How context shapes the appropriateness of a robot's voice

Ilaria Torre[1], Adrian Benigno Latupeirissa[2] and Conor McGinn[3]

***Abstract*— Social robots have a recognizable physical appearance, a distinct voice, and interact with users in specific contexts. Previous research has suggested a 'matching hypothesis', which seeks to rationalise how people judge a robot's appropriateness for a task by its appearance. Other research has extended this to cover combinations of robot voices and appearances. In this paper, we examine the missing connection between robot voice, robot appearance, and deployment context. In so doing, we asked participants to match a robot image to a voice within a defined interaction context. We selected widely available social robots, identified task contexts they are used in, and manipulated the voices in terms of gender, naturalness, and accent. We found that the task context mediates the 'matching hypothesis'. People consistently selected a robot based on a vocal feature for a certain context, and a different robot based on the same vocal feature for another context. We suggest that robot voice design should take advantage of current technology that enables the creation and tuning of custom voices. They are a flexible tool to increase perception of appropriateness, which has a positive influence on Human-Robot Interaction.**

I. INTRODUCTION

Spoken communication is the primary form of interaction between a growing number of social robots and their users. However, the overall effort of designing robot voices is considerably less than the amount of work that goes into designing their physical appearance [1]–[4]. Research in Human-Robot Interaction (HRI) could exploit the flexibility afforded by many available Text-to-Speech systems, voice banks, and custom recordings, enabling designers to gain greater control over how robots are perceived.

Voices contribute to impression formation of newly-met individuals in human-human interactions [5], as well as shaping how impressions develop over time [6]. Besides linguistic content, voices carry a wide variety of information, ranging from indexical characteristics of the speaker such as gender, age, and place of origin, as well as temporary state alterations, like mood, emotions, or health [7], [8].

*The research was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska- Curie grant agreement No. 713567, and by the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2016) and is co-funded by the European Regional Development Fund. The first author is funded by a WASP Expedition Project on Correct-by-design and Socially Acceptable Autonomy (CorSA)

[1]Ilaria Torre is with the Division of Robotics, Perception and Learning, KTH Royal Institute of Technology, Stockholm, Sweden ilariat@kth.se

[2]Adrian Benigno Latupeirissa is with the Division of Media Technology and Interaction Design, KTH Royal Institute of Technology, Stockholm, Sweden ablat@kth.se

[3]Conor McGinn is with the School of Engineering, Trinity College Dublin, Dublin, Ireland mcginnco@tcd.ie

Importantly, listeners are accurate at deciphering this information [9] and use it to decide on their next action [10]. Applying these observations for human-robot interactions, we can hypothesise that an appropriate voice for a robot will influence users' impressions. But what is an 'appropriate' voice for a robot? One suggestion is that voices should match the physical features of robots [11], [12], for example in terms of anthropomorphism [13]. However, as social robots are increasingly employed in a variety of settings, such as in hotel receptions, nursing homes, museums, schools, and hospitals, it is conceivable that the task context in which the robot is employed will also contribute to voice appropriateness.

A voice may provide important indicators to influence the perceived suitability of the robot in a specific deployment context, similar to what has been shown for human voices advertising products in commercials [14]. Voice, physical appearance, and context might shape the communication between humans and social robots, yet we are not aware of any systematic approach to understand this interaction.

The paper is structured as follows: previous work on first impressions and expectations based on voice and appearance in human-robot as well as human-human interaction is discussed in Section II; our contribution is detailed in Section III; results from the experiment are presented in Section IV and discussed in Section V.

II. RELATED WORK

Previous studies have shown that people perceive robots differently depending on the context in which the interaction takes place. This effect of context has mostly been studied in terms of specific physical robot features, most notably anthropomorphism (machine-like vs. human-like), behaviour (playful vs. serious; making mistakes; exhibiting joint attention; etc.), and gender. Regarding anthropomorphism, in [15], people selected their preferred robot companion from a list of different pictures, varying in terms of gender, age, and human-likeness, for different tasks (personal care, social interaction, decision making and house chores). Generally, people preferred the more machine-like robots for all tasks except the decision making ones. [16] demonstrated that anthropomorphism level can also affect the attribution of blame to robots that have to make moral decisions: specifically, human-like robots are blamed more for action in the footbridge dilemma (derailing the train and killing the worker on the tracks), while machine-like robots are blamed more for inaction (not derailing the train and letting the passengers crash). Robot appearance and behaviour were also found to affect robot mind judgments, i.e. what intentions

a robot is perceived to have [17]. Finally, participants in [18] attributed higher social capabilities, including honesty, to robots that looked more sophisticated. This is an important consideration, as over-trusting or under-trusting a robot in a critical situation might lead to serious harm [19], [20].

Regarding gender, there is evidence that people apply gender stereotypes to robots; for example, participants in [21] had to converse with a robot, which was manipulated to look either male or female, about dating norms – a stereotypically feminine topic. Participants spent longer time talking with the female robot, perhaps because they thought that this robot had more knowledge on the topic. Participants in [15] also mentioned that they would like a female robot to assist with chores and personal care, and a male robot to assist with decision making tasks.

Delving into why certain robot features might be perceived as more appropriate than others in a specific context, [22] found support for a 'matching hypothesis': robot appearance and behaviour should match the task that the robot has to carry out, for instance in terms of seriousness. Their participants liked a serious-behaving robot more in a serious task, and a playful-behaving robot more in a playful task. A similar explanation was suggested by [23]: the appearance and behaviour of a robot might evoke some first impressions, or mental model, on what tasks it might be able to perform; these first impressions are then weighted against the actual tasks it needs to perform. For example, participants in [15] might have thought that a mechanical-looking robot lacked the skills to help them make economic decisions, while a human-like robot must have possessed them. In general, researchers agree that a robot should evoke the right first impressions and expectations, because in cases where expectations are higher than performance, future trust and compliance will decrease [6], [23]–[25].

