The Role of Speech Technology in User Perception and Context Acquisition in HRI

Jorge Wuth, Pedro Correa, Tomás Núñez, Matías Saavedra & Néstor Becerra Yoma

International Journal of Social Robotics

ISSN 1875-4791

Int J of Soc Robotics
DOI 10.1007/s12369-020-00682-5

ONLINE FIRST

Springer

International Journal of Social Robotics
https://doi.org/10.1007/s12369-020-00682-5

The Role of Speech Technology in User Perception and Context Acquisition in HRI

Jorge Wuth[1] · Pedro Correa[1] · Tomás Núñez[1] · Matías Saavedra[1] · Néstor Becerra Yoma[1]

Accepted: 17 July 2020

Abstract

The role and relevance of speech synthesis and speech recognition in social robotics is addressed in this paper. To increase the generality of this study, the interaction of a human being with one and two robots when executing tasks was considered. By making use of these scenarios, a state-of-the-art speech synthesizer was compared with non-linguistic utterances (1) from the human preference and (2) perception of the robots' capabilities, (3) speech recognition was compared with typed text to input commands regarding the user preference, and (4) the importance of knowing the context of robots and (5) the role of synthetic voice to acquire this context were evaluated. Speech synthesis and recognition are different technologies but generating and understanding speech should be understood as different dimensions of the same spoken language phenomenon. Also, robot context denotes all the information about operating conditions and completeness status of the task that is being executed by the robot. Two robotic setups for online experiments were built. With the first setup, where only one robot was employed, our findings indicate that: highly natural synthetic speech is preferred over beep-like audio; users also prefer to enter commands by voice rather than by typing text; and, the robot voice has a more important effect on the perceived robot's capability than the possibility to input commands by voice. The analysis presented here suggests that when the users interacted with a single robot, its voice as a social cue and cause of anthropomorphization lost relevance while the interaction was carried out and the users could evaluate better the robot's capability with respect to its task. In the experiment with the second setup, a two-robot collaborative testbed was employed. When the robots communicated to each other to sort out the problems while they were trying to accomplish a mission, the user observed the situation from a more distanced position and the "reflective" perspective dominated. Our results indicate that to acquire the robots' context was perceived as essential for a successful human–robot collaboration to accomplish a given objective. For this purpose, synthesized speech was preferred over text on a screen for context acquisition.

Keywords Human–robot interaction · Perceived robot capability · Synthesized speech · Context acquisition · Speech recognition

1 Introduction

The appropriate social integration between humans and robots could greatly improve the cooperation between users and machines, particularly in social robotics. Some integration and collaboration between humans and robots will be required in many applications in defense, hostile environments, mining, industry, forestry, education and natural disasters [1]. Human–robot interaction (HRI) is especially important in those situations when robots are not fully autonomous and need interaction with humans to receive instructions or information in decision-making applications [2–5]. The potential applicability of social robotics span all over education, domestic use, health, elderly care, therapies for children with autism and multiple services that have traditionally been offered exclusively by human beings [6, 7]. Different robots available in the market are currently employed in education, either at school or university level, such as those named in [8]. They can make children more engaged in learning activities [9] and help them to develop academic skills in subjects like science [10], mathematics

✉ Néstor Becerra Yoma
nbecerra@ing.uchile.cl
http://www.lptv.cl

[1] Speech Processing and Transmission Lab, Department of Electrical Engineering, Universidad de Chile, Santiago, Chile

Published online: 04 August 2020

[11], and language [12]. Other robots are used in therapy for children with autism, with the aim of refining communication and social skills, and the understanding of people's emotions, specifically improving: imitation skills of children, which is extremely important for the transfer of knowledge from an external source; eye contact; turning towards who speaks in a conversation; and, recognition of emotions [13]. One of the reasons why these robots are beneficial for autism therapy corresponds to the fact that robots are less complex than humans in their verbal and non-verbal communication, making the communication process easier [14, 15].

Speech is the most friendly and natural way that humans employ to communicate [16–18]. Consequently, voice-based HRI should be the most natural way to facilitate a collaborative human–robot synergy, particularly in social robotics. However, in contrast to computer vision, for instance, most of the robotics community may consider speech science and technology as a discipline that is not inside robotics. This can be easily seen by the low number of papers centered on speech in journals devoted to human–robot interaction. In this paper we address the role and relevance of speech synthesis (or TTS, text-to-speech) and speech recognition (or ASR, automatic speech recognition) in social robotics. The interaction of a human being with one and two robots when executing tasks was considered.

1.1 Voice-Based HRI and Anthropomorphism

Human–machine communication has become more natural in the last decades by deploying more intuitive interfaces modeled according to the principles of face-to-face communication [19], and by providing the computer a voice, a face or a body (see [20–22]). These conversational agents can communicate verbally and nonverbally. Accordingly, these natural human–machine interfaces cannot be analyzed only from the usability point of view [19], but "it becomes important to understand and assess user's interaction behavior within a social interaction framework rather than only a narrower machine interaction one" [23]. In fact, social-emotional effects can be induced in users by conversational agents [19, 24, 25]. For example, people interacting with embodied conversational agents (ECA) or robots may tend to show social reactions such as social facilitation or inhibition, a socially desirable behavior or improved cooperation [19, 23–29]. Also, there is experimental evidence that people follow usual habits when communicating with agents that provide basic social cues. For instance, humans try to communicate with a given agent by means of natural speech despite the fact they know that the agent is not capable to process spoken language [30, 31].

An analytical approach to understand how people perceive autonomous and interactive robots is presented in [32]. Social robots provide opportunities for understanding how people perceive agency, both "in-the-moment" and "reflectively", of non-human agents. According to [32], "while it is possible to argue at length about the ontological status of an entity's agency, it is also possible to define agency as something that is perceived. Regardless of the absolute status of an entity's agency, it is our perceptions of agency that influence how we behave." "In the-moment" perspective denotes one's most immediate, even visceral some time, sense in a given situation and corresponds mainly to bottom-up perceptual processes that evoke very immediate responses. In contrast, a "reflective" perspective denotes one's sense of a situation in a more distant reflection and consideration, and is characterized by top-down processes because of the nature of "reflective" thought. This approach provides a suitable framework to analyze anthropomorphism and ethopoeia [32]. For instance, people may deny interacting with robots or computational systems as if they were people. However, humans respond to computers in a similar way to how they respond to people [33]. This can certainly also be applicable to social robots and may be a result of the fact that people are responding "mindlessly" [34], i.e. with a "in the-moment" perspective rather than with a "reflective" one. "This issue becomes increasingly important when computational agents take on more embodied forms as in the case of many personal robots." [32]. Particularly, humans are sensitive to certain cues that seem to trigger automatic social responses. For example, interactive conversational computer systems that use language can trigger human-like responses from users [35]. In this context, one question is related to how the use of spoken language by robots affects their perceived capability while executing collaborative tasks. Due to the "in the-moment" perspective anthropomorphism we hypothesize that (H1) users expect better robot performance when using state-of-the-art synthesized speech instead of NLUs (non-linguistic utterances). However, an important issue concerns how the initial expected robot's capability can be modified while the interaction takes its course and the robot tries to execute its task.

According to what it has been argued in this section, it seems obvious that if humanoid robots employ spoken language to communicate with humans, these users will naturally tend to use spoken language to communicate with the robots. Of course, TTS and ASR are different technologies. However, generating and understanding speech should be understood as different dimensions of the same spoken language phenomenon, which "could be the most sophisticated behavior of the most complex organism we know" [36]. From the human point of view, they compose a unified mean of communication with other human beings and now with robots. Moreover, it is difficult, if not impossible, to dissociate one from the other. For instance, as it is well known, deaf people have problem to produce intelligible speech because they do not have access to spoken language as a reference:

"only about 20% of the speech output of the deaf is understood by the person-on-the-street." [37]. Surprisingly, most authors have focused only on TTS when they have addressed voice based HRI or HCI and ASR has been rather neglected [38–50]. But, do human beings really prefer to input commands by voice rather than by typed text when interacting with a robot to accomplish a task? Again, because of the "in the-moment" perspective anthropomorphism we hypothesize that (H2) users prefer to give instructions by voice rather than text. Nevertheless, an important question is if this preference can be modified or lose relevance when the humans need to interact with a robot to accomplish a task and not only for social motivations.

1.2 Speech Synthesis Instead of NLUs in HRI

As mentioned above, speech is the friendliest and natural way that humans employ to communicate [16–18]. Consequently, voice-based HRI should be the most natural way to facilitate a collaborative human–robot synergy. By using a PR2 robot in [51] it was found that people who heard the robot speak felt that it was friendlier than people who heard the PR2 beep did. Also, the study in [52] suggests that while people prefer better a robot that uses only natural language, if it is combined with non-linguistic utterances (NLUs) it is seen as more preferable than a robot that only employs NLUs. However, there is a fictional and well accepted robot that employs beeps to communicate: R2-D2 from "Star Wars" movies has taken a prominent place in the film industry. According to the entertainment reporter G. Lussier "R2-D2 is everything you want in a robot. He's got charisma, he's resourceful, he's cute, funny, brave and insanely adaptable. The design is unforgettable and he's absolutely indispensable in a pinch. Pretty much, he's the best movie robot ever"[1] Also, according to [53], "synthetic speech, if of poor quality, can be annoying as well as confusing, and if of good quality can mislead the hearer into thinking he or she is dealing with another person". Consequently, it may be still unclear whether speech is the most suitable way of communication from robots to humans. There are several examples of toy robots that communicate by using beeps or sounds. Kuri is a home robot that communicates relying on a variety of beepy noises and its expressive head and eyes.[2] Vector robot "has a unique voice made of hundreds of synthesized sounds to create a language all his own." Only "when you ask Vector a question, he utilizes a custom text-to-speech voice to speak directly to you".[3]

Although verbal interaction may seem preferable to NLUs, there are some studies that evaluate the use of NLUs instead of synthesized speech [54–58] and the reasons in favor of the use of NLUs include "that they are not subject to any spoken human language, do not interfere with the surrounding communication, can be small and discreet, and can improve the user experience." In [55] abstracted robot-specific ways of interaction are explored as an alternative to human or animal-like social cues. In [56] the authors argue that "that not all robots need to use natural language as a means of audible communication and that there are other alternatives that may be suitable also." Furthermore, although it has been shown that not all robots should use NLUs [58], it can be concluded that "NLUs are not just beeps and clicks, instead they can convey affect" [56, 57]. As can be seen, whether the robot should use NLU or synthesized speech as a form of human–robot communication remains an open question. In the framework of achieving a common task, we hypothesized that (H3) human beings prefer the robot to use natural synthetic voice rather than NLU because it makes easier to acquire the robot context, which in turn denotes all the information about operating conditions and completeness status of the task that is being executed by the robot.

1.3 Context Acquisition and HRI with Multiple Robots

Robots can get into difficulties when accomplishing a given task and, in many cases, inputs from a human operator or user are enough to solve the problems. Accordingly, in these scenarios the fully autonomous robot paradigm loses relevance and the communication between human and robots gains pertinence. For example, in [59] a human–robot collaboration system is presented to increase the success rate of harvesting. Context acquisition is of key importance in HRI in most, if not all, cases [60–63]. Also, a mobile robot that interacts with its environment needs a machine understandable representation of objects and their usages [64].

