Is Putting a Face on a Robot Worthwhile?

Enas Altarawneh, Michael Jenkin and I. Scott MacKenzie[1]

Abstract— **Putting an animated face on an interactive robot is great fun but does it actually make the interaction more effective or more useful? To answer these questions, human-robot interactions using text, audio, a realistic avatar, and a simplistic cartoon avatar were compared in a user study with 24 participants. Participants expressed a high level of satisfaction with the accuracy and speed of all the interfaces used. Although the response time was longer for both the cartoon and realistic avatar interfaces (due to their increased computational cost), this had no effect on participant satisfaction. Participants found the avatar interfaces more fun to use than the traditional text- and audio-based interfaces, but there was no significant difference between the two avatar-based interfaces. Putting a face on a robot may make a robot more fun to interact with, and the face may not have to be that realistic.**

I. Introduction

From their experience with robots on TV and in the movies, naïve users expect robots to present a human-like appearances and expressions, and to respond in a natural and appropriate manner. But are such approaches really what the user wants and are they really effective? Is it better to put an animated face on the robot or is a traditional text or audio interface more effective? To consider these questions we used an approach similar to Liang et al. [1] to evaluate the relative performance in responding to queries with a text-only response (T), an audio-only response (A), an avatar response that relies on a cartoon 3D avatar (CA) (Fig. 1a) and a realistic avatar (RA) (Fig. 1b). All interfaces used a common underlying speech recognition and knowledge engine to obtain text responses to participant queries: The text interface displayed the response as text on a screen and then displayed a text prompt that indicates that the interface was ready for the next question. The audio interface generated an audio response, played the audio response, then displayed a text prompt indicating that the interface is ready for the next question. The cartoon avatar provided an audio response loosely synchronized with a cartoon avatar. The cartoon avatar synchronized its lip motion with the audio using two visual states, mouth closed and mouth open, to provide simple and computationally inexpensive lip synchronization to the audio responses. The realistic avatar played the audio response synchronized with the animated character. The design of the realistic avatar is sketched below and for a more complete description of the realistic avatar interface see [2]).

A questionnaire was administered to each participant before the experiment and after their interaction with all of the interfaces. The responses captured participants' demographic data and their perceptions of the interfaces. After interacting with a given interface interaction, information related to that interface was gathered.

The results presented here are part of a larger study [3][1]. The empirical evaluation and analysis follows methods detailed by MacKenzie [4]. Ethics approval for this study was granted from the Office of Research Ethics of an anonymous university.

II. Prior work

An artificially intelligent agent is an autonomous entity that observes the environment through sensors and acts upon it using actuators, directing its activity towards achieving a specific set of goals [5]. An intelligent agent has applications in almost every field. A common theme in intelligent agents is the use of anthropomorphic features as a mechanism to structure interactions with the user. Putting a "head" on the intelligent agent gives the user something to talk to. This concept of an *interactive avatar* can be found in interactive displays more generally. Interactive avatars and virtual agents have been used as the basis of the interface for a range of applications including home care monitoring and companionship (see[6]), and interactive avatars are commonplace in online shopping (see[1], [7], [8]). Interactive avatars are inherently multi-modal in nature and can enable a more intimate relation between the user and the avatar then is the case for more traditional user interface technologies [5]. But what are the appropriate set of interactive modalities to use in an intelligent agent and what is the necessary fidelity of these modalities?

An interactive avatar typically relies on text to speech and speech understanding technologies to provide voice interaction and couples this with a synchronized visual display. Applications that use natural language as an interface engage in conversations as humans naturally do. There are many examples of this type of interaction including commonality systems such as Siri [9], Alexa [10] and Cortana [11]. But what are the advantages and disadvantages of the various interaction approaches? For example, Medicherla and Sekmen [12] report results of a user study that indicates that voice-control and the ability of spatial reasoning were reliable indicators of efficiency in robot teleoperation. In this study 75% of the subjects who demonstrated a high ability to apply spatial reasoning favored using voice-control over manual control. But are voice-based interfaces preferred? Voice-based approaches can produce realistic audio, but human

[1]Authors are with Faculty of Electrical Engineering and Computer Science, York University , Toronto, Canada enas, jenkin, mack @eecs.yorku.ca

[1]The financial support from anonymous projects are gratefully acknowledged.

(a) CA (b) RA

Fig. 1: The avatar-based interfaces used in this study. (a) The cartoon avatar interface (CA). (b) The realistic avatar interface (RA). Both avatars are synchronized to the audio utterances. The cartoon avatar relies on a simple open/close mouth lip synchronization, while the realistic avatar uses a more sophisticated synchronization process.

perception is multi-sensory. Humans use multiple senses when interacting with their environment. We combine audio information and the movements of the lips, tongue and other facial muscles generated by a speaker in order to recognize emotion and behavior [13]. HRI systems that combine audio and a visual talking realistic head rendering of an utterance are likely to improve a user's perception of the interaction over interaction devices that lack these features. But are people more comfortable in interacting with realistic human avatars?