A few studies discovered a similar 'matching effect' for integrating robot voice and appearance. In these instances, robot voice was mostly investigated in terms of naturalness, gender, and accent. For example, regarding naturalness, [13] argued that robot voice and body should be matched in terms of anthropomorphism, to avoid feelings of eeriness. [26] suggested that a robot's language skills influence people's behaviour towards that robot: participants gave more commands to a robot that had a voice, whether synthetic or natural, and fewer to a robot that communicated with beeps. Participants likely assumed that speechless robots could not understand language, so they did not speak either. Within the speaking robot condition, however, participants gave more commands to the synthetic-voiced robot than the natural one, perhaps because they thought that a robot with a human voice was more competent and therefore needed fewer commands. [26] also investigated robot deployment context: participants watched videos of a robot in different scenarios (robot damaged, robot in danger, robot requiring more information, robot has located target, robot has completed task). They found that, for example, participants gave more commands to the robot in the videos where the robot needed assistance, and concluded that a robot's voice should be chosen based on task context. In particular, this would allow for the transmission of pragmatic information which may increase the operation success.

In terms of voice gender, it has been found that a NAO robot was assigned a different gender based on its voice alone [27]; however, the voice gender of the robot did not affect participants' perceived robot friendliness, trustworthiness, or likeability [28], [29]; nor did it influence people's interpersonal distance from the robot [30]. However, robots that showed feminine physical traits were matched with a female voice more often than a male voice, and the same happened the other way around [12]. Thus, while perceiving a robot as having a gender might not affect the quality of the interaction, we still find evidence of a 'matching hypothesis', in that people tend to match multimodal features related to the same gender.

Finally, in terms of accent, a few studies have shown that a robot speaking with an accent that matches participants' accent is perceived more favourably [27], [31]. This can be explained in terms of in-group preferences, as perceiving someone as belonging to one's same social group immediately primes favourable first impressions [32], including of robots [33]. Finally, there is one study that examined the interaction between accent and context in Human-Robot Interaction: an experiment focusing on the Arabic language showed that participants believed that robots with the same regional accent as theirs were more credible – when the robots were knowledgeable – than those with a standard accent. On the other hand, robots with standard accents were perceived to be the more credible when the robots had little knowledge [34]. Similar interactions between accents and context are plausible with other languages. Thus, these studies suggest that vocal features such as naturalness, gender, and accent influence people's perception and behaviour towards robots.

Nevertheless, all this evidence seems to not have yet inspired a change in the way researchers design speech-based Human-Robot Interaction studies. [12] conducted an informal survey of researchers whose paper at the HRI 2018 conference featured a speaking robot, asking why they chose a certain voice for this robot. Very few responses mentioned looking for a specific feature in the voice, such as gender or accent, to suit the type of interaction; the majority reported choosing a voice due to convenience. These findings indicate how little attention has been dedicated to robot voice design in the past. The contrast with other fields is striking; robot voices in films and television are carefully crafted to achieve the desired character effect [4], and voice design for bodyless conversational user interfaces is receiving a lot of attention in its research venues [35]–[37].

With this brief review of relevant literature, we have shown how different robot characteristics can help form first impressions of a robot. Critically, these first impressions contribute to forming expectations of what a robot can or cannot do. Robots available nowadays are used for a variety of jobs – e.g. we might have seen a Pepper robot being used for catering, tutoring, elderly care, and more [38]–[40].

Thus, people working with these off-the-shelf robots will be somewhat limited in terms of being able to adapt the physical characteristics of the robot to suit a certain work context better; for example, they could change its eye colour, but might not be able to easily change the degrees of freedom of a joint. However, other, more versatile characteristics, such as voice, could be used to increase the match between robot and context. In this paper, we combine some of the characteristics that have so far been studied mostly in isolation – robot anthropomorphism, gender, voice accent, voice naturalness – and see how robot deployment context influences mental models of robot appearance. We build on our previous study [12] – where people matched a robot voice with a robot picture based on vocal and physical features of the robot – by adding a context variable.

III. METHOD

We examined whether people would associate a robot picture to a certain voice, given a specific Human-Robot Interaction task context. Thus, we collected a set of voices to form the basis of our stimuli. The voices used in the experiment included: (a) voices previously used in recent HRI research (Table I), and (b) human voices we recorded and manipulated in terms of naturalness, gender, and accent. Participants were then randomly assigned to one of these two experimental conditions: robot voices (RV) and human-derived voices (HV).

A. Stimuli

Participants listened to robot voices uttering a short script, which was designed to be plausible in different contexts. The script recited as follows: *"Hello, sorry to bother you. My software needs an update. I just wanted to let you know that I need to be offline for a short period. I will get back to work in around five minutes."*.

The goal of the experiment was to examine whether participants would associate a certain robot voice, in a certain context, to a specific robot. Therefore, we selected 8 possible robots that participants could choose from. These robots were selected because they represented a diverse sample of widely used social robots (i.e. wheeled/legged, digital/mechanical head, two/one/zero arms, etc.), had a similar overall form factor (estimated range 1000-1600mm tall), and because they had been used in conjunction with a specific synthetic voice in the past (Table I).