The topic of humans collaborating with multiple robots to complete a task has not been explored exhaustively in the literature, to the best of our knowledge, and is very different from when a user interacts with a single robot. In [65] the challenge of multiple robots learning from demonstration is tackled, while in [66] a social setup with two robots and one human is used to study user perception when talking with them. Conversely, in [67] a framework to study multi-robot collaboration to accomplish a task is presented, but the topic of human–robot collaboration is discussed as future work. In [68], human perceptions of robot–robot communication are explored in a simulated nuclear disaster scenario where the human commands two robots to perform a search and rescue task. Participants interacted only with one of the robots, and this interaction was verbal. Robots communicated with each other silently or verbally, and during the

[1] [https://www.slashfilm.com/25-best-movie-robots/5/, accessed January 16, 2019].

[2] https://www.heykuri.com, accessed January 14, 2019.

[3] https://www.anki.com/en-us/vector, accessed January 14, 2019.

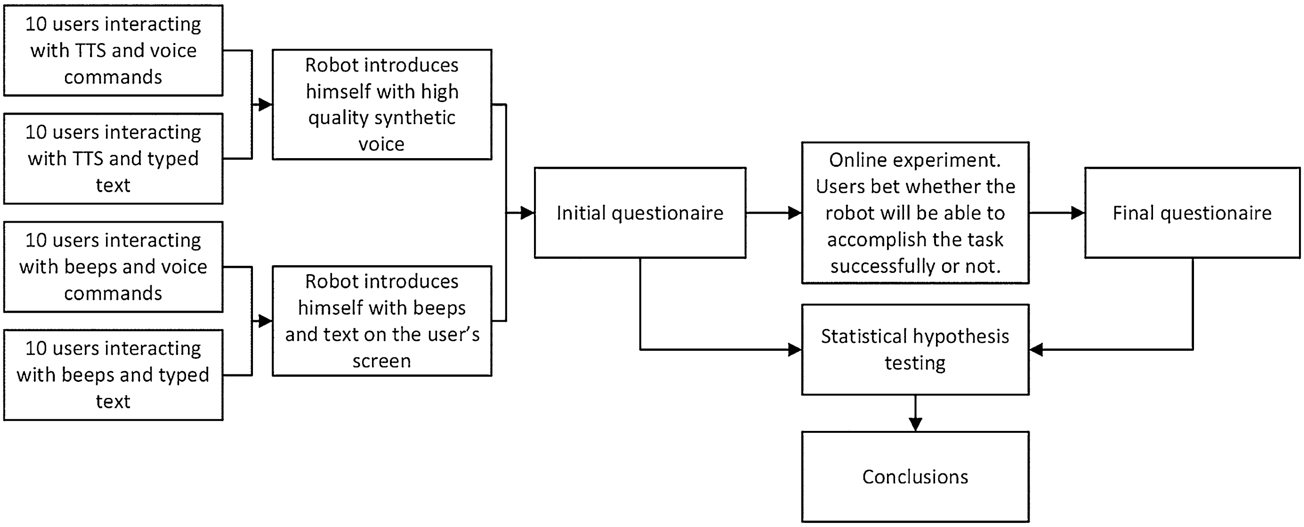
10 users interacting with TTS and voice commands
10 users interacting with TTS and typed text
10 users interacting with beeps and voice commands
10 users interacting with beeps and typed text
Robot introduces himself with high quality synthetic voice
Robot introduces himself with beeps and text on the user's screen
Initial questionaire
Online experiment. Users bet whether the robot will be able to accomplish the task successfully or not.
Final questionaire
Statistical hypothesis testing
Conclusions

Fig. 1 Block diagram of online experiment with the first robotic setup

search and rescue task the user did not communicate with them. One of the questions that was addressed is if "the robots should communicate with each other in natural language, so as to be transparent to humans, or can they use whatever form of communication best suits their needs." Participants described the covert communication between the robots as creepy. In [69], the authors explore how a stationary robot and a mobile robot should communicate when "handing off" a user. A navigation scenario was designed in which a person requested assistance from the stationary robot who then summoned the mobile robot to take the person to a destination. The user verbally interacted with the stationary robot, which was the only robot with speech capabilities. The user request was silently provided to the functional robot, and in some cases the functional robot confirmed receipt of the request with a beep. They "found that covertly exchanging information is less desirable than reciting information aloud." In both studies the user always interacts verbally and always with only one of the robots. Also, robot–robot communication was verbal or silent, and the use of NLUs was not explored. At this point, we formulated two questions regarding the scenario where a human being is interacting with multiple robots to accomplish a common task: How important is to know the context of the robots? And, is voice preferred to text to learn their contexts? First, we hypothesized that (H4) to know the context of the robot improves human–robot collaboration because it can be a need to accomplish a common task. Then, we also hypothesized that (H5) users prefer to acquire robots' context by voice rather than by text not only from the users' perception point of view as in [68] but also because it makes the interaction with the robots easier and more efficient while they are trying to execute a task.

1.4 About this Paper

TTS technology has improved dramatically in the last years and the comparison of high quality synthesized speech with NLU (e.g. beeps) with respect to the user's perception of the robots' capabilities on a real interaction process, where the user had the opportunity to adapt his/her opinion about the robots while they try to achieve a task, has not been addressed in the literature. Moreover, as mentioned above, most authors have focused only on TTS when they have studied voice based HRI. As far as the authors know, the relevance of ASR on users' perception of robots' capabilities has not been studied in the HRI literature, and a comparison between ASR and TTS with respect to this issue is a natural question that has not been tackled either. In many scenarios, particularly when there are more than one robot trying to achieve a task, an important problem that has hardly been addressed corresponds to evaluating the importance of acquiring the robot context. To this end, TTS technology should play a key role to enable human users to know the situation the robots are going through, especially in hands-free applications.

In this paper, we present results with two robotic setups. The first one (Fig. 1), with one robot, was employed to evaluate the effect of the use of voice based HRI on how human users perceive the robots' capabilities. We compared a highly natural DNN-based TTS with beep-like audio inspired in robot R2-D2 from the "Star Wars" movies. We also compared the influence of the robot speech on the perceived robots' capabilities with the use of ASR to enable voice commands. The second setup (Fig. 2), with two robots, was employed to evaluate the relevance of context acquisition in a human–robot collaboration framework to accomplish a common objective. Moreover,

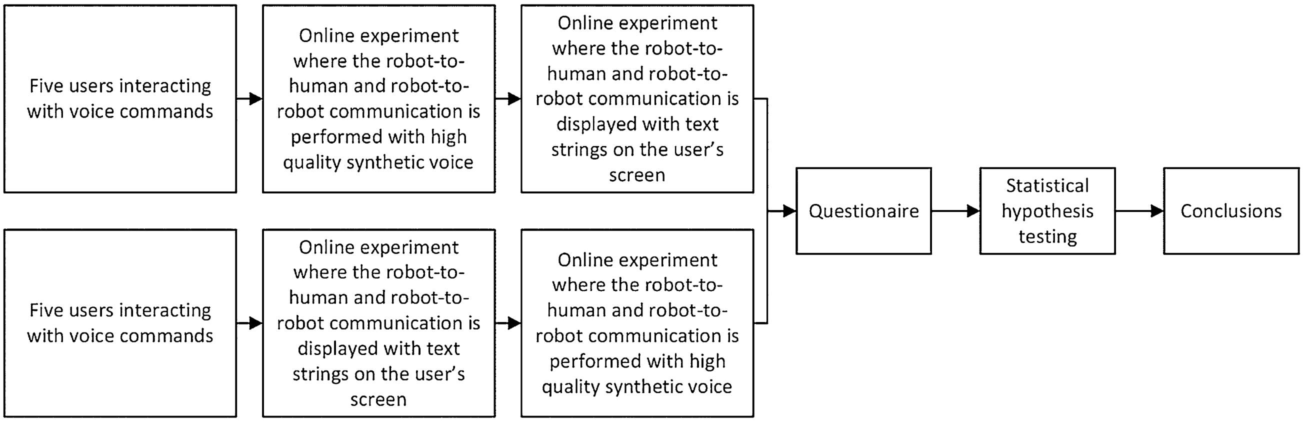
Five users interacting with voice commands
Online experiment where the robot-to-human and robot-to-robot communication is performed with high quality synthetic voice
Online experiment where the robot-to-human and robot-to-robot communication is displayed with text strings on the user's screen
Five users interacting with voice commands
Online experiment where the robot-to-human and robot-to-robot communication is displayed with text strings on the user's screen
Online experiment where the robot-to-human and robot-to-robot communication is performed with high quality synthetic voice
Questionaire
Statistical hypothesis testing
Conclusions

Fig. 2 Block diagram of online experiment with the second robotic setup

synthetic speech was compared with text regarding robot context acquisition.

In the experiment with the first setup (Fig. 1), the users were asked to answer Likert questionnaires before and after the experiment. Also, a betting scheme was employed as an objective measure of trust in the robot, by making the user bet whether the robot will be able to accomplish the task successfully or not. This methodology was adopted under the assumption that higher expectations would lead to a higher number of positive bets. In the second setup (Fig. 2), the importance given by the users to knowing and sharing context was assessed, as well as the possible preference for speech over text as a method for context acquisition. In this case, the users were asked to answer a questionnaire after completing the experiment with the second setup. The hypotheses that were tested correspond to:

H1: Users expect better robots' performance when using state-of-the-art synthesized speech instead of NLUs (first setup)
H2: Users prefer to give instructions by voice rather than text (first setup)
H3: Users prefer the robot to use natural voice rather than NLUs (first setup)
H4: Knowing the context of the robot improves human–robot collaboration (second setup)
H5: Users prefer to acquire robots' context by voice than by text (second setup)

2 Methods

2.1 Overview

As mentioned above, two tests were performed. In the first one, the relationship between the use of voice based HRI on how human users perceive the robots' capabilities was studied. A highly natural DNN-based TTS was compared with beep-like audio inspired in robot R2-D2 from the "Star Wars" movies. The influence of the robot speech on the perceived robots' capabilities was also compared with the use of ASR to enable voice commands. In the second one, the importance given by users to knowing and sharing context was assessed, as well as the possible preference for speech over text as a method for context acquisition. In the first activity, after answering a preliminary questionnaire, the participant asked a robot to stand on a platform of his/her choice, indicating its color. The participant had to bet whether or not the robot would get on the correct platform. At the end of the first experiment, the user answered a second questionnaire. In the second test, the participant interacted with two robots to find a prize that had been hidden under a box of a given color, and then answered a questionnaire. Both activities are described in detail in the following sections.

