IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 17, 2020, accepted October 10, 2020, date of publication November 19, 2020,
date of current version November 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3037724

Robots and Wizards: An Investigation Into Natural Human–Robot Interaction

DOMINYKAS STRAZDAS[1], JAN HINTZ[1], ANNA-MARIA FELßBERG[2], AND AYOUB AL-HAMADI[1]

[1]Neuro-Information Technology, Otto-von-Guericke-University Magdeburg, 39106 Magdeburg, Germany
[2]Department of Experimental Psychology, Otto-von-Guericke Universität Magdeburg, 39106 Magdeburg, Germany

Corresponding authors: Dominykas Strazdas (dominykas.strazdas@ovgu.de) and Ayoub Al-Hamadi (ayoub.al-hamadi@ovgu.de)

This work was supported in part by the Federal Ministry of Education and Research of Germany (BMBF) RoboAssist under Grant 03ZZ0448L, in part by the HuBa under Grant 03ZZ0470, and in part by the Robo-Lab within the Zwanzig20 Alliance 3Dsensation, under Grant 03ZZ04X02B.

ABSTRACT The goal of the study was to research different communication modalities needed for intuitive Human-Robot Interaction. This study utilizes a Wizard of Oz prototyping method to enable a restriction-free, intuitive interaction with an industrial robot. The data from 36 test subjects suggests a high preference for speech input, automatic path planning and pointing gestures. The catalogue developed during this experiment contains intrinsic gestures suggesting that the two most popular gestures per action can be sufficient to cover the majority of users. The system scored an average of 74% in different user interface experience questionnaires, while containing forced flaws. These findings allow a future development of an intuitive Human-Robot interaction system with high user acceptance.

INDEX TERMS Activity recognition, cooperative systems, gesture recognition, human-robot interaction, intelligent robots, interactive systems, robot control, robot learning, telerobotics.

I. INTRODUCTION

The factory of the future, the so-called *smart factory*, relies on intelligent, independently operating and globally networked systems [1]. As part of the fourth industrial revolution (Industry 4.0), a new possibility for production was introduced: Human-Robot Interaction (HRI). Robots are not operated behind a protective barrier as usual, but work in the same space as humans. Three levels of cooperation between humans and robots have been established: coexistence, cooperation and collaboration [2].

To enable collaboration and integrate humans into these smart surroundings, intelligent systems and sensors are required, since humans lack a direct digital interface. Human communication is complex. Only a part of a spoken message depend on the actual words, while the vocal information (tone and other sounds) and nonverbal information (body stance, mimic, gestures), depending on the context, sometimes play the major role, as observed by Albert Mehrabian in his studies [3]. That is why the quality of interaction is crucial and a lot of testing is needed. What is the use of a great expensive system that is not safe or not applicable? The effort of programming as many scenarios as possible would not be economical. One way to overcome this issue is to use the well-established method of Wizard of Oz (WoZ), to evaluate and adapt existing, unfinished systems.

The associate editor coordinating the review of this manuscript and approving it for publication was Maurizio Tucci.

A. WIZARD OF OZ

The WoZ methodology allows the use of technology as a rapid prototype that would be too costly or time consuming to implement otherwise or is generally not available yet [4], [5]. The subjects are tricked into thinking that they interact with an autonomous machine when in reality they are interacting with a human operator instead. For the duration of the experiment, the operator secretly controls the systems actions, unknown to the test subjects. The operator is concealed from the participants by remotely working from another room, or hidden in plain sight, e.g. as an experiment observer, pretending to be taking notes or as a fake participant.

This method has been proven useful throughout many different Human-Machine Interaction (HMI) studies [6]–[10]. It can also be used repetitively to further improve the design of the system [11]. WoZ gives us the opportunity to take a more general approach that facilitates a natural, intuitive HMI.

B. RELATED WORK

Many ideas have been developed to further improve the human interaction with industrial robots [12]–[16]. Among others, they propose augmented reality (AR) concepts to give visual feedback [12], [16] or use external tools as interface [12], [14]. For example, Zaeh & Vogl [12] successfully presented an interactive laser-projection system for programming industrial robots. And although noise in industrial environments is a challenge for automatic speech recognition systems, Pires [13] has shown that it is possible to overcome with the use of headsets.

Serrano and Nigay [10] propose a component-based WoZ-approach for prototyping of multimodal user interfaces. Another approach has been undertaken by Speicher and Nebeling [17] with ''GestureWiz'', a tool that uses the WoZ-technique to quickly model gesture interfaces. Hoffman [18] describes another framework, ''OpenWoZ'', which is specifically made for HRI studies.

We implemented our own WoZ-framework in order to allow a restriction-free, multimodal HRI, since none of the mentioned techniques provided all of the necessary features (speech, gesture, head pose, gaze, posture, touch).