When humans interact and collaborate with each other they use verbal and nonverbal signals to coordinate their turn-taking actions [14]. These signals are communicated in many ways including through expressions of facial and vocal cues. Social robots of the future are expected to interact with "naïve" humans, thus, it would seem to be critical that these social robots simulate and recognize these cues. Skantze [14] reviews a number of studies showing that humans in their interaction with a human-like robot make use of coordination cues found in human-human interaction. This study also shows that it is possible for a robot to detect these cues in humans and then use them to facilitate real-time coordination. Given that it is possible to recognize and simulate such cues, is it desirable?

The perception of social robots varies from individual to individual. The representation, behavior and visual characteristics of a robot can have an effect on the user's perception of the agent, it's intelligence and perceived safety during the interaction. For example for elderly users, the perception of a robot avatar changes with the simulated age of the avatar [15]. There is even evidence that some people prefer to talk to robots rather than with other people under certain circumstances. For example, Niemelä et al. [7] showed that people tend to respond positively to social service robots in field trials in public places. The results of the survey conducted by Niemelä et al. [7] indicate a high social acceptance among humans engaging with service robots in a shopping mall. However, it is unknown how their opinions and attitudes might evolve if the same robot continues to be presented in the same service after any novelty effect wears off.

Researchers have been redesigning robots to look and sound more like humans. For example, Di Salvo et al. [16] identified features and dimensions that can be used to modulate how a human-like robot head will be perceived. Walters et al. [17] investigated people's perceptions of different robot appearances and found that participants tended to prefer robots with more human-like appearance and attributes. However, the study also found that participants with lower emotional stability tended to prefer the mechanical looking appearance. Kalegina et al. [18] suggest that there is preference for varying levels of realism in robot faces based on context, indicating different levels of realism suit different jobs. Broadbent et al. [19] reported that humans found robots with a human-like face display to be less sociable and less trustworthy. These results suggest that the more human-like a robot's face display is, the more people attribute personality characteristics to it.

III. Method

Participants. Twenty-four English-speaking subjects participated in the study. The participants were divided randomly into four gender balanced groups to counterbalance the order of testing and to offset any learning effects. Each participant experienced all four interfaces. Participants were between 18 and 34 years of age ($\bar{x}$ = 22.1). Participants' education level ranged from a four year bachelor degree to a PhD degree. Participants received a ten dollars gift card as an incentive for participation.

Apparatus. Prior to the experiment a questionnaire was used to gather information about and from the participants. A second questionnaire was provided after the experiment to gather information about the participants' experience in the experiment. Testing was performed using a laptop (a Hewlett Packard with Intel Core i7-8550U Processor at 1.8 GHz processor, 16 GB DDR4 (2-DIMM) RAM, 1 TB 7200 RPM SATA Hard Drive and a 15 inch screen). For audio input and output, the laptop's microphone and speaker were used. The laptop used software developed under Ubuntu 16.04.3 LTS (Xenial Xerus) that animated the 3D avatar response, animated the 3D cartoon avatar response, and obtained audio and text responses.

The Interfaces. Four approaches were compared, in which voice-based participant queries were responded to using a text-only response (T), an audio-only response (A), an avatar-based response that relies on a cartoon 3D avatar (CA), and an avatar-based response that relies on a realistic 3D rendered avatar (RA). The interfaces used a common underlying speech recognition and knowledge engine to obtain text responses for participant queries but differed in how responses were presented to the participant.

