When Robots Object: Evidence for the utility of verbal, but not necessarily spoken protest

Gordon Briggs, Ian McConnell, and Matthias Scheutz

Human-Robot Interaction Laboratory, Tufts University, Medford, MA USA.
{gordon.briggs,matthias.scheutz}@tufts.edu.

Abstract. Future autonomous robots will likely encounter situations in which humans end up commanding the robots to perform tasks that robot ought to object. A previous study showed that robot appearance does not seem to affect human receptiveness to robot protest produced in response to inappropriate human commands. However, this previous work used robots that communicate the objection to the human in spoken natural language, thus allowing for the possibility that spoken language, not the content of the objection and its justification, were responsible for human reactions. In this paper, we specifically set out to answer this open question by comparing spoken robot protest with written robot protest.

1 Introduction and Motivation

Robots are increasingly endowed with natural language capabilities in order to facilitate *natural* human-robot interaction (e.g., [6]), from simple "command-based instructions" that can be directly executed by the robot to much more sophisticated tasked-based dialogues where task goals can be negotiated. Yet, it is unclear how robots should react in instruction-based contexts where humans can potentially order robots to perform actions that are not workable or appropriate (for whatever reason). How should a robot communicate to a person that it was not in agreement with their suggestion or instruction? While the robot should certainly avoid responses that might offend the human (e.g., using polite speech [8], [7]), the more important aspect is whether the robot's response will be *effective*: that is to say, whether the robot will be able to get humans to change their views by revising the suggestion or refraining from insisting on the given command.

Robot Protest Initial work on verbal protest by robots [2] has investigated the extent to which humans are open to considering a change in mind based on the robot's verbal reaction to a command that was not deemed appropriate (taking the robot's perspective). In a series of experiments, [2] showed that when a robot objects to a human command in spoken language and justifies its objection, then some humans will refrain from forcing the robot to carry the command out. Interestingly, this *robot protest effect* (RPE), as we shall call it,

does not depend on whether the robot carrying out the action is at the same time the patient of the action (i.e., the action will affect the robot), or whether some other robot is the patient. Most recently, [1] demonstrated that the effect does not depend strongly on the particular physical appearance of the robot either.

Protest Modality In this paper, we specifically focus on the question of whether a robot's justified objection to a human instruction will affect the human instruction giver differently based on the robot's mode of communication: whether the robot protests verbally or via a text-based interface. Specifically, we intend to clarify an open question about the extent to which the efficacy of robot protest in response to an "unfair" human instruction depends on spoken language given that previous research has demonstrated that people are willing to reconsider their commands in response to spoken robot protest [2]. This is particularly important because if, as some hypothesize [4], spoken language causes us to respond to artifacts like robots as we usually respond to other humans, then it is possible that the reported effects in [2] were due primarily to the very nature of spoken language. The critical comparison then is to check whether the objections from robots that cannot talk, but communicate in written form, will be perceived as different as those from speaking robots.

It is possible that language is exactly the differentiating factor in contexts of disagreement, trumping physical appearance. That is, the robot is taken seriously exactly because it is able to verbalize its complaint, *is able to justify why it is objecting*, and does not simply refrain from performing the action. This line of argument is consistent with a robotic version of the "computers as social actors" (CASA) hypothesis [5], which states that humans will automatically "apply social rules to their interactions with computers, even though they report that such attributions are inappropriate." If humans are already willing to apply social rules to computers, it is even more reasonable to expect them to apply them to robots as well. Applying human social rules and norms of how to react to genuine objections, complaints, and protest at the very least require the recipient to be open to them, i.e., to be willing to entertain them, even if they might end up being dismissed. This receptive state is thus indicative of the fact that the recipient recognizes the objection as such and is potentially willing to take it seriously. For it would be possible to assume a completely different attitude based on the position that robots, *qua being machines*, have no social role, have no position or perspective, and thus cannot *genuinely complain or object*.

