Exploring Human Gaze and Visual Attention: Implications of Dynamic Stimuli for Multimodal Humanoid Robot Design

Mark de Bruijn[1]

Faculty of Science, Artificial Intelligence, Social AI Lab, Vrije Universiteit Amsterdam, the Netherlands

Abstract. Social robots can convey information using verbal and non-verbal channels. However, little is known on the effects on how these channel affect human gaze behaviour. The current study investigates how human gaze behaviour works when humans receive congruent and incongruent cues using a combination of visual and auditory directions. To achieve this, a state-of-the-art gaze estimation model and external eye tracking glasses are used to measure gaze fixations in three area of interests: the head, a tablet on the torso, and the arms for gestures. The experiment uses a combination of two of these three modalities, with one being primed for participants to focus on. The results of the study indicate that while humans gaze at the face more than gestures when instructed by a robot, this effect does not occur when they are prompted to look at the gesture. The study also found that there is no difference in the choice between a modality being the prompted or non-prompted modality. Additionally, congruent cues are processed faster than incongruent cues, except when humans are prompted to look at the tablet. This result shows that congruent directions, and the choice of modality to convey information, are useful for improving the design of social robots.

Keywords: Human-Robot Interaction · Gaze Behaviour · Social Robotics

1 Introduction

Originating from human-human studies, the growing focus of gaze dynamics with social robots highlights the importance of research into the psychological mechanisms that underlie human-robot interaction. The mechanisms of gaze are used to estimate our intentions to and from other humans [4]. However, they may not directly translate into the same behaviour when a human interacts with a robot. Capturing, interpreting, and acting on human behaviour during interactions can then lead to enhancements in the design on social robots. These enhancements are useful for turn-taking, which makes conversations with robots more engaging and naturalistic. But what is meant when we talk about interactions, and what influence does gaze have on human-robot interaction?

Hornbæk & Oulasvirta have investigated what constitutes interaction and have concluded that interaction concerns two entities that determine each other's behaviour over time [20]. Interaction can be accomplished by using cues in verbal and nonverbal modalities, such as using gestures and speech. Gestures of humans are mainly performed in the central gesture space, which is the space of the torso and the length of the lower arms [14]. Additionally, the perception of speech during social situations is one of the building blocks of face-to-face interaction. For instance, the addition of visual cues from the face leads to an increase in the intelligibility of speech [18].

Humans use explicit cues (e.g., instructions, direct statements) and implicit cues (e.g., gaze, body language), to convey information. These cues can be congruent (e.g., the word `red' in red letters) and incongruent (e.g., the word `red' but in green letters), where congruent cues are processed

faster than incongruent cues. The performance cost in the mismatch (e.g., incongruent) condition is called the Stroop effect [21]. Additionally, cues can form instructions, which influence correctly executing skills [9].

Gaze is a visual cue using the eyes. The eyes can provide signals to send and capture information about the world [11]. The relative spatial position of the eyes indicate fixations, which are aggregated gaze points on a specified area and time span [5]. Fixations play an important role in the survival of humans by indicating visual attention, with Grossman stating that the cues created by the eyes are a hallmark of social function and are deeply rooted in human biology [13]. Gaze can be estimated using eye tracking technology. These technologies can be on-screen, mobile, and head-mounted eye trackers [3]. Fast commercial eye trackers can reach up to 2000 recordings per second (2000Hz), while wearable eye trackers might only reach 50 recordings per second (50Hz). The latter rate of recordings per second is preferred when researchers are interested in where humans look, whereas a higher rate of recordings per second is preferable when millisecond accuracy is required [7]. Additionally, to measure the distributions of attention to regions, areas of interest (AOIs) can be defined [5], which are useful for determining where people look.

While gaze describes where the eyes are directed, visual attention is about processing and understanding what the eyes see. Visual attention can be described using overt and covert attention. Overt attention is the physical movement of directing the eyes to a stimuli. Covert attention is the mental shift of attention, without the physical movement of the eyes [26]. Visual attention can be measured as an inference from differences in manual reaction times [18]. Visual attention is exactly what was measured in a study executed by Özer *et al.* [23] through capturing accuracy, response time (RT), and gaze. Their study has a human watching a video recording of another human giving a cue of a relative spatial relation (e.g., *left-right* or *on-under*), and serves as the basis for the current study. The logic underlying the psychological processing of human-human interaction can also apply to social robots: if one understands the effect of a human looking at a robot, then the robot can also tailor themselves to the human.

Placing a human in front of a robot, rather than using a video recording brings forth its challenges, such as capturing human gaze in a more naturalistic setting. For this, recent advancements in state-of-the-art gaze estimation models can be employed to enable robot interactions without the need for humans to wear external instruments. These models can be used as a cue for the robot to be able to deduct where a human is looking. These cues might be useful for the robot to know when to, or when not to, engage with a human. Given that humans and robots can communicate using different modalities, knowing how people pay attention to verbal and nonverbal modalities of the robot might help with the design and development of robots.

1.1 Research Question and Hypothesis

Central in the study is the aim to explore the effects that influence human accuracy, response time, and gaze behaviour, as well as the effectiveness of gaze models. The congruency and modality of cues will be changed and the effect will be measured by the response time, accuracy, and gaze behaviour. One modality, called the focus modality, will be the cued modality for participants to focus on. Additionally, whether dedicated hardware is required to track human gaze will be investigated. This leads to the following research question:

`How do congruent and incongruent multi-modal cues from a social robot influence human gaze behaviour?'

In order to investigate this research, the following hypotheses are defined:

Hypothesis 1: Congruent cues will be processed faster than incongruent cues (Stroop effect [21]).
Hypothesis 2: The face will receive more fixations during each trial than the gestures [23].
Hypothesis 3: The focus modality will receive more fixations during each trial compared to when that modality is not the focus modality (task compliance).

1.2 Approach

The aim of the current study is to contribute to the understanding of gaze behaviour during human-robot interaction, with particular focus on human gaze at the robot's body parts. To achieve this, an experiment will be set up with a human facing the social robot Pepper. The robot will instruct the human to focus on one of three modalities: gesture, speech or the tablet. It does so by saying "*Please focus on [modality]*", with `modality' being a directional `left' or a `right' signal, with an example being the left arm being raised. The human has to match this direction by pressing the key on the keyboard with the same direction. Simultaneously, another modality will send a directional signal, which can be the same (congruent) or the inverse (incongruent) direction. The response time and accuracy of the response will be measured. Simultaneously, the gaze of the human will be tracked, twice with two cameras through the state-of-the-art L2CS model [1] and once through the Pupil Labs Invisible eye tracking glasses [25]. This gaze will be mapped to areas of interests, which can then be interpreted to identify where participants gaze during the experiment.

1.3 Overview

First, the work of other researchers will be described in chapter 2 Related work. Next, chapter 3 Technical implementation talks about the technical components such as waking the Pepper, controlling actuators, and setting up the calibration and experimental code. Following the technical implementation are the technical issues encountered in chapter 4 Technical Issues. Following the technical issues is chapter 5 Methods, containing the choice of instruments, participants, the interaction setup, and an in-depth description of the experiment. Chapter 6 Results will then present the visualization and data of the study. The findings of chapter 6 Results will be discussed in chapter 7 Discussion. Chapter 8 LEDs proposes a study idea on the effect of applying different visual features using the LEDs on the head of the Pepper, combined with variations in the head yaw and tilt. The current study will then be summarized by chapter 9 Conclusion, where the research question will be answered. This thesis will then conclude with chapter 10 Acknowledgements, which includes thank-you notes.

2 Related Work

2.1 Reference paper: Özer *et al.* ([23])

The current study is based on a study executed by Özer *et al.* [23], and will be used as the main reference paper. The study by Özer *et al.* [23] places focus on visual attention from humans on gestures in the comprehension of spatial relations between objects in different speech contexts (e.g., "The candle is *here/right*") and gesture condition (e.g., with or without gesture). A particular

focus is placed on the comprehension and allocation of direct visual attention to the gestures using an eye tracking paradigm, such as using accuracy, response time and tracking gaze. Their study investigates whether participants gaze more at gestures when complementary or redundant cues are provided, in relation to the accompanying speech compared to no gestures. The reference study tests the comprehension of two types of spatial relations: a viewpoint-independent spatial relation (e.g., *on-under*) and a viewpoint-dependent spatial relation (e.g., *left-right*).

During the study of Özer *et al.* [23], participants are asked to watch a video recording and choose the picture that best depicted the spatial relation of two objects among four options. Participants first go through familiarization and practice sessions, with the latter providing oral feedback on the correctness of their input. They then go through calibration and the actual experimental task. Participants are presented with a "Get Ready!", a preview of the response screen, a 1,000-millisecond fixation screen, the relative position (*left*, *right*, *on*, *under*), and finally the response screen. They are then asked to choose the correct picture with the mouse as quickly and accurately as possible on the response screen.

The main elements from the reference study that have been used in the current study are the "Get Ready!" message, providing participants with practice session, using eye tracking and their measurements of response time and accuracy. Different from their study, the current study does not specifically look at spatial relations, but at gaze behaviour onto a robot. Additionally, where the study by Özer *et al.* [23] is a human-human study with a video recording of a human, the current study will be executed in a physical setting with a human and a robot.