1) Robot voices: In the RV condition, participants heard voices that had been used on our robots in previous studies (listed in Table I). These voices allowed us to study whether people associated a robot with their previously used voice, which we call 'default' for the purposes of this study. If participants associate these robots to their 'default' voice, this suggests that this was a good choice of robot voice. These voices also allow us to study whether people associate a certain voice gender to a robot, since we know this characteristic from the type of voice used (Table I). Note that, since the manufacturers of Pepper explicitly indicate that its voice is neither male nor female [41], for the purposes of this study we categorised it as 'ambiguous'. These previously used voices also differ in terms of naturalness and accent, but this variation is more difficult to quantify than the gender one, as there are different synthetic voice qualities, depending on the TTS system used. Therefore, we used the voices in the RV condition to observe voice appropriateness along two categories: default (previously used on this robot or not) and gender (female, male, ambiguous). Also note that one reference study had used the same voice for both G5 and iCub. Therefore, in the experiment we considered this voice to be the default for both robots. Thus, we had a total of 7 robot voices for the RV condition.

TABLE I
ROBOTS AND CORRESPONDING VOICES USED IN THE ANALYSIS.

Robot	Speech Engine	Voice Name	Voice gender	Reference
Flash	CereProc	Heather	Female	[42]
G5	Acapela	Rod	Male	[43]
iCub	Acapela	Rod	Male	[43]
Pepper	Pepper	Default	Ambiguous	Developer
Poli	Amazon	Kim	Female	[44]
PR2	Cepstral	David	Male	[45]
SCIPRR	Cepstral	Alison	Female	[46]
Stevie	CereProc	Giles	Male	[47]

2) Human-derived voices: In the HV condition, participants heard voices that were recorded from human speakers, and either resynthesised to sound 'mechanical', or kept as they were. These original voices were recorded from 4 speakers: 2 from California (1 male, 1 female) and 2 from Dublin (1 male, 1 female). We chose these two accents due to their distinctive features, and due to their nature as local (Irish) and global (American) accents of English. The four speakers were recorded in a quiet room using a Zoom H6 Handy Recorder and an AKG C520L condenser microphone, where they read the aforementioned robot script. All the recordings were cleaned with a noise-removal filter in Audacity. The 4 voices were also passed through a vocoder to obtain a synthetic-sounding effect, while retaining the accent and gender features. To obtain this effect, we first flattened the fundamental frequency (f0) of each speaker to that speaker's mean f0 value, and then applied a comb filter using Audacity. The resulting re-synthesised voices were monotonous and had a metallic flare. These voices allowed us to observe voice appropriateness along three categories: gender (male and female), accent (American and Irish), and naturalness (natural or resynthesised). Thus, we had a total of 8 human-derived voices.

3) Context creation: We identified 4 plausible contexts where our robots could work: home, hospital, restaurant, and school. These were chosen because at least some of our robots of interest have already been used in these contexts [38]–[40], and because they represent venues that are currently being explored for robot deployment. To create a contextual illusion, we added some background noise to each of the voices, to immerse participants in the different contexts. We added a living room ambience noise to simulate a home context, an echoing corridor noise to simulate a hospital, chatting and clinking to simulate a restaurant, and

children running and laughing to simulate a school. Thus, we had 4 versions for each voice (one per context). The total number of sound files was therefore 32 in the HV condition, and 28 in the RV condition.

B. Participants

The experiment took place over the course of several days during a museum/gallery space, located adjacent to Trinity College Dublin. In total, 60 participants (age 19-74, mean = 27, sd = 9) volunteered to take part in the experiment. There were 35 women and 25 men. Of the participants, 26 were originally from Ireland, 8 from France, 6 from the USA, 4 from Germany, and the remaining came from 13 other countries (Australia, Austria, Azerbaijan, Brazil, Canada, Germany, Greece, Italy, Netherlands, Nigeria, Spain, Sweden, UK). Their self-reported English language fluency was as follows: 31 native speakers; 2 native-like; 20 fluent; 7 basic. The majority of participants (n = 37) were not affiliated with the University, 12 were University employees and 11 were students. To avoid any potential confounds due to familiarity with the robots used in the study, participants were also asked if they had interacted with a robot before, and if so, with which robot. The majority of participants had either never seen a robot before (N = 11), or had only seen robots in media (N = 27). Some people had interacted with a robot before (N = 21) and only one person declared that they interacted with robots on a regular basis. Of the robots of interest in the current study, Pepper was mentioned 4 times, Stevie 5 times, iCub and Flash one time each.

C. Procedure

The experiment was conducted in a quiet space, where participants were first asked to read an information sheet and provide written informed consent, in accordance with ethics requirements. Then, they filled in a short demographics questionnaire about their age, gender, English language fluency, country and city of origin, and degree of familiarity with robots. Participants were randomly assigned to one of the two experimental conditions (RV or HV), and were positioned at a computer desk wearing good quality over-ear headphones, and ran one practice trial with the experimenter. This practice trial used voices and robots from popular culture, which were not shown again during the actual experiment. Then, they were left to complete the experiment. Condition HV consisted of 32 trials (one per sound file) and condition RV of 28. Depending on the experimental condition, participants were presented with voices (in random order) from either the HV or the RV pool. Every trial proceeded as follows: for each voice, participants listened to the scripted utterance, while a fixation cross appeared in the centre of the screen. After the voice sequence had terminated, they were shown the pictures of the 8 robots of interest (Fig. 1), equally positioned on the screen, and were asked to select the picture of the robot that best suited the voice they just heard. The relative location of each picture was fixed throughout the experiment. A 'restart' button was placed adjacent to the pictures, to allow participants to hear the voice again, or change their selection. The experiment lasted approximately 15 minutes.