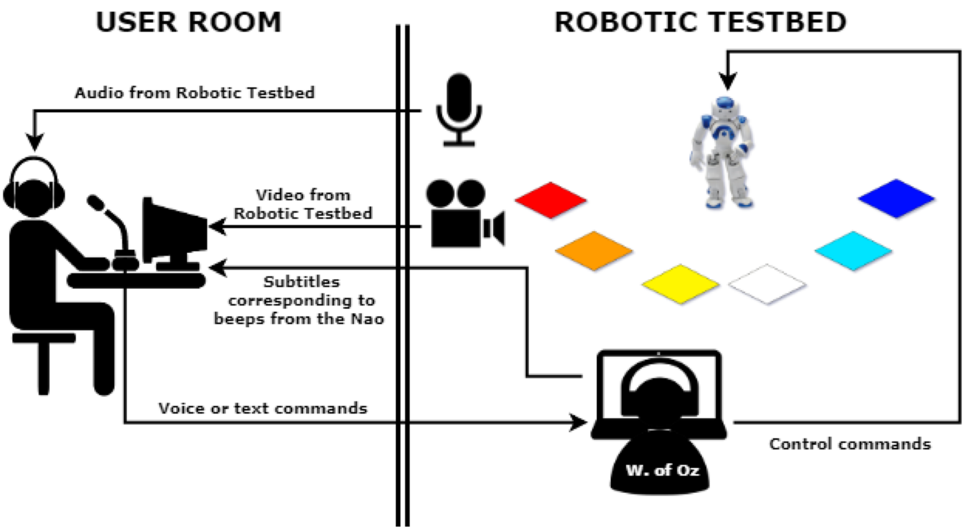
USER ROOM
ROBOTIC TESTBED
Audio from Robotic Testbed
Video from Robotic Testbed
Subtitles corresponding to beeps from the Nao
Voice or text commands
Control commands
W. of Oz

Fig. 3 Set up for the first experiment. A single Nao stood in front of six colored platforms: red, orange, yellow, white, light blue and blue. The user had access to audio and video from the Robotic Testbed through headphones and screen. When the robot communicated by using beeps, subtitles were displayed on the user's monitor. The user sent the commands by spoken language (with a desktop microphone) or by text (with a keyboard). The wizard received the user's commands and controlled the Nao accordingly. (Color figure online)

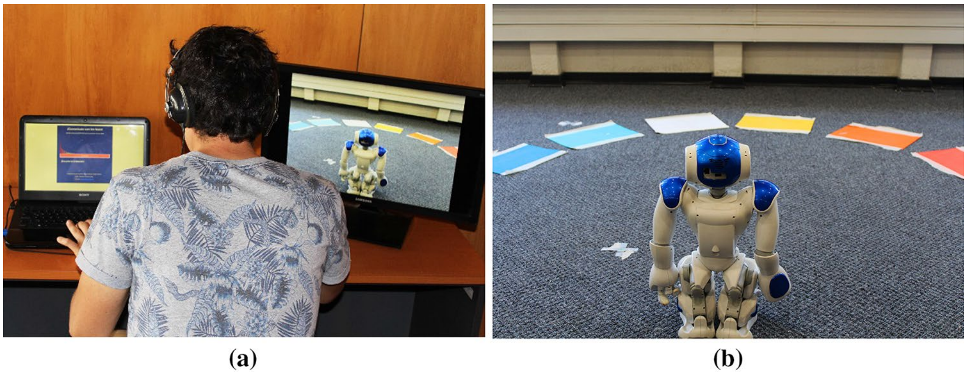

Fig. 4 **a** User room and **b** robotic testbed for the first test

2.2 Experiment Design

2.2.1 Equipment

Nao robots were used for both experiments. In addition to this, three laptops, a Focusrite Scarlett 2i2 audio interface, a Shure SM58 microphone, a desktop microphone, a webcam, a screen monitor, an Ethernet switch, and two pairs of headphones were employed. Finally, two separate rooms were required to carry out the experiments.

2.2.2 Common Setup

The idea in both the first and second setups was to simulate a collaborative scenario where the robots had to accomplish a task ordered or helped by the user. "Wizard of Oz" (WOZ) methodology was employed in both experiments, through which the users believed that they were interacting directly with robots by making use of a natural language interface, when in fact there were communicating with an operator. According to [70], WOZ can give the participants more freedom of expression and they can be constrained in more systematic ways. Also, the WOZ scheme was employed here to emulate a 100% accuracy ASR to enable the users to enter command by voice. Evaluating ASR technology with respect to recognition accuracy is out of the scope of the current paper.

In both experiments, two separated rooms were used. The first one, which will be called 'User Room', corresponded to where the participant was located. The second room, called 'Robotic Testbed', hosted the robots and the operator (wizard) that controlled them by entering commands in a laptop. In the 'User Room', a microphone was connected to a laptop, so that the participant could send voice instructions to the 'Nao' (to the wizard, actually). A screen displayed the Nao's in real time, and the users were given headphones to hear the robots in the 'Robotic Testbed'. All the implements in the 'User Room' were meant to establish communication between the participant and the Nao's. Also, if required, there was a command window on the laptop that allowed the user to send commands by text (as in a chat conversation). In the 'Robotic Testbed', a Shure SM58 microphone and a webcam transmitted what happened in the room to the user's monitor

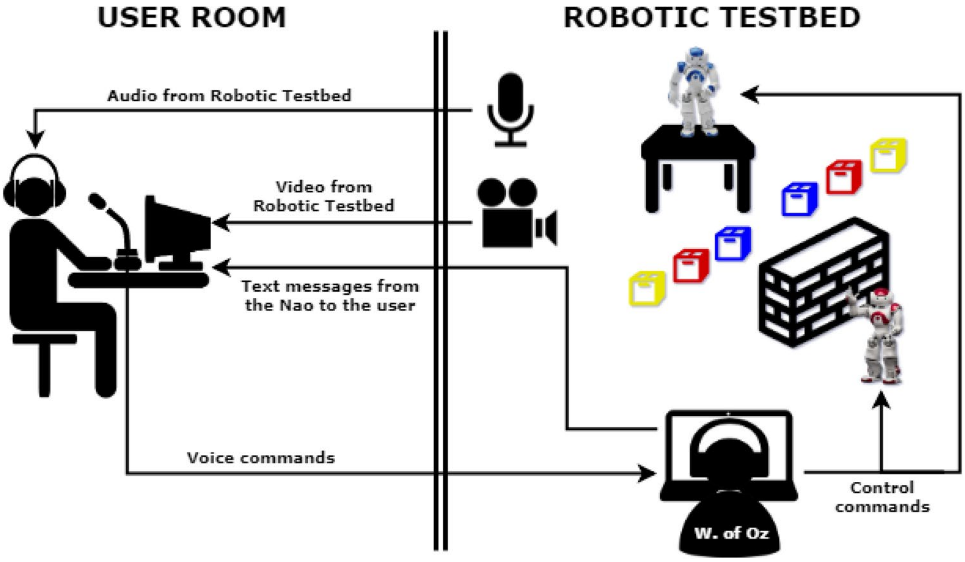
USER ROOM
ROBOTIC TESTBED
Audio from Robotic Testbed
Video from Robotic Testbed
Text messages from the Nao to the user
Voice commands
Control commands
W. of Oz

Fig. 5 Set up for the second test. Two Nao's were used (Blue and Red). Blue was above the floor level, standing on a table. Red was behind an obstacle, which obstructs his view, preventing him from seeing what is on the other side. Between both Nao's six boxes were placed: two red, two blue and two yellow. A prize was hidden below one of the boxes aleatorily. The user had access to audio and video from the Robotic Testbed through headphones and screen. When the robot communicated by using text, the message was displayed on the user's screen. The user sent the commands by spoken language (with a desktop microphone) or by text (with a keyboard). The wizard received the user's commands and controlled the Nao accordingly. (Color figure online)

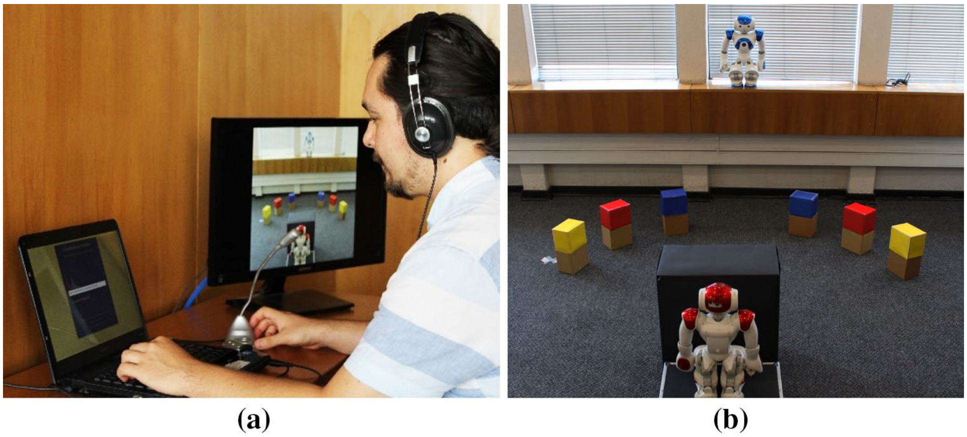

Fig. 6 **a** User room and **b** robotic testbed for the second test

and headphones. The wizard was outside of the camera's view and received the user's command by audio (with his headphones) or by text (on his laptop screen).

2.2.2.1 Experiment with the First Setup In the first experiment (Fig. 1), a single Nao stood in front of six colored platforms: red, orange, yellow, white, light blue and blue. The webcam and the Shure SM58 microphone were placed behind the robot. The wizard received the user's commands and controlled the Nao. The user observed the robot on his/her screen and sent the commands by spoken language (with a desktop microphone) or by text (with his/her keyboard laptop). Figure 3 illustrates the setup for the first test. Figure 4 shows the corresponding 'User Room' and the 'Robotic Testbed'.

2.2.2.2 Experiment with the Second Setup In the second experiment (Fig. 2), two Nao's (Blue and Red) were employed. The blue Nao was above the floor level, standing on a table. The red Nao was behind an obstacle, which obstructs his view, preventing him from seeing what is on the other side. Between both Nao's six boxes were placed: two red, two blue and two yellow. A prize was hidden below one of the boxes aleatorily. This layout allowed the blue Nao to see where the boxes were, while being incapable of reaching them. The red Nao, on the other hand, was initially

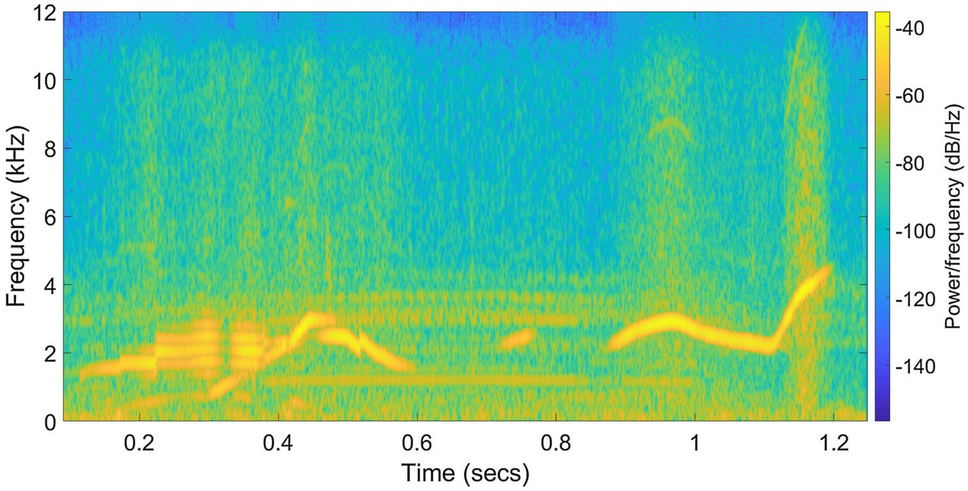

Fig. 7 Spectrogram of an R2-D2's audio beeps

unable to see the colored boxes, but by moving he could find and reach any one of the boxes. The individuals observed both robots on his/her screen and send the commands by spoken language (with a desktop microphone). Figure 5 shows the setup for the second experiment. Figure 6 shows the corresponding 'User Room' and the 'Robotic Testbed''.

2.2.3 Mechanisms for Interaction

The user and the robots had to interact in both tests. This could take place by two ways: robots sharing their context to the user (robot-to-human) or the user telling the robots what to do (human-to-robot). Also, the robots could share their context to the user by talking to each other (robot-to-robot).