II. STUDY DETAILS

Experiments were conducted by an experimenter who provided guidance to the subjects for the duration of the experiment and a second experimenter who operated the robot, from now on referenced as ''wizard''. The wizard sat in another, well isolated chamber, invisible to the subjects, equipped with two surveillance monitors and headphones, along with a control unit. This was necessary to preserve the illusion of an independently, intelligently acting robot. To improve uniformity and standardization, the same people and roles were used for all experiments. The control unit consisted of a handheld controller (*Dual Shock 4*), a robot control-panel, a keyboard for shortcuts and a second keyboard for direct speech output. Shortcuts were speech feedback, projector control or pre-saved robot-coordinates for fast and precise interaction. The wizard's chamber is shown in fig. 1.

The experimental setup is based on an *UR5e* industrial robot with a *RG6* gripper. The robot was developed by *Universal Robots* with the explicit goal of enabling safe HRI. It was placed on a table in front of a TV-screen, displaying the depth view from a time-of-flight camera together with a gesture recognition which were — apart from the intention to strengthen the illusion — not used.

A projector facing down from the ceiling was used to highlight blocks or positions. Five cameras recorded the experiment and streamed the view in real time to the wizard. The instructions were presented on a separate monitor. The entire system was introduced to the subject as ''RoSA'' (Robot System Assistant).

12 cubes with letters from A-Z and numbers from 0-9 were placed in front of the robot. Half of the cubes were black with white letters and the other half white with black letters. The setup as described is shown in fig. 2.

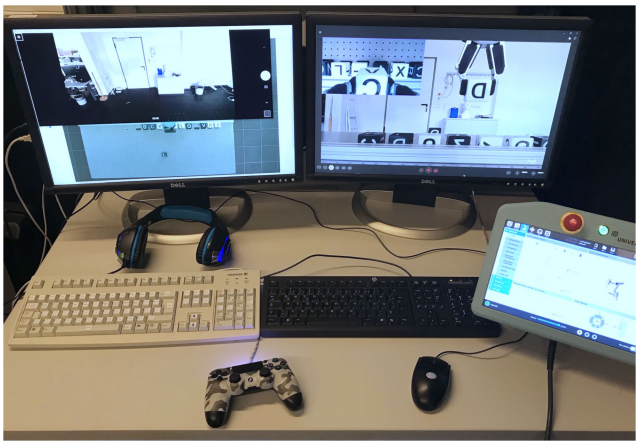


FIGURE 1. Wizard's chamber: monitors with a handheld controller, *UR5e* control-panel, one keyboard for shortcuts and another for direct speech output.

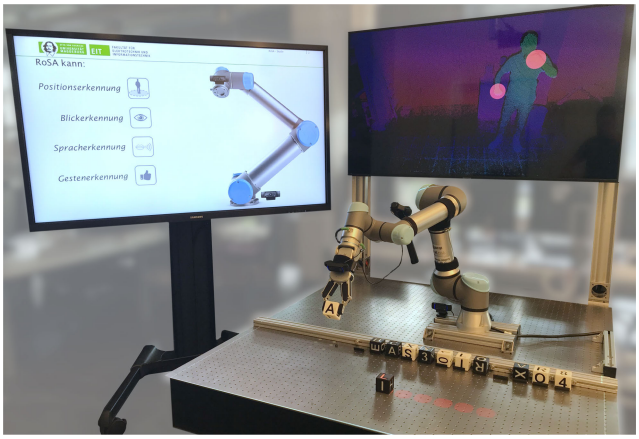


FIGURE 2. RoSA system setup: *UR5e* industrial robot, ''mock-up''-gesture recognition (right), instructions (left).

A. CAPABILITIES AND RESTRICTIONS OF RoSA

The skills and restrictions described below were used to create a consistent set of rules for the wizard when acting as RoSA. The skills include high-level speech recognition, covering generic terms: direction (left, right, front, etc.), position (home, ''here''), distances (cm, mm), angles (degrees) and adapting to custom user vocabulary after several repetitions (''cube'' can be referred to as ''block'' or ''brick''). Furthermore RoSA is able to interpret any shown gesture at human level, filtering obvious non-HRI gestures (e.g. scratch head, check wristwatch). When speech and gestures are used simultaneously, the gestures are interpreted secondarily as an amplifier for the spoken words and their intention, as understood by the wizard. RoSA is not able to recognise the letters or numbers on the given cubes, but is able to recognise and identify the object, knowing what colors and shapes are. RoSA is able to route to specific coordinates, hand-over, pick or place objects, collision-free.

In addition to the recognition capabilities, the system is able to learn from the subject. The learning process can

be separated into *passive teaching* by repetition and *active teaching* by direct commands.

If a task requires more than one step, these could be sequenced and the program could be reused later. This is referred to as *advanced programming*.