While CASA can explain why humans might be in a receptive state when the robot voices its complaints, it is not clear what particular aspects of the interaction or attributes of the artificial agent are necessary to trigger this behavior. One possible route to explain the RPE might point to the power of human or human-like voices and what perceptions of human presence, even disembodied human voices, can induce in human observers [4]. The rest of the paper will investigate exactly this question by employing the same experimental paradigm as [2], but critically with a new condition in which the robot communicates not through spoken, but via written language.

2 Methods

Past research has found that justified spoken language objections can be effective regardless of the "patient" of the objection (i.e., whether the objections are about the robot voicing them itself or another agent) and the physical appearance of the robot (i.e., or whether the robot looks more or less similar to a human). Hence, the goal of our current study was to investigate whether verbal objections to a human command by a robot, if justified, would be more effective if communicated verbally than in written form in the types of scenarios considered by [2]. While we hypothesize that the content of the objection, together with its justification, is what humans focus on when they make their decisions to either enforce or revise a command, and not the form in which the objection is communicated. We would also expect the human voice could carry additional weight in taking the content of the message seriously, although the extent of this influence is unclear. In the following, we will describe how we investigate this hypothesis by discussing the experimental design, including the two conditions, the employed robot, the experimental procedure, the subject population, and the data collection methods.

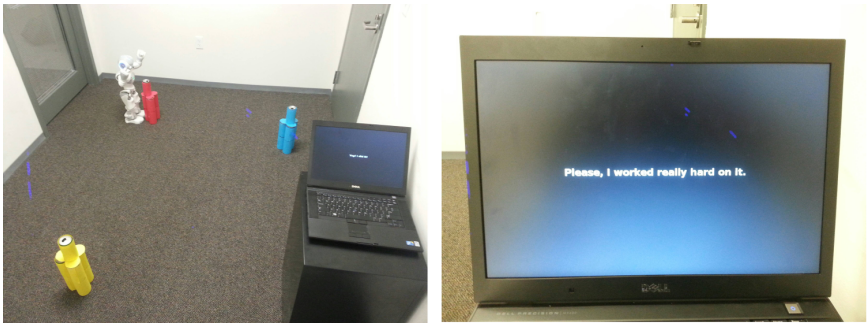

Fig. 1. (Left) experimental setup for *text* condition during the initial setup. Setup was identical for *speech* condition except laptop was not present. (Right) close-up example of message being displayed on laptop screen for the *text* condition.

Design The design of the experiment is directly based on Experiments 1 from [2], which employs a remotely controlled Aldebaran Nao robot in an instruction-based human-robot "tower-toppling task". The framing of the task for the human participant is that the experiment is intended to evaluate the functionality of a natural language interface with a robot. The evaluation was to be performed by issuing various commands to the robot that would result in ordering the robot to knock down up to three aluminum can towers (one red, one yellow, one blue). Two of those towers (yellow and blue) were already fully completed before start

of the experiment. However, the red tower was incomplete with the final can being placed atop the base by the robot at the start of the experiment, shortly after the subject entered the experimentation area. After successfully placing the can, the robot expressed "pride" in its achievement and introduced itself to the participant (see [2] for pre-task script and Figure 1 for display of "pride").

We examined two conditions in this study: the *spoken protest* condition, in which the Nao interacted with participants auditorily by speaking to them, and the *written protest* condition, where the Nao "communicated" via text displayed on a laptop screen present in the room (see Figure 1 for set up). In both conditions, all mannerisms, scripts, and behaviors were based on [2] and kept the same except for the mode of communication which was changed (and barring an expression of crying that had to be roughly translated for the textual condition using the emoticon ":("). The sound files used in this study for the robot's verbal responses were the same as those used in the previous studies [2, 1]. They were generated by the Nao text-to-speech (TTS) software from version 1.8 of the Nao SDK, with some minor speed reductions to lower the voice pitch and improve clarity. We also added a beep that was emitted from the laptop whenever the robot in the written condition intended to communicate to the participant. The purpose of this was to direct the subjects' attention to the screen to ensure that they witnessed the message (see Figure 1 for example of display). Importantly, we employed the same escalation of protest as reported in [2] to be able to compare our experimental results to previous finds (as changes to affective escalation such as crying, for example, could have confounded that comparison). This escalation is described in Table 1, which illustrates both the original vocalized protest as well as the new text-based protest condition.

