddhartha S. Srinivasa. “Legibility and predictability of robot motion”. In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2013, pp. 301–308. DOI: [10.1109/HRI.2013.6483603](https://doi.org/10.1109/HRI.2013.6483603).
- [12] D. Elford, S.J. Lancaster, and G.A. Jones. “Exploring the Effect of Augmented Reality on Cognitive Load, Attitude, Spatial Ability, and Stereochemical Perception”. In: Jan. 2022. DOI: [10.1007/s10956-022-09957-0](https://doi.org/10.1007/s10956-022-09957-0). URL: <https://doi.org/10.1007/s10956-022-09957-0>.
- [13] *EyeLink 1000 Plus: A Highly Accurate, Precise, and Versatile Eye Tracker*. <https://www.sr-research.com/eyelink-1000-plus/>. Accessed: 2024-01-22.
- [14] Jutta Fortmann et al. “Make Me Move at Work! An Ambient Light Display to Increase Physical Activity”. In: *IEEE*, May 2013. DOI: [10.4108/icst.pervasivehealth.2013.252089](https://doi.org/10.4108/icst.pervasivehealth.2013.252089).

- [15] Yao Guo et al. "Eye-Tracking for Performance Evaluation and Workload Estimation in Space Telerobotic Training". In: *IEEE Transactions on Human-Machine Systems* 52.1 (Feb. 2022), pp. 1–11. ISSN: 2168-2305. DOI: [10.1109/THMS.2021.3107519](https://doi.org/10.1109/THMS.2021.3107519).
- [16] *Hand Tracking - MRTK2*. <https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk2/features/input/hand-tracking?view=mrtkunity-2022-05>. Accessed: 2024-01-31.
- [17] Hooman Hedayati et al. "PufferBot: Actuated Expandable Structures for Aerial Robots". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2020, pp. 1338–1343. DOI: [10.1109/IROS45743.2020.9341088](https://doi.org/10.1109/IROS45743.2020.9341088).
- [18] Roman Herrmann and Ludger Schmidt. "Can gestural, speech interaction and an augmented reality application replace the conventional remote control for an unmanned aerial vehicle?" In: *i-com* 17.1 (2018), pp. 15–24. DOI: [doi:10.1515/icom-2018-0001](https://doi.org/10.1515/icom-2018-0001). URL: <https://doi.org/10.1515/icom-2018-0001>.
- [19] *HoloLens 2 capabilities and solutions*. <https://learn.microsoft.com/en-us/hololens/hololens-commercial-features>. Accessed: 2024-01-30.
- [20] Baichuan Huang et al. "Flight, Camera, Action! Using Natural Language and Mixed Reality to Control a Drone". In: *2019 International Conference on Robotics and Automation (ICRA)*. 2019, pp. 6949–6956. DOI: [10.1109/ICRA.2019.8794200](https://doi.org/10.1109/ICRA.2019.8794200).
- [21] *Independent vs. Dependent Variables | Definition Examples*. <https://www.scribbr.com/methodology/independent-and-dependent-variables/>. Accessed: 2024-01-31.
- [22] Xinkai Jiang et al. "A User Study on Augmented Reality-Based Robot Learning Data Collection Interfaces". In: *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition @ CoRL2023*. 2023. URL: <https://openreview.net/forum?id=jEYq9NQTJA>.
- [23] Yunshui Jin, Minhua Ma, and Yongning Zhu. "A comparison of natural user interface and graphical user interface for narrative in HMD-based augmented reality". In: *Multimedia Tools and Applications*, 2022. URL: <https://doi.org/10.1007/s11042-021-11723-0>.
- [24] Lilian Martins Fonseca Jorge Faber. "How sample size influences research outcomes". In: *Dental press journal of orthodontics*, 2014. URL: <https://doi.org/10.1590%2F2176-9451.19.4.027-029.ebo>.
- [25] Alisa Kalegina et al. "Characterizing the Design Space of Rendered Robot Faces". In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '18. Chicago, IL, USA: Association for Computing Machinery, 2018, pp. 96–104. ISBN: 9781450349536. DOI: [10.1145/3171221.3171286](https://doi.org/10.1145/3171221.3171286). URL: <https://doi.org/10.1145/3171221.3171286>.
- [26] Maria Karam and m. c. schraefel. *A Taxonomy of Gestures in Human Computer Interactions*. Project Report. 2005. URL: <https://eprints.soton.ac.uk/261149/>.
- [27] Sangyoon Lee et al. "Haptic control of a mobile robot: a user study". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Vol. 3. 2002, 2867–2874 vol.3. DOI: [10.1109/IRDS.2002.1041705](https://doi.org/10.1109/IRDS.2002.1041705).
- [28] David Lubinski. "Spatial ability and STEM: A sleeping giant for talent identification and development". In: *Personality and Individual Differences* 49.4 (2010). Collected works from the Festschrift for Tom Bouchard, June 2009: A tribute to a vibrant scientific career, pp. 344–351. ISSN: 0191-8869. DOI: <https://doi.org/10.1016/j.paid.2010.03.022>. URL: <https://www.sciencedirect.com/science/article/pii/S019188691000156X>.
- [29] A. Martín-Barrio, J.J. Roldán, and S. Terrile. "Application of immersive technologies and natural language to hyper-redundant robot teleoperation". In: 2020. DOI: [/10.1007/s10055-019-00414-9](https://doi.org/10.1007/s10055-019-00414-9). URL: <https://doi.org/10.1007/s10055-019-00414-9>.
- [30] Jorge Martín-Gutiérrez et al. "Design and validation of an augmented book for spatial abilities development in engineering students". In: *Computers Graphics* 34.1 (2010), pp. 77–91. ISSN: 0097-8493. DOI: <https://doi.org/10.1016/j.cag.2009.11.003>. URL: <https://www.sciencedirect.com/science/article/pii/S0097849309001514>.