IV. RESULTS

We conducted chi-square tests for independence on each of the variables of interest – gender, naturalness, accent, context – to see if there was a causal relationship between the robots being selected and the variables. Post-hoc analyses – to see whether a robot was selected more often than the others for each variable of interest – were conducted by testing the χ^2 residuals for each robot against a critical $zvalue$ and adjusting the α level for multiple comparisons (Bonferroni correction). The full contingency tables can be found in the supplementary materials[1]. Given that this study deals with the effect of context, here we will not describe in detail the individual effects of gender, naturalness and accent – which mostly replicate our previous findings [12] – but we will focus on their interaction with context. All the results can still be found in the supplementary materials.

A. Robot voices

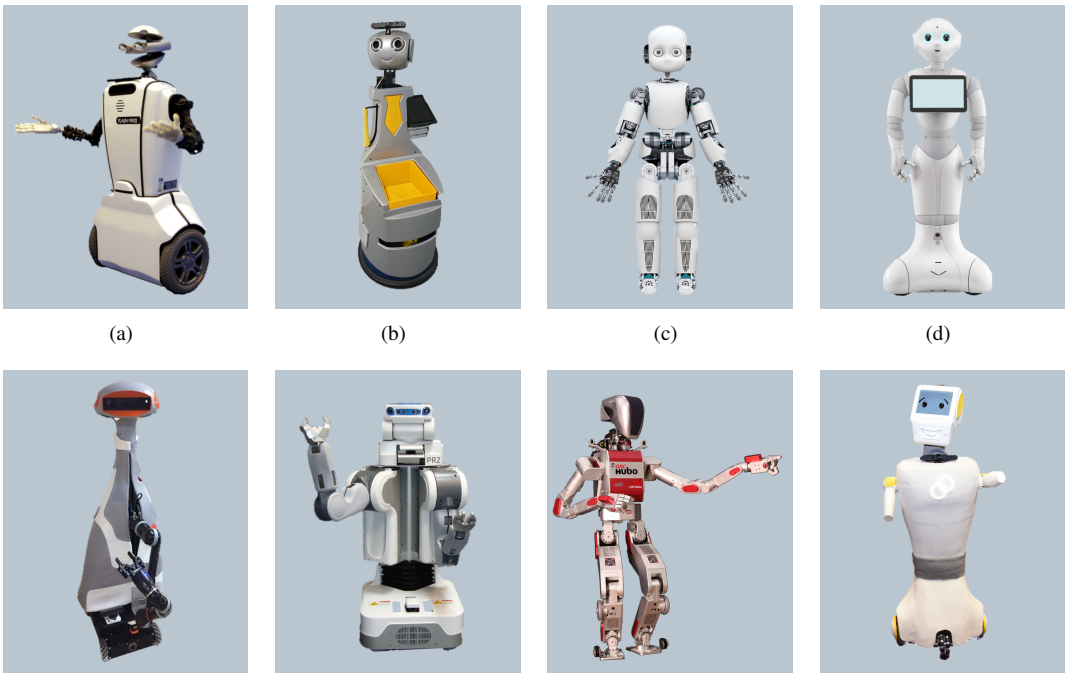
First of all, we examined whether participants selected the 'default' robot upon hearing the voice that robot had in previous HRI studies (Table I). As can be seen from Fig. 2, people did not generally associate a robot image with its 'default' voice. The notable exception was PR2, whose voice was deemed appropriate much above chance level, thus replicating our previous results [12]. Pepper's and Poli's voices were also recognised above chance level.

Then, we looked at whether voice gender played a role in robot selection. Here, we had categorised the voices as either male, female, or ambiguous (Table I). We performed χ^2 test of independence on the whole contingency table (see Table 1 in the supplementary materials), and we found a significant association between gender and the robot being selected ($\chi^2(14, N = 29) = 237.18, p < .001$). We then looked at each robot selection in the 4 different contexts. As can be seen from Fig. 3, Flash was selected significantly more often with a male voice in the home context, and significantly less often with a female voice in the school context; G5 was selected less often with a male voice in the home context, and more often with an ambiguous voice in the hospital context; iCub was selected less often with a male voice in the home, hospital, and restaurant contexts; Pepper was selected more often with a female voice in the home and school contexts, and less often with a male voice in the home and restaurant contexts; PR2 was selected more often with a male voice in all 4 contexts, and less often with a female voice in the home and school contexts; SCIPRR was selected more often with a male voice in the home and school contexts.

B. Human-derived voices

For the human-derived voices, we had 3 parameters that could interact with context: voice gender, voice accent, and voice naturalness.

[1]The supplementary materials can be found here: `https://doi.org/10.5281/zenodo.3776826`



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 1. Images of the robots used in this study: (a) *Flash*, (b) *G5*, (c) *iCub*, (d) *Pepper*, (e) *Poli*, (f) *PR2*, (g) *HUBO-SCIPRR*, (h) *Stevie*.