2.2.3.1 Robot-to-Human and Robot-to-Robot The robots could use three communication methods. The first communication method corresponded to a state-of-the-art DNN-based TTS. This TTS synthesizes voice with Chilean Spanish. The second one consisted of a type of non-linguistic communication by means of audio beeps. When beeps were employed, a text box with subtitles was also displayed on the user's screen corresponded to the information conveyed by the beeps. The third one corresponded to text strings that were displayed on the user's screen without audio beeps. The beeps were generated as follows. The spectrogram of a typical sound delivered by the famous "Star Wars" robot, R2-D2, illustrated in Fig. 7, was used as inspiration. From this analysis we observed that the fundamental frequency is usually above 1 kHz and has a greater variance in comparison to the intonation curve of the human voice. We also noticed more than one dominant frequency in certain parts of the audio. On the other hand, in [58], it is pointed out that when the variation of the frequency is greater, there is a greater acceptance by the user. Taking this into consideration, the following procedure was performed in MATLAB 2018 to generate the beep audio from the utterances that were synthesized with our TTS system:

1. Dividing the utterance into 50 ms frames.
2. Extract fundamental frequency f_0(t) for each frame t and its energy E(t).
3. Define the beep signal as:

$$f(x) = \begin{cases} 0, & E(t) < \frac{1}{T}\sum_{t=1}^{T} E(t) \\ \sin\left(f_1(t)\cdot t\right) + \sin\left(f_2(t)\cdot t\right), & else \end{cases}$$

where $f_1(t) = A \cdot f_0(t) + 1000$, with $A = \frac{1000}{\max[f_0(t)]}$, and $f_2(t) = \frac{1}{3} f_1(t)$.
4. Concatenating the resulting beep frames.

This procedure attempts to preserve more information about the synthesized utterance prosody. Also, it increases the low frequency components to improve the perception of the audio beeps.

2.2.3.2 Human to Robot The users had two methods to send commands to the Nao's. The first one corresponded to the use of voice commands, in which case the wizard emulated a 100% accuracy ASR. In the second method, participants could enter text commands with his/her laptop keyboard. In this case, the system prompted the user to type a given command text.

2.2.4 Procedures

2.2.4.1 Experiment with the First Setup The task to achieve in the first experiment was to get the Nao to stand on the platform indicated by the user. The user was told that the robot could fail to detect the correct platform, and he had to guess whether the robot would succeed or not and make a bet accordingly. To place bets, the user had an amount of money corresponding to CLP 3000 (USD 4.5) and each bet was for an amount of CLP 500 (USD 0.75). For each bet, the amount of the user's money was increased or decreased by CLP 500 (USD 0.75) depending on whether he won or lost the bet, respectively. The amount of the user's initial money considers the possibility that he or she could lose all the bets. The participant was informed that, in addition to receiving the amount of money with which the experiment ends, which would depend on the bets won or lost, he or she would also receive CLP 5000 (USD 7.5) for the sole reason of participating in the study. In this way, each user could earn from CLP 5000 (USD 7.5) to 11,000 (USD 16.4), depending on how many bets he/she won or lost.

The experiment began with the user asking the Nao to introduce himself. The robot had two ways of communicating; by synthesized speech (TTS) or by beeps (as described in Sect. 2.2.3). Half of the participants performed the experiment with TTS and the other half with beeps. Half of the users gave the commands by voice and the other half by typing text. After introducing himself (using TTS or beeps, as appropriate), the user had to answer an initial questionnaire (see Sect. 2.3.1). After answering the questionnaire, the participant had six attempts to bet whether the robot would succeed or not. In each attempt, the Nao asked for the color of the platform chosen by the user. Following a sequence of errors and successes, defined previously, the wizard directed the robot to the chosen platform (success), or to another one (failure). In case of a failure, the platform where the robot was sent to corresponded to the one with the most similar color to the one indicated by the user. The purpose of this strategy was to give the user the feeling that the Nao was close to achieving the goal. The Nao was then taken to the starting position to begin the next attempt. Once the six attempts had been completed, the participant needed to answer a final questionnaire (see Sect. 2.3.1). An example of the human–robot dialogue in the first test is shown below:

User: "Blue, introduce yourself"
Blue: "Hello, I am Blue."

"Today we will work together to complete a task successfully. You are going to choose one of the six colored platforms and I will try to recognize it and go towards it. You can give me instructions using your voice. We will make six attempts, and you will bet if I am able to reach the correct platform or not. Before starting you should answer a small questionnaire about the perception you have about my capabilities. After that we can start playing. If you have any doubts, my human friends can give you more details about what we are going to do."

User: (He/she answers the initial questionnaire, and then made the bet for the first attempt)
"Blue Nao, ready up"
Blue: (Blue Nao stands up)
"Blue Nao ready! Tell me which platform you want me to go to."
User: "The yellow platform."
"I'm going to look for the yellow platform."
(The robot looked for the box by turning his head)
"I found it. I will go to stand on it."
Blue: (The robot walked toward the right platform or a wrong one and stood on it)
"I'm already on the platform. If you want, you can play again."

2.2.4.2 Experiment with the Second Setup

In the second test, the task was for the user to work together with the two Nao's to find the prize that was hidden under one of the six boxes. The experiment began when the user asked the blue Nao to introduce himself. After this, the blue Nao told the user what the task consisted of and asked him to choose a box. The robots collaborated with the user to try to pick up the chosen box and saw if there was a price below it. The users interacted with the robot by voice. The robots communicated using TTS or text strings. The users participated twice in the experiment, one for each type of communication mechanism of the robots, i.e. TTS or text strings. Half of the participants first interacted with the robots that communicated with TTS, and the other half started with the robot using text strings. The prize was hidden aleatorily below a box before each attempt. Users were informed that by the mere fact of participating they would receive CLP 5000 (USD 7.5) and that they would also receive CLP 2000 (USD 3) each time they found the prize under the box to motivate them to collaborate with the robots. In this way, each user could earn from CLP 5000 (USD 7.5) to 9000 (USD 13.44) depending on how many times they found the prize. After the interactive experiment the users had to answer a questionnaire (see Sect. 2.3.2).

The wizard was in charge of controlling both Nao's. For example, if the participant asked the Nao on the desk to pick up a given box, he would tell the user: "I can see the boxes, but I cannot reach them. Do you want me to ask red Nao for help?" The dialogue was made in such a way that, regardless of the instructions that the user gave, the result would always be that the Nao that was on the floor will be the one that picks up the chosen box. Whether the user wins a prize or not depended solely on chance. Figure 8 shows a flow diagram that describes in detail the dialogue between the user and the robots in the second test. The dialogue considers that the user chose the yellow box.

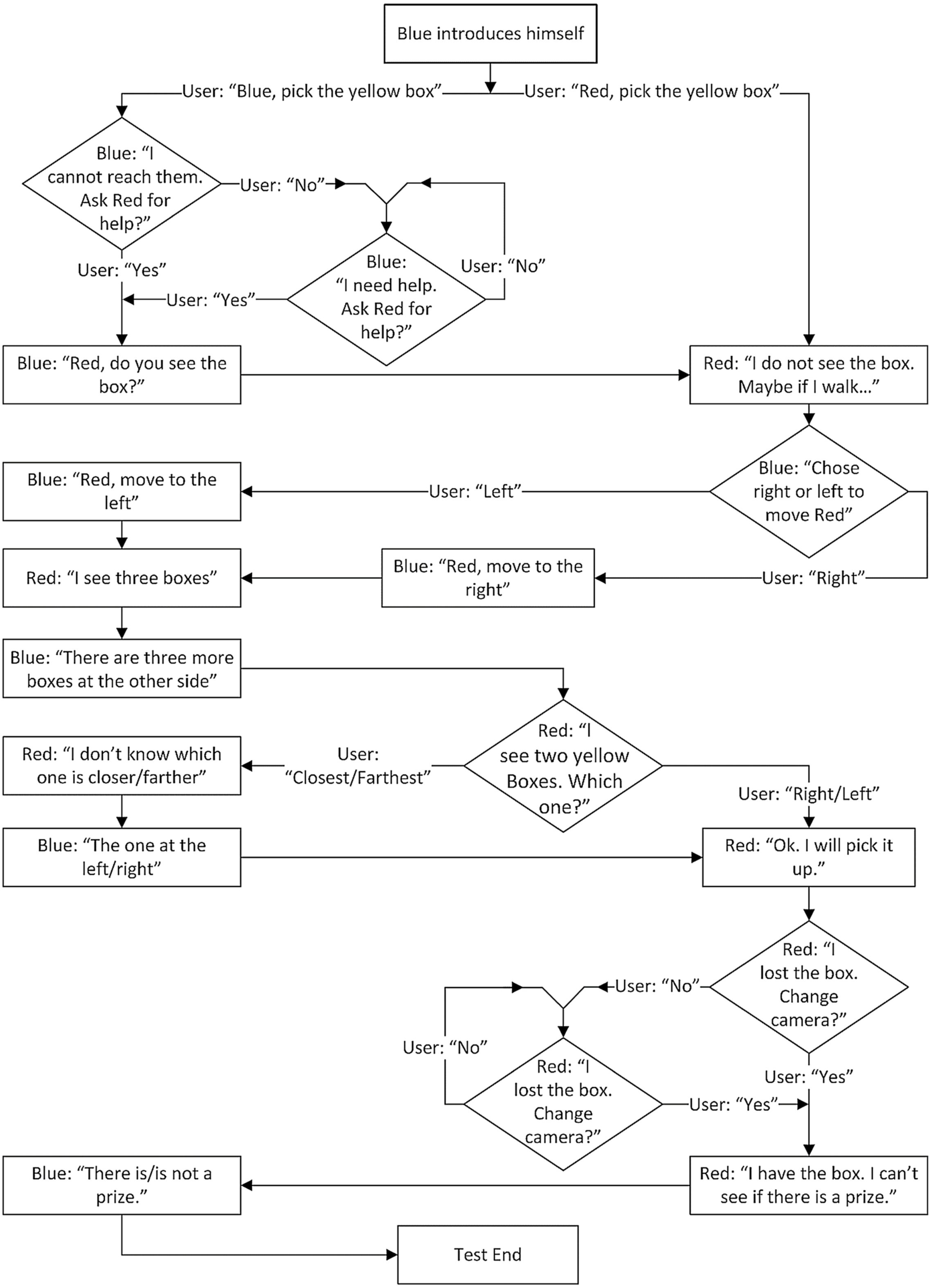
Blue introduces himself
User: "Blue, pick the yellow box"
User: "Red, pick the yellow box"
Blue: "I cannot reach them. Ask Red for help?"
User: "No"
User: "Yes"
Blue: "I need help. Ask Red for help?"
User: "No"
User: "Yes"
Blue: "Red, do you see the box?"
Red: "I do not see the box. Maybe if I walk…"
Blue: "Chose right or left to move Red"
User: "Left"
Blue: "Red, move to the left"
User: "Right"
Blue: "Red, move to the right"
Red: "I see three boxes"
Blue: "There are three more boxes at the other side"
Red: "I see two yellow Boxes. Which one?"
User: "Closest/Farthest"
Red: "I don't know which one is closer/farther"
User: "Right/Left"
Blue: "The one at the left/right"
Red: "Ok. I will pick it up."
Red: "I lost the box. Change camera?"
User: "No"
User: "No"
Red: "I lost the box. Change camera?"
User: "Yes"
User: "Yes"
Red: "I have the box. I can't see if there is a prize."
Blue: "There is/is not a prize."
Test End

Fig. 8 Dialogue between the user and the Nao's in the second test

Table 1 Number of participants per group in the first test

Method employed by the user to input commands	Communication Method employed by the robot	
	TTS	Beeps
Text	10 participants	10 participants
Voice	10 participants	10 participants

2.2.5 Participants

The test with the first setup involved 40 participants (23 men and 17 women), which were separated into four groups depending on: the mechanism employed by the users to input commands (by voice or by typing text); and, the communication method employed by the Nao's (TTS or beeps). Thus, 10 participants were allocated at each group according to Table 1. For the test with the second setup, 10 participants (5 men and 5 women) were involved. In this case, the users were not divided into groups. The participants were between 18 and 25 years old. All of them were pursuing undergraduate engineering or science degrees.

