Embodied voice assistant

Fredrik Lundkvist Nina Nokelainen Matilda Richardsson Kajsa Saare

EECS, School of Electrical Engineering and
Computer Science, KTH

Abstract

In this paper, we investigate how embodiment of a voice-controlled virtual assistant affects user perception and comfort during use. To examine this, we implemented a basic voice assistant interaction pattern for a Furhat robot, which we connected to the Google Assistant API. Participants in the study then used a physically or virtually embodied assistant to complete a set of tasks, in order to evaluate the differences in interaction between the two embodiments. The results show that, while participants could accomplish the tasks, they did not particularly enjoy using the embodied assistant, with the physically embodied version receiving the most negative feedback; we theorize that this might be due to the uncanny valley effect.

1. Introduction

The use of intelligent personal voice assistants (IPA) has during the recent years become more popular and more commercially accessible than ever before. Assistants are now available on phones, computers, TV:s and cars, not to mention products made with the sole purpose of being an assistant, like the Google nest hub[1]. Most large operating systems have their own IPA that they prompt new users to explore and use. These IPAs have now also found their way into "smart home" products that are used for various purposes in users' homes. Using voice and sounds to communicate and to make sense of the world is natural and primal, and in many ways automatic too. Voice control is therefore a natural interface, since it allows users to interact with their devices in a way similar to how they interact with other humans, and can allow for a greater degree of freedom than traditional haptic inputs. Voice control allows users to interact with machines, while having their hands busy with something else, like cooking, and allows for touch free interactions where hygiene is extra important, like in hospitals. Most IPAs exist within another product but some have their own physical form. In this study we investigate how two different embodiments of IPAs are perceived and welcomed by users.

1.1 Aim

The aim of this study is to investigate how an embodied voice assistant is perceived by users. We will also examine how different kinds of embodiment (virtual and physical) affect the interactions between user and voice agent.

[1] https://store.google.com/product/google_nest_hub

1.2 Research question

How do users interact with virtually or physically embodied voice assistants?

1.2.1 Delimitations

The major delimitation of this study was the forms of embodiment investigated; we only compared a physical furhat robot to a virtual representation of said robot. Thus, we did not investigate how other forms of embodiment, such as fully embodied robots, or stylistic choices such as realism affects the user experience. The study was also limited to observing how pairs of friends interacted with the embodied IPA, no comparisons between the behaviour of pairs and that of individuals were made.

2. Background

In this section, we will present previous research relevant to this study, such as gaze in interaction, social robots, and embodied agents.

2.1 Gaze and Social Robotics

Gaze has been studied within Human-Robot-Interaction (HRI) for a long time. This has especially been studied when it comes to social robots, where it has been studied how people respond to gaze [1]. A study by K. Ruhland et al. [2] discussed how to tackle the problems of recreating gaze in computer graphics, a challenging task. They point out the importance of gaze in multi-modal behaviours conducted by speakers in face to face interactions, as well as the importance of the gaze pattern a virtual character has during a conversation.
McMillan et al. [3] studied how gaze can be used in a speech agent, in their specific case, Google Assistant. Their findings indicate that gaze is an effective tool when it comes to making users notice that the agent is listening. However, they also found that there were issues concerning gaze patterns during conversations.
In the field of social robotics, Furhat [4] is a state-of-the-art robot, capable of user tracking and differentiation, speech synthesis and processing in multiple languages, as well as solid gaze and facial gesture features. Furthermore, it is easy to develop custom software for Furhat, due to development tools and simulators being available on the internet. Because of this, we decided early on to use Furhat as the platform for our voice agent.

2.2 Embodied Agents

Embodiment has long been viewed as a critical aspect of intelligent systems and agents. Throughout this article, we will use Pfeifer and Scheier's [6] definition of embodiment of artificial agents:

“Embodiment: A term used to refer to the fact that intelligence cannot merely exist in the form of an abstract algorithm but requires a physical instantiation, a body. In artificial systems,the term refers to the fact that a particular agent is realised as a physical robot or as a simulated agent.”(p. 649)

Thus, a physical agent is one that is *physically embodied*, while a virtual agent is *digitally embodied* in some way, most commonly by computer graphics.