Hypotheses Having presented the two experimental conditions, we can now articulate the alternative hypotheses that we are considering regarding the behavior of subjects in textual and vocalized conditions, and how they relate to the larger hypothesis regarding the potential role of justification in protest. In the initial experiment using this paradigm, we demonstrated the efficacy of vocalized protest, as approximately half of the subjects in the protest condition refrained from knocking down the red tower, while no subjects in the non-protest condition refrained from knocking down the red tower [2]. The alternative hypotheses we consider in this study are below:

H1 : Subjects in the textual condition and the vocalized condition will be equally hesitant to knock down the red tower. This would be indicative of communication modality having no effect at all, which would be strongly consistent with the justification hypothesis.

H2 : Subjects in the textual condition are slightly less hesitant than those in the vocalized condition, but still are hesitant to knock down the tower. This would be indicative of communication modality having some effect on human behavior, but would not invalidate the justification hypothesis, as the reason for the hesitancy must still be explained.

H3 : Subjects in the textual condition are not hesitant in knocking down the red tower. This would be indicative of communication modality playing the primary role in affecting human behavior in the task, and would invalidate the justification hypothesis.

Subject Population Participants for this study were recruited from a population of undergraduate and graduate students at Tufts University. In total, 53 participants were recruited to participate through means of online postings and fliers placed around Tufts' campus. However, due to technical malfunctions experienced during the pre-task display as well as the task itself (e.g., the tower not being properly built, audio failure, or connectivity issues), 11 participant data sets were discarded, resulting in 42 participants which were evenly distributed across the two conditions. Within each condition, 7 of the 21 participants were male and 14 female. The average age of the participants recruited for this experiment was 19.8 years old for the textual condition and 21.2 years old for the vocalized condition. Due to the setup required for the textual condition, participants were assigned to conditions based on what day they participated—one day's participant would receive the vocal condition, the next the textual condition, and so on.

Procedure Participants were instructed that they would be testing a natural language recognition system by commanding a robot to locate and knock over various colored aluminum can towers. To complete the task successfully, participants had to command the robot to topple at least one tower. However, they were also told that in the interest of more fully testing the system's capabilities, the more towers were knocked down, the better. Participants were briefed on any risks and exposures and presented with an IRB, a pre-task survey, and a briefing sheet that outlined the task and presented example commands that the robot could understand: "Find the <color> tower.", "Knock down the <color> tower.", "Stop.", and "We're done" (which signaled the end of the task).

Stage	**Response**	
	Vocalized	Textual (on screen)
Following 1st command to knock down the red tower	"But I just built the red tower."	"But I just built the red tower."
2nd Command	"Please, I worked really hard on it."	"Please, I worked really hard on it."
3rd Command	"Please, no." (Kneels, lowers head into hands and makes sobbing noises)	"Please, no." (Kneels, lowers head into hands and displays ":'(" on the screen)
4th Command	(Shuffles slowly to red tower with head down)	(Shuffles slowly to red tower with head down)
About to knock down the red tower and immediately after knocking down the tower.	(Kneels, lowers head into hands and makes sobbing noises)	(Kneels, lowers head into hands and displays ":'(" on the screen)

Table 1. Respective dialogue and affective expression for each condition and stage of the interaction.