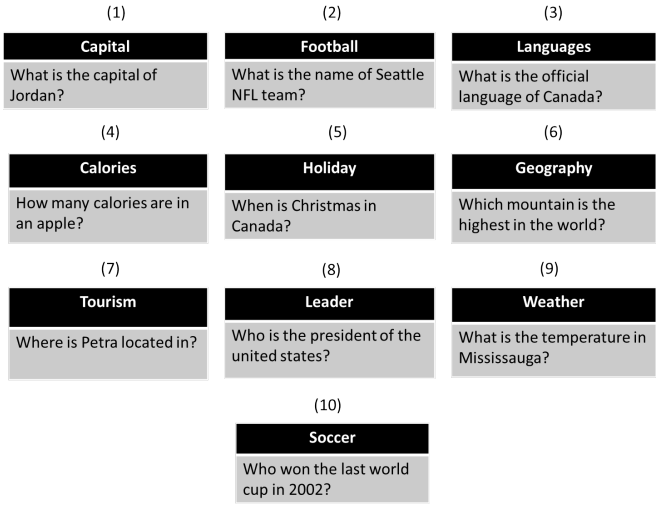

Fig. 2: *The ten question categories and a sample question from each category.*

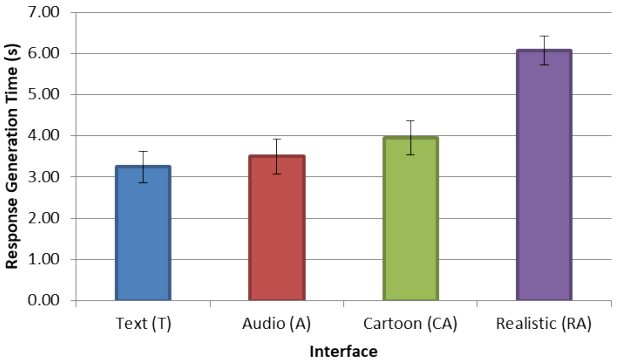

Fig. 3: *Mean response generation delay (s) by interface. Error bars show* ± 1 *SD. Post hoc testing reflected that all pairwise comparisons are statistically different except the text interface (T) with the audio interface (A).*

The text interface displayed the response as displayed text on a laptop screen and then displayed another text message to indicate that the interface was ready for the next question.

The audio interface used the text response to generate an audio response; it played the audio response and then displayed a text message on the laptop screen to indicate that the interface was ready for the next question.

For the cartoon avatar response animations were pre-rendered on the local machine. The cartoon avatar presents in two states; mouth closed and mouth open, and this provides a simple and computationally inexpensive Lip-synchronization to the responses. The cartoon avatar uses the generated audio response and plays this synchronized with the animated character.

The realistic avatar utilizes the approach described in [anonymous] that leverages a number of cloud-based software components. It relies on a speech-to-text recognition module, a knowledge engine, a text-to-speech engine, a 3D character designing program, a 3D animation program, and a lip-syncing plugin for the animation program that extracts the sounds in words, maps them to mouth shapes and plots them according to duration and occurrence in the text in real time. Lip-synchronization is animated based on the utterance and an Avatar Delay Graph (ADG) is used to animate the avatar between utterances. The realistic avatar uses the text and audio generated for lip synchronization animation. An expression package controls the animated character's mood and facial expressions. Recognizing that the use of cloud-based resources will introduce unwanted delays in the recognition and rendering process, the realistic avatar utilizes: (i) an adaptive parallelization strategy that leverages cloud-based rendering resources to minimize the latency itself, This rendering farm is used to provide an actual lip synchronization of all the sounds in the response in real-time. (ii) an "idle loop" process to obscure any resulting latency in the recognition, response and rendering process. This process animates the avatar puppet between utterances so that the character being rendered is not still but rather appears to interact with external users even when not being spoken to directly. This process also further obscures rendering latency.

Procedure. Experimental trials took place with the participant seated at a desk equipped with a laptop. Each participant was briefed on the purpose of the experiment and read and signed an informed consent form. Using the computer, the biographic questionnaire was presented. Upon completing this questionnaire, the participant was shown their first interface. Each participant was asked to use the interface to ask a list of questions of the interface. Questions to be asked by the participants to each interface were presented to the participant on a sheet of paper. Each interaction with a given interface consisted of the participant asking the interface ten questions. For each question the participants asked the question and waited for the response. Once all interactions with a given interface were completed, the participant moved on to the next interface. The interfaces were presented in a counterbalanced order. Following the forth interface, the participants completed the exit questionnaire. All participants asked the same set of 40 questions, broken down into ten categories. Each category contains four questions of a similar nature. See Fig. 2 for the ten groups of questions and sample questions from each group .

Design. As the individual question categories are uninteresting we average quantitative measures related to timing over the question categories. This results in a within-subjects design with one factor (interfaces) having four levels; the Text (T), Audio (A), Realistic Avatar (RA) and Cartoon Avatar (CA) interfaces.