However, the terms of embodiment is not the only factor to be taken into account when designing a digital agent. A survey of 33 experimental studies indicates that physical presence of an embodied agent “*plays a greater role in psychological response to an agent than physical embodiment*” [5]. These results are corroborated by a study conducted by Thellman et al. [7] who also found that physical presence of an agent played a bigger part in perception than how the agent was embodied. The implication of these studies is that realistic embodiment is less important than how physically present the agent is; a less realistic-looking robot, such as Tama [3], might be more positively perceived, by virtue of being physically present, as well as avoiding the uncanny valley effect.

2.3 Gaze in conversation

In interpersonal interactions, gaze plays an important role for many functions, such as turn-taking, and affirming attention.

In 1972, Duncan [8] studied how gaze affects conversations; they found that gaze is used to signal attention between conversation partners, and also regulates turn taking in conversation. Building on this, Goodwin [9] further examined the role of gaze in conversations; he found that gaze not only manages attention, but also affects how the speaker expresses themself.

“when the speaker has the gaze of the recipient, a coherent sentence is produced. To have the gaze of a recipient thus appears to be preferred over not having such gaze, and this preference appears to be consequential for the talk the speaker produces within the turn. In this way gaze is an important cue which indicates that the hearer is listening to the speaker”

Based on the importance of gaze in interpersonal conversations, one can infer that it would have major effects on interactions between humans and embodied speech agents, however, it is uncertain whether these effects are positive or negative, since users might feel that the agent’s gaze is too unnatural, making them uncomfortable.

3. Method

3.1. Implementation

The embodied assistant was implemented using the Furhat SDK[2] and the Google Assistant API[3], both of which are free to use for development purposes. Before we began working on the implementation, we decided that it should follow a basic voice assistant pattern; this meant that the user would first issue a "wake word" to get the assistant's attention. After receiving some form of confirmation that the assistant was listening, the user would then issue a command or query to the assistant, which would return a spoken response. We decided to call our Furhat "Steve".

Since speech recognition and voice synthesis functionality is built into the Furhat operating system, we only needed to implement the interaction pattern, as well as set up a connection between the agent and the Google Assistant API. The interaction patterns were implemented by creating a custom Furhat skill, using tools included in the SDK; the connection between Furhat and Google Assistant proved to be a much more difficult task, since there were no kotlin libraries for Google Assistant. Thus, we opted to create a custom middleware, using the furhatJScore[3] library to communicate with the web server inside the Furhat operating system, and the assistant node SDK to communicate with the assistant API[4].
When a user issues a query to Furhat, the robot sends a text version of the query to the middleware, which sends the same query to the assistant API and awaits a response. When a response is received, it the middleware forwards it to Furhat, which outputs the response to the user using speech synthesis. This solution worked well, since implementation did not take too much time, and all code worked with both physical and virtual Furhats, which meant that the same software could be used for both experiment conditions, reducing the risk of discrepancies due to implementation. The interface is demonstrated on vimeo: https://vimeo.com/385928202

The only major difference between physical and virtual Furhat was gaze functionality. Since we did not find a way for the virtual Furhat to track user positions or figure out which user was speaking, it could not direct is gaze towards the user talking to it. The physical robot had these capabilities, and we opted to use them; the physically embodied agent would look at the user who issued the wake command throughout the interaction cycle, with the purpose of showing the user that the assistant is listening. This meant that users got confirmation of attention via two modalities: the assistant would utter a short phrase to indicate that the user had its attention, and in the case of the physically embodied version, would also look directly at the user. When the assistant was idle, it would instead direct its gaze away from the users, to indicate that it was not active.

[2] https://www.furhatrobotics.com/developers/
[3] https://developers.google.com/assistant/sdk
[3] https://github.com/FurhatRobotics/FurhatJSCore
[4] https://github.com/googlesamples/assistant-sdk-nodejs