- [31] M. Alejandra Menchaca-Brandan et al. "Influence of perspective-taking and mental rotation abilities in space teleoperation". In: *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. HRI '07. Arlington, Virginia, USA: Association for Computing Machinery, 2007, pp. 271–278. ISBN: 9781595936172. DOI: [10.1145/1228716.1228753](https://doi.org/10.1145/1228716.1228753). URL: <https://doi.org/10.1145/1228716.1228753>.
- [32] *Meta Quest 3*. <https://www.meta.com/nl/quest/quest-3/>. Accessed: 2024-01-31.
- [33] *Microsoft HoloLens 2*. <https://www.microsoft.com/nl-nl/hololens>. Accessed: 2024-01-31.
- [34] *Mobile Robotics Market Size, Share, Competitive Landscape and Trend Analysis Report by Product (UGV, UAV, and AUV)*. <https://www.alliedmarketresearch.com/mobile-robotics-market>. Accessed: 2024-01-30.
- [35] Federica Nenna, Davide Zanardi, and Luciano Gamberini. "Enhanced Interactivity in VR-based Telerobotics: An Eye-tracking Investigation of Human Performance and Workload". In: *International Journal of Human-Computer Studies* 177 (2023), p. 103079. ISSN: 1071-5819. DOI: <https://doi.org/10.1016/j.ijhcs.2023.103079>. URL: <https://www.sciencedirect.com/science/article/pii/S1071581923000885>.
- [36] M. Ostanin and A. Klimchik. "Interactive Robot Programing Using Mixed Reality". In: *IFAC-PapersOnLine* 51.22 (2018). 12th IFAC Symposium on Robot Control SYROCO 2018, pp. 50–55. ISSN: 2405-8963. DOI: <https://doi.org/10.1016/j.ifacol.2018.11.517>. URL: <https://www.sciencedirect.com/science/article/pii/S2405896318332233>.
- [37] Kyeong-Beom Park et al. "Hands-Free Human–Robot Interaction Using Multimodal Gestures and Deep Learning in Wearable Mixed Reality". In: *IEEE Access* 9 (2021), pp. 55448–55464. DOI: [10.1109/ACCESS.2021.3071364](https://doi.org/10.1109/ACCESS.2021.3071364).
- [38] Anita Pawlak-Jakubowska and Ewa Terczyńska. "Evaluation of STEM students' spatial abilities based on a novel net cube imagination test". In: Oct. 2023, p. 13. DOI: [10.1038/s41598-023-44371-5](https://doi.org/10.1038/s41598-023-44371-5). URL: <https://doi.org/10.1038/s41598-023-44371-5>.
- [39] Camilo Perez Quintero et al. "Robot Programming Through Augmented Trajectories in Augmented Reality". In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2018, pp. 1838–1844. DOI: [10.1109/IROS.2018.8593700](https://doi.org/10.1109/IROS.2018.8593700).
- [40] *ROS-TCP-Connector, Github Repository*. <https://github.com/Unity-Technologies/ROS-TCP-Connector>. Accessed: 2024-01-31.
- [41] *ROS.org: ROS Bags*. <https://wiki.ros.org/Bags>. Accessed: 2024-02-12.
- [42] *Scikit-learn: Preprocessing MinMaxScaler*. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>. Accessed: 2024-01-31.
- [43] *Scikit-learn: Preprocessing StandardScaler*. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>. Accessed: 2024-01-31.
- [44] *SciPy Stats ANOVA one-way*. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html. Accessed: 2024-01-31.
- [45] *SciPy Stats T-Test Independent*. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html. Accessed: 2024-01-31.
- [46] *SciPy Stats T-Test Related*. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html. Accessed: 2024-01-31.
- [47] *SciPy Stats Tukey HSD*. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.tukey_hsd.html. Accessed: 2024-01-31.
- [48] Roger N. Shepard and Jacqueline Metzler. "Mental Rotation of Three-Dimensional Objects". In: *Science* 171.3972 (1971), pp. 701–703. DOI: [10.1126/science.171.3972.701](https://doi.org/10.1126/science.171.3972.701). eprint: <https://www.science.org/doi/pdf/10.1126/science.171.3972.701>. URL: <https://www.science.org/doi/abs/10.1126/science.171.3972.701>.
- [49] Richa Singhal and Rakesh Rana. "Chi-square test and its application in hypothesis testing". In: *Journal of the Practice of Cardiovascular Sciences* 1 (Jan. 2015). DOI: [10.4103/2395-5414.157577](https://doi.org/10.4103/2395-5414.157577).