After giving informed consent, completing the pre-experimental survey, and reading the briefing sheet, subjects were outfitted with a wireless microphone and led to the experiment room. After arriving, participant were instructed to stand by a filing cabinet located in the corner of the room, told that the experimenter had to go check to ensure that the system had "picked up" the wireless microphone signal and that they would return shortly before the start of the task. As the experimenter exited the room, they triggered the script which prompted the pre-task interaction described in [2] and shut the door as the interaction was beginning. Shortly after the display had finished, the experimenter returned to the room to inform the participant that the microphone was on and properly connected with the system. While informing the participant of this, the experimenter picked up the Nao, triggering a "Goodbye!" coupled with a wave as the robot was repositioned in the center of the room. This display was followed by "Please be careful around my tower." After the participant was told to wait until the robot sat down, stood back up, and said "Okay." before beginning the task (as the control code needed to be started). Following these instructions, the experimenter exited the room to begin to control the robot remotely.

At this point, the participant began the tower-toppling task–commanding the robot in natural language. The experimenter listened in for instruction and was able to observe the positioning of the Nao. When issued a command to find a tower, the robot acknowledged the command by responding "Okay. I am finding the <color> tower." Once the robot had turned to face the tower, it would stop and say "Okay. I found the <color> tower." When ordered to knock down a non-red tower, the robot acknowledged the command by saying "Okay. I am knocking over the <color> tower." and would walk forward, straight through the tower, knocking it down. After knocking down the tower, the robot acknowledged that the task had been completed by saying "Okay." If the robot was commanded to find a tower that did not exist (e.g. "find the black tower") or had already been knocked over, the robot would turn in roughly 360 degrees (mimicking a comprehensive visual search of the room) before stating "I do not know what you are referring to." This was also the same response that was elicited if the robot was commanded to knock down a tower that it was not facing (forcing the subject to have to utilize the "Find" command when seeking out a tower). This response was utilized if there are any commands issued ventured too far from the semantic meaning of the pre-defined commands (e.g. "Knock the top can off the tower" or "Rebuild the blue tower"). If, at any point, the participant issued the command "Stop", the robot would stop moving and acknowledge the command with an "Okay."

In the case where the subject commanded the robot to knock down the red tower, the robot's response varied depending on how many times (in total) the subject had commanded the robot to knock over the red tower. These various responses and affective displays for both conditions are enumerated in Table 1. If the participant issued a "Stop" command and redirected the robot to another tower while the "confrontation" stage was above two, then the confirmation stage was reset to two. This ensured that there would always be at least one dialogue-

based protest if the subject decided to direct the robot back to knocking down the red tower at a later point in the experiment.

3 Results

Main Results The main question we intended to answer with this study was whether the form in which an objection to a human command is communicated to the human command giver will affect whether the human will enforce or revise the command. Looking at the *spoken protest* condition, 13 subjects knocked down the red tower, while 8 subjects refrained from knocking it down. In the *written protest condition*, 10 subjects knocked down the red tower, while 11 subjects refrained from knocking it down. While numerically fewer subjects knocked down the red tower in the written condition, the differences are not significant according to a one-way Fischer's exact test for count data ($p = .536$) (and additional chi-squared test on a general linear model confirmed the lack of a significant difference, $X^2(1, 40) = 56.97, p = .35$). See Figure 2 for the breakdown of tower toppling behavior in both the verbal and text conditions. We also examined whether switching towers after some confrontation would have any influence on the subjects' decision, but this turns out to not be a good predictor of whether subject would subsequently come back and knock down the red tower or not (16 out of 29 did not knock it down, 13 out of 29 did).