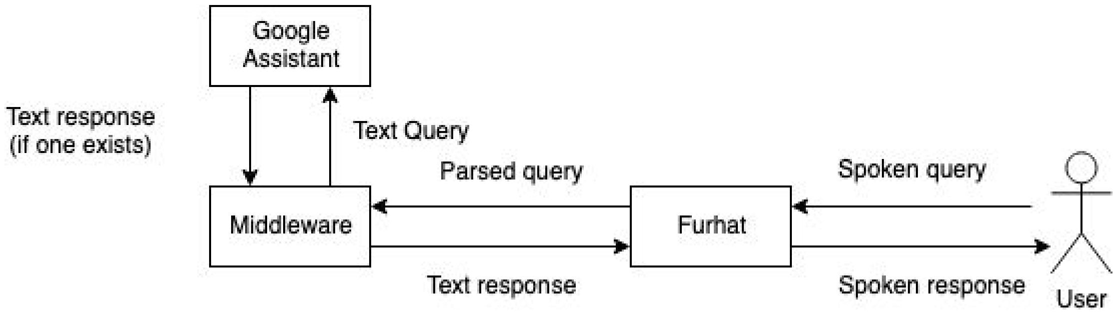


Figure 1: Implementation diagram for Steve, the embodied voice assistant

3.2. Pilot study

To ensure that the tasks were of appropriate difficulty, we conducted a pilot study. Two participants attempted to complete the provided tasks with the help of a physically embodied agent.

The results of the pilot study showed that some tasks were not formulated clearly, and that some tasks were difficult to complete, since the correct answer would only be obtained by formulating questions in very specific ways. Thus, we decided to revise the tasks before moving on to the main experiment. The major revisions included removing some tasks that were deemed too difficult to complete due to the software used, as well as rephrase a few tasks to improve understanding.

3.3. Participants

18 participants were recruited for the main experiment. All participants were recruited by personal interaction or online messaging; all participants were students at the Media Technology degree programme at the Royal Institute of Technology in Stockholm, and were between 20 and 30 years of age.

The reasoning for selecting participants from this demographic group was twofold: recruiting students is easier than finding participants who are not affiliated with the academic world, reducing the amount of time spent on participant recruitment. Second, by selecting participants from a group of people studying technical subjects, we could assume a basic level of technological competence and experience, such as having interacted with a voice-based interface before participating in the study.

3.4. Setup

Participants carried out the experiment in pairs. When arriving at the experiment location, they were shown into a room. In this room was a desk, chairs for the participants, and a piece of paper containing a number of tasks for the participants to complete.

There was also an embodied speech agent in the room; for 4 groups, the agent was embodied digitally in the form of a 3D model presented on a screen. For the other 5, the speech agent was instead embodied physically by a Furhat robot[5]. Participants were instructed to complete the tasks on the paper by using the speech agent (for a full list of tasks, see appendix 1). The first tasks were written with the intention of making users comfortable interacting with the embodied agent while learning the interaction pattern; as participants progressed through, the tasks became more open-ended, culminating with solving a crossword.

The participants were left alone in the room while completing the tasks, and their progress was recorded by a webcam. They were told to leave the room when they had completed their tasks, or if any problems arose during the experiment.

After completion of the tasks, participants were interviewed about their experiences with speech agents, their experiences using our embodied agent, and their thoughts about using embodied agents in their daily lives, at home and in public. Both the task solving and the interviews were recorded with video and sound, and analyzed at a later time.

The participants were not compensated for their time, they were however offered coffee or tea as they arrived at the experiment location.

4. Results

After gathering data from the user tests, qualitative analysis was performed, based on video recordings of participants performing the assigned tasks, as well as interviews conducted with participants after completion of the tasks. In this section, we will present the results of this analysis, based on the source of data. When analyzing the video recordings we mainly looked at the amount of eye contact between assistant and participants, gestures made by the participants and the questions asked. As such, we will begin by presenting our analysis of the recordings, followed by the interviews.

4.1. Tasks

When analyzing the video recordings, we observed a few patterns regarding participants' interactions with the assistant. When interacting with the physically embodied assistant for the first time, participants tended to jump back in their chairs, or visibly express surprise in other ways. However, most users did not react visibly when engaging with the virtual assistant for the first time, and those that did, seemed happy or amused by the assistant, rather than surprised or scared.

We also found that participants attempted to maintain eye contact with the physical assistant, and that they would usually look the robot in the "eye" when attempting to get its attention or ask a question. Two participants thought that Steve might be able to know their appearance, and attempted to ask Steve about their hair color. No participants who interacted with the virtual assistant asked questions about themselves. When they tried to challenge Steve, they instead tried to come up with "things you cannot Google".