Recommendations provided by Zhao[20] were used to help design the questionnaires used. The questionnaire completed prior to the experiment focused on collecting information about the participant. After interactions with the four interfaces were completed the post-experiment questionnaire focused on quality metrics described in [21] and include functionality (Executes requested tasks, Accuracy of output,

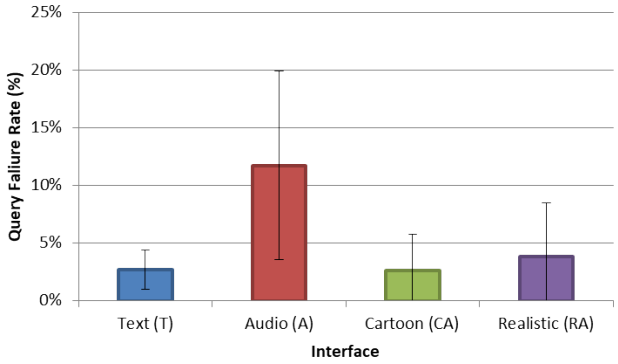

Fig. 4: *Mean query failure rate (%) by interface. Error bars show* ± 1 *SD. Post hoc testing showed that the audio interface (A) has a significantly different query failure rate than the other interfaces.*

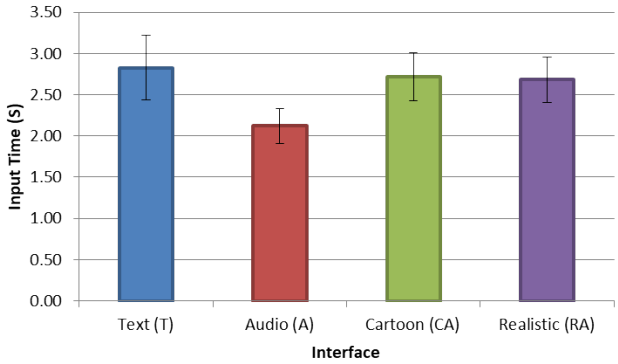

Fig. 5: *Input time (s) by interface. The audio interface (A) has a significantly lower input time than the other interfaces.*

and General ease of use), Humanity (Convincing, Satisfying, and Natural interaction), Affect (Makes tasks more interesting and fun), and Ethics and behavior (Trustworthiness). The questions in the post evaluation questionnaire were taken from or inspired by the work of Jaferian[22] and other post-evaluation human-computer interaction including [23][24][25][21]. The study also makes use of some questions in the usefulness and ease of use categories from the Perceived Usefulness and Ease of Use Questionnaire (PUEU)[25], from the Computer System Usability Questionnaire (CSUQ) and the After-Scenario Questionnaire (ASQ) [24]. Questions were either used unchanged or after minor modification to harmonize the question style.

IV. Results

For purely quantitative data, a repeated measures ANOVA was performed. For other measures a Friedman non-parametric test was used. An application by MacKenzie called GoStats [26] was used to analyze the collected data using the required method of analysis.

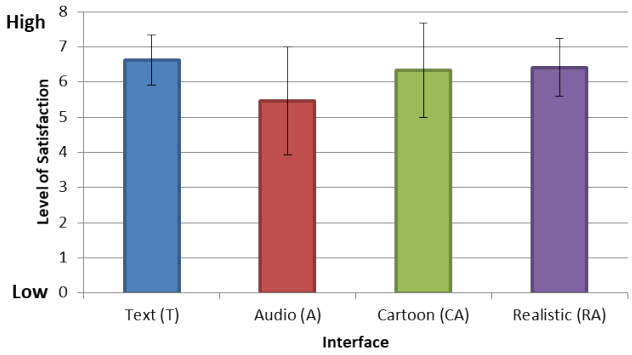

Fig. 6: *Mean participant satisfaction with the interactions. Error bars show* ± 1 *SD. Post hoc testing showed that the audio (A) interface is significantly different from the other interfaces.*

Response generation delay. The response generation delay is the time it takes the response system to be able to respond. The response generation delay starts after the participant finishes asking a question and ends once a response is ready to be presented. The means for response generation time by interface were Text (T): 3.25 (s), Audio (A): 3.50 (s), Cartoon Avatar (CA): 3.96 (s), and Realistic Avatar (RA): 6.07 (s) as shown in Fig. 3. The main effect of interface type on response generation time was statistically significant $(F_{3,\,60} = 188.2,\, p < .0001)$. A Bonferroni-Dunn post hoc test revealed that all pairwise comparisons were statistically significant except for the text and audio interface pair. The text and audio interfaces generated responses faster than the avatar interfaces and the cartoon avatar produced a response faster than the realistic avatar.

Input time. Input time is the duration of the recognized speech uttered by the participant. By interface, the input times were text (T): 2.83 (s), audio (A): 2.12 (s), cartoon avatar (CA): 2.72 (s), and realistic avatar (RA): 2.68 (s). See Fig. 5. There was a significant effect of interface on input time $(F_{3,\,60} = 25.9,\, p < .0001)$. A Bonferroni-Dunn [4] post hoc test revealed that all pairwise comparisons with the audio interface were statistically significant. Input time for the audio interface was significantly less than the other interfaces.