Since RoSA is controlled by an operator using different presets and a custom controller, the system is not able to move in exact numerical steps when prompted (''move left 5 cm''). In such cases the speech input is still accepted, but the resulting distance differs due to manual control.

RoSA does not know the position between two blocks and is not able to route to this position without prior teaching. Slight ''system imperfections'' were applied in order to force diversity in the use of interaction modalities.

B. EXPERIMENT

The study was held at the Otto-von-Guericke-University Magdeburg, Germany. The experiments were conducted in German language. The subjects were asked to fill in their sociodemographic information, as well as their experience with artificial intelligence (AI), industrial robots and programming skills. In the following the experimenter provided material with the instructions. These vaguely described the capabilities of the system (''RoSA is capable of: speech- and gesture-recognition''). The subjects were then given three tasks:

- 1. Have RoSA give you a block.
- 2. Spell a specific word with alternating color of blocks.
- 3. Build a 3-2-1-Pyramid with black-white-black layers (as seen in figure 3).

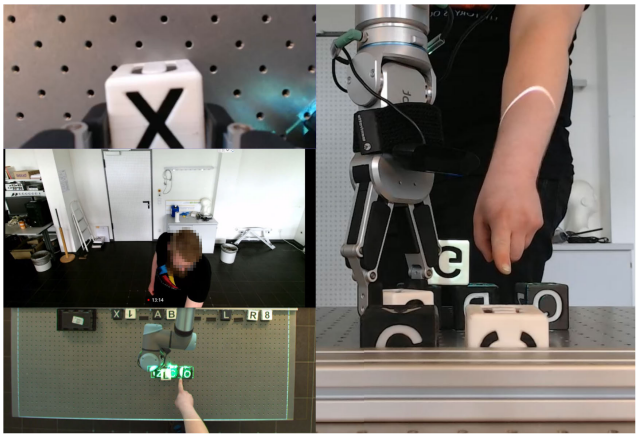


FIGURE 3. Composition of view from the streaming cameras; subject using *pointing gesture*.

After fulfilling all tasks the subjects were asked to complete several questionnaires to evaluate their user experience (see chapter IV-B) [19]–[22].

Later the subjects were told to show a set of gestures that would further improve the systems capabilities. The set contained gestures to *sign in, start, stop, move XYZ and open/close/rotate gripper*. During the procedure there were two enforced system failures at certain events. The first one occurred during the second task: when asked to put a block on a projected field the system would purposely stack the cube onto another one until the subject actively intervenes. The second failure occurred during the third task: right before finishing the task the system would crash the last cube into the pyramid. Doing so would set back the progress, so steps would have to be redone. These deliberate errors were introduced in order to trigger reactions to errors or unpredictable behaviour. At the end of the experiment the wizard was revealed to the subjects.

III. ANALYSIS AND DATA SEGMENTATION

After the experiment the video data was reviewed and the different modalities of interaction summed up. The possible interactions were categorized into the following groups: *speech direction [continuous, steps, units], speech instruction [acknowledgment, numeric, descriptive], gesture direction [continuous, macro, micro], gesture instruction [acknowledgment,pointing]*. Examples for the modalities can be seen at table 1. The intention of the interaction determines the categories, which are mutually exclusive.

TABLE 1. Interaction modalities with examples.

	Speech	Gesture
Instruction	**Acknowledge** *"Yes" / "No"*	**Acknowledge** **Thumbs-Up**
	Numeric *"...the third block"*	**Numeric** **Tree-fingers-up**
	Descriptive *"...block with letter O"*	**Pointing** **Points at an object**
Direction	**Unit** *"...5 cm left"*	**Micro** **Twitches with fingertips**
	Step *"...a bit right"*	**Macro** **swipes left**
	Continuous *"left... ...stop"*	**Continuous** **points right**

A *Pointing gesture* to show a direction is a different intention as the same gesture to show a location. It was sufficient for the subject to use a modality once for it to be registered. Furthermore the use of: *teaching features, advanced programming, active gaze, touching and leading the robot*, as well as time needed for a task, robot problems and number of *unclear instructions* were noted. The term ''instructions unclear'' was a response from RoSA, when knowledge restrictions were surpassed.

To ensure inter-rater reliability, the same footage was examined by the raters with a resulting 93% of agreement leading to a Cohen's kappa coefficient κ of 0.86 which corresponds to ''almost perfect'', according to Landis and Koch [23].

The recorded data was then expanded with the data from the questionnaires and the full data set examined with the help of descriptive statistics and Pearson product-moment correlation coefficients [24]. This method was chosen because Pearson correlation coefficient estimated for two binary variables computationally will return the phi-coefficient, used for describing binary data [25].