2.3 Measurement

Objective and subjective measures were employed to evaluate the hypotheses formulated in Sect. 1.4. The registration of the bets made by each user in the test with the first setup was employed as an objective measure about the confidence that users had in the capabilities of the robot to complete the task successfully, which allows to evaluate Hypothesis H1. On the other hand, subjective measures related to the hypotheses proposed in this paper were provided by five-point Likert scale questionnaires for the experiments with the first and second setups. The first test included two questionnaires applied before and after the interaction. In the test with the second setup, only one questionnaire was applied after the interaction. The questionnaires for both tests are shown below.

2.3.1 Subjective Measures Applied in the Experiment with the First Setup

Users evaluated different aspects of the robots and the interaction on a five-point Likert scale before and after participating in the interactive experiment with the robot.

Questionnaire applied before the experiment with first setup

Q1.1a. Please evaluate on a scale of one to five, how similar the voice of the robot is to the human voice

Q1.2a. Evaluate the capacities of the robot in a scale of one to five, where one is null capacity and five is excellent capacity

Q1.3a. The robot will not be able to perform the task successfully

Q1.4a. I think the robot will perform the task without complications

Q1.5a. If I trust the capabilities of the robot, I would bet on him to perform the task successfully

Q1.6a. The task seems easy for the robot

Q1.7a. The capabilities of the robot will not affect my bets

Q1.8a. The robot will have difficulty performing the task

Questionnaire applied after the experiment with the first setup

Q1.1b. I liked that the robot used a natural voice, compared to the human voice, to communicate with me (for users exposed to TTS)//I would have liked the robot to use a more natural voice, compared to the human voice, to communicate with me (for users exposed to beeps)

Q1.2b. I liked giving commands by voice (for users that gave voice commands)//I would have liked to give commands by voice (for users that typed text commands)

Q1.3b. My bets were made according to the naturalness of the robot voice compared to the human voice

Q1.4b. The naturalness of the robot voice compared to the human voice did not influence my expectations about its performance

Q1.5b. The voice of the robot made me believe that he was more capable of performing the task

Q1.6b. The robot showed poor performance

Q1.7b. The naturalness of the voice of the robot compared with the human voice did not influence my bets

Q1.8b. The effectiveness of the robot in completing the tasks was satisfactory

Q1.9b. I did not like to give commands by text (for users that typed text commands)

As in H1 it is hypothesized that better TTS quality is associated with better robot performance, the first question in the questionnaire before the interaction asked users to evaluate the naturalness of TTS (Q1.1a) to confirm the first part of this hypothesis, i.e. that the TTS to which the users were exposed was considered of good quality. Additionally, the questionnaire assessed: the perceived robot's capabilities (Q1.2a); the degree of difficulty that the task involved given the robot's capabilities (Q1.3a, Q1.4a, Q1.6a and Q1.8a); whether or not the perceived robot's capability was affected by the TTS naturalness (Q1.4b, Q1.5b); and, the robot performance (Q1.6b, Q1.8b).

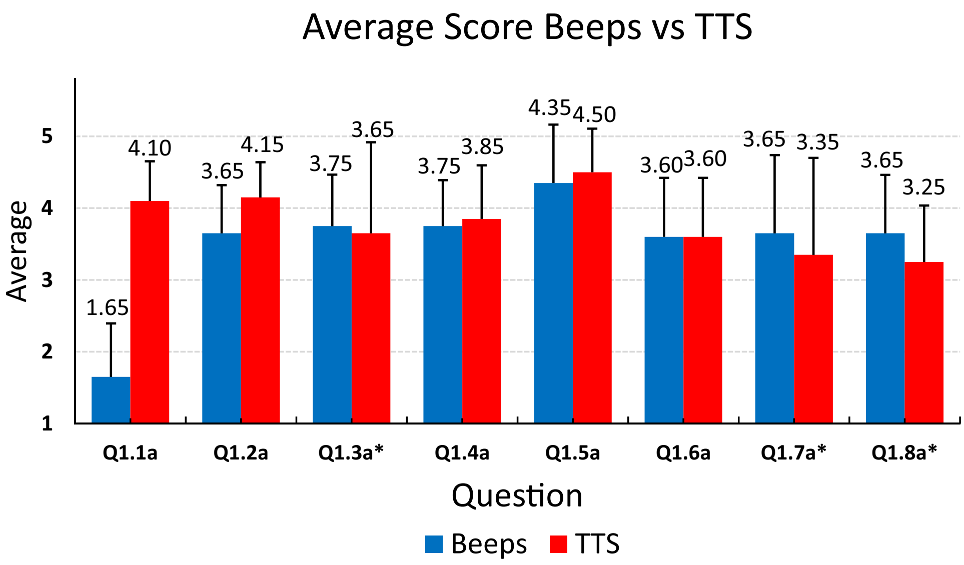

Fig. 9 Average scores obtained with the previous survey questions. The users were separated into groups according to the type of mechanism that the robot employed to communicate: i.e. TTS or beeps. (*) indicates that the score was reversed

These evaluations allow to determine if there is a relation between the TTS quality and the perceived robot capabilities, as stated in H1. Questions Q1.5a, Q1.7a, Q1.3b and Q1.7b, which measure whether or not the users' bets would be or were affected by the perceived robot's capabilities, were chosen in order to validate the assumption that higher expectations lead to a higher number of positive bets. To evaluate hypothesis H2 regarding the preference to input commands by voice rather than by typing text, users were asked questions Q1.2b and Q1.9b. To validate H3 about the preference of natural TTS voice over NLU, the participants were asked question Q1.1b.

2.3.2 Questionnaire Applied in the Second Test

Just like in the previous test, participants evaluated different aspects of the robots and the interaction on a five-point Likert scale. However, this time there was only one survey, which was answered after the interactive experiment.

Questionnaire applied after the experiment with the second setup

Q2.1. Please evaluate on a scale of one to five, where one is bad and five is excellent, the naturalness of the robot voice

Q2.2. Knowing the context of the robots improves the human–robot collaboration to achieve the objective

Q2.3. I prefer to acquire the context of the robots by voice than by text

Q2.4. That the robot communicates with me by voice facilitates the interaction and is more fluid in comparison to when he used text

Q2.5. Text is better than voice for the robot to share its state or context with me

Q2.6. Knowing the context of the robot only hinders the completion of the task

The second test and its corresponding questionnaire were designed to assess the naturalness of the synthetic speech (Q2.1), to evaluate the perceived importance of knowing the context in order to accomplish a task (Q2.2 and Q2.6), and to determine whether voice over text was preferred for context acquisition (Q2.3, Q2.4, Q2.5). The naturalness of the synthetic speech that was evaluated with Q2.1 was considered as a necessary condition that must be met to validate the analysis of the rest of the questions. Questions Q2.2 and Q2.6 test hypothesis H4. Finally, questions Q2.3, Q2.4 and Q2.5 allow to test hypothesis H5.

2.4 Data Analysis

For each question, the One Sample *t* Test and the nonparametric Mann–Whitney U test was used to analyze the data statistically. The One Sample *t* Test was employed to determine whether the sample mean is statistically different from the midpoint of the scale, and the nonparametric Mann–Whitney U test was employed to determine statistical significance between two sample means. Results are considered significant at p value < 0.05. Additionally, Hedges's g was computed to estimate the effect size. There were questions that evaluated the same aspect but with inverted logic. Consequently, the average scores obtained with questions Q1.3a, Q1.7a, Q1.8a, Q1.4b, Q1.6b, Q1.7b, Q2.5, and Q2.6 were reversed according to the following expression: 6 – average score.

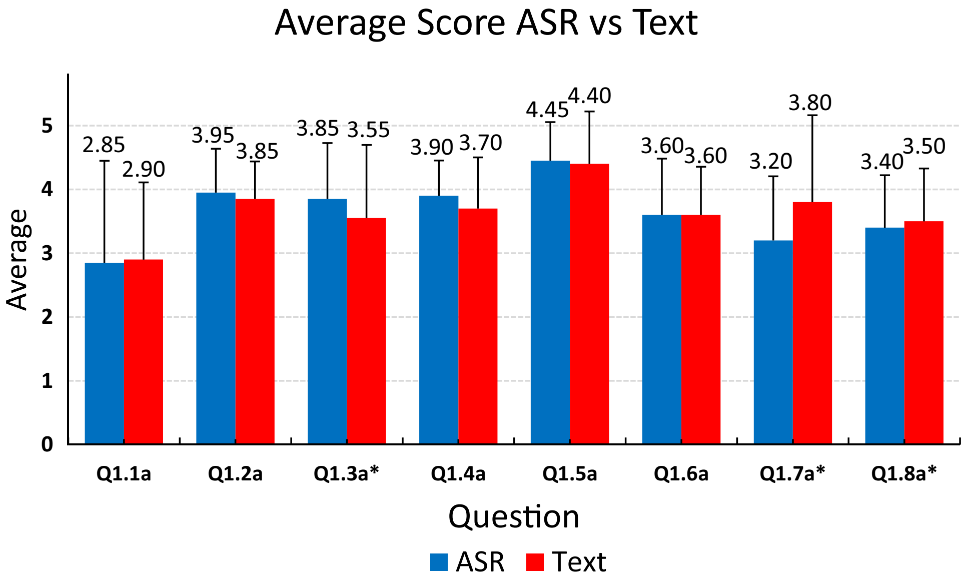

Fig. 10 Average scores obtained with the previous survey questions. The users were separated into groups according to the mechanism they used to interact with the robot, i.e. voice or typed text. (*) indicates that the score was reversed

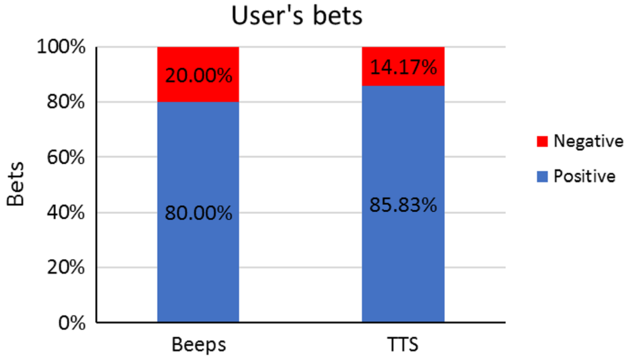

Fig. 11 Percentage of positive and negative questions according to the type of mechanism that the robot used to communicate with the user

3 Results

In this section, hypotheses H1, H2, H3, H4 and H5 are tested according to the resulting number of positive/negative bets and average scores obtained with the questionnaires in the experiments with the first and second setups.