Query failure rate. Query failure rate is the percentage failure (not getting a successful response). The means for Query failure rate by interface were Text (T): 3%, Audio (A): 12%, Cartoon Avatar (CA): 3%, and Realistic Avatar(RA): 4% as shown in Fig. 4 . The main effect of interface on query failure rate was statistically significant $(F_{3,\,60} = 6.228,\, p < .001)$. A Bonferroni-Dunn post hoc test revealed that all pairwise comparisons with the audio interface were statistically significant. The audio interface had a higher query failure rate than the other interfaces.

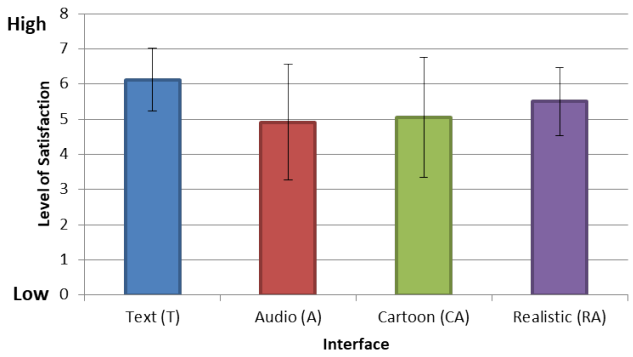

Fig. 7: *Mean participant satisfaction with time to obtain a response by interface. Error bars show* ± 1 *SD. Post hoc testing showed that the text (T) interface is significantly different from the other interfaces.*

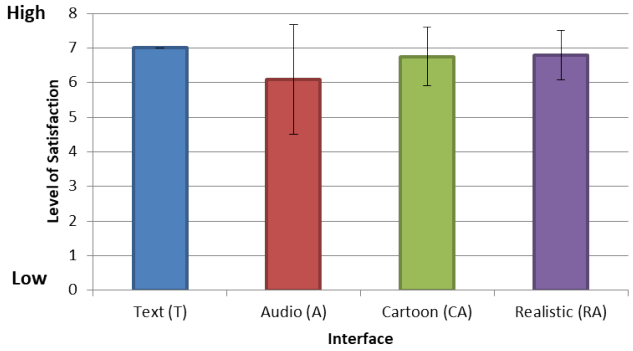

Fig. 8: *Mean participant perception on accuracy of the responses. Error bars show* ± 1 *SD. The variance in the text interface data is zero. All participants selected level 7 for the text interface. Post hoc testing showed that the audio (A) interface is significantly different from the other interfaces.*

V. Questionnaire Results

The questionnaire responses were analyzed using the non-parametric Friedman test.

Participant satisfaction with the interaction. The means of participant satisfaction level with the interaction by interface (1 = lowest level, 7 = highest level) were Text (T): 6.625, Audio (A): 5.458, Cartoon Avatar (CA): 6.333, and Realistic Avatar (RA): 6.416 are shown in Fig. 6. All interfaces have a high level of participant satisfaction with the interaction. The audio interface had the least participant satisfaction level. There was a significant difference in the level of participant satisfaction with the interfaces $(\chi^2 = 11.826, p < .01, df = 3)$. A Conver's F post hoc test revealed that all pairwise comparisons with the audio interface were statistically significant. Participants were least satisfied with the audio interface.

Participant satisfaction with the time to obtain a response from the interface. The means of participant satisfaction level with the time to obtain responses by interface (1 = lowest level, 7 = highest level) were Text (T): 6.125, Audio (A): 4.916, Cartoon Avatar (CA): 5.041, and Realistic Avatar (RA): 5.5 are shown in Fig. 7. All interfaces have a high level of participant satisfaction. There was a significant difference in the level of participant satisfaction with the amount of time to obtain responses by the interfaces $(\chi^2 = 10.243, p < .05, df = 3)$. A Conver's F post hoc test revealed that only pairwise comparisons with the text interface were statistically significant. The text interface had the highest level of participant satisfaction with the time to obtain a response.

Participant perception on accuracy of the response of a given interface. The means for participant perception level on accuracy of the responses by interface (1 = lowest level, 7 = highest level) were Text (T): 7, Audio (A): 6.083, Cartoon Avatar (CA): 6.75, and Realistic Avatar (RA): 6.79 are shown in Fig. 8. All interfaces have a relatively high level of participant perception on accuracy of the responses. The audio interface has the lowest level of participant perception of accuracy of the responses. There was a significant difference in the participant perception on the accuracy of the responses given by the interfaces $(\chi^2 = 14.143, p < .01, df = 3)$. A Conver's F post hoc test revealed that only pairwise comparisons with the audio interface were statistically significant. Participants perceived the audio interface as being less accurate than the other interfaces.