Due to problems with memory cards and the unexpected failure of cameras or gimbals, we suffered data loss. Fortunately these gaps could partly be covered by other camera angles.

IV. RESULTS

We have a sample size of 36 subjects and assume the data to approximately follow a normal distribution. There were no outliers. The demographic questionnaire reveals 14 females, 22 males and 0 diverse subjects with age ranges from [20]–[24] to [55]–[59] with the median being the group of [25]–[29] years. All subjects completed the tasks with a time between 13 and 37 minutes, averaging around 19.58 minutes.

A. HUMAN-ROBOT INTERACTIONS

Some biases (''this is just like *Alexa/Google/Jarvis*'') could be observed, when subjects were completing the demographic questionnaire regarding their experience with artificial intelligence systems. This resulted in 97.2% of the subjects using any form of speech for interaction at least once. The subjects used different interaction modalities to complete the tasks. Most of the modalities were used independently, while some depended on auxiliary input to function efficiently. *Speech description*, for example, was used exclusively in combination with *pointing gestures*, whereas both of the modalities alone, while not being as effective, would have been sufficient. Most of the subjects used a single working strategy and maintained it throughout the experiment, until encountering restrictions. Only a few subjects experimented and explored the possibilities.

For instance: If the subjects (S) that primarily used *descriptive speech* for HRI faced the restriction of RoSA (R) not being able to read the letters on the blocks. The subjects either retried the same input or changed strategy. This could be switching to *numeric instructions*:

S: ''RoSA, take the cube with the letter A.''
R: ''I did not understand, the input was unclear.''
S: ''Letter A, [. . .].''
R: ''I did not understand, the input was unclear.''
S: ''RoSA, take the first cube and place it on the first projected position.''

Other users who ran into the same problem tried adding new interaction modalities. These ''hybrids'' (72.2 % of the subjects) used *pointing gestures* in combination with *speech* to fulfill the given tasks:

S: ''Take this **points** and put it here.'' **points** In rare cases the subjects switched modalities to gestures only, using either *pointing gestures* or *directional gestures* to operate the system.

An in-build system flaw of not understanding the instructions ''put on two cubes'' led to another obstacle that would come up when trying to stack the pyramid. Before that, all instructions, which were compliant with the rule-set, given to RoSA by the user, were executed flawlessly. When asked to stack on two cubes, RoSA would choose one of the cubes and put the third one directly on top.

S: ''Put the cube in the gripper on the cube A *and* the cube B. On both of the cubes.''

The instruction was accepted if the letters A and B were already taught to RoSA. Alternatively ''this'' accompanied with a *pointing gesture* instead of letters would also have been accepted. Still, instead of placing the cube at the desired position (on top of space between the cubes) RoSA would execute the command literally: placing the third cube on the first and then quickly picking it up and placing it on the second.

At this point most subjects switched to a more manual approach using only *directional voice commands* (91.7%) or *directional gestures* (36.1%). Subjects not relying on automatic path planning, i.e the pre-programmed positions known to RoSA, but using directional commands from the beginning, did not notice the position deviation when stacking, since they were in direct control of the robot path.

The popularity for each interaction modality can be seen in the table 2.

TABLE 2. Frequencies of use for interaction modalities and other factors.

	Factor	%
Teaching	Overall	72.2
	Passive	58.3
	Active	27.8
	Using the new knowledge	55.6
	Sequential programming	16.7
Miscellaneous	AI experience	33.3
	Ind. robots experience	16.7
	Two-handed gestures	16.7
	Robot freedrive mode	8.3
	Asking RoSA questions	22.2
	Automatic pathfinding	88.9

	Modality	%
Gesture	Acknowledge	8.3
	Point	75.0
	Micro	30.6
	Makro	25.0
	Continuous	13.9
Speech	Step	86.1
	Unit	44.4
	Continuous	94.4
	Descriptive	33.3
	Numeric	77.8
	Acknowledge	66.7

It was possible to actively teach RoSA new commands by explaining the action beforehand or by summarising the actions already taken:

S: "RoSA, take the cube and place it here." **points**
S: "This cube is next to the last one."
R: "I understand, I learned: *next to the last one*."
S: "RoSA, take another black cube and place it *next to the last one*."
RoSA places the cube as expected.
R: "Is this correct ?"
S: "Yes, RoSA, very good."

The passive teaching took place when RoSA was given additional information not necessarily required for the execution of the command:

S: "RoSA, give me the third cube, the one with the letter A."
RoSA moves to the cube.
R: "Processing data. I learned: *letter A*. Is this correct ?"
S: "Yes, give me the cube [...]."
RoSA hands-over the correct cube.

Although 72.2% of the subjects used teaching in a passive or active manner (58.3% / 27.8%), only 55.6% of the subjects actually (re-)used the commands learned by RoSA.