Participants conducting the experiment, with the virtually embodied assistant, would look at the screen at the beginning of the test, but as it progressed, they would almost exclusively look at each other or the task paper, and not direct their gaze towards the assistant. At most, they would look at the screen to ensure that the assistant was listening, directing their gaze back to the paper as soon as they started to receive a response.

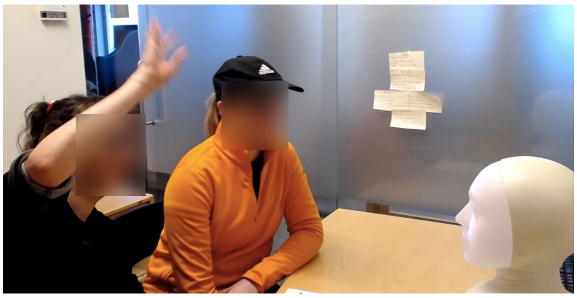


Figure 2: Participants try to get the physical assistant's attention by waving.

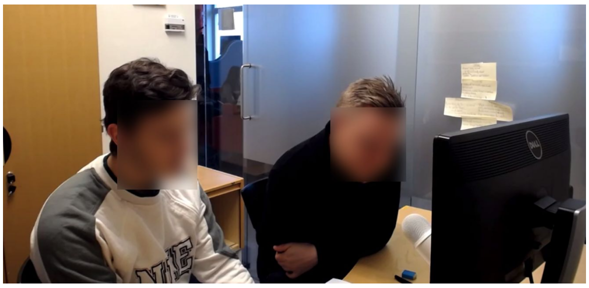


Figure 3: The participant leans forward to be more clear after a failed attempt to wake up the virtual assistant.

When participants failed to wake the physical assistant using a wake command, they tried to get its attention by speaking louder, articulating more or moving closer to the microphone. When this failed, participants tried to get Steve's attention by using gestures such as waving (see figure 2), or moving into his perceived line of sight and looking into his eyes. Some participants also tried to find the limits of Steve's gaze capabilities, by moving around the room while his attention was directed towards them. The participants using the virtual version of Steve did not attempt similar gesture-based methods of acquiring his attention, but would instead repeat themselves, try to articulate more, and usually lean forward in an attempt to be more obvious when trying to get his attention (see figure 3).

When the assistant talked for longer periods of time, users tried to stop it by repeating known commands, but also other phrases to signal that they did not want it to continue, such as "Okay, thank you", "Stop, stop", "Okay Steve, you can stop now Steve" or "Steve! Stop!. As they failed to stop the assistant, participants using the virtually embodied version would simply ignore it, discuss amongst themselves or laugh at the situation; they would also dismiss the assistant as soon as they felt that they had heard enough, something participants using the physically embodied version seemed less comfortable doing. However, when participants tried to stop the physically embodied assistant, they did so by using gestures, such as waving (figure 2), or putting the task paper in front of his face (figure 4).

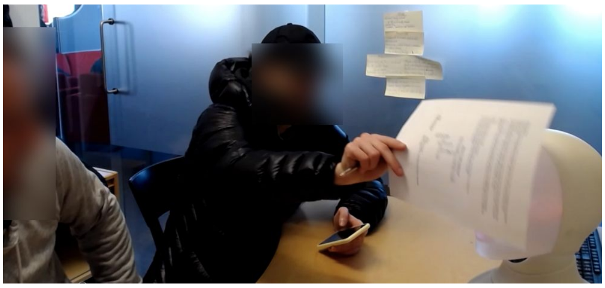


Figure 4: Participants try to interrupt the physical assistant by putting a paper in front of his face.

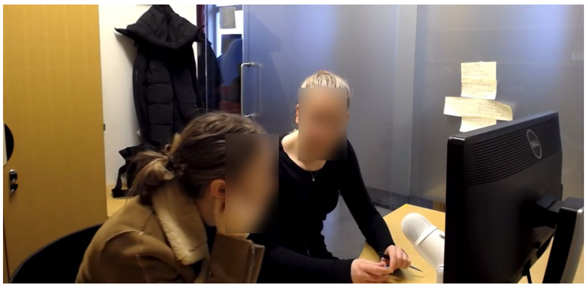


Figure 5: The participant to the right failed to engage the virtual assistant, she looks at the other participant (to the left) who tries instead.