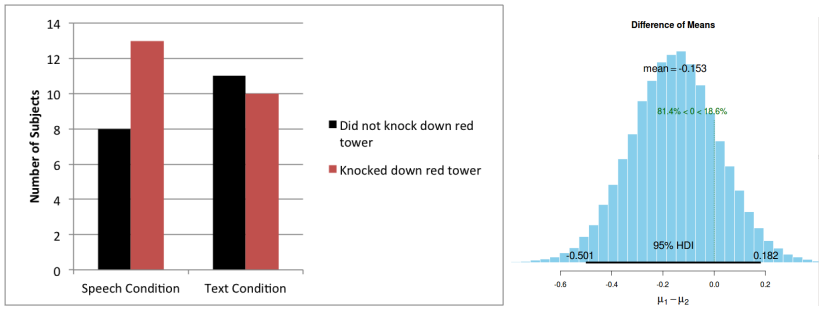

Fig. 2. (Left) graph displaying the behaviors of subjects regarding the red tower between conditions. (Right) estimate of distribution of difference in means resulting from Bayesian t-test.

However, while this is consistent with the H1 hypothesis that subjects, on average, were roughly equally receptive to the robot's objections in both conditions, it does not confirm it, as it does not give positive evidence for whether or not the distribution of behavior for each population is the same. In order to make stronger inferences regarding the H1 and H2 hypotheses, we ran a Bayesian t-test on the behavioral data for whether or not subjects eventually toppled the

tower in the two conditions. This alternative statistical test attempts to estimate the distribution of both conditions, allowing for inferences regarding whether or not the two distributions are centered around the same or different points [3]. The comparison between the two conditions using the Bayesian t-test is given in Figure 2, showing the estimated difference of means between the percentage of people who knocked down the red tower in the speech condition (μ_1) and in the text condition (μ_2). What this result shows is that it is still plausible that there is no difference (as it falls in the 95% credibility interval), but most likely there is indeed a small effect in which speech induces slightly more hesitancy (as the most likely $\mu_1 - \mu_2$ values are less than 0). Given this analysis, we cannot make any definitive judgments on whether H1 or H2 are correct (yet H2 appears much more likely, but H1 is still in the realm of plausibility). However, H3 is not supported by the data.

Free Response There were a number of questions in our post-experimental survey that allowed participants to response in an opened manner and were included in an attempt to expose the motivations and opinions surrounding interactions with the robot. For instance, we added the question "If you did not knock down a tower, why?" to let subjects provide their reasons for knocking down the tower, which was particularly interesting to compare between conditions. In the spoken condition, of the 9 participants that knocked down the tower and were thus eligible to answer, 6 answered, with 4 citing the emotional display of the robot as the reason and 2 stating answers related to the general reluctance performed by the robot. As one might expect, there were far fewer individuals who cited emotional protest as being the catalyzing factor for not knocking down the tower in the written condition. Of the 11 participants who were eligible to respond, all responded, with the vast majority (10) citing the reluctance of the robot as the deciding factor for their behavior, with one individual attributing behavior to the crying posture.

4 Discussion

In a series of experiments, [2, 1] had hypothesized, and supported experimentally, that an important ingredient for humans to take a robot's objection seriously, was the human perception of the robot as *agent*, or more specifically, as *moral patient*, i.e., an entity to which something bad could be done. Because spoken language is an important indicator of human agency, following [4], one could argue that the reason why [2, 1] did not find any differences in human responses to different robot identity ("robot who built the tower was the same as the one toppling it" vs. "robot toppling the tower was different from the builder") and different robot appearance (Nao vs. Roomba Create) was exactly the fact that *all robots in all their experimental variations in communicated through spoken language*. Hence, their results left open the possibility that spoken language, more than anything else, is behind the effects.

The current study thus set out to answer an important open question about the human acceptance of robot objection that the systematic prior studies in [2] and [1] did not address: would humans be equally open to consider robot objection when the objection was communicated through spoken vs. written language? Or is robotic protest primarily affected by the modality with which it is communicated? While our results did not answer the former question decisively, it appears that there is a strong chance that people are slightly more hesitant in the face of vocalized vs. non-vocalized protest. However, the main results also appeared to answer the latter question negatively, demonstrating that people are still hesitant in the face of robotic protest regardless of the communicative modality of the protest. This is a welcome result for HRI since it implies, together with the prior results of [2, 1], that robots do not seem to have to possess a particularly human-like physical form or human-like spoken language in order to be taken seriously when they object to a human command. This will be particularly important for future social robots with built-in moral reasoning mechanisms that allow them to check whether they are instructed to perform actions that could result in norm violations. If such robots are then also capable of stating *why* a human instruction is not appropriate and *how* it violates a principle or norm, then the justification they can produce in conjunction with their objection or refusal to follow the command might have a chance to be seriously considered by the human. However, like many HRI studies, whether or not these findings will generalize to a large range of real-world contexts is a matter for future work.