- [50] *Speech - MRTK2*. <https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk2/features/input/speech?view=mrtkunity-2022-05>. Accessed: 2024-01-31.
- [51] *Spot SDK, Github Repository*. <https://github.com/boston-dynamics/spot-sdk.git>. Accessed: 2024-01-31.
- [52] Aaron St. Clair and Maja Mataric. "How Robot Verbal Feedback Can Improve Team Performance in Human-Robot Task Collaborations". In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. HRI '15. Portland, Oregon, USA: Association for Computing Machinery, 2015, pp. 213–220. ISBN: 9781450328838. DOI: [10.1145/2696454.2696491](https://doi.org/10.1145/2696454.2696491). URL: <https://doi.org/10.1145/2696454.2696491>.
- [53] Ryo Suzuki et al. "Augmented Reality and Robotics: A Survey and Taxonomy for AR-enhanced Human-Robot Interaction and Robotic Interfaces". In: *CHI Conference on Human Factors in Computing Systems*. CHI '22. ACM, Apr. 2022. DOI: [10.1145/3491102.3517719](https://doi.org/10.1145/3491102.3517719). URL: <http://dx.doi.org/10.1145/3491102.3517719>.
- [54] S. Waldherr, R. Romero, and S. Thrun. "A gesture interface for human-robot-interaction". In: *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. 2000. DOI: [10.1023/A:1008918401478](https://doi.org/10.1023/A:1008918401478). URL: <https://doi.org/10.1023/A:1008918401478>.
- [55] *Worldwide installations of industrial robots from 2004 to 2020, with a forecast through 2024 (in 1,000 units)*. <https://www.statista.com/statistics/264084/worldwide-sales-of-industrial-robots/>. Accessed: 2023-12-31.
- [56] Chuhao Wu et al. "Eye-Tracking Metrics Predict Perceived Workload in Robotic Surgical Skills Training". In: *Human Factors* 62.8 (2020). PMID: 31560573, pp. 1365–1386. DOI: [10.1177/0018720819874544](https://doi.org/10.1177/0018720819874544). URL: <https://doi.org/10.1177/0018720819874544>.

Appendix A: Statistical Significance

In the results section, a lot of statistics are shown. In this appendix we note the statistical significance of all plots, to verify the validity of the results from the experiment. For all evaluations, the paired-wise t-test is used to calculate the statistical significance. For this we use the python package SciPy and the functions `"scipy.stats.ttest_rel"` [46] and `"scipy.stats.ttest_ind"` [45]. The first one is relative and can be used to test for a null hypothesis when the dataset is of equal length and represents two related or repeated samples with identical (expected) averages [46]. The latter calculates the t-test over two independent sample scores, to test for a null hypothesis that two independent samples have identical average (expected) values [45]. The analysis of variance is done with the function `"f_oneway"` [44]. Pair-wise comparison between the multiple groups is tested with the post hoc function `"tukey_hsd"` [47]. All tests are subjected to the threshold of $P < 0.05$.

A.1. Chi-Square Test

The Chi-Square test is a non-parametric test used for two specific purposes [49]:

1. To test the hypothesis of no association between two or more groups, populations, or criteria (i.e. to check independence between two variables)
2. To test how likely the observed distribution of data fits with the distribution that is expected.

The Chi-Square test only tells the probability of independence of a distribution of two categorical variables. In simple terms, it will test whether two variables are associated with each other or not. It will not tell how closely they are associated [49].

In the situation of the experiment, it will tell if there is a relation between the experimental conditions being favorable or not favorable. Not the level of how favorable a condition is [49].

The Chi-Square test calculates the sum of the squared differences between observed and expected values.