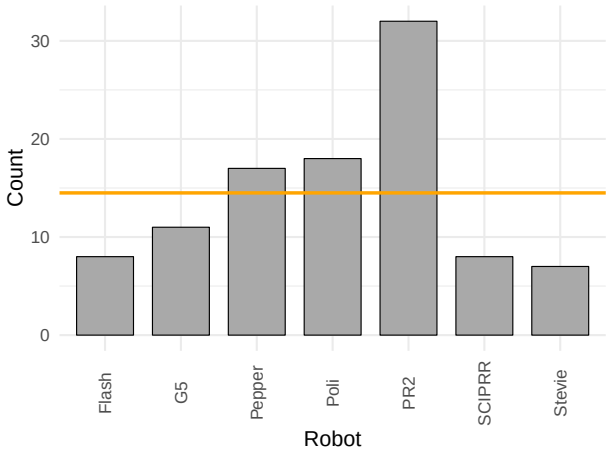


Fig. 2. Number of people who selected a robot upon hearing its 'default' voice, assuming independence between trials. The orange horizontal line indicates the 12.5% chance level of selecting the 'correct' robot at each trial.

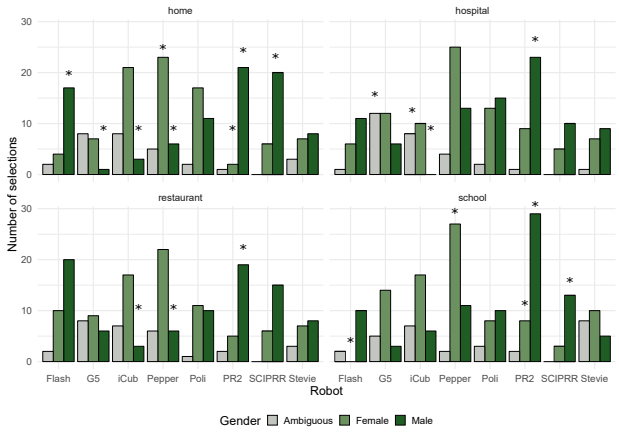


Fig. 3. Robot selection based on voice gender (robot voices) in the 4 context conditions. '*' indicates that a robot was selected significantly more or less often upon hearing a certain voice (at the 95% significance level).

First of all, a χ^2 test of independence on the whole contingency table (see Table 2 in the supplementary materials), found a significant association between gender and the robot being selected ($\chi^2(7, N = 31) = 178, p < .001$). As can be seen from Fig. 4, context influenced robot selection for the human voices as well. Flash was selected more with a male voice than a female voice in all 4 contexts; Pepper was selected significantly more often with a female than a male voice only in the school context; Poli was also selected significantly more often with a female than a male voice only in the school context; Stevie was selected more often with a male than a female voice in the home, restaurant, and school contexts.

There was also a significant main effect of accent ($\chi^2(7, N = 31) = 15.35, p = .03$). However, after adjusting for multiple comparisons, only one comparison approached significance: there was a tendency for Stevie to be selected more often with an American accent in the restaurant context (see Table 4 in the supplementary materials).

Finally, there was a main effect of voice naturalness ($\chi^2(7, N = 31) = 119.65, p < .001$). Individual residual comparisons can be found in Table 3 in the supplementary

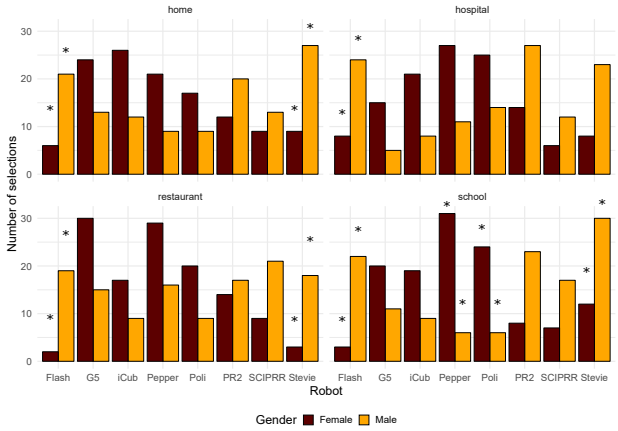


Fig. 4. Robot selection based on voice gender (human-derived voices) in the 4 context conditions. '*' indicates that a robot was selected significantly more or less often upon hearing a certain voice (at the 95% significance level).

materials. As can be seen from Fig. 5, context influenced robot selection based on voice naturalness. iCub was selected more often with a natural voice in the home context; Pepper was selected more often with a natural voice in the hospital and restaurant contexts; Poli was selected more often with a synthetic voice in all 4 contexts; PR2 was selected more often with a synthetic voice in the home and hospital contexts.

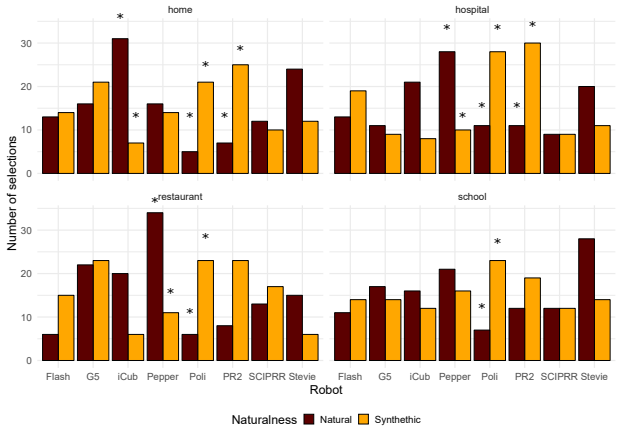


Fig. 5. Robot selection based on voice naturalness (human-derived voices) in the 4 context conditions. '*' indicates that a robot was selected significantly more or less often upon hearing a certain voice (at the 95% significance level).