2.2 Human-Robot Interaction

Admoni and Scassellati divide three broad categories of research on the current state of gaze in human-robot interaction, all distinguished by their goals and methods [2]. The three categories they describe are the human-focused, design-focused and technology-focused research. The focus of the current study is on the side of human-focused research, with the understanding of how humans behave when they are placed in front of a robot. Simultaneously, the technological side of the research is investigated with attention to using the L2CS model, which might be usable for further improvements in robot design. While the deeper technological implementation of the model will not be explored, the current study does investigate the effectiveness of using the model on two cameras, compared to eye tracking glasses. The design-focused research is on the physical appearance of the robot and is not in the focus of the current study.

2.3 Eye tracking

Advancements in eye tracking instruments, such as improved head-mounted, glass, table-mounted, and embedded systems, have shown rapid developments in accuracy, stability, and sampling rates [12]. Additionally, adding a head-rest during eye tracking improves the overall precision [22] but does reduce the natural setting in which an experiment takes place.

A mathematical issue that arises with eye tracking is a systematic shift called calibration error. This calibration error is an offset in the values of the coefficients that determine where a human is looking. In order to reduce this calibration error, it is adamant that calibration is performed to provide corrective values for the values of coefficients in the mathematical equation [28]. Using one or multiple markers is a common approach for calibrating human gaze. Alternatives to this simple calibration are the standard 9-Point calibration (where the human has to look at 9 markers), smooth

pursuit calibration (where the human has to follow a predefined path), and vestibulo-ocular reflex (VOR) calibration (where the human has to look at a static marker and rotate and turn their head) [16]. However, these methods require more complex procedures to validate that the calibration was followed precisely.

For the current study, the Pupil Labs Invisible eye tracker will be used, which are wearable glasses. The Pupil Labs Invisible eye tracking glasses requires external hardware to measure human gaze [25]. Contrary to requiring hardware are the deep learning models, which instead use a neural network to predict the human gaze. One deep learning model that has been developed to allow human gaze to be tracked through the eyes, is the L2CS model. The L2CS model has shown to outperform several state-of-the-art models on the MPIIGaze dataset [1]. Contrary to the Pupil Labs Invisible, the L2CS model does not require dedicated hardware, but instead uses a Convolutional Neural Network (CNN) to predict 3D gaze vectors. By not requiring dedicated hardware, the L2CS model takes a step to a future gaze system that is low cost and provides good gaze estimation accuracy under natural head movement [15].

3 Technical implementation

The following chapter describes the core technical components used in the experiment. The main part of the code used for this experiment is built using the Social Interaction Cloud (SIC) framework. This framework is created by the Social AI Lab and contains functions to control the robot based on the naoqi code. The naoqi code is used by developers to control the actuators of the Pepper [27]. The code for the current study is all written in Python 3.9 [6].

3.1 Waking the Pepper

The Pepper robot has a sleeping and an awake state. During the sleeping-state, the robot is tucked in, looks down and does not react to commands. When the Pepper robot goes into the awake-state, the robot looks forward and stands straight. During the awake state, a set of motions that reenact functionality of autonomous life appear, like tracking the head of a human, breathing, and blinking. The current study requires the robot to be in the awake-state and to have the autonomous life functionalities turned off. The control of the actuators and tablet from the Pepper are fully controlled by the custom code used for the current study. The implementation of waking the Pepper and disabling the autonomous life functionalities can be seen in Algorithm 1.

Algorithm 1 Wake up Pepper and disable autonomous life functions

```
1: Wake Pepper: the robot wakes up: sets Motor on and, if needed, goes to initial position.
2: pepper.autonomous.request(NaoWakeUpRequest())
3:
4: Autonomous life: disable autonomous life
5: pepper.autonomous.request(NaoBasicAwarenessRequest(False))
6: pepper.autonomous.request(NaoBackgroundMovingRequest(False))
```

3.2 Controller actuators and tablet

While the robot is now awake, it still requires a set of initial handling of all the joints to ensure that all motions, speech and displays are working as intended. This includes setting the tablet to a neutral white screen, stretching and retracting both arms, moving the yaw and tilt of the head to `look' at the participant, and confirming that the speech module works by creating an utterance. During the calibration and the experiment, the joints of the robot will be controlled using a set of actuators on the robot. How the actuators, tablet and the speech are controlled can be seen in Algorithm 2.

Algorithm 2 Control actuators and tablet

***Speech**: `utterance' is the text to be spoken by the robot, can be all utterances*
pepper.tts.request(NaoqiTextToSpeechRequest(utterance))

***Tablet**: `picture' is the image to be displayed on the tablet; can be left arrow, right arrow or white screen.*
pepper.tablet_display_url.send_message(UrlMessage(picture))

***Actuator**: control the head and arms*
pepper.motion_record.request(PlayRecording(NaoqiMotionRecording(recorded_angles=[0, angle, 0], recorded_joints=[head_yaw/left_arm], recorded_times=[[0, 1, 2.5]])))

3.3 Parallelization of main thread and cameras

During the calibration and experimental trials, the code captures and saves the frames from two cameras, and simultaneously controls the actuators of the robot. The following complex multi-threading code arises from issues with internal RAM memory overflowing. For more details on this issue, see Section 4.2.

The main thread first wakes and prepares the robot, after which a new thread is started to save the camera frames. This new thread sets a frame rate and resolution, initializes and receives frames from the cameras, and writes them to the disk, accompanied with a unique id. Simultaneously, the new thread writes a timestamp to an internal list. When this thread is ready, it starts to record the frames and sends a signal to the main thread that the calibration or experiment can start. When the calibration or experiment is over, both threads will stop, and the main thread will write the timestamps of the ids to the disk as a .csv. The workings of this algorithm can be seen in Algorithm 3.

3.4 Controlling Calibration

To capture calibration error, participants are instructed to look at a certain part of the body of the robot for four seconds. During this time, camera frames are saved to the disk with timestamps. This enables tracking which frames belong to which focus point during the calibration. The technical implementation of the calibration can be seen by the pseudocode in Algorithm 4. A more thorough explanation of the calibration can be found in Section 5.4.

Algorithm 3 Parallelization: Multi-thread-initialization and Saving of Frames.

```
Receive Signal: Receive the signal to start saving frames
[externally -> ] start_recording()

Record Frames: capture and save frames
while needs to record do

    Capture Frame: Capture the frame from the camera
    frame, timestamp = capture_frame()

    Process timestamp: Place timestamp in list
    process_timestamp(timestamp)

    New Thread: create a new thread to save the frame to the disk
    thread = create_new_thread()
    thread.save_frame(frame)
end while
```

Algorithm 4 Calibration

```
Robot preparations: raise arms and show white screen
pepper.motion_record.request(PlayRecording(NaoqiMotionRecording( recorded_angles= [0, angle],
recorded_joints= [arms], recorded_times= [[0, 1]])))
pepper.tablet_display_url.send_message(UrlMessage(white_screen))

Record cameras: start a new thread to save camera frames and timestamps
record_cameras()

Human preparations: get the participant in front of the robot
nao.tts.request(NaoqiTextToSpeechRequest(`please get in front of the Pepper'))

Algorithmic preparations: prompts of focus points during calibration
focus_points = [
    `This calibration is for [with/without] eyetracker',
    `please look at my head camera',
    `please look at my nose',
    `please look at my tablet',
    `please look at left elbow',
    `please look at right elbow',
    `calibration finished! thank you very much!']

Execute calibration: instruct humans to focus on a focus point
for each point in focus_points do
    nao.tts.request(NaoqiTextToSpeechRequest(point))
    if four seconds passed then start the next one
end for

Finish up: write timestamps to .csv
save_data_to_csv()
```

3.5 Controlling Experiment

Controlling the experiment is more complex than the calibration due to the order of the trials having to be determined beforehand. At the same time, a participant also has to be able to respond to a trial, so a method of capturing a key press has to be devised. The main idea of controlling the experiment is that 120 randomized trials have to be prepared and executed, with a small 20 second break after completing every 20 percent. After the trials have been created, a participant can start with the trials. Since the trials will been randomized, the code can simply iterate over all the trials, which are then carried out by the robot. Algorithm 5 contains the pseudocode for the randomization of the trials, while Algorithm 6 shows how the experiment is processed.

Algorithm 5 Create all 120 randomized trials

```
1: Get combinations: append all combinations into an array; do this five times
2: items = [`speech', `tablet', `gesture']
3: trials = []
4: for each 24 combinations in range(5) do
5:     combinations = get_possible_combinations()
6:     trials.add(combinations)
7: end for
8:
9: Randomize: Shuffle all trials to prevent order bias
10: random.shuffle(trials)
```

Algorithm 6 Experiment

```
1: Get trials: Retrieve the randomized trials
2: trials = create_random_trials()
3:
4: Execute trials: control and execution of the trials
5: for each trial in trials do
6:     if 20/40/60/80% of trials passed then
7:         pepper.tts.request(NaoqiTextToSpeechRequest(pause))
8:         time.sleep(20seconds)
9:
10:    Execute modalities: have the robot show the modalities
11:    execute_robotic_modalities()
12: end for
```

3.6 Breaks after 20 percent

Anecdotal evidence from pilot testing suggests that participants suffer from fatigue when all 120 trials are executed back to back. In order to prevent this fatigue, a participant gets a 20-second break after completing every 20 percent (or 24 trials), which are announced by the robot. These breaks can be seen in Table 1.