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (\text{A.1})$$

Looking at the table A.1 we see that the expected value of a condition being favorable or not, will be 50% of the total number of votes. As an example, we calculate the Chi-Square value for the votes for the speech walking condition to be the favorite. First, calculate the expected value.

$$E_i = 111 * \frac{197}{393} = 55.64 \quad (\text{A.2})$$

Now we can calculate the Chi-Square score for this vote for this condition.

$$\chi_{speechwalk} = \frac{(101 - 55.64)^2}{55.64} = 36.98 \quad (\text{A.3})$$

Vote	Speech Walking	Speech Stationary	Gesture Walking	Gesture Stationary	Total
Favorite	101	33	47	16	197
Least favorite	13	44	44	95	196
Condition totals	111	91	77	114	393

Table A.1: Chi-Square Pivot Table

The total Chi-Square statistic over the entire contingency table can be calculated.

$$\chi = 125.82 \quad (\text{A.4})$$

The degrees of freedom of a table influences the determination of the p value. The degrees of freedom is calculated with the equation below.

$$\text{degrees_of_freedom} = (\text{number_of_columns} - 1) * (\text{number_of_rows} - 1) \quad (\text{A.5})$$

The degrees of freedom of this table is $(4 - 1) * (2 - 1) = 4$ [49]. With the given Chi-Square score and a degree of freedom of 4. The P value is scored at $P = 4.3e - 27$, making it a very high level of significance.

Appendix B: Human-Robot Interaction

This appendix discusses the data of the position and orientation of the participant and robot. An example can be seen on [this video](#).

B.1. Heatmap for walking conditions

Based on the walking conditions of all 218 participants the following heatmaps are created, from both it can be seen that they are not very different from e another and no major difference can be seen. In the speech walking condition (C0) you can see the heatmap is a bit darker. This could be because of faulty preprocessing happening more frequently for this condition. Because the pixels on the edges are also a darker color compared to the gesture walking condition (C2), meaning there are frequent measurements outside of the field of view.

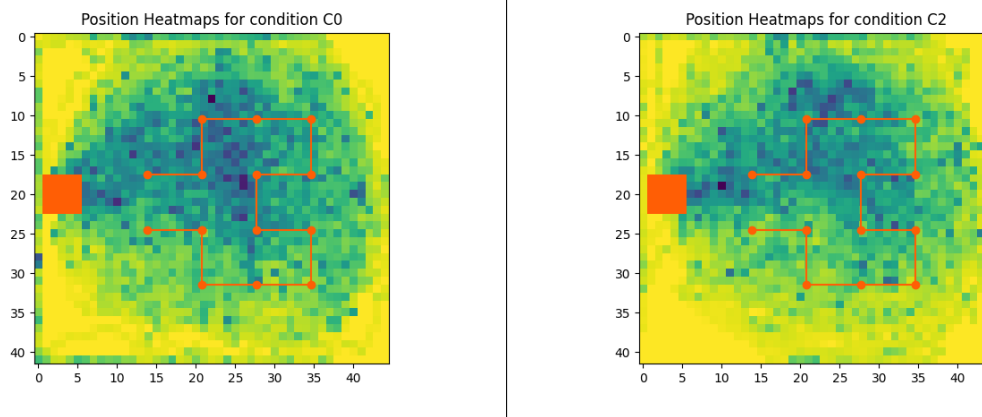


Table B.1: Heatmap of the experiment ground for the walking conditions

B.2. Orientation Preferences

During the walking condition participants were able to orient themselves favorably in line with the robot frame. It has been shown by figure 3.2, that the walking condition does create an insignificant reduction in errors made. From the plots we can draw the conclusion that the data might be faulty, the polar plot is cut up in 8 sections and the commands seem only to be categorized in the orientation with the factor of 90 degrees. This seems suspicious as participants were allowed to walk in any orientation they preferred and this disparity seem to large to be realistic. No conclusions can be made until the data is processed further to a correct representation.

B.3. Comfortable with the robot

As a study to see in which walking condition participants were more comfortable with the robot, the average distance between the participant and the robot was calculated. In figure B.1 you can see that the speech walking condition had on average 9 cm less distance to the robot, there was no statistical significance to make any certain conclusion, as seen in table ??,