22.2% asked RoSA basic "yes/no" questions about the system capabilities or the actual situation:

S: "RoSA, "Can you see the projected fields?"
R: "Processing data. Yes."

-or-

S: "RoSA, "Can you give me a black cube ?"
R: "Processing data. Yes."
RoSA is not moving.
S: "RoSA, "give me a black cube."
RoSA hands-over a black cube.

All the users admitted to not noticing that RoSA was operated manually when greeted by the wizard after the study and thereby confirming the WoZ setup.

B. USER EXPERIENCE

To evaluate the user experience, subjects were asked to complete the four questionnaires, SUS, PSSUQ, UMUX and ASQ after the third task, still thinking that RoSA is an AI system. To plot and compare the PSSUQ and ASQ Scores range of [1 to 7] on the same graph as SUS and UMUX range of [0 to 100], a conversion of

$$Score_{[0:100]} = (Score_{[1:7]} - 1) \cdot \frac{100}{6}$$

was used. The results can be seen below in table 3.

The different questionnaires show a moderate to strong positive correlation (UMUX 0.71, PSSUQ 0.81, ASQ 0.43) when compared to SUS. The moderate correlation of ASQ can be explained by the low count of questions this questionnaire uses, which leads to higher quantisation of possible score values. A total system score can be calculated by averaging the four questionnaire scores, resulting in a total system score of 74.02 out of 100, which will be further referred to as the *score*.

TABLE 3. Results of user-experience questionnaires SUS [19], UMUX [20], PSSUQ [21], ASQ [19].

Variables	SUS	UMUX	PSSUQ	ASQ
Answer Range	1 to 5	1 to 7	1 to 7, NA	1 to 7, NA
Score Range	0 to 100	0 to 100	1 to 7	1 to 7
Nr. of Questions	10	4	16	3
Avg. Score	79.24	71.53	5.42	5.30
Normalised Score	79.24	71.53	73.70	71.60
Std. deviation	12.36	18.00	17.78	20.86
% RSD Score	15.60	25.16	24.13	29.13
Correlation to SUS	-	0.71	0.81	0.43

Total average score: 74.02

C. DATA CORRELATION

The correlation coefficients in table 4 and 5 describe the statistical probability of two or more modalities being used together.

TABLE 4. Pearson's correlation coefficients between interaction modalities.

			Speech						Gesture				
			Instruction			Direction			Direction			Instruction	
			Ack.	Num.	Desc.	Unit	Step	Cont.	Cont.	Makro	Micro	Point	Ack.
Gesture	Instruction	Ack.	0.1	0.1	0.1	0.0	0.2	0.2	-0.1	0.3	0.2	0.2	-
		Point	0.1	-0.1	0.4	-0.1	0.0	0.0	-0.3	-0.1	0.0	-	0.2
	Direction	Micro	-0.1	-0.1	-0.1	-0.2	-0.1	0.0	0.3	0.7	-	0.0	0.2
		Makro	-0.1	0.0	-0.1	0.0	0.0	0.1	0.5	-	0.7	-0.1	0.3
		Cont.	-0.1	0.0	-0.3	0.1	0.0	-0.1	-	0.5	0.3	-0.3	-0.1
Speech	Direction	Cont.	-0.1	0.2	0.3	0.1	0.2	-	-0.1	0.1	0.0	0.0	0.2
		Step	0.2	0.2	0.5	0.2	-	0.2	0.0	0.0	-0.1	0.0	0.2
		Unit	0.1	0.1	0.2	-	0.2	0.1	0.1	0.0	-0.2	-0.1	0.0
	Instruction	Desc.	0.3	0.0	-	0.2	0.5	0.3	-0.3	-0.1	-0.1	0.4	0.1
		Num.	0.2	-	0.0	0.1	0.2	0.2	0.0	0.0	-0.1	-0.1	0.1
		Ack.	-	0.2	0.3	0.1	0.2	-0.1	-0.1	-0.1	-0.1	0.1	0.1
Time			0.1	-0.2	0.2	0.4	0.2	0.3	0.3	0.0	-0.2	-0.3	0.1
Age			0.1	0.2	0.2	-0.2	0.3	0.1	-0.2	-0.1	0.2	0.3	0.1
Score			-0.2	0.0	-0.3	0.0	-0.1	-0.2	0.1	0.2	0.0	-0.3	-0.1

A strong positive correlation can be seen between *pointing gesture* and *descriptive voice commands*. There is also a strong correlation between macro and continuous gestures. The pervasive use of *descriptive voice commands* in combination with *iterative voice commands*, but not with *numerical voice commands* is worth noting.

A subject using the speech domain is more likely to stay in it and less likely to use *directional gestures*. This is especially seen in the overall negative correlation between *speech instruction* and *gesture direction*.