Percentage Completed	Utterance
20%	Great! You have finished 20%. Now you have a 20 seconds break before next trial!
40%	Good job! You have finished 40%. Now you have a 20 seconds break before next trial!
60%	Continue! You have finished 60%. Now you have a 20 seconds break before next trial!
80%	Wow! You have finished 80%. Here is the last round. And you have a 20 seconds break before next trial!

Table 1: Utterances based on the breaks provided after every 20%.

3.7 Capturing keypress

During the experiment, a participant has to respond to the direction of the trial using the keyboard. A participant pressing a key also has to be captured, simultaneously to all other parallelization. The participant has a maximum of four seconds to press a key, which starts simultaneously to the vocalization of `Please' during *"Please focus on [modality]"*. The response time will then start counting as soon as the focus modality is vocalized. If a key is pressed during the four second timer, that key is reported back to the main thread, as well as that it was a valid key press. With the same logic, if no valid key was pressed, or if no key was pressed at all, an invalid response will be send back to the main thread. The pseudocode for capturing key presses starts during `execute_robotic_modalities()' in Algorithm 6, and the inner workings of capturing the key presses can be seen in Algorithm 7.

Algorithm 7 Capturing keypress

***Listener**: initiate a keypress listener*
listener = keyboard.Listener()
start_time = now()

***In parallel**: await for maximum of four seconds whilst listening to a keypress*
while now() − *start_time* < 4 **do**
    sleep(0.001 seconds)
    listener.listens()
end while

***Return response**: If any key was pressed, capture and report output; report False if four seconds passed*
listener.report(
    `valid': True,
    `reason': [*left/right*] keypress,
    `duration': current_time - start_time)

listener.report(
    `valid': False,
    `reason': no_keypress,
    `duration': current_time - start_time)

3.8 Pupil Labs Invisible

Where the previous sections contained information on how the Python code controlled the calibration and experiment, the Pupil Labs Invisible are instead controlled by the Pupil Labs Companion app on the Companion device. A scan video is taken of the robot in order to facilitate calibration to map the 3D environment. After placing the Pupil Labs Invisible on the face of a participant, they are instructed to look at the nose of the Pepper. The Companion device then maps this point as a reference image for calibration for each participant, together with their eye-height measured during the briefing [1]. The experimenters will then click on `start recording' on the Companion device to track and record the human gaze until the calibration and experiment are over. The recording of the human gaze will then stop once the experimenters click on `stop recording'.

4 Technical issues

This chapter contains some of the issues that are encountered during the setup of the experiment.

4.1 Robot failing to execute working modalities

The first issue encountered is that the robot sometimes failed to execute programmed behaviours (e.g., raising the arms). Interestingly, the exact same code executed before and after, without any clear indication. Anecdotal evidence showed that moving the robot around in a circle allowed the arms to be raised. Concurrently, another issue encountered was a weak network connection, causing the code to get stuck while awaiting a response from the robot, leading to the robot not executing the modality at all. This is a problem because the calibration and experiment then stalls, leading to loss of data. Initially the distance from the robot to the router was quite large, and reducing the distance to the router seemed to improve the stability of the execution of the code.

4.2 Internal RAM memory overflowing

A second issue encountered is that capturing and saving frames causes the computer's internal RAM memory to overflow. This overflow causes the computer to freeze and requires a reboot by holding down the power button. Additionally, creating a queue for saving frames in the main thread causes a delay in the execution of other parts of the code. This delay is caused because saving a frame to the disk takes a few milliseconds, which locks the main thread from executing code. The issue of overflowing the internal RAM memory arises due to the code being designed sequentially in the main thread. Using a queue in a separate thread to save frames did cause the problem of locking the thread to disappear, but the overflow of internal RAM memory remained. In order to find a method to also prevent overflowing the internal RAM memory, a creative solution is used in which each frame is saved in a new separate thread. This new thread will then start in parallel to the main thread and save the frame to the disk. It will then terminate itself, clearing up the internal RAM memory for new threads.

[1] For further details on calibration, see Section 5.4.

4.3 Aligning modality execution

Finally, the third issue encountered during the setup of the experiment is a lack of behavioural consistency. An example of this is that displaying the left-arrow takes longer than the right-arrow. Additionally, moving the arms takes longer than displaying an arrow on the tablet. Interestingly, there seems to be a systematic delay in the execution of the modalities. An example of this systematic delay is that the speech occurs near instantly, but that gestures will only occur after about 615-milliseconds [2]. Given that the current study is about multi-modality, the start time of the modalities has to be aligned. Reaching precise delays proved challenging due to the instability of the execution of the modalities, as shown by the mean and standard deviation in Table 2.

A systematic delay has to be implemented to ensure that the experimental trials can be measured equally. The added delay is implemented by recording the modalities and then manually fine-tuning the intensity of the delay. Delays are added to the tablet and the speech, as these modalities appeared fastest, as shown in Table 2. Manual annotation shows that the tablet occurs slightly faster when combined with either speech or gesture, while the timing is evenly shared for the gesture-speech condition [3].

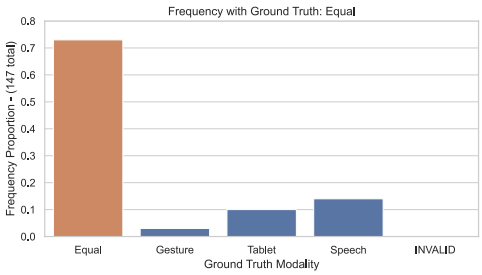
With the systematic delays added, a test is designed to measure whether participants are able to distinguish if one modality occurs faster than the other, or if they appear equal. The trials for the test are taken from 48 trials of a randomized 120 trial experiment, and are tested on seven participants who did not partake in the main experiment. Participants are given the options `Equal', `Gesture', `Tablet', and `Speech' for each trial. The start of the modalities is the ground truth of the trials and is measured using a 120 frames per second video recording. The ground truth is defined as the onset of the sound-wave (for speech), the first visual change in the video frame (for tablet and gesture), or both.

Modality	**Mean (ms)**	**Sample Standard Deviation (ms)**	**Added Delay (ms)**
Left arm	615.00	14.84	-
Right arm	616.67	21.59	-
Left arrow	475.00	35.21	140
Right arrow	527.22	24.89	090
Speech	037.78	16.63	577

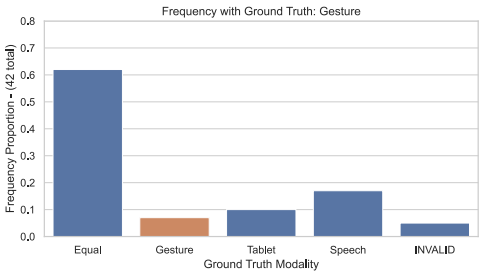
Table 2: Mean and sample standard deviation of fifteen samples of the execution of the modality (in milliseconds)

The measurements of the test are displayed based on their relative proportions in Figure 1. Five trials are marked as `invalid' due to a reported response that is not shown in the trial. The measurements show that `Equal' is the most prominent response, between 60% and 75% of the time in all conditions (1a, 1b, 1c). Participants seem unable to distinguish the gesture and the tablet being faster than the other modality, and instead mostly report `Equal' (1b, 1c). Additionally, participants appear unable to distinguish between an equal and a non-equal trial, indicating that the misalignment in the execution of the modalities has been resolved by adding the systematic delays.

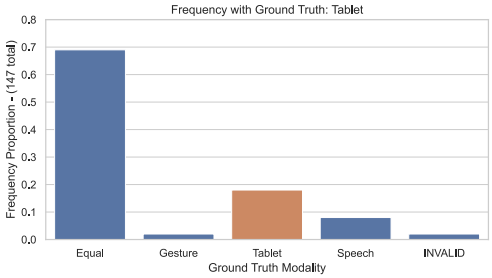
[2] For the collected data, see Appendix C.
[3] For the manual annotation, see Appendix D.



(a) Proportions of `Equal' ground truth.



(b) Proportions of `Gesture' ground truth.



(c) Proportions of `Tablet' ground truth.

Fig. 1: Responses to each of the modalities. Orange bars represent matched ground truth.

5 Methods

5.1 Participants

Thirty four participants are recruited for the current study, mostly at the Vrije Universiteit Amsterdam, as well through an international chat-group of Chinese people living in Amsterdam. While no demographic data has been recorded, all participants were estimated to be between 18 and 65. The main requirement posed upon the suitability of participants is that they cannot wear vision correcting glasses. These participants were excluded since an undistorted view onto the eye is more accurate [24]. The use of contact lenses was allowed.