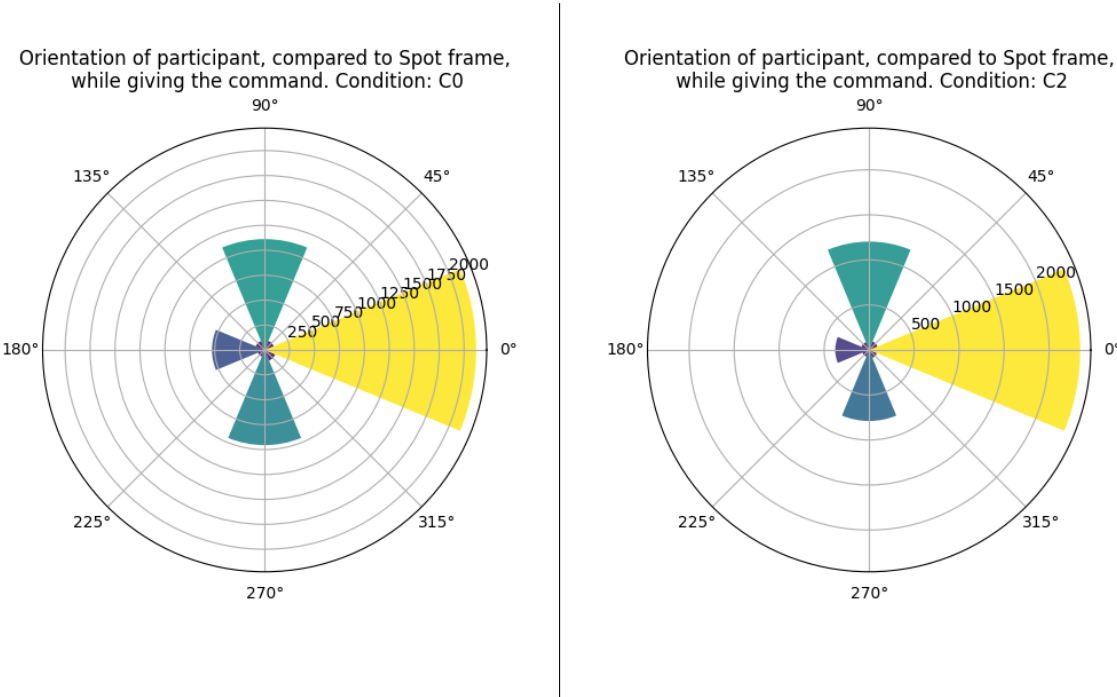


Table B.2: The orientation of participants compared to the robot frame, when giving a control command

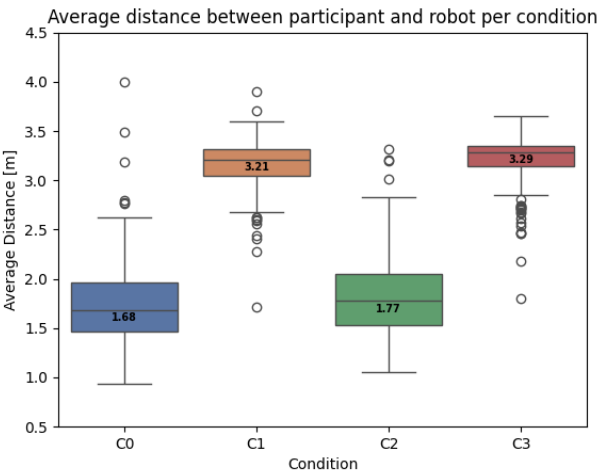


Figure B.1: Average distance between the participant and robot for all conditions

Appendix C: Most mentioned why's for (least) favorite conditions

In this appendix, we find the biggest reasons why a condition was a favorite or least favorite for the participant. The mentioned reasons are ordered from most mentioned to least mentioned.

C.1. Speech Walking

C.1.1. Why Favorite

1. **Intuitive:** Many participants expressed that combining voice commands with walking alongside the robot felt the most natural and required less effort to think about commands; it was akin to walking and instructing a pet.
2. **Ease of Use:** Participants frequently mentioned that voice control during walking resulted in a straightforward experience, with commands being easy to deliver and understand, as opposed to the precision required for hand gestures.
3. **Fast Response:** The quick recognition of voice commands compared to gestures made this condition preferable, as participants felt the robot picked up on voice faster, leading to more efficient interactions.
4. **Directional Orientation:** Walking in the same direction as the robot simplified the cognitive load for participants, as they didn't need to translate the robot's orientation to give accurate left or right commands.
5. **Natural Movement:** Participants appreciated the naturalness of moving with the robot, pointing out that walking and talking felt more enjoyable and less artificial than static gestures.
6. **Less Fatigue:** Participants liked not having to hold up their arm for gestures and found verbal commands less physically demanding.
7. **Interactive Experience:** Walking with the robot provided an interactive dimension that participants found engaging, making the experience more enjoyable.
8. **Visual Perspective:** By walking along, participants could better visualize the robot's path and potential turns, improving their ability to issue accurate commands.
9. **Efficiency:** The combination of walking and speaking was often highlighted as the most efficient method, enabling participants to control the robot effectively without repeated commands or undue attention.
10. **Comfort:** Participants expressed a sense of comfort and reduced stress when walking and using voice commands, as opposed to stationary gestures, which sometimes caused discomfort or required too much concentration.