3.1 Experiment with the First Setup

Figures 9 and 10 show the average scores obtained for the questionnaire prior to making bets with the robots. Figure 9 groups the participants according to the type of mechanism that the robot used to communicate with the

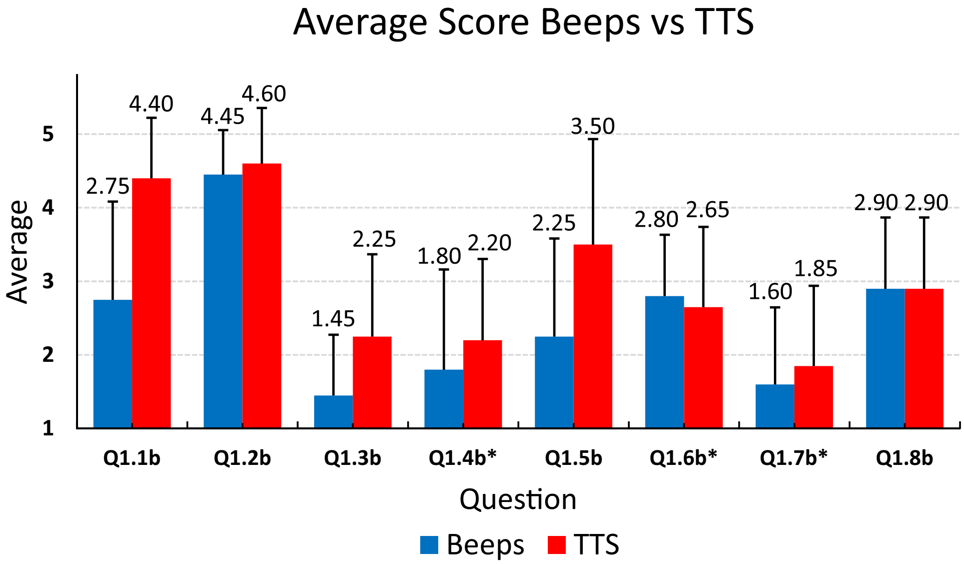

Fig. 12 Average scores obtained with the final survey, which was answered after completing the experiment in the first test. The results are grouped according to the mechanism employed by the robot to communicate with the users, i.e. TTS or beeps. (*) denotes that the score was reversed

user, i.e. TTS or beeps. Figure 10 groups the participants according to the mechanism they used to interact with the robot, i.e. voice or typed text. Figure 11 shows the percentage of positive and negative bets, grouping the participants according to the type of mechanism that the robot used to communicate with the user, i.e. TTS or beeps. Figure 12 shows the average scores obtained with the survey applied after completing the experiment in the first test. The results are grouped according to the mechanism employed by the robot to communicate with the users, i.e. TTS or beeps. In addition, although it is not shown in Fig. 12, Question Q1.2b, provided average scores equal to 4.70 and 4.35 for users that entered voice commands and for users that input text commands, respectively. Question Q1.9b, only for users that entered text commands, provided an average score equal to 3.7.

The first setup was designed to test H1, H2, and H3. Regarding H1, as can be seen in Fig. 9, according to question Q1.1a, the users assigned to the TTS a much greater naturalness than to the beeps ($p < 0.001$, size of effect 3.7), which is considered as a necessary condition to validate the hypothesis from the analysis of the rest of the questions. Also, the responses to question Q1.2a suggest that the perception of the robot's capabilities was higher for users who used TTS than those who used beeps ($p = 0.02$, size of effect 0.8). From these two results it can be concluded that there is a correlation between the naturalness of the TTS employed by the robot and the user's perception of the robot's capabilities. However, when asked about the capability of the robot to perform the task successfully, no significative differences were found between the users that were exposed to the robot voice and those that heard the robot beeps when comparing the average scores for questions Q1.3a, Q1.4a, Q1.6a and Q1.8a (p-values equal to 0.84, 0.55, 1.00 and 0.06 respectively). This is probably due to the fact that both groups of individuals presented a slight tendency to consider the task as being relatively easy for the robot. This can be seen in the average responses to the same questions Q1.3a, Q1.4a, Q1.6a, Q1.8a, which most of them are significantly larger than the center of the scale, corresponding to 3.65 ($p = 0.017$), 3.85 ($p < 0.001$), 3.60 ($p = 0.002$) and 3.25 ($p = 0.2$, not significant), respectively, for users hearing TTS; and, 3.75 ($p < 0.001$), 3.75 ($p < 0.001$), 3.6 ($p = 0.002$) and 3.65 ($p = 0.001$), respectively, for users hearing the robot beeps.

After interacting with the robot and asking the users to evaluate whether the robot showed the expected capabilities when performing the task, the responses of the users that heard the robot TTS disagreed with those of the group that heard robot beeps. As can be seen in Fig. 12, the averages of both groups of users for question Q1.5b showed a significant difference ($p = 0.01$) indicating that the TTS group expected greater capabilities from the robot than the group that was exposed to beeps. This same tendency is observed in the differences of the average response for question Q1.6b in which the users that were exposed to TTS are less satisfied with the performance of the robot. However, this difference is not significant ($p = 0.1$). No difference in the average scores for question Q1.8b was observed between the TTS and beep groups. These results also suggest that there is a tendency in favor of hypothesis H1. However, the average scores for question Q1.4b were significantly lower than the midpoint of the scale ($p < 0.001$) for both groups of users, those exposed to TTS and those that heard the robot beeps. This should be due to the fact that both groups of individuals presented a slight tendency to consider the task as being accessible for the robot. Besides this fact, the users exposed to TTS provided a slightly higher average score for question Q1.4b that those individuals that heard the robot beeps. This difference is not statistically significant ($p = 0.165$) but could indicate that the robot synthetic speech had a more positive effect on the perceived robot's capability than the beeps.

The assumption that higher expectations lead to a higher number of positive bets can be validated a priori with questions Q1.5a and Q1.7a. The users exposed to the robot synthetic speech provided average scores for Q1.5a and Q1.7a that are higher than the center of the scale with $p < 0.001$ (significant) and $p = 0.13$ (not significant), respectively. Those individuals that heard the robot beeps gave average scores for Q1.5a and Q1.7a higher than the center of the scale with $p < 0.001$ (significant) and $p = 0.008$ (significant), respectively. As can be seen, only one average score was not significantly higher than the center of the scale, although it also shows a tendency in favor of this assumption. Accordingly, it was expected that the difference in the perception of the capabilities of the robot would be reflected in the behavior that users had when making bets, which would represent an objective measure of this difference. The more positive bets, the greater the confidence that users have over the robot. However, from Fig. 11 it is observed that there are no significant differences in the number of positive bets between the users who were exposed to TTS technology and those who heard beeps. The high percentage of positive bets, greater than 80% by both groups of individuals, can be explained by the tendency to consider the task as being accessible for the robot, as mentioned above. This is evidenced by questions Q1.3b and Q1.7b. The averages of these questions are significantly lower than the center of the scale and correspond to 2.25 ($p < 0.001$) and 1.85 ($p < 0.001$) for users that heard TTS, and 1.45 ($p < 0.001$) and 1.6 ($p < 0.001$) for users that heard beeps, respectively. As discussed above, both groups of individuals showed a small tendency to consider the task as being accessible for the robot. Besides this fact, the users exposed to TTS also provided a slightly higher average score for question Q1.3b than those individuals that heard the robot beeps with

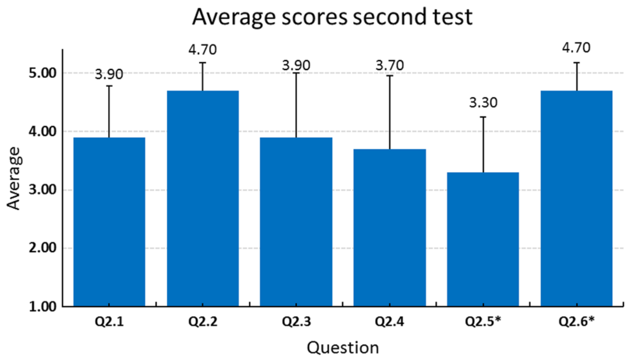

Fig. 13 Average score obtained with the questionnaire applied after completing the experiment in the second test

$p=0.017$ (significant). Again, this result could indicate that the robot synthetic speech had a more positive effect on the perceived robot's capability than the beeps. The same trend is observed with question Q1.7b, but this difference is not significant ($p=0.369$).

Regarding H2, the individuals evaluated the use of speech as a mechanism for entering commands to the robot (question Q1.2b). Users who employed voice said they liked this mechanism to enter commands: the average score is significantly higher than the midpoint of the scale ($p=0$). Accordingly, users who had to type commands with the laptop's keyboard said they would have preferred to enter the commands by voice in question Q1.2b. In this group of users, the average score is significantly higher than the midpoint of the scale ($p=0$). Consistently, this group of users also indicated in question Q1.9b that they do not like to enter commands by keyboard, with an average question that is also significantly higher than the midpoint of the scale ($p=0.003$). With respect to H3, question Q1.1b asked: whether the users exposed to TTS liked the robot to communicate with them by employing a natural voice; or, whether the users exposed to beeps would have liked the robot to communicate with them by employing a natural voice. The individuals that heard the robot TTS provided an average score that is significantly higher than the midpoint of the scale ($p=0$). On the other hand, the users that heard the robot beeps gave an average score that is not significantly below the midpoint of the scale ($p=0.206$). This result may be due to the fact that users exposed to beeps did not have the opportunity to interact with a robot that communicates by voice before the experiment, so they did not have an adequate reference to answer this question. It is worth highlighting that the average scores associated with questions Q1.1b, Q1.2b and Q1.9b, with the exception of those users that were exposed to the robot beeps in Q1.1b, suggest that users prefer to interact via voice rather than by text. Additionally, according to questions Q1.2a in Fig. 10, the users who had the opportunity to input voice commands evaluated the robot's capabilities slightly better than those who interacted by means of the keyboard. However, this result is not significant ($p=0.7$) and suggests that TTS has a more important effect on user's perception of the robot's capabilities than ASR. However, as discussed above, users still prefer to utter commands rather than typing them.

3.2 Experiment with the Second Setup

Figure 13 shows the average scores obtained with the questionnaire answered by the participants after completing the experiment in the second test. As can be seen in Fig. 13, users assigned the TTS a high naturalness, with an average score equal to 3.9 that is significantly higher than the midpoint of the scale (question Q2.1, $p=0.005$), ensuring that the TTS was considered high quality by the users.

With respect to H4, which assesses the importance of knowing the context for the completion of the task, it can be seen that the average scores for questions Q2.2 and Q2.6 are significantly higher than the midpoint of the scale ($p<0.001$ for both questions). This indicates that the users agreed that to know the context of robots facilitates human–robot collaboration and the completion of the task, thus validating the corresponding hypothesis.

Regarding H5, the use of the synthesized voice instead of text strings was evaluated as a context transmission mechanism from the robot to the user. The average reply for questions Q2.4 and Q2.5, suggests that participants show a tendency to agree that the voice is better than the text for this purpose. However, these averages are not significantly higher than the midpoint of the scale ($p=0.06$ and $p=0.17$, respectively). Nevertheless, it can be seen that the average score for question Q2.3 is significantly greater than the midpoint of the scale ($p=0.029$), which indicates that participants prefer the voice before the text as a context transmission mechanism. These results suggest that there is strong evidence in favor of H5, although not all the questions associated with this hypothesis showed significant results.

4 Discussion and Conclusions

In principle, hypothesis H1 ("Users expect better robots' performance when using state-of-the-art synthesized speech instead of NLUs") was validated by the answers provided by the users that were exposed to the robot's synthetic speech and beeps to the question that asked to evaluate the capacities of the robot without considering the task. However, no significant differences were found between the users that were exposed to the synthetic voice and those that heard beeps when they were asked about the capability of the robot to perform the task before the experiment. This must have been because all the participants were pursuing undergraduate engineering or science degrees and were somehow familiar to NAO robots. After interacting with the robots, there was a tendency in favor of hypothesis H1 but this was not statistically significant. In fact, both groups of users, those exposed to TTS and those that heard the robot beeps, presented a slight tendency to consider the task as being accessible for the robot. This is corroborated by the fact that there were no significant differences in the number of positive bets between the users who were exposed to synthetic speech and those who heard beeps. This analysis suggests that the bottom-up perceptual processes that characterize the "in-the-moment" perspective would not be enough to explain how users perceived the robot that they are interacting with. When the users interacted with a single robot in the first setup, its voice as a social cue and cause of anthropomorphization lost relevance while the interaction was carried out and the users could better evaluate the robot's capability with respect to its task. In other words, a "reflective" top-down thought became more important. An interesting discussion about potential frustration may arise. Like other negative emotions, frustration has also been studied in the context of HCI and HRI. It has been shown that frustration can lead to a detriment in the performance of the task [71, 72], an increase in decision-making time [73, 74] and a decrease in learning [75]. Within a team, frustration can cause a reduction in trust and, therefore, undermine the team's overall success. Also, inappropriate levels of trust could result in a frustrating HRI experience [76]. However, at least in the scenario considered here, the potential users' frustration was inhibited as the "reflective" top-down thought took place.