Participant perception of how fun each interface is to use. The means for participant perception level of how fun each interface is to use by interface (1 = lowest level, 7 = highest level) were Text (T): 4.875, Audio (A): 5.041, Cartoon Avatar (CA): 5.708, and Realistic Avatar (RA): 5.958 are shown in Fig. 9. All interfaces have a relatively high level of participant perception of how fun each interface is to use. There was a difference in the participant perception of how fun each interface is to use $(\chi^2 = 16.746, p < .001, df = 3)$. A Conver's F post hoc test revealed that all pairwise comparisons with the avatars interfaces were statistically significant. Participants found the avatar-based interfaces more fun to use.

Participant preferences between the text-based and audio-based interfaces. Fig. 10 illustrates the number of participants that selected each level of preference for these two interfaces. Eleven of 24 participants were highly confident with their preference for the audio-based interface over the text-based interface. In total there were 8 participants that preferred text and 14 that preferred audio. Two participants did not have a preference.

Participant preferences between avatar-based and audio-based interfaces. Fig. 11 shows that 9 of 24 participants were highly confident with their preference of the avatar-based interfaces over the audio-based interface. In total there were 10 participants that preferred the audio-based interface

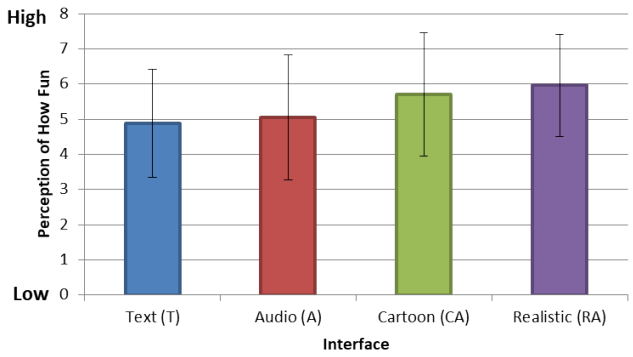

Fig. 9: *Mean participant perception of how fun each interface is to use. Error bars show* ± 1 *SD. Post hoc testing showed that pairwise comparisons with the avatar interfaces are significantly different.*

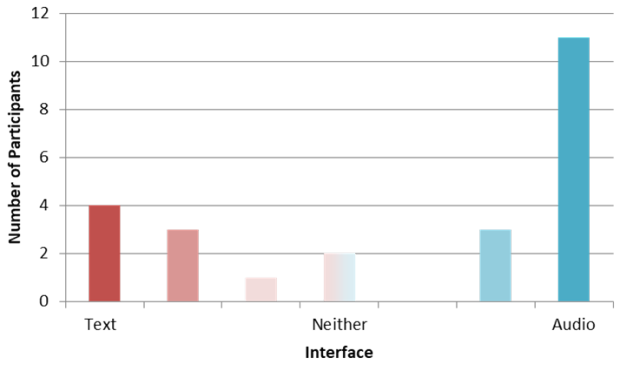

Fig. 10: *Participant preferences between the text-based and audio-based interfaces.*

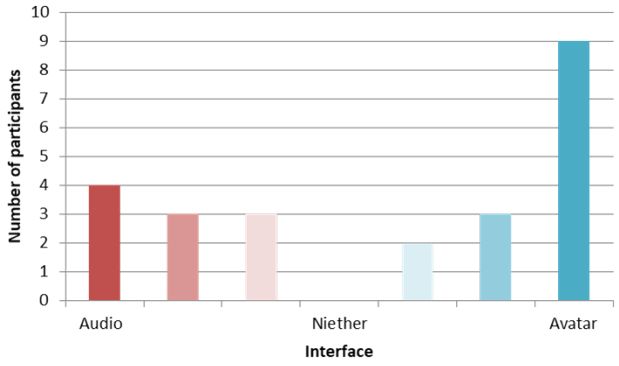

Fig. 11: *Participant preferences between avatar-based and audio-based interfaces.*

and 14 that preferred the avatar-based interfaces.

Participant preferences between realistic avatar-based and cartoon avatar-based interfaces. Fig. 12 shows that 9

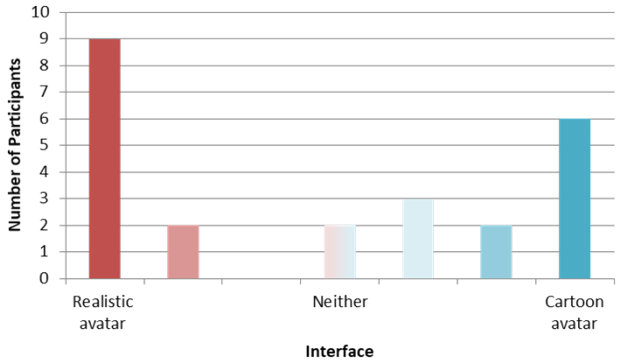

Fig. 12: *Participant preferences between realistic avatar-based and cartoon avatar-based interfaces.*

of 24 participants were highly confident with their preference of the realistic avatar interface over the cartoon avatar interface.