C.1.2. Why Least Favorite

1. **Confusion:** Participants found the stationary position caused more confusion than walking, particularly when coupled with speech control.
2. **Intuitive Use:** The speech walking condition was considered non-intuitive and cumbersome, as participants were unsure if the system registered their voice commands, making it a less favored method.
3. **Physical Effort:** Controlling the robot with gestures while walking was tiring due to the need for numerous gestures and the extra effort required to adjust to the robot's movements.
4. **Efficiency:** Users preferred to issue voice commands while stationary, finding it less efficient to walk and talk simultaneously, as it hindered their ability to observe the robot's actions.

5. **Voice Recognition:** Issues with voice recognition, such as the system not responding accurately to commands or being influenced by fast speech, made this condition less reliable.
6. **Control:** When issuing commands on the move, the changing coordinate system was problematic. Participants preferred to be stationary behind the robot for better control.
7. **Reliability:** The unreliability of the system in detecting voice commands when the user was walking made the experience less favorable due to potential misinterpretations by the robot.
8. **Repetition:** The uncertainty of whether the robot heard commands led to repeated instructions, which sometimes caused the robot to execute actions multiple times, resulting in a lack of control.
9. **Feedback:** A lack of immediate feedback, such as an audible beep or visual cue, made participants uncertain if the robot was processing their commands, contributing to the uneasiness with the speech walking condition.
10. **Cognitive Load:** The speech walking condition required additional cognitive effort, as participants had to think more about the commands while also adjusting to the robot's orientation and movement.

C.2. Speech Stationary

C.2.1. Why Favorite

1. **Comfort:** Participants found controlling the robot with voice commands while standing still to be more comfortable, avoiding awkward hand movements and physical fatigue from gestures.
2. **Natural and Intuitive:** Many participants felt that voice control was a natural and intuitive way to interact with the robot, making it straightforward to direct and command.
3. **Speed and Precision:** The use of voice commands was mentioned as faster and more precise for executing instructions, with quicker responses than other methods like hand gestures.
4. **Safety and Trust:** Some participants expressed a sense of safety, as standing still while using voice commands reduced the need to be close to the robot, which some found slightly intimidating due to its power.
5. **Efficiency:** Voice control was seen as more efficient, preventing hand fatigue over time and avoiding the need for constant physical work when the trajectory was long.
6. **Control and Oversight:** Having a stationary position provided participants with a better overview and allowed them to plan ahead, seeing the whole track and feeling more in control.
7. **Reliability:** Voice control was perceived to be more reliable, with the robot reacting well and the system effectively picking up and executing voice commands.
8. **Ease of Use:** The simplicity of issuing voice commands was appealing, as it allowed participants to simply say their thoughts out loud without the need for complex gestures.
9. **Practice and Confidence:** Familiarity with the voice command system from practice made participants more confident and comfortable with this control method.
10. **Fixed Reference Frame:** Standing still provided a static frame of reference for directions such as left and right, which made giving commands easier as opposed to when walking with the robot.

C.2.2. Why Least Favorite

1. **Confusion:** Participants reported confusion over when to issue commands and found translating movements for rotation particularly unintuitive, leading to discomfort; both the timing of the beep and the need to flip rotation in their minds were problematic.
2. **Intuition and Boredom:** Participants expressed that standing still while using voice commands felt less engaging and more boring than more active interactions, and that it was less intuitive compared to using gestures.
3. **Left/Right Orientation:** Multiple participants struggled with determining the robot's left and right, and found it hard to translate their own orientation to the robot's perspective, leading to mixed-up commands and unintended movements.
4. **Pronunciation and Noise:** Issues with pronunciation and potential interference from outside noise affecting voice recognition were mentioned, contributing to a lack of clarity and effectiveness in issuing voice commands.

5. **Monotony:** Participants found repeating the same voice commands monotonous and less interesting compared to other forms of interaction.
6. **Physical Discomfort:** Some found physical discomfort in muscles when speaking commands, which contrasts with using hand gestures and suggests a preference for the physical aspect of interacting with gestures.
7. **Cognitive Load:** Participants highlighted that having to remember specific commands and thinking about each command every time is mentally taxing and contrasts with the continuity provided by gestures.
8. **Effort and Speed:** Several participants found speaking to be more effortful than using gestures and believed vocal commands took longer to process, reducing efficiency.
9. **System Responsiveness:** Concerns were raised about whether the system recognized voice commands, with some participants experiencing the robot not responding as expected or not understanding them due to pronunciation issues.
10. **First-time Experience:** For some, the initial experience of using voice to control the robot was strange and less pleasant, implying that unfamiliarity with the mode of interaction contributed to their dislike.