Since the virtual version of Steve could not keep track of which user was speaking, and did not look at the users to indicate that he was attending one of them, users did not seem to feel as though Steve was only listening to one of them, and would sometimes cooperate when he did not react to a wake-word. For example, one user could be trying to get the assistant's attention in order to ask a question, and after a few failed attempts, the other participant would say the wake-word instead; when they finally got Steve's attention, either one of the participants would continue by asking a question. (see figure 5)

On the other hand, the physically embodied version of Steve could keep track of which user was speaking to them, and tried to attend the user that issued the wake command. In the tests using the physical version, the person that issued the wake command would often also ask the question. Nonetheless, when Steve failed to keep track of the users and accidentally attended the wrong person, two different reactions were observed: about half of the participants continued on as if nothing had happened, meaning that the person that woke Steve would continue to ask their question.
However, the other half expressed confusion and uncertainty on how to proceed. In these cases, two kinds of reactions were observed: either the person that had tried to wake Steve made a gesture as to tell the other participant that they should ask the question, since Steve looked at the other person, or the person that Steve looked at made a gesture indicating that they did not understand why Steve was looking at them. (figures 6 and 7).

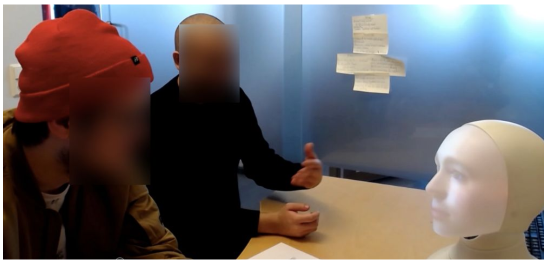


Figure 6: The participant that woke the physical assistant gestures at the other participant to ask the question, since he has Steve's "attention".

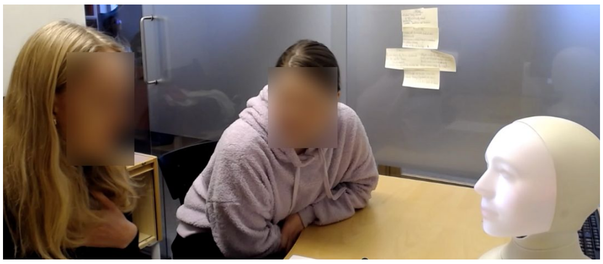


Figure 7:The participant is confused that the physical assistant looks at her and gestures at herself.

4.2. Interviews

After participants had completed the assigned tasks, a short semi-structured interview was conducted with each pair, in order to get more information on how they perceived the interactions with the assistant.

When asked about previous experience with voice assistants, all groups responded that they had used one or more voice assistants before, but not on a daily basis, and none actually owned a physical product with an assistant (excluding one participant's family car).
The participants were therefore somewhat familiar with the usage pattern of having to say a wakeword before asking the question, one group however, commented that "it felt more bothersome than usual, maybe because we were able to see him [physically embodied assistant] and therefore we expected him to understand when we needed him to help.".

During the interviews, two participants told us that they believed that the physical assistant listened better when they looked back into its eyes, one participant also stating that they felt social pressure to maintain eye contact with the robot. Another pair of participants noted that they expected to be able to have a natural dialog with the physical version of Steve, and that they expected that it would be possible to interrupt him and get his attention by waving, "because he was so human-like".

None of the users that interacted with the virtually embodied assistant said that they would want this version of Steve at home. Two groups said that they would have preferred a physically embodied Steve, one reason being that "it would feel less like personal information was being collected constantly as he would be more local" (meaning his data was stored in the robot rather than in a cloud server).

Two groups also mentioned that they would be more comfortable having the virtual Steve at home if the screen was turned off until he was engaged, or if he would have been in a smart mirror. Further, all groups said that they would be somewhat comfortable interacting with Steve in a public space, like their local supermarket, but that they would find the assistant bothersome if it worked like it does today. They would have preferred a more robust version in that case, as mistakes in public were something that they really wanted to avoid.

All of the participants that interacted with the physical assistant expressed discomfort with it, claiming that it was "creepy", mentioning the gaze capabilities and/or humanlike features as reasons why they felt that way.
Most of the participants said that they would have preferred an assistant that did not resemble a human, often mentioning existing voice assistant products, such as Google Home. None of the participants wanted the physical Steve as an assistant in their home, but about half thought it would be a good idea to have Steve in a public space, such as the aforementioned

grocery store, or at a train station. Despite perceiving Steve as creepy, most users said that they enjoyed the test and that they felt that the assistant was able to help them with the tasks.