V. DISCUSSION

We asked participants to match the picture of a robot working either at home, hospital, restaurant, or school, with a voice. The voices were manipulated in terms of gender, naturalness, and accent – features that have previously been examined in studies on the effect of robot voice in HRI [26], [27], [34]. In addition, we used TTS voices that had been used on the featured robots in previous studies (Table I). This experimental design allowed us to see if people formed a mental model of how a robot should look like, based on its voice and its deployment context. We confirmed our previous finding [12] that people choose a robot's voice to match certain characteristics – such as male voices with square-angled robots, and synthetic voice with mechanical-looking robots. In addition, we observed that context influences which robot is chosen for a voice.

In our previous study, we found that people formed consistent associations between robot images and voices based on voice gender [12]. The same strong associations are maintained here (see Supplementary materials for more details), but they are also mediated by context. Specifically, for both the robot voice (RV) and human-derived voice (HV) conditions, the knowledge that the robot works in a certain context dictates whether different associations due to gender emerge or not. For example, in the RV condition, Flash was selected more often with a male voice in the home context, and less often with a female voice in the school context, while voice gender was not a discriminant of how often Flash was selected in the other contexts (Fig. 3). This suggests that, while voice gender might elicit strong first impressions of how a robot should look like in some contexts (e.g. from the Flash example, for working in a home), voice gender might not be as important in other contexts.

The same was true for naturalness and, to a limited extent, accent, in the HV condition. Regarding naturalness, participants also formed consistent associations between robot images and voices based on voice naturalness, the same as our previous study (see supplementary materials for details). Again, in the current experiment we delved more into these associations, and found that they are also mediated by context. For example, iCub was selected more often with a natural voice, but only in the home context; while Poli was selected more often with a synthetic voice in all four contexts (Fig. 5). This suggests that another of our manipulated voice characteristics, naturalness, might induce context-specific associations with a robot appearance. Regarding accent, our results show that only Stevie tended to be selected more often with an American accent, only in the restaurant context. This is surprising, since accents greatly influence impression formation in human-human interaction [32], [48]. However, the phenomenon has not been studied extensively in HRI yet. Also, until not long ago the main concern for designers of robot voices was intelligibility, and TTS systems have only recently started offering high quality accent synthesis (e.g. Cereproc). Thus, it is possible that people are not yet used to thinking of robots as having human accents, resulting in no consistent associations between accent and robot in our sample. It is also possible that accent appropriateness for a robot might manifest itself in terms of stereotypical competence over a specific topic, similar to advertising agencies' use of accents [14]. Finally, our participants came from a wide variety of countries, and it is possible that their own accent of origin (e.g. Irish vs. American) might have influenced their voice-robot pairings. However, with 26 participants from Ireland and 6 from the USA, our participant sample was too small to test this hypothesis. Thus, the concept of accent appropriateness for

a robot warrants further investigation (see also [49] for a recent survey on people's explicit preferences towards a robot accent).

We also found more evidence that previously used voices might not have been a good match for their robots. With the notable exception of PR2, which was consistently selected upon hearing its 'default' voice in both our studies, the other robots were matched less frequently with their 'default' voice. This suggests that these previous matches were ill-chosen, and that greater attention should be placed on robot voice design.

As previous studies have shown, voice characteristics can influence how agents, including robots, are perceived, and they contribute to user behaviour towards these agents [5], [6], [10]. With our results, we add that voice characteristics do not elicit an absolute mental model of how a robot should look like; rather, they interact with the context of the robot's deployment. Thus, when designing a voice for a robot that will be used in a restaurant, researchers should bear in mind that this voice might have to be different than if this robot was working in a school.

Voices are made of a wide variety of features that contribute to creating a mental model of the speaker: gender, age, emotional state, personality, etc. Most studies, including our own, have investigated the effect of certain features as independent factors. However, it is important to note that these characteristics never exist on their own. It is not possible to look at the effect of voice gender *per se*, as this will be mixed with all these other voice characteristics. Perception of an agent based on voice is likely due to an interaction of all these different features, and future studies should also look at these holistically, rather than treat them as independent factors.

Another consideration concerning gender and context is that people matched gendered voices to corresponding gendered physical features (e.g. curved vs. squared shapes) and to stereotypically corresponding occupations (e.g. school and home for Pepper). This confirms the findings of our previous experiment [12] and of previous studies where gender stereotypes were assigned to robots that needed to perform a certain task [50]. As [50] point out, this is a double-edged sword: on the one hand, making robots match gender stereotypes by, for instance, employing female-looking robots for traditionally female tasks seems to result in higher interaction success; on the other hand, it could be an opportunity to contribute to eliminating these stereotypes. In this sense, creating robots that do not match stereotypes will 'force' users in an interaction that is potentially awkward and unsuccessful. It is an ethical judgment that our community should address. Similar considerations have been made for the design of Conversational User Interfaces [36].

Finally, a limitation of this study is the use of pictures of robots, without a live interaction. This is a typical issue for studies that seek to compare multiple robots. Thus, it is important to stress that our results concern first impressions and expectations of a robot, but do not inform about how these first impressions might influence a live interaction with these robots. Several studies in Human-Robot Interaction have highlighted how direct experience of a robot's (mis)behaviour affects trust building when the initial expectations are not met [6], [23]–[25]. The interaction of voice-based first impressions with task type and behavioural experience remains to be explored.