Time was not a significant factor in the experiment which is supported by the evaluation (table 5) since there is no correlation with the system *score*. The biggest time saver, suggested by the data, was using automatic path-planning. On the contrary, using voice commands or two handed gestures had a strong correlation with time. Subjects that mainly used *instructional numeric* voice commands or *instructional*

TABLE 5. Pearson's correlation coefficients for score, time and age.

Factor	correlation to		
	Score	Time	Age
Time	0.0	-	0.1
Age	-0.4	-0.1	-
Nr. of unclear instructions	-0.4	0.2	-0.2
Nr. of robot problems	0.0	0.5	-0.2
Teaching overall	-0.2	0.0	0.2
Teaching passive	-0.4	0.0	0.1
Teaching active	0.1	0.1	0.1
Using the new knowledge	0.0	0.0	0.2
Sequential programming	-0.1	-0.1	0.1
AI experience	-0.4	0.2	0.1
ind. robots experience	-0.2	0.1	-0.2
Two-handed gestures	0.1	0.6	-0.2
Robot freedrive mode	0.2	0.0	-0.2
Asking RoSA questions	0.0	0.2	0.1
Automatic pathfinding	-0.2	-0.6	0.2

TABLE 6. Results gesture catalogue: 1st and 2nd most popular choices*.

Action	Gesture	%
Sign-in	Wave hand	40.0
	Rise hand	14.3
Start	Thumb-up	31.4
	Virtual touch-pad	14.3
Stop	Stop gesture (1H)	48.6
	Stop gesture (2H)	17.1
Move XYZ	Handtracking	51.4
	Point direction	37.1
Open gripper	Open one hand	62.9
	Spread arms apart	31.4
Close gripper	Close one hand	71.4
	Put arms together	28.6
Rotate gripper	Rotate hand (wrist axis)	42.9
	Rotate finger (draw a circle)	37.1

*data from 35 out of 36 subjects was used due to data loss

pointing gestures were usually the fastest to finish all three tasks. Subjects who were keen to experiment with different modalities needed more time to complete the assignments.

System failures had no contribution to the *score*, in contrast to ''instruction unclear'' feedback which had a negative correlation of 0.4.

The factors that impacted the *score* negatively the most are: age, experience with AI, *passive teaching*, number of *unclear instructions* and modalities *speech description* and *gesture pointing*. The factors that had positive influence on the *score* are: Robot ''*Freedrive-mode*'', when the robot arm can be manually moved by the subject, and the use of modalities *Gesture Micro* and *Macro*. There is a weak positive correlation (0.22) between experience with AI and active teaching, as well as a moderate correlation (0.35) between experience with industrial robots and not using the knowledge taught to RoSA. Passive learning shows a negative correlation with the *score*.

D. GESTURE CATALOGUE

The gestures given by the subjects to ''improve the system'' were relatively consistent for the actions: *stop, move* and *operate the gripper*; while the actions for *start* and *sign in* were more individual. The summary of the two most popular gestures per action can be seen in the table 6.

V. CONCLUSION

This study provides deep insight into user experience and interaction with industrial robotic assistants. In giving the subjects the illusion of a restriction-less system, we had the possibility to observe truly intuitive HRI.

As described in chapter II-A, RoSA, although being operated by a wizard, had several knowledge-based restrictions in order to comply to the experiment setup.

The feedback ''instruction unclear'' could have led to a high degree of frustration. This especially applies to subjects with more experience in AI, robotics and programming, leading to lower scores. Subjects from this group are more likely to use such a system on a daily basis and are naturally more critical.

The negative correlation between *passive teaching* and *score* indicates that RoSA actively saying: ''Processing data. I learned: [...]'', or teaching basics (''This is letter A.''), could have been perceived as annoying due to disturbance in the workflow. Some subjects seemed to expect the system to be already programmed to complete the given assignments and only wanted to give orders and not program sub-tasks that would seem basic.

The average *score* over all questionnaires being 74%, shows that the average user was pleased with the system, unaware that some of the errors were enforced on purpose and the system was remotely operated.

Since the industrial environment of robots tends to be noisy, we did not expect the high impact of speech. As the experiment shows, under laboratory conditions, this seems to be an intuitive way to communicate with such a system.

We assume the high use of dialogue-like speech to be an effect of the subjects being biased by already existing AI systems (*Alexa*, *Google Assistant*, *Jarvis*, etc.) and expecting RoSA to behave in a similar manner (''OK, RoSA.'' vs ''OK, Google'').

In many cases the system was expected to be able to calculate forward kinematics and path-planning, fewer subjects operated the robot with *directional speech/gestures*. The high

correlation between *macro* and *continuous gestures* might be due to the fact that they are very similar in execution and can be used in an auxiliary manner.