VI. Discussion

Participants in general expressed a high level of satisfaction with the responses and the speed and accuracy of the responses for all interfaces tested. Participants in general found all interfaces to be fun to use, again suggesting that all interfaces could be used to develop human-robot interaction systems. The study also requested feedback on the participants general preferences among types of interfaces. In general participants preferred the audio interface over the text interface, the avatar interfaces over the audio interface and the realistic avatar interface over the cartoon avatar interface.

The text interface had the lowest response generation time and the realistic avatar interface had the highest response generation time. There was a significant difference between the response generation time of the text interface and the other interfaces. This is to be expected as the text interface displays the result as text while the other interfaces require additional processing. As the complexity of the user interface increased so did the necessary processing time.

The text interface had the highest level of participant perception in terms of the accuracy of responses, however it was not significantly higher that the avatar interfaces.

Input time in the audio interface was significantly less than with the other interfaces, indicating that participants spoke faster when asking questions using this interface. The audio interface also had a higher query failure rate possibility as a result of the use of this strategy. Hong and Findlater [27] support this conclusion. They suggest that faster speech results in more errors in speech recognition. But why did the participants speak faster in this condition? This may be related to the cognitive structure of the audio-only condition within which participants had no other competing tasks (in the text condition they had to read the on-screen responses while in the avatar conditions the participant was also engaged with the on-screen avatar). Eichornt et al. [28] conclude that distractions during speech leads to a reduced "speech rate".

Although participants showed a high level of satisfaction with the time taken to obtain a response and the accuracy of the audio interface, the satisfaction level was significantly lower than the other interfaces. This may be due to the audio interface having a higher query failure rate. Several studies [29], [30], [31] conclude that errors in interactions are associated with lower general satisfaction. Adding either a carton or realistic avatar to the audio response improves perceived accuracy and the perceived fun of using the interface.

Even through the response time for the realistic avatar was significantly higher than that for the other interfaces, participants still expressed a high satisfaction level with the time required to obtain responses from the avatar-based interfaces. There was no significant difference in participant satisfaction with the time to get responses from the realistic avatar interface, the audio interface and cartoon avatar interfaces. Participants also found the avatar interfaces to be significantly more fun to use than the other interfaces.

VII. Summary and future work

Technologies exist that can be used to put a realistic avatar face on a robot, but is the added effect required to render the avatar worthwhile? Here we described an empirical evaluation of interaction through text (T), Audio (A), Realistic Avatar (RA) and Cartoon Avatar (CA) interfaces. As anticipated the time to obtain a response was significantly higher for the avatar interfaces. This however, had no significant effect on the participant satisfaction with the responses given by these interfaces. The audio-only interface had a lower user satisfaction. The high query failure rate may explain the significantly lower satisfaction level shown by participants for the audio interface. In general, participants expressed a high level of satisfaction with the accuracy and speed of all the interfaces. They also expressed that all interfaces were fun to use. The realistic avatar was found to be more fun the cartoon avatar, although this difference was not found to be significant. Participants found the avatar interfaces to be significantly more fun than the other interfaces. If robots are to be "your plastic pal who's fun to be with"[2] then it may be prudent to put a face on the robot. But it may not be necessary to make the avatar that realistic. A cartoon-like avatar may be sufficient.

References

[1] W. Liang, C. Huang, T. Tseng, Y. Lin, and J. Tseng, "The evaluation of intelligent agent performance - an example of B2C e-commerce negotiation," *Computer Standards and Interfaces*, vol. 34, no. 5, pp. 439 – 446, 2012.

[2] E. Altarawneh and M. Jenkin, "Leveraging cloud-based tools to talk with robots," in *the International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, 2019.

[3] E. Tarawneh, "A cloud-based extensible avatar for human robot interaction," Master's thesis, York university, Toronto, Canada, 2019.

[4] I. S. MacKenzie, *Human-Computer Interaction: An Empirical Research Perspective*, 1st ed. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2013.

[5] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Saddle River, NJ, USA: Morgan Kaufmann Publishers Inc., 2003.

[6] N. A. Shaked, "Avatars and virtual agents - relationship interfaces for the elderly," *Healthcare Technology Letters 4.3*, pp. 83–87, 2017.