5.2 Experimental Design

There are two halves of the experiment (glasses and no glasses), two congruency conditions (congruent and incongruent), three modalities (tablet, speech, and gesture), two orders of modality (focus and secondary), and five repetitions per unique trial, leading to 24 unique experimental trials. The current study is designed as an within-subjects groups with a randomized order. To prevent the order effect, all trials are repeated five times and shuffled per half of the experiment. This results in 240 trials per participant. Additionally, half of the the order of the two halves of the experiment with and without eye-tracking glasses are counter-balanced.

5.3 Materials

In order to complete these experiments, a physical Pepper robot (version v1.8) was used. The robot used its internal 2D camera, and had a separate high resolution Logitech Brio camera mounted on top of its head, which provided a higher resolution and frame rate compared to the Pepper's internal 2D camera. Additionally, external Pupil Labs Invisible eye-tracking glasses were used (version Pupil Labs p1-ow1). These three cameras all ran using a specific frame rate and resolution. Where all cameras ran a mostly stable frame rate, these were all unique, with the frame rate of the Pepper's internal 2D camera stabilizing at fourteen frames per second, with a resolution of 640x480 pixels. The Logitech Brio was placed at a mostly stable thirty frames per second with a higher resolution of 1920x1080 pixels. The Logitech Brio was running at a lower frame rate and resolution than it can handle due to the size and storage of the frames, as well as being more similar to the frame rate and resolution of the Pepper's internal 2D camera. This lower frame rate reduces the gap between the two cameras.

In contrast to the Pepper's internal 2D camera and the Logitech Brio, the data from the Pupil Labs Invisible was extracted at a frame rate of two hundred frames per second using the Pupil Invisible Companion app from Pupil Labs, which was running on the Companion phone (OnePlus) [25]. The Pupil Labs Invisible used a small resolution of 192x192 pixels and only captured the eyes, which is possible due to the close proximity to the eyes.

5.4 Procedure

The procedure can be divided into seven parts: a briefing, a practice session, a calibration set, the first half of the experiment, another calibration set, the remaining half of the experiment, and finally the debriefing. The division of two halves of the experiment was due to half of the experiment being with and half without eye tracking glasses, which are randomized to control for order-bias.

Interaction setup The experiments were executed in a secluded room with the participant standing 1.5-meter away from the robot, as shown in Figure 2a, with the point-of-view from the robot in Figure 2b, and the point-of-view of the human in Figure 3. The participant was standing in front of the robot behind a small desk with a keyboard on top. This keyboard was used by the participant to respond to every trial.

Briefing Right before the experiment starts, a participant arrived to the Social AI Lab at the 11th floor of the NU building at the Vrije Universiteit Amsterdam. At the Social AI Lab, the participant was taken to the room in which the experiment took place, and provided with the instructions on how to perform the experimental tasks. To record that they agree with the experimental setup, the participant is asked to read pre-experiment instructions, and was asked to read and sign a consent form, which allows the researchers to collect the necessary data. Additionally, the vertical distance, also called height, from the eyes of a participant to the floor was measured. During this time, participants were invited to ask questions if they needed clarification.

The current study instructs the participants that the participants' egocentric perspective is `left' or `right' (`your left/right'). This is important for the calibration as otherwise it may cause the calibration to be inverse for the two markers on the elbows.

(a) Interaction setup from the side.

(b) Point-of-view of the robot.

Fig. 2: Side and frontal view of the interaction.

Practice session In order to aid participants with understanding their task in this experiment, a participant first goes through a small, randomized set of five trials. Participants will be told by the experimenters and the robot that the robot will provide feedback on whether the correct or incorrect key is pressed or if the participant took too long to respond during this practice session. Anecdotal evidence from participants to the researchers suggests that several participants benefited from this practice session, mainly through the participants gaining a better understanding of their task when they provided incorrect responses.

Calibration After completion of the practice session, calibration was required to reduce the calibration error. A participant was instructed to not remove the eye tracking glasses unless instructed, as doing so would invalidate the calibration, and the trials up until that point. This would then lead to requiring another set of calibration and restarting the trials. When a participant confirmed that they were ready for calibration and the experiment, the calibration started. During the calibration, the robot `woke up' and started a calibration sequence in which the robot raised both arms, displayed a neutral white screen on the tablet and asked the participant to stand behind the desk. The visualization of this interaction setup can be seen in Figure 2a.

When the participant was standing behind the desk, the robot asked the participant to look at five calibration markers on the robot, each separated by four seconds: the head camera, the nose, the tablet, the left elbow and the right elbow. These five points were marked on the robot with a one-centimeter square black marker, serving as the focus point for calibration to calculate the ground truth of their gaze, which was used for shift correction of the calibration error. The black markers and stance of the robot can be seen in Figure 3. This part was executed for both halves of the experiment. When completed, the experimenters confirmed that the data was saved successfully.

Experiment After calibration, the experimenters asked a participant if they were ready for half of the experiment. When the participant confirmed this, the experimenters started the experiment. During each half of the experiment, 120 trials were generated, shuffled and placed in a queue.

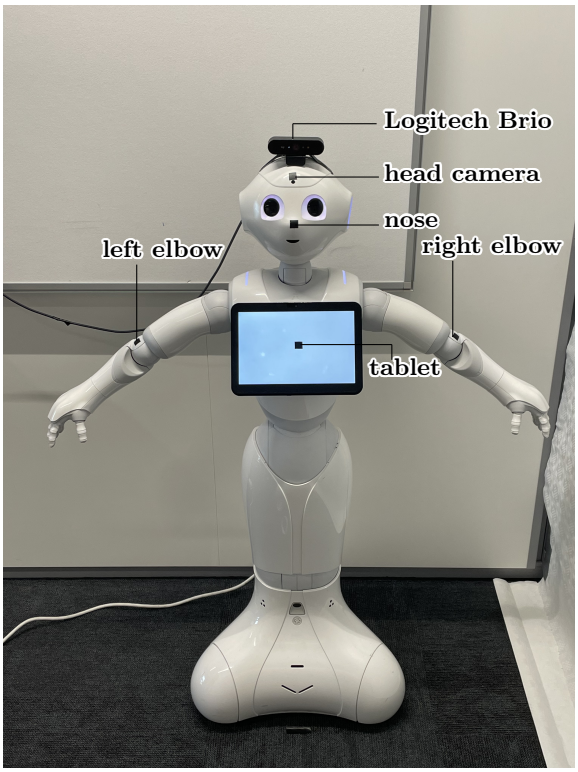
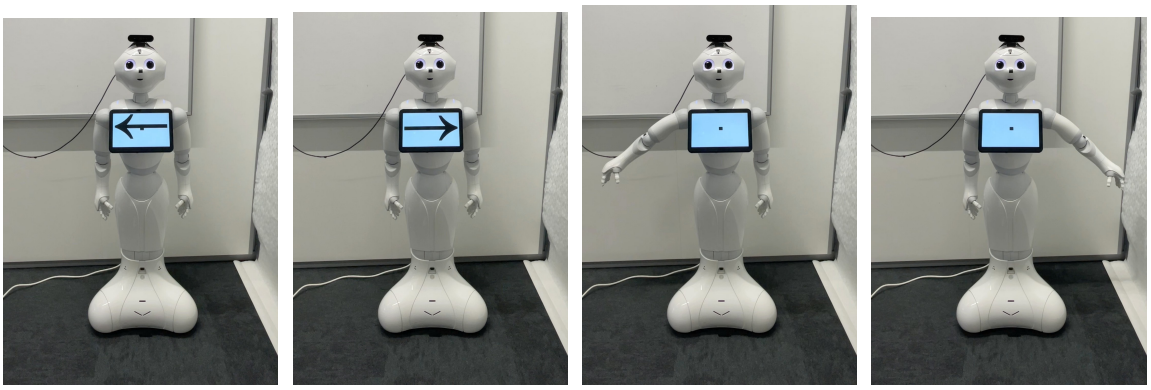


Fig. 3: Point of view of a participant on the robot during calibration; includes the five calibration markers and the externally head-mounted Logitech Brio.

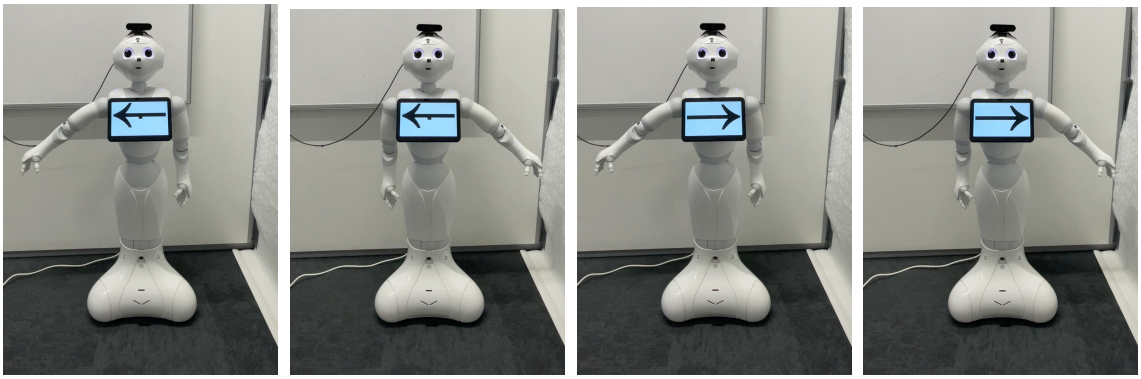


(a) Tablet Left + Speech Left/Right

(b) Tablet Right + Speech Left/Right

(c) Gesture Left + Speech Left/Right

(d) Gesture Right + Speech Left/Right



(e) Tablet Left + Gesture Left

(f) Tablet Left + Gesture Right

(g) Tablet Right + Gesture Left

(h) Tablet Right + Gesture Right

Fig. 4: All visual combinations of modalities for the trials.