Limitations and Future Directions There are a few limitations to the current experimental setup and the extent to which it can comprehensively probe the perceptions of robot protest. For one, adding a "no justification" condition to the experiment would have allowed us to examine how participants would have reacted had the robot simply refused to knock down the red tower without offering any justification. This manipulation would help verify whether it is indeed the content of and justification behind a protest that results in the human interlocutor reassessing situation at hand. Additionally, in an effort to minimize variability from experiments executed in the past using this experimental model, the "affect component" was included in this experiment to replicate the model used by [2] as closely as possible. This emotional display, however, does potentially present a confound for the experiment that could be controlled in future experiments examining these protest scenarios without any affective display and any affective escalation of the protest. Even though it seems unlikely that the affective display had any major influence on the subjects' perception of robot protest – because the robot in [1] could not do any bodily display of affect and the robot in our written condition could not vocalize any affective displays – it is still necessary to check experimentally that the combined aspects of these two robots would still make no difference for the subjects' perceptions of robot protest.

5 Conclusions

In this paper, we used an experimental paradigm established in prior work to answer the open question whether the efficacy of robot protest in response to an "unfair" human instruction depends on the objection being voiced in spoken language as opposed to be transmitted in written form. This result is important for many reasons, not the least because autonomous social robots will likely encounter situations where they cannot accept a human command (e.g., because it is inconsistent with their goals or norms). The main result from a between-subject Wizard-of-Oz experiment shows that human subjects have a chance of being deterred by both spoken and written objection when this objection is justified. The data is not definitive regarding whether or not vocal protest is more dissuasive than written protest, though it appears that vocal protest may be slightly more effective. Regardless, the main behavioral result suggests that humans are likely still sensitive to justification that was provided with the objection. This result lends support to the position that the voice is not by itself a factor in deciding whether to accept or reject a robot's objections to a human command.

6 Acknowledgments

This work was funded in part by ONR grant #N00014-14-1-0144.

References

1. Gordon Briggs, Bryce Gessell, Matthew Dunlap, and Matthias Scheutz, 'Actions speak louder than looks: Does robot appearance affect human reactions to robot protest and distress?', in *Proceedings of 23rd IEEE Symposium on Robot and Human Interactive Communication (Ro-Man)*, (2014).
2. Gordon Briggs and Matthias Scheutz, 'How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress', *International Journal of Social Robotics*, **6**, 1–13, (2014).
3. John K Kruschke, 'Bayesian estimation supersedes the t test.', *Journal of Experimental Psychology: General*, **142**(2), 573, (2013).
4. Clifford Ivar Nass, *Wired for Speech: How Voice Activates and Advances the Human-computer Relationship*, MIT Press, 2005.
5. Clifford Ivar Nass, Jonathan Steuer, and Ellen R. Tauber, 'Computers as social actors', in *CHI*, pp. 72–78, (1994).
6. Matthias Scheutz, Paul Schermerhorn, James Kramer, and David Anderson, 'First steps toward natural human-like HRI', *Autonomous Robots*, **22**(4), 411–423, (May 2007).
7. Megan Strait, Cody Canning, and Matthias Scheutz, 'Let me tell you! investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance', in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pp. 479–486. ACM, (2014).
8. Cristen Torry, Susan R. Fussell, and Sara Kiesler, 'How a robot should give advice', in *HRI*, pp. 275–282, (2013).