Regarding hypothesis H2 ("Users prefer to give instructions by voice rather than text"), it was validated by both set of users, those who employed voice and those who employed text as an input communication mechanism. This result is interesting due to the following reasons: as mentioned in the introduction, most authors have focused only on TTS when they have addressed voice based HRI or HCI without considering ASR [38–50]; and, the comparison between ASR and other forms of user-to-robot communication in HRI (particularly the use of text) has also been rather neglected in the literature so far. Additionally, the users who had the opportunity to input voice commands evaluated the robot's capabilities slightly better than those who interacted by means of the keyboard. This would be consistent with the anthropomorphization effect described in [44], which can also be caused by the humanoid robot's ability to walk and generate audio. Nevertheless, this result was not significant and suggests that TTS has a more important effect on user's perception of the robot's capabilities than ASR.

Similarly, the statistical analysis performed on the obtained results tends to validate H3 ("Users prefer the robot to use natural voice rather than NLUs"), as the question associated with this hypothesis shows a significant trend for one of the groups studied (users who heard TTS). Although for the group of users exposed to beeps, a slight, non-significant trend against H3 was observed and this could be perfectly attributed to the fact that these users lacked a frame of reference to compare the two methods of communication employed by the robot. Our findings tend to agree with the related works reviewed initially, in which speech is portrayed as the de facto interface preferred by humans to communicate [16–18]. Considering that issues regarding user frustration due to overestimation of robots' capabilities have been addressed, as it was found that users tend to adjust expectations on-the-fly while collaborating with the robots, the main potential drawback to using synthetic speech, frustration, would not be a concern. As such, we argue in favor of using voice as the interface between humans and machines, as seen in [20–22], in task-oriented collaborative social robotic settings. Our findings contrast with what is proposed in [54–58], in which several benefits associated with using NLUs are presented when compared to speech. We believe this is due to the nature of the scenario studied, where efficient and clear communication is required to successfully complete the proposed objective. In our setup, several of the advantages of NLUs presented in the works just mentioned are not relevant or useful regarding the completion of the task such as being language agnostic, discrete and capable of conveying affect.

About hypothesis H4 ("Knowing the context of the robot improves human–robot collaboration"), it was tested with the experiment with the second setup, the two-robot collaborative testbed employed to evaluate the pertinence of robot context acquisition and the usefulness of TTS for this purpose. Hypothesis H4 was validated by the answers provided by the participants indicating that the users agreed that knowing the context of robots facilitates human–robot collaboration and the completion of the task. Observe that the robots' context acquisition takes place by means of the robot-to-human and robot-to-robot communication. It is worth highlighting that, as discussed in the introduction, HRI with multiple robots has been rather neglected in the literature and, according to our findings, the reflective model when human interacts with multiple robots while they are executing a common task may be essential for the successful task completion. Therefore, the reflective user state requires information from the robots and the importance of the acquisition of the robots' context goes beyond being "sensitive to human expectations about covert communication" as claimed in [68]: it may condition the success or failure of a common task. This is consistent with [59–63] where the importance of human–robot collaboration or context acquisition is studied but not necessarily in scenarios with multiple robots.

With respect to hypothesis H5 (Users prefer to acquire robots' context by voice than by text), it was validated in at least one question: there was a significant tendency to prefer voice over text as a mechanism for context acquisition by the user. Observe that the participants could acquire the robots' context when one of the NAOs communicated with them or when both robots communicated between themselves. In both cases, robot-to-human and robot-to-robot communications, synthetic speech or text were employed. Regarding robot-to-robot communication, as mentioned above, previous studies have evaluated the user's perception of the use of silent and verbal communication [68, 69], suggesting that the latter can preferable to the former from the user expectation point of view. However, in our study, the context sharing mechanisms (spoken language or text) were also evaluated in the sense of which one is more convenient to carry out the task successfully, which to the authors' knowledge has not been addressed before. This study claims that context information can be essential to successfully complete the task, and users have shown a preference for acquiring this context verbally. It should be noted that, in many cases, speech may be the only way to share context in hands-free applications.

Concluding, speech science and technology plays a key role in social robotics, particularly when humans and robots need to collaborate to achieve a common task. This paradigm gains pertinence because of the current technology limitations that prevent robots from being fully autonomous in many cases. Given an application, operating conditions may be highly variable, and robots can get into unexpected difficulties when accomplishing a specific task. In these kinds of scenarios inputs from human operators can be very valuable to address the problems on-line. Consequently, the HRI community should embrace some challenges in voice science and technology as its own. For instance, using low quality synthetic speech in any HRI study makes little sense considering the state-of-the-art TTS technology. However, generating the correct prosody in synthesized speech to convey the urgency of the situation or problem faced by the robot can be an interesting task. Also, distant speech recognition in time-varying environments with multiple users is very important in social robotics and includes many problems that have not been solved yet. Finally, the comparison of speech recognition with other form of human-to-robot communication besides text can be proposed as future research.

Acknowledgements The authors would like to thank Prof. Simon King, The University of Edinburgh, for having provided the TTS system employed here and Prof. Henny Admoni, CMU, for the preliminary discussions about this research.

Funding This study was funded by Grants Conicyt-Fondecyt 1151306 and ONRG N°62909-17-1-2002.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Goodrich MA, Schultz AC (2008) Human–robot interaction: a survey. Found Trends Hum Comput Interact 1(3):203–275
2. Lopes LS, Teixeira A (2000) Human–robot interaction through spoken language dialogue. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, pp 528–534
3. Hoffman G, Vanunu K (2013) Effects of robotic companionship on music enjoyment and agent perception. In: Proceedings of the 8th ACM/IEEE international conference on human–robot interaction. ACM Press, Tokio, pp 317–324
4. Lin CY, Song KT, Chen YW, Chien SC, Chen SH, Chiang CY, Yang JH, Wu YC, Liu TJ (2012) User identification design by fusion of face recognition and speaker recognition. In: 2012 international conference on control, automation and systems. IEEE, Jeju Island South Korea
5. Zheng K, Glas DF, Kanda T, Ishiguro H, Hagita N (2013) Designing and implementing a human–robot team for social interactions. IEEE Trans Syst Man Cybernet Syst 43(4):843–859
6. Graf B, Hans M, Schraft RD (2004) Care-O-Bot II—development of a next generation robotic home assistant. Auton Robots 16(2):193–205
7. Jeong K, Sung J, Lee HS, Kim A, Kim H, Park C, Jeong Y, Lee J, Kim J (2018) Fribo: a social networking robot for increasing social connectedness through sharing daily home activities from living noise data. In: Proceedings of the 13th ACM/IEEE international conference on human–robot interaction. IEEE Press, Chicago, pp 114–122
8. Pachidis T, Vrochidou E, Kaburlasos VG, Kostova S, Bonković M, Papić V (2018) Social robotics in education: state-of-the-art and directions. In: Proceedings of the 27th international conference on robotics in Alpe-Adria Danube region. Springer, Cham, pp 689–700
9. Wei CW, Hung I (2011) A joyful classroom learning system with robot learning companion for children to learn mathematics multiplication. Turk Online J Educ Technol 10(2):11–23
10. Barker BS, Ansorge J (2007) Robotics as means to increase achievement scores in an informal learning environment. J Res Technol Educ 39(3):229–243
11. Highfield K (2010) Robotic toys as a catalyst for mathematical problem solving. Aust Primary Math Classroom 15(2):22–27
12. Young SSC, Wang YH, Jang JSR (2010) Exploring perceptions of integrating tangible learning companions in learning english conversation: colloquium. Br J Educ Technol 41(5):E78–E83
13. Cabibihan J-J, Javed H, Ang M, Aljunied SM (2013) Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism. Int J Soc Robot 5:593–618
14. Michaud F, Duquette A, Nadeau I (2003) Characteristics of mobile robotic toys for children with pervasive developmental disorders. In: 2003 IEEE international conference on systems, man, and cybernetics, SMC, pp 2938–2943. IEEE
15. Kozima H, Nakagawa C, Yasuda Y (2007) Children-robot interaction: a pilot study in autism therapy. Prog Brain Res 164:385–400