The robot instructed a participant by saying: *"Get Ready! Please focus on [modality]"*, after which one of three modalities was pronounced. This pronounced modality is called the focus modality. The goal of a participant was to respond with the same direction as the focus modality instructed by the robot. At the same time, the robot retrieved a second modality with either a congruent or incongruent direction. The robot then had a 615-millisecond delay to prepare and execute the modalities simultaneously. The task of the participant was to then press the direction of the modality as instructed by the robot, without being distracted by the other modality.

An example of a trial was that a participant received the focus modality `Tablet', the secondary modality `Gesture' and the incongruent condition. The robot then did the following: Instruct a participant by saying *"Get Ready! Please focus on tablet"*, after which the robot showed a left arrow on the tablet, and simultaneously activated the gesture modality to raise the `*Right*' arm. What the participant visually saw, can be the trial seen in Figure 4f. A participant was then expected to press on the left key on the keyboard, as that was the direction shown on the focus modality (the tablet). Different to the practice session received earlier by the participant, the robot would not provide feedback to the participant about the correctness of their key pressed.

Debriefing When participants finished the second half of the experiment, the experimenters debriefed the participants by thanking and providing them with a reward of a ten euro Bol.com gift card. If requested by the participants, the goal of the study was elaborated.

6 Results

As a data preprocessing step, the data of responses to trials from four participants had to be excluded due to their data becoming invalid. The data from one participant was excluded due to the researchers forgetting to click `start recording' on the Companion device (participant 8), and the other three due to errors during saving and merging .csv data (participants 28, 31, and 34). From the remaining thirty participants, the responses of four trials were marked as invalid and removed due their response time being before the execution of the modality. This is due to the parallelization and the added delay, which enables a response to be registered before the modality is actually executed.

Due to technical failures, the data of the Pupil Labs Invisible from four participants is unusable, with the gaze estimation of participant 8 and 13 being excluded due to overheating of the eye tracking glasses, participant 11 for failing in mapping, and participant 15 for forgetting to press record.

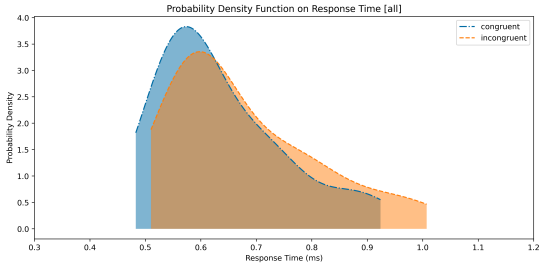
6.1 Congruent cues versus incongruent cues

As visually seen in Figure 5a, the response time for the congruent and incongruent conditions do not appear to differ much from each other, with congruent cues ($M = 482.4$ms, $SD = 161.5$ms, $Mdn = 416.8$ms) processed slightly faster compared to incongruent cues ($M = 527.7$ms, $SD = 183.9$ms, $Mdn = 450.2$ms). Shapiro-Wilk test has been executed on the congruent and incongruent datasets to check the assumption of normality. The Shapiro-Wilk test showed that neither the congruent ($W = .837$, $p < .001$) nor incongruent ($W = .829$, $p < .001$) dataset were normally distributed. As such, a Wilcoxon test is executed, resulting in a significant difference between the congruent and incongruent condition ($z = 10.0$, $p < .001$).

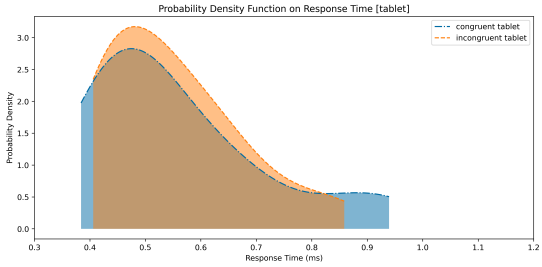
Modality	Congruency	Mean ms	Std. Dev. ms	Median ms
Tablet	Congruent	401.6	436.5	290.6
	Incongruent	420.3	379.8	305.7
Speech	Congruent	528.8	231.2	475.3
	Incongruent	584.7	284.2	519.5
Gesture	Congruent	517.0	255.9	450.5
	Incongruent	578.2	286.5	497.4

Table 3: Statistical Data of Response Time over Congruency and Modality

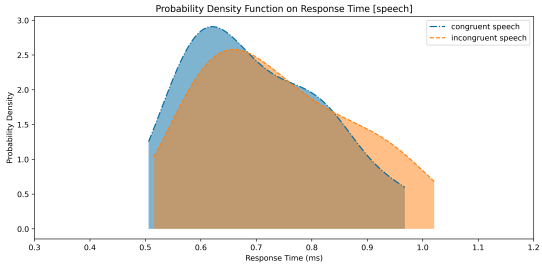
When comparing the response time of each individual modality, Table 3 shows that a systemic faster response time also occurs regardless of the modality. To confirm that this is also statistically accepted, a test for normality over all modalities first has be executed. Shapiro-Wilk has been executed and that normality cannot be assumed for any of the conditions as all p-values are $< .001$, except for the congruent-speech condition ($W = .936$, p .075). A Wilcoxon test is executed, with the results showing that for the tablet ($z = 143.0$, $p = .067$) indicate that there is no significant difference in response time between the congruency of the tablet. For the speech ($z = 24.0$, $p < .001$) and gesture ($z = 6.0$, $p < .001$), there is a statistical significant between the congruency for these modalities.



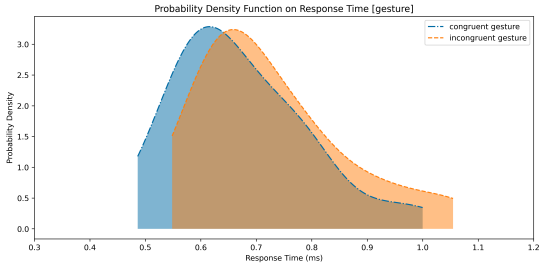
(a) Probability Density Function of All Modalities Combined.



(b) Probability Density Function of the Tablet Modality.



(c) Probability Density Function of the Speech Modality.



(d) Probability Density Function of the Gesture Modality.

Fig. 5: Probability Density Functions over the Response Time per Modality

6.2 Gaze attention to area of interest

The Pupil Labs Invisible eye tracking glasses are used to investigate gaze onto areas of interest on the tablet, head and gestures. These areas of interest have been defined in conjunction with Linlin Cheng to represent areas of interest similar to the reference paper of Özer *et al.* [23]. The area of interest of the head is the same as the area of interest of the reference study. The central gesture space from the reference study was divided into three vertical parts. The outer areas of interest define the gestural space, and the middle part is cropped to only fit the outer bounds of the tablet. The fixations from the Pupil Labs Invisible are mapped and placed inside the coordinates of an area of interest, or marked as `Other'.

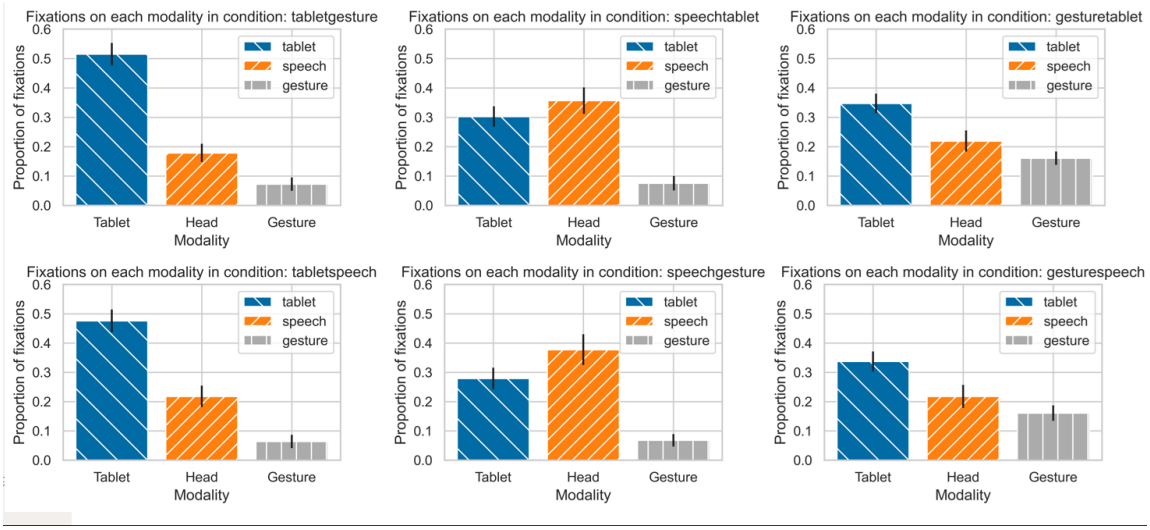


Fig. 6: Proportions of gaze fixations over all modality conditions.