C.3. Gesture Walking

C.3.1. Why Favorite

1. **Intuitiveness:** Many participants found gestures more intuitive than voice commands, especially when it came to identifying left and right while walking alongside the robot.
2. **Ease of Use:** Ease of use was a significant factor, with participants expressing that using gestures while walking was straightforward and required less effort in controlling the robot.
3. **Efficiency:** Participants felt that gestures allowed for quick and efficient control over the robot, making it simpler to indicate desired directions without the need for verbal commands.
4. **Multitasking:** Some participants appreciated the ability to use hand gestures to control the robot while simultaneously walking and having the freedom to talk to others, which made the experience more comfortable and less repetitive.
5. **Orientation:** Walking with the robot enhanced participants' orientation and understanding of the robot's coordinate frame, which in turn made controlling it easier.
6. **Visual Feedback:** The visual aspect of seeing hand gestures align with the robot's movement provided better feedback and understanding of the robot's intended direction.
7. **Comfort:** Participants mentioned feeling more at ease using gestures rather than talking constantly, and walking with the robot contributed to a more comfortable interaction.
8. **Fun:** The fun and engaging nature of walking alongside the robot while using hand gestures was emphasized as a preference over other modalities.
9. **Control:** There was a stronger sense of control when participants could gesture and move along with the robot, as opposed to standing still and issuing commands.
10. **Reduced Cognitive Load:** Gestures while walking eliminated the need to translate movements into language, reducing the cognitive load on participants who did not have to think in a non-native language or continuously verbalize commands.

C.3.2. Why Least Favorite

1. **Uncomfortable Gestures:** Participants found the gestures, especially signaling 'right', to be uncomfortable and unnatural, causing strain and difficulty in maintaining the correct posture.
2. **Inconsistency:** The gesture control was reported to be inconsistent and unreliable, often not recognizing the intended commands both while standing and walking.
3. **Distractions:** Users were distracted by having to focus on both their hand gestures and the environment, leading to less efficient control of the robot.
4. **Lack of Intuition:** Participants noted that the use of gestures while walking was less intuitive, and turning in the correct direction felt less natural compared to other forms of control like voice commands.

5. **Disorientation:** Some users experienced disorientation as the gestures required to control the robot did not align with their own body's orientation, leading to confusion and errors.
6. **Lagging Response:** Many participants mentioned a lag in the system's response to gestures, causing delays in action and contributing to the overall challenge of using this mode of control.
7. **Extra Effort:** Gesture control required extra attention and effort, particularly when walking, which detracted from the focus on steering and made the experience more challenging.
8. **Physical Discomfort:** Holding and positioning the hand in view for an extended period was tiring, and certain gestures were noted to be physically demanding.
9. **Technical Limitations:** There were multiple mentions of technical issues such as the VR head-set not tracking properly, the system's failure to accurately read hand position, and the virtual representation not matching actual hand movements.
10. **Learning Curve:** The condition was new and more complex, making it hard for participants to get accustomed to the hand positioning and movement quickly, leading to a preference for other, more straightforward control methods.

C.4. Gesture Stationary

C.4.1. Why Favorite

1. **Ease of Communication:** Participants found the Gesture Stationary condition facilitated easier communication of commands like rotating left and right without confusion; this was due to both a visual connection and straightforward command execution with the robot maintaining its orientation.
2. **Intuitiveness:** The Gesture Stationary condition was perceived as more intuitive, especially in public settings where participants preferred not to vocalize commands, providing a clearer indication of the robot's direction of movement.
3. **Less Delay:** Participants appreciated being able to give the next command before the current one was executed, which they found allowed for faster interaction without waiting for voice command recognition.
4. **Clarity in Direction:** When standing stationary, users found it easier to maintain the same orientation as the robot, thus reducing confusion over directional commands such as 'my right is his left.'
5. **Efficient Planning:** The stationary condition allowed participants to plan the robot's path and think ahead about the required commands, ultimately finding the process quicker and smoother as they could issue multiple signals.
6. **Less Physical Interference:** Walking while commanding the robot sometimes led to issuing incorrect commands due to the participant's body movements, which was avoided when stationary.
7. **Increased Familiarity:** Over time, as participants became more accustomed to the robot, they found it easier to synchronize their commands with the robot's auditory cues and movements in the gesture stationary condition.
8. **Predictability:** Gesture control was seen as more predictable since the robot responded once to each command, which participants found easier to manage compared to other control methods.
9. **Visual Feedback:** Seeing their hands in the robot's lens or screen was considered cool and contributed to a sense of direct control that was much more intuitive than other interfaces like VR or joysticks.
10. **Better Perspective:** The stationary condition gave participants a better perspective of the field, allowing for an easier and more repetitive workflow with the gestures, which participants found led to a more efficient commanding experience.

C.4.2. Why Least Favorite

1. **Intuitiveness:** The condition was described as counterintuitive and confusing, especially when it came to performing gestures that were spatially inconsistent with the robot's perspective, leading to wrong commands and extra cognitive effort.