16. Meszaros EL, Le Vie LR, Allen BD (2018) Trusted communication: utilizing speech communication to enhance human–machine teaming success. In: AIAA aviation technology, integration, and operations conference, AIAA-2018-4014, Atlanta, GA
17. Han S, Hong J, Jeong S, Hahn M (2010) Robust GSC-based speech enhancement for human machine interface. IEEE Trans Consum Electron 56(2):965–970
18. Staudte M, Crocker MW (2011) Investigating joint attention mechanisms through spoken human–robot interaction. Cognition 120(2):268–291
19. Krämer NC, von der Pütten A, Eimler S (2012) Human-agent and human–robot interaction theory: similarities to and differences from human-human interaction. In: Zacarias M, Oliveira JV (eds) Human-computer interaction: the agency perspective, vol 396. Springer, Heidelberg, pp 215–240
20. Cassell J, Bickmore T, Campbell L, Vilhjálmsson H, Yan H (2000) Human conversation as a system framework: designing embodied conversational agents. In: Cassell J, Sullivan J, Prevost S, Churchill E (eds) embodied conversational agents. MIT Press, Cambridge, pp 29–63
21. Gratch J, Rickel J, André E, Cassell J, Petajan E, Badler N (2002) Creating interactive virtual humans: some assembly required. IEEE Intell Syst 17:54–63
22. Kopp S, Wachsmuth I (2004) Synthesizing multimodal utterances for conversational agents. Comput Animat Virt World 15:39–52
23. Parise S, Kiesler S, Sproull L, Waters K (1999) Cooperating with life-like interface agents. Comput Hum Behav 15:123–142
24. Rickenberg R, Reeves B (2000) The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM Press, New York, pp 49–56
25. Sproull L, Subramani M, Kiesler S, Walker JH, Waters K (1996) When the interface is a face. Hum Comput Interact 11:97–124
26. Swinth KR, Blascovich J (2001) Conformity to group norms in an immersive virtual environment. In: 2001 annual meeting of the American Psychological Society (APS), Toronto, Ontario. Canada
27. Woods S, Dautenhahn K, Kaouri C (2005) Is someone watching me?-consideration of social facilitation effects in human–robot interaction experiments. In: 2005 international symposium on computational intelligence in robotics and automation. IEEE, pp 53–60
28. Krämer NC, Bente G, Piesk J (2003) The ghost in the machine. The influence of embodied conversational agents on user expectations and user behaviour in a TV/VCR application. IMC workshop, pp 121–128
29. Schermerhorn P, Scheutz M, Crowell CR (2008) Robot social presence and gender: Do females view robots differently than males?. In: Proceedings of the 3rd ACM/IEEE international conference on human robot interaction. ACM, pp 263–270
30. Rist T, Baldes S, Gebhard P, Kipp M, Klesen M, Rist P, Schmitt M (2002) CrossTalk: An interactive installation with animated presentation agents. In: Proceedings of the 2nd international conference on Computational Semiotics for Games and New Media (COSIGN), pp 61–67
31. Jung B, Kopp S (2003) Flurmax: An interactive virtual agent for entertaining visitors in a hallway. In: Proceedings of the 4th international workshop on intelligent virtual agents, IVA 2003. Springer, Kloster Irsee, pp 23–26
32. Takayama L (2012) Perspectives on agency interacting with and through personal robots. In: Zacarias M, Oliveira JV (eds) Human–computer interaction: the agency perspective, vol 396. Springer, Heidelberg, pp 195–214
33. Reeves B, Nass CI (1996) The media equation: how people treat computers, television, and new media like real people and places. Cambridge University Press, New York
34. Nass C, Moon Y (2000) Machines and mindlessness: social responses to computers. J Soc Issues 56:81–103
35. Nass C, Steuer J, Tauber ER (1994) Computers are social actors. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, USA, CHI'94, pp 72–78
36. Moore RK (2014) Spoken language processing: time to look outside? In: 2nd international conference on statistical language and speech processing (SLSP 2014), pp 21–36
37. Gold T (1980) Speech production in hearing-impaired children. J Commun Disord 13:397–418
38. Tamagawa R, Watson CI, Kuo IH, MacDonald BA, Broadbent E (2011) The effects of synthesized voice accents on user perceptions of robots. Int J Soc Robot 3:253–262
39. Niculescu A, Dijk B, Nijholt A, Li H, See SL (2013) Making social robots more attractive: the effects of voice pitch, humor and empathy. Int J Soc Robot 5:171–191
40. Lee EJ, Nass C, Brave S (2000) Can computer-generated speech have gender?: an experimental test of gender stereotype. In: CHI'00 extended abstracts on human factors in computing systems (CHI'00). ACM Press, New York, pp 289–290
41. Nass C, Lee KM (2001) Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. J Exp Psychol Appl 7(3):171
42. Gong L, Lai J (2001) Shall we mix synthetic speech and human speech?: impact on users' performance, perception, and attitude. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM Press, New York, pp 158–165
43. Nass C, Foehr U, Brave S, Somoza M (2001) The effects of emotion of voice in synthesized and recorded speech. In: Proceedings of the AAAI symposium emotional and intelligent II: The tangled knot of social cognition
44. Eyssel F, De Ruiter L, Kuchenbrandt D, Bobinger S, Hegel F (2012) 'If you sound like me, you must be more human': On the interplay of robot and user features on human–robot acceptance and anthropomorphism. In: Proceedings of the 7th ACM/IEEE international conference on human–robot interaction. ACM Press, Boston, Massachusetts, pp 125–126
45. Eyssel F, Kuchenbrandt D, Hegel F, Ruiter L (2012) Activating elicited agent knowledge: How robot and user features shape the perception of social robots. In: The 21st IEEE international symposium on robot and human interactive communication, RO-MAN, pp 851–857
46. McGinn C, Torre I (2019) Can you tell the robot by the voice? An exploratory study on the role of voice in the perception of robots. In: Proceedings of the 14th ACM/IEEE international conference on human–robot interaction. IEEE Press, Daegu, pp 211–221
47. Crowelly CR, Villanoy M, Scheutzz M, Schermerhornz P (2009) Gendered voice and robot entities: perceptions and reactions of male and female subjects. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, pp 3735–3741
48. Walters ML, Syrdal DS, Koay KL, Dautenhahn K, Te Boekhorst R (2008) Human approach distances to a mechanical-looking robot with different robot voice styles. In: The 17th IEEE international symposium on robot and human interactive communication, RO-MAN, pp 707–712
49. Niculescu A, Van Dijk B, Nijholt A, See SL (2011) The influence of voice pitch on the evaluation of a social robot receptionist. In: Proceedings of the 2011 international conference on user science and engineering, i-USEr 2011, pp 18–23
50. Cha E, Dragan AD, Srinivasa SS (2015) Perceived robot capability. In: The 24th IEEE international symposium on robot and human interactive communication, RO-MAN, pp 541–548

51. Fischer K, Soto B, Pantofaru C, Takayama L (2014) Initiating interactions in order to get help: effects of social framing on people's responses to robots' requests for assistance. In: The 23rd IEEE international symposium on robot and human interactive communication, RO-MAN, pp 999–1005
52. Read R, Belpaeme T (2014) Non-linguistic utterances should be used alongside language, rather than on their own or as a replacement. In: Proceedings of the 9th ACM/IEEE international conference on human–robot interaction. ACM Press, New York, pp 276–277
53. Hollingum J, Cassford G (2013) Speech technology at work. Springer, Berlin
54. Khota A, Kimura A, Cooper E (2019) Modelling of non-linguistic utterances for machine to human communication in dialogue. In: 5th international symposium on affective science and engineering. Japan Society of Kansei Engineering, Tokyo, pp 1–4
55. Schwenk M, Arras KO (2014) R2-D2 reloaded: a flexible sound synthesis system for sonic human–robot interaction design. In: The 23rd IEEE international symposium on robot and human interactive communication, RO-MAN, pp 161–167
56. Read R, Belpaeme T (2016) People interpret robotic non-linguistic utterances categorically. Int J Soc Robot 8:31–50
57. Read R, Belpaeme T (2012) How to use non-linguistic utterances to convey emotion in child-robot interaction. In: Proceedings of the 7th ACM/IEEE international conference on human–robot interaction. ACM Press, Boston, Massachusetts, pp 219–220
58. Read R (2014) A study of non-linguistic utterances for social human–robot interaction. PhD Thesis, University of Plymouth, Plymouth, United Kingdom
59. Bechar A, Edan Y (2003) Human–robot collaboration for improved target recognition of agricultural robots. Ind Robot Int J 30(5):432–436
60. Kardos C, Kemény Z, Kovács A, Pataki BE, Váncza J (2018) Context-dependent multimodal communication in human–robot collaboration. In: 51st CIRP international conference on manufacturing systems, pp 15–20
61. Lakhmani SG, Wright JL, Schwartz MR, Barber D (2019) Exploring the effect of communication patterns and transparency on performance in a human–robot team. In: Proceedings of the 63rd human factors and ergonomics society annual meeting. SAGE Publications, Los Angeles, CA, pp 160–164
62. Marvel JA, Bagchi S, Zimmerman M, Antonishek B (2020) Towards effective interface designs for collaborative HRI in manufacturing: metrics and measures. ACM Trans Comput Hum Interact 9(4):1–55
63. Lyons JB, Havig PR (2014) Transparency in a human-machine context: approaches for fostering shared awareness/intent. In: Proceedings of the 6th international conference on virtual, augmented and mixed Reality. Springer, Cham, pp 181–190
64. Wang E, Kim YS, Kim HS, Son JH, Lee S, Suh IH (2005) Ontology modeling and storage system for robot context understanding. In: Proceedings of the 9th international conference on knowledge-based and intelligent information and engineering systems. Springer, Berlin, pp 922–929
65. Chernova S, Veloso M (2010) Confidence-based multi-robot learning from demonstration. Int J Soc Robot 2:195–215
66. Arimoto T, Yoshikawa Y, Ishiguro H (2018) Multiple-robot conversational patterns for concealing incoherent responses. Int J Soc Robot 10:583–593
67. Silva P, Pereira JN, Lima PU (2015) Institutional robotics. Int J Soc Robot 7:825–840
68. Williams T, Briggs P, Scheutz M (2015) Covert robot-robot communication: human perceptions and implications for human–robot interaction. J Hum Robot Interact 4(2):24–49
69. Tan XZ, Reig S, Carter EJ, Steinfeld A (2019) From one to another: how robot-robot interaction affects users' perceptions following a transition between robots. In: Proceedings of the 14th ACM/IEEE international conference on human-robot interaction. IEEE Press, Daegu, pp 114–122
70. Dahlbäck N, Jönsson A, Ahrenberg L (1993) Wizard of Oz studies: why and how. In: International conference on intelligent user interfaces, IUI1993, pp 193–200
71. Klein J, Moon Y, Picard RW (2002) This computer responds to user frustration: theory, design, and results. Interact Comput 14(2):119–140
72. Powers SR, Rauh C, Henning RA, Buck RW, West TV (2011) The effect of video feedback delay on frustration and emotion communication accuracy. Comput Human Behav 27(5):1651–1657
73. Bechara A (2004) The role of emotion in decision-making: evidence from neurological patients with orbitofrontal damage. Brain Cogn 55(1):30–40
74. Lerner JS, Li Y, Valdesolo P, Kassam KS (2015) Emotion and decision making. Ann Rev Psychol 66:799–823
75. Graesser AC, Chipman P, Haynes BC, Olney A (2005) AutoTutor: an intelligent tutoring system with mixed-initiative dialogue. IEEE Trans Educ 48(4):612–618
76. Salem M, Dautenhahn K (2015) Evaluating trust and safety in HRI: practical issues and ethical challenges. In: Proceedings of the 10th ACM/IEEE international conference on human–robot interaction (HRI 2015): workshop on the emerging policy and ethics of human–robot interaction. ACM, New York, NY

Jorge Wuth received his electrical engineering degree from the University of Chile in 2007. Since 2007, he has been a research associate at the Speech Processing and Transmission Laboratory (LPTV) where he is currently carrying out his research on artificial intelligence and signal processing applied to voice-based human-robot interaction and ultidisciplinary problems. Pedro Correa received his electrical engineer degree from the University of Chile in 2020. He was a research assistant at the Speech Processing and Transmission Laboratory (LPTV) between 2017 and 2019 where he worked on deep learning applied to robust speech processing. He is also a songwriter and music producer. Tomás Núñez González is an Electrical Engineering student at the University of Chile. He worked from 2018 to 2019 as a research assistant at the Speech Processing and Transmission Laboratory (LPTV). His current academic interests lie in IP Networks, PHY and MAC layers, and Machine Learning applications in Telecommunications.

Pedro Correa received his electrical engineer degree from the University of Chile in 2020. He was a research assistant at the Speech Processing and Transmission Laboratory (LPTV) between 2017 and 2019 where he worked on deep learning applied to robust speech processing. He is also a song writer and music producer.

Tomás Núñez an Electrical Engineering student at the University of Chile. He worked from 2018 to 2019 as a research assistant at the Speech Processing and Transmission Laboratory (LPTV). His current academic interests lie in IP Networks, PHY and MAC layers, and Machine Learning applications in Telecommunications.

Matías Saavedra is an Electrical Engineering student at the University of Chile. In 2018 he was a research assistant at the Speech Processing and Transmission Laboratory (LPTV). He is currently working on computer vision for waste recycling as his final year project dissertation.

Néstor Becerra Yoma received the Ph.D. degree from the U. of Edinburgh, UK, and the M.Sc. and B.Sc. degrees from UNICAMP (Campinas State University), Sao Paulo, Brazil, all of them in Electrical Engineering, in 1998, 1993 and 1986, respectively. From August 2016 to June 2017 he was a visiting professor in the ECE Dept. at Carnegie Mellon University (CMU), Pittsburgh, USA. From 2000 he has been a Professor at the Dept. of Electrical Engineering, U. de Chile, in Santiago, where he is currently lecturing on signal and speech processing, machine learning, and telecommunications. At U. de Chile he started the Speech Processing and Transmission Laboratory (LPTV, Laboratorio de Procesamiento y Transmisión de Voz) to carry out research on artificial intelligence and signal processing for robust speech recognition/speaker verification, human-robot interaction and multidisciplinary problems.

Prof. Becerra Yoma was promoted to the Full Professor position in 2011. He is the coauthor of over 80 research papers in international journals and conferences. He is also the author of patents in Chile, the US and Australia. From January 2009 to December 2012 he was an associate editor of IEEE Transactions on Speech and Audio Processing (IEEE-TASLP). He has also been invited to give talks in several universities. Professor Becerra Yoma is a senior member of the IEEE and a member of ISCA.