The plots in Figure 6 show that the face receives more fixations in any type of modality condition over the gestures. Additionally, there is a big increase in the area of interest of the focus modality. The fixations on the area of interest on the tablet are between 28% and 35% when the tablet is not the focus modality. When the tablet is the focus modality, this proportion of fixations jumps up to 50% of the time. The area of interest of the head also moves from 20% to around 38%. The gestures also go through a similar increase, from around 5% to 15%. The stable factor in these results is an increase of 10% to 20%. Additionally, the remaining proportions, which are fixations that do not fall within the tablet, head, and gesture areas of interest, is also steady, with approximately 27% of fixations.

A Shapiro-Wilk test was conducted to check for the assumption of normality to determine an appropriate statistical method. The results from the Shapiro-Wilk test show that the speech and gesture modalities depart significantly from normality, with the exception of speech in the speech-tablet condition. Due to these results, a Wilcoxon test was conducted to check if the speech and gesture modalities are indeed statistically significantly different. The results from the Wilcoxon test in Table 4 show that when the tablet or the head is the focus modality, there is a significant

difference. However, when the gesture is the focus modality, there is no significant difference.

To compare the fixations per trial on a modality being the focus modality to when that modality is not the focus modality, a Shapiro-Wilk test was conducted to test for the assumption of normality. The Shapiro-Wilk test indicated that normality cannot be assumed for any of the modalities (all $p < .001$). To test whether there is a significant difference, a Wilcoxon test was used to see if a modality has more fixations when this is the focus modality compared to when it is not. The results of the Wilcoxon test can be seen in Table 5, all resulting in the null hypothesis not being rejected.

Condition	**Z-statistic**	**p-value**
tablet-speech	57	$< .001$
tablet-gesture	98	.004
speech-tablet	48	$< .001$
speech-gesture	53	$< .001$
gesture-tablet	197	.477
gesture-speech	197	.477

Table 4: Wilcoxon Test Results for fixations on head and gesture

AOI	**Compared to**	**Z-statistic**	**p-value**
Tablet	Focus - secondary	8065	.909
	Focus - not in trial	7633	.465
Head	Focus - secondary	7772	.594
	Focus - not in trial	7430	.307
Gesture	Focus - secondary	7782	.604
	Focus - not in trial	8003	.839

Table 5: Results of Wilcoxon tests of a modality on different AOIs as focus modality, non-focus modality, and as an unused modality.

7 Discussion

The current study investigated how humans process and gaze at instructions given by a robot when prompted with congruent and incongruent cues using different combinations of modalities. Specifically, the current study looked at a) the Stroop effect on the combination of modalities used and b) whether humans gaze more at a robot's body part when instructed to.

7.1 Stroop effect

Hypothesis 1: Congruent cues of the focus modality will be processed faster than incongruent cues (Stroop effect).

In line with expectations, humans mostly respond faster to congruent cues compared to incongruent cues given by a robot. However, while the speech and gesture have a faster response time, the tablet does not have a significant faster response time. Interestingly, the tablet does have response times around 100ms faster compared to the gesture and speech modality. This may follow from participants fixating at the tablet, as this modality received most fixations at most of the trials. Due to the tablet not being statistically significantly faster in the congruent condition as compared to the incongruent condition, this hypothesis cannot be accepted.

7.2 Gaze at the body parts

Hypothesis 2: The face will receive more fixations during each trial than the gestures.

The face receives more fixations than gestures during each trial. This result matches with the results found in the reference paper by Özer *et al.* [23]. However this is only statistically different when not instructed to focus on the gestures. Additionally, when the focus modality was the tablet, the head also received more fixations than the gestural area. Since speech cues comes from the head, the tablet is a magnet for attraction, and gestures are mainly performed in the central gesture space [14]. This may explain the decreased amount of fixations to the gesture space compared to the head. Interestingly, the results shown in Chapter 6 indicate that the gaze of participants was at the head of the robot in 20% of all trials, even when no visual cues were provided there, even reaching 35% when speech was the focus modality. This may indicate that humans gaze at the head of Pepper to receive cues.

Hypothesis 3: The focus modality will receive more fixations during each trial compared to when that modality is not the focus modality (task compliance).

While the results of the current study shows that when humans are cued to look at specific robot's body parts, they seem to follow these instructions, without being distracted by the secondary modality. The focus modality sees an increase of more than 10% for each modality compared to when this area of interest is a non-focus modality. The non-cued modality only increases, decreases and stabilizes within 5% of the secondary modality. However, the Wilcoxon tests indicate that no significant differences are shown between the proportion of fixations, regardless of whether a modality is the focus modality or not the focus modality. This hypothesis hence can also not be accepted.

7.3 Pupil Labs Invisible & L2CS

At the time of writing this thesis, the data from L2CS is still being processed, leading to the results shown for Hypothesis 2 and Hypothesis 3 being only from the Pupil Labs Invisible. While the performance of the L2CS model is not part of the current thesis, it is interesting to see if the fixations from this model align with the ground truth of the Pupil Labs Invisible.

8 LEDs

In addition to checking for differences in the combination of auditory and visual changes, a research setup is created to check if the type of facial cue matters. This setup follows earlier work that shows that humans primarily gaze at the human face during initial contact [19], that the design and expressions of the face of a robot can be used to display emotions [17], and that the perception of speech improves through visual cues from facial features [8]. Additionally, Edirisinghe *et al.* have implemented six basic emotions using eye brows, eye lids and a mouth using 3D prototyping of a robot head [10]. The main idea behind the proposed study is to investigate whether using the LEDs of the Pepper robot will have an influence in behavioural changes in human perception of the robot. A study by Gullberg can be taken as inspiration, which uses gestures and oral cues to provide matching and incongruent conditions [14]. The underlying research question here is whether the Pepper can use its LEDs in its eyes, ears, and turning its head, to convey additional or improved cues during an interaction. The study proposed in this chapter can be tested by having a robot tell a story and then changing its head and LEDs during certain key-points. During these key-points, the robot gives a signal to the human by changing three conditions (eye-LEDs, ear-LEDs, and head-turning). Examples of a condition can be a different color, turning them on or off, or moving

the head away from the human. These conditions can all be tested using the LEDs of the Pepper robot, for example with a gaze following task in which the Pepper uses its LEDs to suggest gaze shifts. However, due to the complex combination of auditory and visual cues already generated by the main study described in this thesis, as well as due to time and budgeting constraints, this work was unable to be executed during this thesis. Nonetheless, this setup can still serve as a basis for future work. This proposed study can be another improvement for the design of social robots.

9 Concluding

In the current study, I have investigated human gaze behaviour, and in particular gaze behaviour onto various robot body parts using different modalities. The research question `*How do congruent and incongruent multi-modal cues from a social robot influence human gaze behaviour?*' has been answered by demonstrating that congruent cues are processed faster than incongruent cues when prompted to focus on speech or gestures, and that the face receives more fixations than the gestures, when not prompted to focus on the gestures. Additionally, no significant difference can be found between when a modality is the focus modality or not the focus modality. The method of how to convey information or give instructions then seems to impact how humans respond to a social robot. Researchers should thus consider how to design social robots on how they want the robot to perceived.

In addition to these findings, it can be fruitful to explore the avenue of adding more facial features to see if the type of facial cue has an impact on gaze behaviour. Additionally, the data from the L2CS model is still being processed, which may enable eye tracking without requiring external hardware.

10 Acknowledgements

The work described in this thesis is fully created by the author, Mark de Bruijn. The only exceptions to this, are the definition of the AoI's, as well as the data preprocessing step of the Pupil Labs Invisible, which was done by Linlin Cheng.

I want to express my deepest gratitude to my professors and supervisors Koen Hindriks and Artem Belopolskiy, as well as Linlin Cheng. Their support and tools have enabled this thesis to be possible. I also want to thank my colleagues at the Social AI Lab for their invaluable assistance. Additionally, I want to thank my dear family and friends for their support and being there for me along this incredible journey.

Als ze willen, komen ze er wel.

References

[1] Ahmed A. Abdelrahman et al. "L2CS-Net: Fine-Grained Gaze Estimation in Unconstrained Environments". In: *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*. IEEE, Oct. 2023. DOI: 10.1109/icfsp59764.2023.10372944. URL: http://dx.doi.org/10.1109/ICFSP59764.2023.10372944.

[2] Henny Admoni and Brian Scassellati. "Social Eye Gaze in Human-Robot Interaction: A Review". In: *J. Hum. Robot Interact.* 6.1 (Mar. 2017), p. 25. URL: https://doi.org/10.5898/JHRI.6.1.Admoni.