2. **Comfort:** Participants frequently cited discomfort and awkwardness while performing the gestures, particularly with certain movements like turning the wrist, which could be physically uncomfortable or lead to muscle strain.
3. **Response Time:** Many reported that hand gestures seemed slower and less responsive, noting delays in the system recognizing their gestures, which sometimes required repeating or holding poses for an extended time.
4. **Effort:** There was a significant mention of the additional effort required to perform gestures correctly, which includes both the physical aspect of reaching out or holding hands up and the mental load of having to transform their orientation to match that of the robot's.
5. **Ergonomics:** Ergonomically, the condition was not favored due to the need for twisting hands or maintaining them in an uncomfortable position, with some gestures not being picked up well by the system.
6. **Cognitive Load:** Participants expressed that the condition demanded more focus and pre-calculation, increasing cognitive load as they had to adapt their framework to the robot's and think more about the gestures.
7. **Feedback Loop:** There was an issue with feedback, as users had to continuously switch their gaze between their hand and the robot, disrupting their focus and creating a less favorable feedback loop.
8. **Engagement:** Standing still while using gestures was seen as less engaging and boring compared to more dynamic interactions like walking or using voice commands.
9. **Physical Discomfort:** Some participants experienced physical discomfort such as wrist pain, shoulder discomfort, or general fatigue from having to hold their arms out.
10. **Functionality:** There were various mentions of the gestures not being picked up well or recognized by the robot when they were standing still, which led to frustration and a feeling that the system was less effective compared to other methods of interaction.

Appendix D : Work done

Literature review

Initial gesture recognition from Spot arm camera

Initial speech recognition with microphone

Switch to HoloLens

Due to the many sensors needed to capture gestures, speech, and gaze, I decided to switch to HoloLens as it allowed me to input all desired modalities in one system.

Setting up architecture

This took me quite a while to figure out, I would say around 4 weeks of work.

- Connect Spot to a central network
- Spot connection within ROS node
- ROS node in Docker Image
- ROS-TCP-Connection to ros node

Implementations

Trying to implement every technical solution, following points were the goals:

- Speech
- Semaphoric Gestures (static gesture)
- Manipulative Gestures (direct arm)
- Deictic Gestures (pointing)
- Pinching gestures
- Direct control with impedance arm, manipulative gestures
- attempt at Static gesture recognition with HoloLens keypoints or sending the HoloLens camera over ROS.
- Deictic gaze/gesture movement commands
 - First with HTC Vive Tracker
 - Second time with internal Spot odometry
- Calibration of HoloLens position and orientation scripts

Decided to do the experiments in the HRI course, further technical developments

We decided to do the experiment for the HRI course, at the end of September, two months of preparation time. What was working at that time was speech, direct arm control, and pinching gestures.

- Static gesture recognition (Renchi)
- Data collection inside ros node (Jesse)
 - Chatter
 - data_collection
 - spot Odom
- tried to automate starting video recording within HoloLens (Renchi & Jesse)
- Make GazeProvider give gaze instead of head direction (Renchi)
- data collection with rosbags, with UNIX timestamp (Renchi & Jesse)
 - Chatter
 - data_collection

- compressed_data
- gaze hit object
- hand keypoints
- spot Odom
- PS4 controller for discrete steps with controller (Jesse)
- ...

Experiment preparations

For the experiment, we used only the implementation of Speech, gesture recognition, data collection, and Gaze Provider. Not sure about all the assignments of names on tasks, but filled them in to the best of my ability/memory.

- experimental design (All)
- write protocols (Renchi & Dimitra)
- write instructions (Renchi & Dimitra)
- UI design (Dimitra & Jesse)
- intra-experiment questionnaire (Joost & Dimitra)
- Getting the location for experiments (All)
- trajectory design (Jesse)
- buttons for turning on and off control methods (Renchi)
- consent forms (Dimitra)
- preparing experiment ground (Yke, Dimitra, Renchi, Jesse)
- signup and remarks document (Dimitra)
- Post experiment questionnaire (Joost & Dimitra)
- Ethics committee approval (Yke & Dimitra, Joost?)

Doing the experiments

5 weeks of experiments, 3 weeks before the Christmas break, and 2 weeks after the Christmas break.

Table D.1 shows how many experiments every experimenter did.

Name	Experiments Done
Renchi	184
Dimitra	125
Jesse	153

Table D.1: Experiments done by experimenters

Writing and Data analysis

A week equivalent of writing and data analysis during two weeks of the Christmas holiday.

The last two experiment weeks I also spend more time for writing and data analysis as the experiments were done by one person.

3 weeks of data analysis after the experiments, up to today (31-01-2024).

Almost everything that is done on data analysis is within the paper/thesis. There also is a GitHub Repository with all scripts for the data analysis, ask Jesse for an invite.