VI. CONCLUSIONS

Many human-robot interactions are speech-based. In this paper we have shown that people have an idea of what an appropriate voice for a robot should be. Specifically, voice gender, naturalness, and accent interact with the human-robot interaction context to inform this appropriateness. However, contrary to the robots depicted in popular culture, real robot voices are often chosen out of convenience, without taking advantage of their full potential [4]. Following the 'matching hypothesis' speculated by [22], we suggest that voice features, robot features, and context interact to form an impression of appropriateness. When one of the variables is unchangeable, for example due to context constraints, the other two can be used to create this impression.

REFERENCES

[1] F. Hegel, F. Eyssel, and B. Wrede, "The social robot 'flobi': Key concepts of industrial design," in *Proceedings of the 19th International Workshop on Robot and Human Interactive Communication*, ser. RO-MAN '10. IEEE, 2010, pp. 107–112.

[2] J. Saldien, K. Goris, S. Yilmazyildiz, W. Verhelst, and D. Lefeber, "On the design of the huggable robot probo," *Journal of Physical Agents*, vol. 2, no. 2, pp. 3–11, 2008.

[3] A. Kalegina, G. Schroeder, A. Allchin, K. Berlin, and M. Cakmak, "Characterizing the design space of rendered robot faces," in *Proceedings of the 13th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18, 2018, pp. 96–104.

[4] S. Wilson and R. K. Moore, "Robot, alien and cartoon voices: Implications for speech-enabled systems," in *Proceedings of the 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*, ser. VIHAR '17, 2017, pp. 40–44.

[5] P. McAleer, A. Todorov, and P. Belin, "How do you say Hello? Personality impressions from brief novel voices." *PLoS ONE*, vol. 9, no. 3, p. e90779, 2014.

[6] I. Torre, J. Goslin, L. White, and D. Zanatto, "Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience," in *Proceedings of APAScience '18: Technology, Mind,and Society (TechMindSociety '18)*.

[7] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, pp. 189–212, 2003.

[8] J. D. M. Laver, "Voice quality and indexical information," *British Journal of Disorders of Communication*, vol. 3, no. 1, pp. 43–54, 1968.

[9] B. L. Brown, W. J. Strong, and A. C. Rencher, "Acoustic determinants of perceptions of personality from speech," *Linguistics*, vol. 13, no. 166, pp. 11–32, 1975.

[10] C. C. Tigue, D. J. Borak, J. J. M. O'Connor, C. Schandl, and D. R. Feinberg, "Voice pitch influences voting behaviour." vol. 33, pp. 210–216.

[11] R. K. Moore, "Appropriate voices for artefacts: Some key insights," in *Proceedings of the 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*, ser. VIHAR '17, 2017.

[12] C. McGinn and I. Torre, "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*, 2019, pp. 211–221.

[13] W. J. Mitchell, K. A. Szerszen, A. S. Lu, P. W. Schermerhorn, M. Scheutz, and K. F. MacDorman, "A mismatch in the human realism of face and voice produces an uncanny valley," *i-Perception*, vol. 2, no. 1, pp. 10–12, 2011.

[14] A. K. Lalwani, M. Lwin, and K. L. Li, "Consumer responses to english accent variations in advertising," *Journal of Global Marketing*, vol. 18, no. 3-4, pp. 143–165, 2005.

[15] A. Prakash and W. A. Rogers, "Why some humanoid faces are perceived more positively than others: Effects of human-likeness and task," *International Journal of Social Robotics*, vol. 7, no. 2, pp. 309–331, 2015.

[16] B. F. Malle, M. Scheutz, J. Forlizzi, and J. Voiklis, "Which robot am i thinking about? the impact of action and appearance on people's evaluations of a moral robot," in *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '16. IEEE, 2016, pp. 125–132.

[17] A. Abubshait and E. Wiese, "You look human, but act like a machine: Agent appearance and behavior modulate different aspects of human–robot interaction," *Frontiers in Psychology*, vol. 8, p. 1393, 2017.

[18] F. Hegel, "Effects of a robot's aesthetic design on the attribution of social capabilities," in *Proceedings of the 21st International Workshop on Robot and Human Interactive Communication*, ser. RO-MAN '12. IEEE, 2012, pp. 469–475.

[19] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust," in *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 141–148.

[20] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '16, 2016, pp. 101–108.

[21] A. Powers, A. D. I. Kramer, S. Lim, J. Kuo, S.-l. Lee, and S. Kiesler, "Eliciting information from people with a gendered humanoid robot," in *Proceedings of the 14th International Workshop on Robot and Human Interactive Communication*, ser. RO-MAN '05. IEEE, 2005, pp. 158–163.

[22] J. Goetz, S. Kiesler, and A. Powers, "Matching robot appearance and behavior to tasks to improve human-robot cooperation," in *Proceedings of the 12th International Workshop on Robot and Human Interactive Communication*, ser. RO-MAN '03. IEEE, 2003, pp. 55–60.

[23] S.-l. Lee, I. Y.-m. Lau, S. Kiesler, and C.-Y. Chiu, "Human mental models of humanoid robots," in *Proceedings of the 2005 IEEE international conference on Robotics and Automation*, ser. ICRA '05. IEEE, 2005, pp. 2767–2772.

[24] R. van den Brule, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, and P. Haselager, "Do robot performance and behavioral style affect human trust?" *International Journal of Social Robotics*, vol. 6, no. 4, pp. 519–531, 2014.

[25] S. Kiesler, "Fostering common ground in human-robot interaction," in *Proceedings of the 14th International Workshop on Robot and Human Interactive Communication*, ser. RO-MAN '05. IEEE, 2005, pp. 729–734.