[3] Onur Asan and Yushi Yang. "Using Eye Trackers for Usability Evaluation of Health Information Technology: A Systematic Literature Review". In: *JMIR Human Factors* 2.1 (Apr. 2015), e5. ISSN: 2292-9495. DOI: 10.2196/humanfactors.4062. URL: http://dx.doi.org/10.2196/humanfactors.4062.

[4] Anna Belardinelli. "Gaze-based intention estimation: principles, methodologies, and applications in HRI". In: *ACM Transactions on Human-Robot Interaction* (Apr. 2024). ISSN: 2573-9522. DOI: 10.1145/3656376. URL: http://dx.doi.org/10.1145/3656376.

[5] T. Blascheck et al. "Visualization of Eye Tracking Data: A Taxonomy and Survey". In: *Computer Graphics Forum* 36.8 (Feb. 2017), pp. 260–284. ISSN: 1467-8659. DOI: 10.1111/cgf.13079. URL: http://dx.doi.org/10.1111/cgf.13079.

[6] Mark de Bruijn. *Gaze Following*. https://github.com/FredAlfabetAdmin/Gaze-Following. 2023.

[7] Benjamin T. Carter and Steven G. Luke. "Best practices in eye tracking research". In: *International Journal of Psychophysiology* 155 (Sept. 2020), pp. 49–62. ISSN: 0167-8760. DOI: 10.1016/j.ijpsycho.2020.05.010. URL: http://dx.doi.org/10.1016/j.ijpsycho.2020.05.010.

[8] Zubin Datta Choudhary, Gerd Bruder, and Gregory F. Welch. "Visual Facial Enhancements Can Significantly Improve Speech Perception in the Presence of Noise". In: *IEEE Transactions on Visualization and Computer Graphics* 29.11 (Nov. 2023), pp. 4751–4760. ISSN: 2160-9306. DOI: 10.1109/tvcg.2023.3320247. URL: http://dx.doi.org/10.1109/TVCG.2023.3320247.

[9] Sabrina Dunham, Edward Lee, and Adam M. Persky. "The Psychology of Following Instructions and Its Implications". In: *American Journal of Pharmaceutical Education* 84.8 (Aug. 2020), ajpe7779. ISSN: 0002-9459. DOI: 10.5688/ajpe7779. URL: http://dx.doi.org/10.5688/ajpe7779.

[10] E.A.N.S. Edirisinghe et al. "Design of a human-like robot head with emotions". In: *2016 Moratuwa Engineering Research Conference (MERCon)*. 2016, pp. 421–426. DOI: 10.1109/MERCon.2016.7480178.

[11] N.J. Emery. "The eyes have it: the neuroethology, function and evolution of social gaze". In: *Neuroscience amp; Biobehavioral Reviews* 24.6 (Aug. 2000), pp. 581–604. ISSN: 0149-7634. DOI: 10.1016/s0149-7634(00)00025-7. URL: http://dx.doi.org/10.1016/S0149-7634(00)00025-7.

[12] Hong Fu et al. "Advances in Eye Tracking Technology: Theory, Algorithms, and Applications". In: *Computational Intelligence and Neuroscience* 2016 (2016), pp. 1–2. ISSN: 1687-5273. DOI: 10.1155/2016/7831469. URL: http://dx.doi.org/10.1155/2016/7831469.

[13] Tobias Grossmann. "The Eyes as Windows Into Other Minds: An Integrative Perspective". In: *Perspectives on Psychological Science* 12.1 (Jan. 2017), pp. 107–121. ISSN: 1745-6924. DOI: 10.1177/1745691616654457. URL: http://dx.doi.org/10.1177/1745691616654457.

[14] Marianne Gullberg and Kenneth Holmqvist. "Keeping an eye on gestures: Visual perception of gestures in face-to-face communication". In: *Pragmatics amp; Cognition* 7.1 (1999), pp. 35–63. ISSN: 1569-9943. DOI: 10.1075/pc.7.1.04gul. URL: http://dx.doi.org/10.1075/pc.7.1.04gul.

[15] Dan Witzner Hansen and Qiang Ji. "In the Eye of the Beholder: A Survey of Models for Eyes and Gaze". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.3 (2010), pp. 478–500. DOI: 10.1109/TPAMI.2009.30.

[16] Almoctar Hassoumi, Vsevolod Peysakhovich, and Christophe Hurter. "Improving eye-tracking calibration accuracy using symbolic regression". In: *PLOS ONE* 14.3 (Mar. 2019). Ed. by Jinjun Tang, e0213675. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0213675. URL: http://dx.doi.org/10.1371/journal.pone.0213675.

[17] Frank Hegel, Friederike Eyssel, and Britta Wrede. "The social robot Flobi: Key concepts of industrial design". In: *19th International Symposium in Robot and Human Interactive Communication*. IEEE, Sept. 2010. DOI: 10.1109/roman.2010.5598691. URL: http://dx.doi.org/10.1109/ROMAN.2010.5598691.

[18] Roy S. Hessels. "How does gaze to faces support face-to-face interaction? A review and perspective". In: *Psychonomic Bulletin amp; Review* 27.5 (May 2020), pp. 856–881. ISSN: 1531-5320. DOI: 10.3758/s13423-020-01715-w. URL: http://dx.doi.org/10.3758/s13423-020-01715-w.

[19] Johannes Hewig et al. "Gender Differences for Specific Body Regions When Looking at Men and Women". In: *Journal of Nonverbal Behavior* 32.2 (Mar. 2008), pp. 67–78. ISSN: 1573-3653. DOI: 10.1007/s10919-007-0043-5. URL: http://dx.doi.org/10.1007/s10919-007-0043-5.

[20] Kasper Hornbæk and Antti Oulasvirta. "What Is Interaction?" In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. ACM, May 2017. DOI: 10.1145/3025453.3025765. URL: http://dx.doi.org/10.1145/3025453.3025765.

[21] Colin M. MacLeod. "The Stroop Effect". In: *Encyclopedia of Color Science and Technology*. Springer Berlin Heidelberg, 2015, pp. 1–6. ISBN: 9783642278518. DOI: 10.1007/978-3-642-27851-8_67-1. URL: http://dx.doi.org/10.1007/978-3-642-27851-8_67-1.

[22] Marcus Nyström et al. "The influence of calibration method and eye physiology on eyetracking data quality". In: *Behavior Research Methods* 45.1 (Sept. 2012), pp. 272–288. ISSN: 1554-3528. DOI: 10.3758/s13428-012-0247-4. URL: http://dx.doi.org/10.3758/s13428-012-0247-4.

[23] Demet Özer et al. "Gestures cued by demonstratives in speech guide listeners' visual attention during spatial language comprehension". en. In: *J. Exp. Psychol. Gen.* 152.9 (Sept. 2023), pp. 2623–2635. URL: https://doi.org/10.1037/xge0001402.

[24] papr. *Comment on the issue "Eye camera fails to initialize #724"*. Accessed: 2024-06-21. May 2017. URL: https://github.com/pupil-labs/pupil/issues/724#issuecomment-301704364.

[25] *Pupil Invisible - Eye tracking glasses for the real world*. en. https://pupil-labs.com/products/invisible. Accessed: 2024-6-13.

[26] Yashas Rai and Patrick Le Callet. "Chapter 3 - Visual attention, visual salience, and perceived interest in multimedia applications". In: *Academic Press Library in Signal Processing, Volume*

6. Ed. by Rama Chellappa and Sergios Theodoridis. Academic Press, 2018, pp. 113–161. ISBN: 978-0-12-811889-4. DOI: https://doi.org/10.1016/B978-0-12-811889-4.00003-8. URL: https://www.sciencedirect.com/science/article/pii/B9780128118894000038.

[27] SoftBank Robotics. *SOFTBANK ROBOTICS DOCUMENTATION - PEPPER DOCUMENTATION"*. Accessed: 2024-06-22. Aug. 2020. URL: http://doc.aldebaran.com/2-8/home_pepper.html.

[28] A. Villanueva and R. Cabeza. "A Novel Gaze Estimation System With One Calibration Point". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 38.4 (Aug. 2008), pp. 1123–1138. ISSN: 1083-4419. DOI: 10.1109/tsmcb.2008.926606. URL: http://dx.doi.org/10.1109/TSMCB.2008.926606.

A Recruitment Poster

Fig. 7: Recruitment Poster placed within the Vrije Universiteit Amsterdam and shared in the international chat-group of Chinese in Amsterdam.

B Recruitment Message

Gaze Tracking Experiment
Description: In this experiment, you will interact with a cute social robot that tracks your eye gaze.
Duration: 35 minutes
Place: Room 11A56, NU building #11
Payment: 10 euro
Requirement: Do not wear spectacles; contact lenses are welcome.
Sign-Up Information: If you are interested in joining this experiment, please visit the following website: *https://form.everestwebdeals.co/?form=2775697f4ec1d6112f38edaffe7771ee* for info to sign up.