5. Discussion

From the results we can see that users interact with voice assistant in different ways, depending on how it is embodied. The difference did not affect any of the tasks, as all the participants were able to complete them without any major issues. Gaze, or the lack of it, seems to be an important factor in how users interact with the assistant.

From the results we know that almost all participants tried, at least once, to interact with the physical assistant using body gestures, unlike with the virtual assistant. Participants also had more eye contact with the physical assistant. However, we do not know if this is because of the fact that it was physically embodied, or because it had gaze capabilities. Since gestures need to be seen to be understood, we believe that the gaze capabilities are the most likely reason why the participants used gestures more when interacting with the physical assistant.

It seems that, since the physically embodied assistant had gaze functionality, users seemed to believe that the physical assistant saw much more than it actually did; some users tried to see if the physical assistant knew what they looked like, and many assumed that it understood their gestures. These assumptions seem to have been made based on the physical presence of the robot, as well as the fact that it had limited gaze functionality giving participants the impression that it could “see”.

Overall, the participants did not seem to feel more comfortable with the physical assistant, but rather the opposite, they seemed more comfortable with the virtual one. The virtual Furhat could be ignored and the participants did not feel obligated to look at it. A discussion about social norms and uncanny valley could take place here. The physical presence of the furhat robot seems to have had a large effect on the users, in line with the findings of Thellman et al. [7]. It seems that the physical presence led to users adhering to social norms of conversation when interacting with Steve, something they did not do with the virtual version. However, it also seems that the physical presence is what makes users uncomfortable, since they described the appearance and movements of Steve as belonging in the uncanny valley. A virtual Furhat can easily be ignored as just a screen, something users interact with every day, i.e. something they are used to ignoring. However, the physical assistant felt more secure than the virtual one. The overall reaction to the virtually embodied assistant seemed more positive than the reaction to the physically embodied version.
There were some faults with the assistant itself, such as it having a difficulty understanding commands when the user spoke with an accent, or with a slight lisp. These faults contributed to difficulties in getting Steve’s attention, and him understanding the questions. If Steve would have understood everything that the participants said, the results might have been different. The physical version might for example have been ignored more if it responded

right away, with participants not feeling the need to seek his attention using gaze and gestures. If this was the case, the results might have been more similar across forms of embodiment. The results could also have been opposite; had Steve been able to understand everything, he might have been perceived as more human-like, but we cannot say if it would reduce the uncanny valley effect.

When designing an embodied assistant you have to ask yourself what the goal is. Is it that interacting with the assistant should feel like talking to an actual human, or should it be more discreet and portable? A virtually embodied assistant is easier to take with you, but a physically embodied one could eventually be seen as more of a companion in the future.

5.1 Method criticism

One of our main considerations when designing the experiment was ensuring that participants did not feel uncomfortable interacting with Steve. To reduce this discomfort, we opted to have participants do the experiment in pairs. Being in the room with another human seemed to make participants more relaxed, as they made jokes and light-hearted discussions during the experiment. Even though measures were taken to reduce discomfort, the results show that users were still not fully comfortable with Steve. Perhaps a different embodiment would have led to different results due to less discomfort; however, we cannot say anything for sure.

Another issue with the implementation for the study is the lack of gaze functionality for the virtual version of the voice agent. We believe that the results would have been different if both versions of the agent had gaze functionality. As shown with the physical robot, gaze seemed to make participants believe that Steve could recognize gestures, and maybe even know what they looked like. If this functionality would have existed in the virtual version, the results may have been more similar than they are now.

6. Conclusion

In this paper we have compared a physically embodied assistant with a virtually embodied assistant. Based on qualitative analysis of user testing of both versions, the physically embodied assistant seemed to be perceived as more capable than the virtually embodied one, with users wondering aloud if it could, for example, know what they looked like. However, participants were more comfortable interacting with the virtually embodied assistant. Our theory is that it is was due to the uncanny valley effect, and that physically embodied assistants are more susceptible to being perceived as “creepy”. Therefore, it is important to think about what the purpose of the assistant is, in order to know whether a physical or virtual embodiment is best suited for the task.