We learned that the questionnaires were likely inadequate for the given scenario as the *scores* for several questions suggest. For example one of the questions being: ''The system helped to quickly fulfill the given task'' – considering the task being ''stack a pyramid'', which could have been done faster by the subject – received a low score. If looking at the task being ''[program this robot to] stack a pyramid'' the *score* could have been different. Alternatively a modification of an existing questionnaire or a set of new questions, especially concerning HRI, could lead to a better and more detailed user evaluation.

With the data from the ''system improvement'' catalogue we present a set of gestures that can be applied to many contact-less user interfaces. A system using the first two most popular choices would allow for an intuitive way of interaction for the actions *Move XYZ, Open/Close/Rotate Gripper* and *stop*. The multitude of sign-in and start gestures suggest a need for a customizable interface in order to fulfill user expectations.

The interactive laser-projection system for programming industrial robots by Zaeh & Vogl could be used to provide the user with feedback to improve the precision of pointing gestures [12]. A low-cost passive stylus ''the DodecaPen'' which allows 6 degrees of freedom input using a single camera as proposed by Wu *et al.* could work as an alternative to pointing gestures when higher precision is required [26].

For the next steps, we suggest development and testing of a system based on *pointing gestures* complemented by *object orientated speech* and *micro/macro gestures* for fine-tuning as these modalities show high potential and time-efficiency, as these modalities presented as intuitive in our experiments.

It is worth noting that the use of multi modal (speech, gestures, gaze, body pose) interaction for contact-less communication adds an additional layer of redundancy and thus safety, if the different modes are processed as logical conjunctions. Although our system was ''safe'' as the robot operated in a collaborative mode (with permanently reduced speed, force, and power) a combination of the interaction modalities could allow a safe integration of contact-less communication with industrial robots. Our future work will focus on applying the findings to state of the art technologies.

REFERENCES

[1] D. Buhr. *Soziale Innovationspolitik für Die Industrie 4.0*. Accessed: Sep. 1, 2020. [Online]. Available: http://library.fes.de/pdf-files/wiso/11494.pdf

[2] R. Behrens and N. Elkmann, ''Workshop: Sicherheit in der menschroboter-kollaboration,'' in *Proc. Robot. Kongress Hannover*, Feb. 2016. Magdeburg, Fraunhofer, pp. 1–5. [Online]. Available: http://www.robotics-kongress.de/wp-content/uploads/sites/12/2016/11/Fraunhofer-IFF.pdf

[3] A. Mehrabian and M. Wiener, ''Decoding of inconsistent communications,'' *J. Pers. Social Psychol.*, vol. 6, no. 1, p. 109, 1967.

[4] J. Wilson and D. Rosenberg, ''Rapid prototyping for user interface design,'' in *Handbook Human-Computer Interaction*. Amsterdam, The Netherlands: Elsevier, 1988, pp. 859–875.

[5] T. K. Landauer, ''Psychology as a mother of invention,'' *ACM SIGCHI Bull.*, vol. 17, no. SI, pp. 333–335, May 1986.

[6] J. F. Kelley, ''An iterative design methodology for user-friendly natural language office information applications,'' *ACM Trans. Inf. Syst.*, vol. 2, no. 1, p. 26–41, Jan. 1984, doi: 10.1145/357417.357420.

[7] D. Maulsby, S. Greenberg, and R. Mander, ''Prototyping an intelligent agent through wizard of Oz,'' in *Proc. Conf. Bridges Between Worlds (INTERCHI)*, 1993, pp. 277–284.

[8] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, ''Wizard of Oz studies: Why and how,'' in *Proc. 1st Int. Conf. Intell. User Interfaces (IUI)*, 1993, pp. 193–200.

[9] S. E. Hudson, J. Fogarty, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. C. Lee, and J. Yang, ''Predicting human interruptibility with sensors: A wizard of Oz feasibility study,'' in *Proc. New Horizons CHI*, 2003, pp. 257–264.

[10] M. Serrano and L. Nigay, ''A wizard of oz component-based approach for rapidly prototyping and testing input multimodal interfaces,'' *J. Multimodal User Interface*, vol. 3, no. 3, pp. 215–225, Apr. 2010.

[11] S. Dow, B. MacIntyre, J. Lee, C. Oezbek, J. D. Bolter, and M. Gandy, ''Wizard of oz support throughout an iterative design process,'' *IEEE Pervas. Comput.*, vol. 4, no. 4, pp. 18–26, Oct. 2005.

[12] M. Zaeh and W. Vogl, ''Interactive laser-projection for programming industrial robots,'' in *Proc. IEEE/ACM Int. Symp. Mixed Augmented Reality*, Oct. 2006, pp. 125–128.

[13] J. Norberto Pires, ''Robot by voice: Experiments on commanding an industrial robot using the human voice,'' *Ind. Robot, Int. J.*, vol. 32, no. 6, pp. 505–511, Dec. 2005.