C Modality Alignment Raw Data

Category	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Left arm	650.00	616.67	616.67	608.33	608.33	608.33	633.33	616.67	616.67	600.00	608.33	633.33	616.67	600.00	591.67
Right arm	633.33	633.33	608.33	608.33	675.00	600.00	600.00	583.33	616.67	625.00	616.67	633.33	608.33	600.00	608.33
Left arrow	516.67	491.67	458.33	466.67	475.00	483.33	466.67	550.00	525.00	483.33	433.33	425.00	433.33	450.00	466.67
Right arrow	541.67	500.00	508.33	533.33	541.67	566.67	508.33	500.00	566.67	500.00	500.00	508.33	550.00	550.00	533.33
Speech	016.67	041.67	025.00	041.67	033.33	033.33	058.33	075.00	066.67	033.33	033.33	025.00	033.33	025.00	025.00

Table 6: Raw data of the duration of the execution of a modality in milliseconds. Cells indicate a single sample.

D Aligning modality execution - Manual experiment

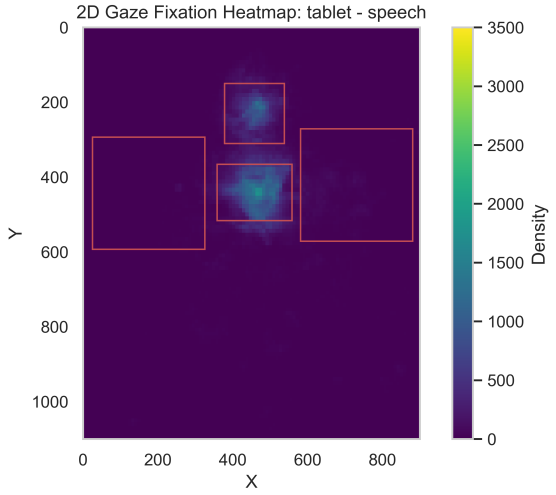
Trial Number	Modality A	Modality B	Fastest Modality
1	Speech right	Gesture right	Gesture
2	Tablet right	Gesture right	Gesture
3	Tablet left	Gesture right	Gesture
4	Speech left	Gesture left	Equal
5	Speech right	Gesture right	Equal
6	Tablet left	Gesture right	Tablet
7	Tablet left	Gesture left	Tablet
8	Speech left	Gesture left	Equal
9	Speech left	Tablet right	Equal
10	Tablet left	Gesture right	Tablet
11	Speech right	Tablet right	Equal
12	Tablet right	Gesture right	Equal
13	Speech right	Gesture left	Equal
14	Tablet left	Gesture right	Tablet
15	Speech right	Tablet left	Tablet
16	Speech left	Tablet right	Tablet
17	Tablet left	Gesture left	Tablet
18	Speech right	Tablet left	Tablet
19	Speech right	Tablet left	Tablet
20	Speech left	Tablet right	Equal
21	Tablet left	Gesture left	Tablet
22	Speech left	Tablet right	Tablet
23	Speech right	Tablet left	Tablet
24	Speech left	Gesture left	Equal
25	Tablet left	Gesture left	Equal

Table 7: Results of trials showing the modalities and their fastest performance (Part 1).

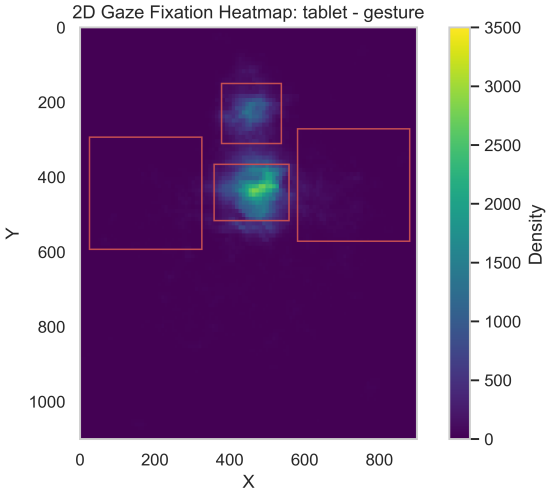
Trial Number	Modality A	Modality B	Fastest Modality
26	Tablet right	Gesture right	Tablet
27	Speech right	Gesture left	Equal
28	Speech left	Gesture right	Equal
29	Speech right	Tablet right	Tablet
30	Speech right	Gesture left	Gesture
31	Speech right	Gesture left	Equal
32	Speech right	Tablet left	Equal
33	Tablet right	Gesture right	Equal
34	Tablet left	Gesture left	Tablet
35	Speech left	Gesture right	Equal
36	Tablet right	Gesture right	Tablet
37	Speech right	Gesture left	Equal
38	Speech left	Tablet right	Equal
39	Speech left	Tablet right	Tablet
40	Speech left	Gesture right	Equal
41	Speech right	Gesture left	Equal
42	Speech right	Tablet left	Tablet
43	Speech left	Tablet left	Tablet
44	Speech right	Gesture left	Gesture
45	Speech right	Tablet right	Tablet
46	Tablet right	Gesture right	Tablet
47	Speech left	Gesture left	Gesture
48	Speech right	Gesture right	Equal
49	Tablet left	Gesture left	Equal
50	Speech right	Gesture right	Equal

Table 8: Results of trials showing the modalities and their fastest performance (Part 2).

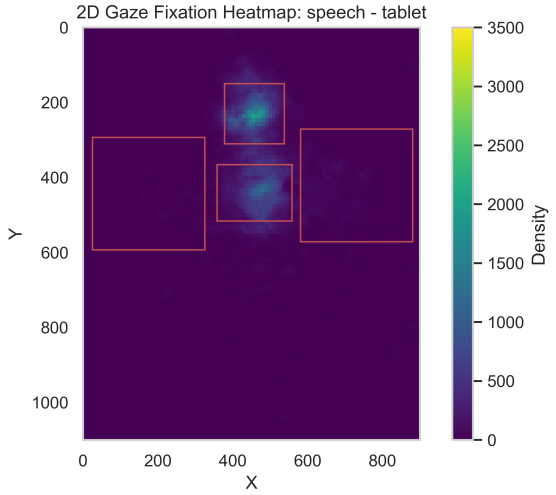
E Gaze Fixation Plots



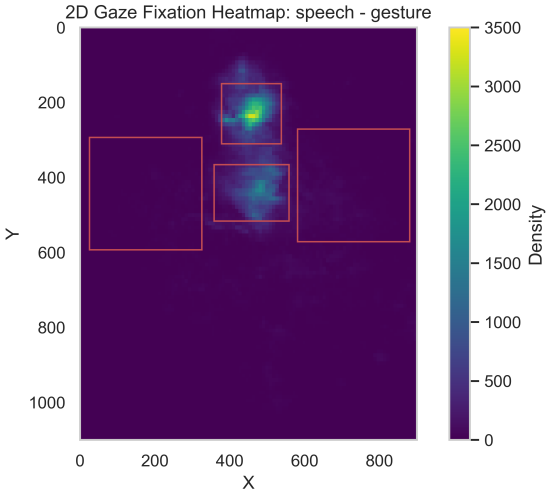
(a) Heatmap Tablet Speech



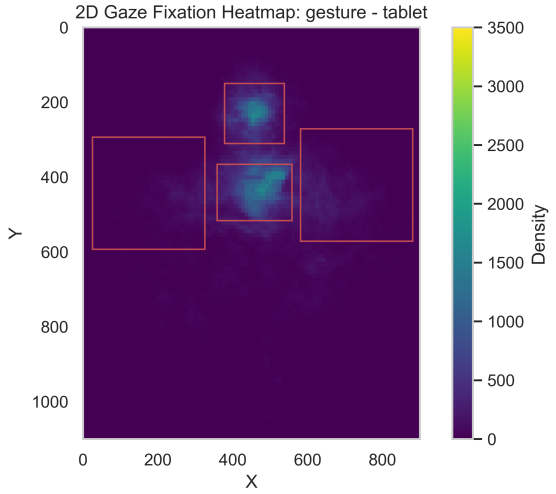
(b) Heatmap Tablet Gesture



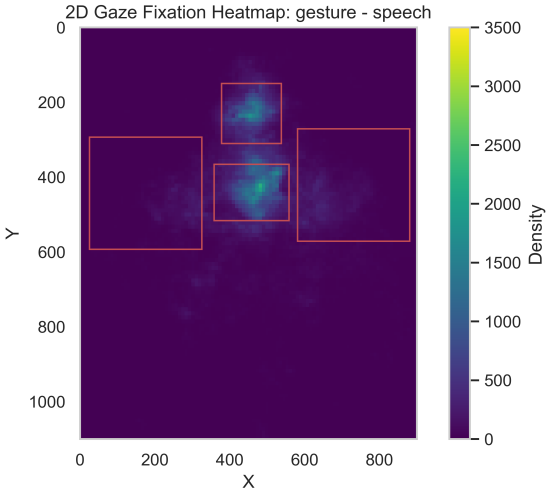
(c) Heatmap Speech Tablet



(d) Heatmap Speech Gesture



(e) Heatmap Gesture Tablet



(f) Heatmap Gesture Speech

F Area of Interest coordinates

AOI	**x min**	**x max**	**y min**	**y max**
head	378	538	150	310
tablet	358	558	366	516
gesture left	25	325	293	593
gesture right	581	881	271	571

Table 9: AOI Boundary Boxes.