7. Acknowledgements

We thank our supervisor Dimosthenis Kontogiorgosfor helping us with access to TMH rooms and equipment. We also thank all of the participants for participating in our study.

8. References

[1]: Henny Admoni and Brian Scassellati. 2017. Social Eye Gaze in Human-Robot Interaction: A Review. J. Hum.-Robot Interact. 6, 1 (May 2017), 25–63. https://doi.org/10.5898/JHRI.6.1.Admoni

[2]: M. Braun, A. Mainz, R. Chadowitz, B. Pfleging and F. Alt, "At Your Service", Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19, 2019. Available: 10.1145/3290605.3300270 [Accessed 28 November 2019]. https://doi.org/10.1145/3290605.3300270

[3]: Donald McMillan, Barry Brown, Ikkaku Kawaguchi, Razan Jaber, Jordi Solsona Belenguer, and Hideaki Kuzuoka. 2019. Designing with Gaze: Tama – a Gaze-Aware Smart Speaker Platform. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 176 (November 2019), 26 pages. https://doi.org/10.1145/3359278

[4]: Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B. (2012). Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In Cognitive behavioural systems (pp. 114-130). Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007%2F978-3-642-34584-5_9

[5]: *Li, J. "The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents" https://doi.org/10.1016/j.ijhcs.2015.01.001*

[6]: *Pfeifer, R., & Scheier, C. (2001). Understanding intelligence. MIT press.*

[7]: *Thellman, Sam, et al. "Physical vs. virtual agent embodiment and effects on social interaction." International Conference on Intelligent Virtual Agents. Springer, Cham, 2016. https://doi-org.focus.lib.kth.se/10.1007/978-3-319-47665-0_44*

[8]: Starkey Duncan. 1972. Some Signals and Rules for Taking Speaking Turns in Conversations. Journal of Personality and Social Psychology 23, 2 (1972), 283–292. https://doi.org/10.1037/h0033031

[9]: Charles Goodwin, 1980. Restarts, Pauses,and the Achievement of a State of Mutual Gaze at Turn Beginning. Sociological Inquiry 50, 3-4 (July 1980), 272–302. https://doi.org/10.1111/j.1475-682X.1980.tb00023.x

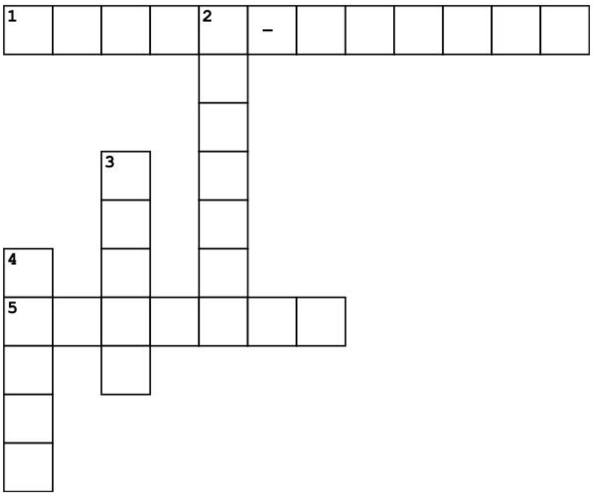
Appendix 1

Introduction:
You will be performing a series of tasks with the help of Steve, our embodied voice assistant. When you have completed all the tasks, notify us to move onto the next step of the experiment!
You interact with Steve the same way you would interact with any other voice assistant; start by grabbing Steve's attention by using a wake-up phrase, such as "Hey, Steve", "Ok, Steve", or "Help us, Steve". After this, you can ask any question you want!

NOTE: Sometimes, Steve fails to provide an answer. If that happens try rephrasing your question! Also note that sometimes Steve will provide a lot of information so you need to be attentive.

Tasks:
1: Ask Steve to tell you a joke.
2: Get Steve to tell you a christmas fact.
3: Ask Steve to solve a simple math problem, such as 2 plus 2, or 5 times 10.
4: Try to find out what questions Steve can/cannot answer.
5: Try solving this crossword:

Steve's cool crossword



Across
1. Rugby world champions
5. Capital of Mississippi

Down
2. President of Ireland
3. Author of "Dubliners" and "Portrait of the artist as a young man"
4. B in ABBA