[26] V. K. Sims, M. G. Chin, H. C. Lum, L. Upham-Ellis, T. Ballion, and N. C. Lagattuta, "Robots' auditory cues are subject to anthropomorphism," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 53, no. 18. SAGE Publications, 2009, pp. 1418–1421.

[27] A. Sandygulova and G. M. P. O'Hare, "Children's perception of synthesized voice: Robot's gender, age and accent," in *Social Robotics*, A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi, Eds. Springer International Publishing, 2015, pp. 594–602.

[28] C. R. Crowell, M. Scheutz, P. Schermerhorn, and M. Villano, "Gendered voice and robot entities: perceptions and reactions of male and female subjects," in *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, ser. IROS '09. IEEE, 2009, pp. 3735–3741.

[29] D. Bryant, J. Borenstein, and A. Howard, "Why should we gender? the effect of robot gendering and occupational stereotypes on human trust and perceived competency," in *Proceedings of the 15th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '20, 2020, pp. 13–21.

[30] M. L. Walters, D. S. Syrdal, K. L. Koay, K. Dautenhahn, and R. Te Boekhorst, "Human approach distances to a mechanical-looking robot with different robot voice styles," in *Proceedings of the 17th International Workshop on Robot and Human Interactive Communication*, ser. RO-MAN '08. IEEE, pp. 707–712.

[31] R. Tamagawa, C. I. Watson, I. H. Kuo, B. A. MacDonald, and E. Broadbent, "The effects of synthesized voice accents on user perceptions of robots," *International Journal of Social Robotics*, vol. 3, no. 3, pp. 253–262, 2011.

[32] P. E. G. Bestelmeyer, P. Belin, and D. R. Ladd, "A neural marker for social bias toward in-group accents," *Cerebral Cortex*, vol. 25, no. 10, pp. 3953–3961, 2014.

[33] D. Kuchenbrandt, F. Eyssel, S. Bobinger, and M. Neufeld, "When a robot's group membership matters," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 409–417, 2013.

[34] S. Andrist, M. Ziadee, H. Boukaram, B. Mutlu, and M. Sakr, "Effects of culture on the credibility of robot speech," in *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '15, ACM. ACM Press, pp. 157–164.

[35] S. J. Sutton, P. Foulkes, D. Kirk, and S. Lawson, "Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19, 2019, pp. 1–14.

[36] J. Cambre and C. Kulkarni, "One voice fits all? social implications and research challenges of designing voices for smart devices," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–19, 2019.

[37] L. Clark, P. Doyle, D. Garaialde, E. Gilmartin, S. Schlögl, J. Edlund, M. Aylett, J. Cabral, C. Munteanu, and B. R. Edwards, Justin Cowan, "The state of speech in hci: Trends, themes and challenges," *Interacting with Computers*, vol. 31, no. 4, pp. 349–371, 2019.

[38] C.-J. Lai and C.-P. Tsai, "Design of introducing service robot into catering services," in *Proceedings of the 2018 International Conference on Service Robotics Technologies*, 2018, pp. 62–66.

[39] F. Tanaka, K. Isshiki, F. Takahashi, M. Uekusa, R. Sei, and K. Hayashi, "Pepper learns together with children: Development of an educational application," in *Proceedings of the 15th IEEE-RAS International Conference on Humanoid Robots*, ser. Humanoids '15. IEEE, 2015, pp. 270–275.

[40] T. Tanioka, "Nursing and rehabilitative care of the elderly using humanoid robots," *The Journal of Medical Investigation*, vol. 66, no. 1.2, pp. 19–23, 2019.

[41] A. K. Pandey and R. Gelin, "A mass-produced sociable humanoid robot: Pepper: The first machine of its kind," *IEEE Robotics & Automation Magazine*, vol. 25, no. 3, pp. 40–48, 2018.

[42] H. Hastie, K. Lohan, A. Deshmukh, F. Broz, and R. Aylett, "The interaction between voice and appearance in the embodiment of a robot tutor," in *International Conference on Social Robotics*. Springer, 2017, pp. 64–74.

[43] D. Zanatto, M. Patacchiola, J. Goslin, and A. Cangelosi, "Priming anthropomorphism: Can the credibility of humanlike robots be transferred to non-humanlike robots?" in *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '16, 2016, pp. 543–544.

[44] E. S. Short, M. L. Chang, and A. Thomaz, "Detecting contingency for hri in open-world environments," in *Proceedings of the 13th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18, 2018, pp. 425–433.

[45] M. Cakmak and L. Takayama, "Teaching people how to teach robots: The effect of instructional materials and dialog design," in *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '14, 2014, pp. 431–438.

[46] A. M. Harrison, W. M. Xu, and J. G. Trafton, "User-centered robot head design: A sensing computing interaction platform for robotics research (sciprr)," in *Proceedings of the 13th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18, 2018, pp. 215–223.

[47] C. McGinn, E. Bourke, A. Murtagh, M. F. Cullinan, and K. Kelly, "Exploring the application of design thinking to the development of service robot technology," in *ICRA2018 Workshop on Elderly Care Robotics-Technology and Ethics (WELCARO)*, 2018.

[48] J. N. Fuertes, W. H. Gottdiener, H. Martin, T. C. Gilbert, and H. Giles, vol. 42, no. 1, pp. 120–133.

[49] I. Torre and S. L. Maguer, "Should robots have accents?" in *Proceedings of the 29th International Workshop on Robot and Human Interactive Communication*, ser. RO-MAN '20. IEEE, 2020.

[50] B. Tay, Y. Jung, and T. Park, "When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction," *Computers in Human Behavior*, vol. 38, pp. 75–84, 2014.