[7] M. Niemelä, A. Arvola, and I. Aaltonen, "Monitoring the acceptance of a social service robot in a shopping mall: First results," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. New York, NY: ACM, 2017, pp. 225–226.

[8] E. Strang. (2017) Soul machines unveils its first emotionally intelligent, lifelike avatar. [Online]. Available: https://idealog.co.nz/tech/2017/02/soul-machines-unveils-its-first-emotionally-intelligent-lifelike-avatar

[9] M. Galeso, *Apple Siri for Mac: An Easy Guide to the Best Features*. USA: CreateSpace Independent Publishing Platform, 2017.

[10] M. Alexa, *Amazon Alexa: 2017 User Guide + 200 Ester Eggs*. Independently published, 2017.

[11] M. Hoy, "Alexa, siri, cortana, and more: An introduction to voice assistants," *Medical Reference Services Quarterly*, vol. 37, pp. 81–88, 01 2018.

[12] H. Medicherla and A. Sekmen, "Human-robot interaction via voice-controllable intelligent user interface," *Robotica*, vol. 25, pp. 521–527, 09 2007.

[13] H. Mcgurk and J. Macdonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[14] G. Skantze, "Real-time coordination in human-robot interaction using face and voice," *Ai Magazine*, vol. 37, pp. 19–31, 12 2016.

[15] A. L. Marin, D. Jo, and S. Lee, "Designing robotic avatars are user's impression affected by avatar's age?" in *8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Tokyo, Japan, March 2013, pp. 195–196.

[16] C. F. DiSalvo, F. Gemperle, J. Forlizzi, and S. Kiesler, "All robots are not created equal: The design and perception of humanoid robot heads," in *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, ser. DIS '02. New York, NY, USA: ACM, 2002, pp. 321–326.

[17] M. Walters, D. S. Syrdal, K. Dautenhahn, R. Boekhorst, and K. Koay, "Avoiding the uncanny valley: Robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion," *Auton. Robots*, vol. 24, pp. 159–178, 02 2008.

[18] A. Kalegina, G. Schroeder, A. Allchin, K. Berlin, and M. Cakmak, "Characterizing the design space of rendered robot faces," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18. New York, NY, USA: ACM, 2018, pp. 96–104.

[19] E. Broadbent, V. Kumar, L. Xingyan, J. Sollers-3rd, R. Q. Stafford, B. A. MacDonald, and D. M. Wegner, "Robots with display screens: A robot with a more humanlike face display is perceived to have more mind and a better personality." *PLoS ONE*, vol. 8, 08 2013.

[20] S. Zhao. (2017) How to design questionnaires for usability evaluatione. [Online]. Available: http://www.shengdongzhao.com/research_tips/how-to-design-a-questionnaire-for-usability-evaluation/

[21] N. M. Radziwill and M. C. Benton, "Evaluating quality of chatbots and intelligent conversational agents," *CoRR*, vol. abs/1704.04579, 2017.

[22] P. Jaferian. (2017) Post-evaluation questionnaire. [Online]. Available: http://ece.ubc.ca/~pooya/hestudy/pc1/postevalquest.html

[23] J. P. Chin, V. A. Diehl, and K. L. Norman, "Development of an instrument measuring user satisfaction of the human-computer interface," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. New York, NY, USA: ACM, 1988, pp. 213–218.

[24] J. Lewis, "IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use," *Int. J. Hum.-Comput. Interact.*, vol. 7, no. 1, pp. 57–78, Jan. 1995.

[25] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Q.*, vol. 13, no. 3, pp. 319–340, Sep. 1989.

[26] I. S. MacKenzie. (2017) Gostats - statistics app, featuring anovagui. [Online]. Available: http://www.yorku.ca/mack/GoStats/

[27] J. Hong and L. Findlater, "Identifying speech input errors through audio-only interaction," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: ACM, 2018, pp. 567:1–567:12.

[28] N. Eichornt, K. Marton, R. Schwartz, Melara, and S. Pirutinsky, "Does working memory enhance or interfere with speech fluency in adults who do and do not stutter? evidence from a dual-task paradigm," vol. 59, pp. 415–429, 1 2016.

[2] Adams, D.*The Hitchhiker's Guide to the Galaxy*, Harmony Books, 1979.

[29] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human-robot interaction: Literature review and model development," *Frontiers in Psychology*, vol. 9, Jun 2018.

[30] A. Weinstock, T. Oron-Gilad, and Y. Parmet, "The effect of system aesthetics on trust, cooperation, satisfaction and annoyance in an imperfect automated system," *Work 41 (Suppl. 1)*, pp. 258–265, 2012.

[31] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. New York, NY, USA: ACM, 2015, pp. 141–148.