[14] G. Reinhart, W. Vogl, and I. Kresse, ''A projection-based user interface for industrial robots,'' in *Proc. IEEE Symp. Virtual Environ. Hum.-Comput. Interface Meas. Syst.*, Jun. 2007, pp. 67–71.

[15] T. Brogãrdh, ''Present and future robot control development—An industrial perspective,'' *Annu. Rev. Control*, vol. 31, no. 1, pp. 69–79, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1367578807000077

[16] R. Bischoff and A. Kazi, ''Perspectives on augmented reality based human-robot interaction with industrial robots,'' in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Dec. 2004, pp. 3226–3231.

[17] M. Speicher and M. Nebeling, ''GestureWiz: A human-powered gesture design environment for user interface prototypes,'' in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–11.

[18] G. Hoffman, ''Openwoz: A runtime-configurable wizard-of-oz framework for human-robot interaction,'' in *Proc. AAAI Spring Symp. Ser.*, 2016, pp. 1–5.

[19] J. Brooke. *SUS-A Quick and Dirty Usability Scale*. Accessed: Sep. 1, 2020. [Online]. Available: www.TBIStaffTraining.info

[20] K. Finstad. (2010). *The Usability Metric for User Experience*. [Online]. Available: https://academic.oup.com/iwc/article-abstract/22/5/323/682940

[21] J. R. Lewis, ''Psychometric evaluation of the PSSUQ using data from five years of usability studies,'' *Int. J. Hum.-Comput. Interact.*, vol. 14, nos. 3–4, pp. 463–488, Sep. 2002. [Online]. Available: https://www.tandfonline.com/action/journalInformation?journalCode=hihc20

[22] J. Lewis, ''Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The ASQ,'' *SIGCHI Bull.*, vol. 23, p. 78–81, Jan. 1991.

[23] J. R. Landis and G. G. Koch, ''The measurement of observer agreement for categorical data,'' in *Proc. Biometrics*, Dec. 1977, pp. 159–174.

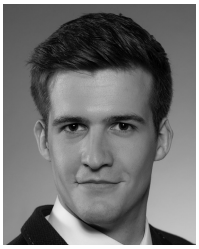
[24] P. Schober, C. Boer, and L. A. Schwarte, ''Correlation coefficients: Appropriate use and interpretation,'' *Anesthesia Analgesia*, vol. 126, no. 5, pp. 1763–1768, May 2018.

[25] J. P. Guilford, *Psychometric Methods*. New York, NY, USA: McGraw-Hill, 1936.

[26] P.-C. Wu, R. Wang, K. Kin, C. Twigg, S. Han, M.-H. Yang, and S.-Y. Chien, ''DodecaPen: Accurate 6DoF tracking of a passive stylus,'' in *Proc. 30th Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2017, pp. 365–374.

DOMINYKAS STRAZDAS was born in Vilnius, Lithuania, in 1989. He received the B.Sc. and M.Sc. degrees in mechatronics from Otto-von-Guericke University Magdeburg, where he is currently pursuing the Ph.D. degree in electrical engineering.

Since 2017, he has been a Research Assistant with the Research Group Neuro-Information Technology, Otto-von-Guericke University Magdeburg. His research interests include human–machine-interaction especially natural, intuitive, and contact-less communication between humans and robots.



JAN HINTZ was born in Brunswick, Lower-Saxony, Germany, in 1996. He received the B.Sc. degree in electrical engineering and information technology from Otto-von-Guericke University Magdeburg, where he is currently pursuing the M.Sc. degree in electrical engineering and information technology.

Since 2018, he has been a Research Assistant with the Research Group Neuro-Information Technology, Otto-von-Guericke University Magdeburg. His research interests include computer vision, image processing, machine learning, and human–machine interaction.

ANNA-MARIA FELßBERG was born in Eisenach, Thuringia, Germany, in 1988. She received the B.Sc. degree in psychology from the Otto-von-Guericke University, Magdeburg, Germany, in 2017, where she is currently pursuing the M.Sc. degree in cognitive neuroscience (psychology).

Since 2020, she has been a member of the Elena Azañón's Sensory Laboratory. Her research interests include visual and sensory perception as well as emotion recognition.

AYOUB AL-HAMADI received the Ph.D. degree in technical computer science, the Habilitation degree in artificial intelligence, and the Venia Legendi degree in pattern recognition and image processing from Otto-von-Guericke University Magdeburg, Germany, in 2001 and 2010, respectively. He is currently an Adjunct Professor and the Head of the Neuro-Information Technology Group, Otto-von-Guericke University Magdeburg. He is the author of more than 300 papers in peer-reviewed international journals, conferences, and books. His research interests include computer vision, pattern recognition, and image processing.