DiVA
http://www.diva-portal.org

Postprint

This is the accepted version of a paper presented at *HRI 2019*.

Citation for the original published paper:

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: *14th ACM/IEEE International Conference on Human-Robot Interaction, HRI 2019, Daegu, South Korea, March 11-14, 2019* (pp. 211-221).
https://doi.org/10.1109/HRI.2019.8673305

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-270674

Can you tell the robot by the voice? An exploratory study on the role of voice in the perception of robots

Conor McGinn
School of Engineering
Trinity College Dublin
Dublin, Ireland
Email: mcginnco@tcd.ie

Ilaria Torre
School of Engineering
Trinity College Dublin
Dublin, Ireland
Email: torrei@tcd.ie

***Abstract*—It is well established that a robot's visual appearance plays a significant role in how it is perceived. Considerable time and resources are usually dedicated to help ensure that the visual aesthetics of social robots are pleasing to users and helps facilitate clear communication. However, relatively little consideration is given to how the voice of the robot should sound, which may have adverse effects on acceptance and clarity of communication. In this study, we explore the mental images people form when they hear robots speaking. In our experiment, participants listened to several voices, and for each voice they were asked to choose a robot, from a selection of eight commonly used social robot platforms, that was best suited to have that voice. The voices were manipulated in terms of naturalness, gender, and accent. Results showed that a) participants seldom matched robots with the voices that were used in previous HRI studies, b) the gender and naturalness vocal manipulations strongly affected participants' selection, and c) the linguistic content of the utterances spoken by the voices does not affect people's selection. This finding suggests that people associate voices with robot pictures, even when the content of spoken utterances was unintelligible. Our findings indicate that both a robot's voice and its appearance contribute to robot perception. Thus, giving a mismatched voice to a robot might introduce a confounding effect in HRI studies. We therefore suggest that voice design should be considered more thoroughly when planning spoken human-robot interactions.**

***Index Terms*—Robot design; Voice; Speech; Mental model**

I. Introduction

To provide greatest utility to human users, autonomous service robots must be capable of performing useful tasks and adapting to existing, human-oriented environments. By extension, it is desirable that they are easy to communicate with and can conform to established social norms and behaviours.

Since spoken language plays a critical role in many HRI scenarios, it is desirable that robots too can communicate through this medium. However, despite the important role it plays, robot designers and HRI practitioners have tended to pay relatively little consideration to the voices used on robots in HRI experiments. To test this assertion, we contacted a sample of first authors from papers involving spoken interaction that were published at last year's HRI conference. Eighteen of the twenty-five contacted authors responded, and we learnt that: six had used the NAO robot built-in voice; seven had used a voice generated with a Text-To-Speech (TTS) system, with the motivation that it was freely available or that it was the voice that the robot came with; three had pre-recorded the voice using actors; and two provided the description of how the voice sounded like (e.g. "androgynous, child-like voice"). One author additionally mentioned robot voice accent, and one other mentioned looking for a voice that would specifically suit the task that the robot had to carry out in the experiment. Regarding the reason behind their choice of voice, two authors specified that: "it was the only good one" and: "it was open source". Eleven authors mentioned the voice gender aspect, and five mentioned other voice characteristics such as pitch or speech rate. Thus, it seems that a unifying, agreed-upon standard criterion or choosing a robot voice is lacking, and that in many cases robot voices are chosen out of convenience.

Indeed, human voices contain a rich amount of information. In addition to communicating explicit speech content, they provide information pertaining to the background of the speaker (i.e. age, nationality, gender), affective state, and identity [4]. Research has also revealed that voices can provide clues on factors pertaining to the speaker's perceived attractiveness [17], personality [56], sexual orientation [50], intelligence [12], and health [44].

Apart from conveying traits such as gender or personality, hearing someone's voice for the first time can also make us form a mental image of how that person might look like. For example, it is a common experience to observe "a speaker whose voice is familiar ... and being surprised by that person's appearance" [22]. This confusion, caused by the mismatch between expectation and reality, may adversely effect the ability to for people to establish common ground with robots since, according to Kiesler, a key ingredient for achieving common ground is "to create in people's minds an appropriate

Ilaria Torre is 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.

doi:10.1109/HRI.2019.8673279 $33.00 copyright 2019 IEEE

mental model of the robot automatically" [21].

The aim of this paper is to investigate whether voices have a tangible effect on the expectations people form of what a robot will look like. We hypothesise that, when asked to identify the most suitable robots given a range of voices, factors such as voice gender, degree of naturalness, and accent will influence people's choices. We aim to further examine whether voices previously used on eight different robots, gathered from recently published HRI studies, could be considered 'good voices', namely whether naive participants would be able to correctly identify the same robot from the voice in a blind test.

In the following section, relevant prior art is discussed. The methodology used in the study is outlined in section III. Results and key findings of the study are presented in section IV, and discussed in section V. Final remarks and future work are outlined in the concluding section.

II. Related Work

If people attribute traits and intentions to robots as if they were humans [cf. 33, 32], then it is likely that the design of a robot's appearance and voice should go hand in hand to create a cohesive image of this agent. Issues to take into consideration, for example, could be: should more human-like robots have natural human voices? Should robots have gendered voices? Should large robots have lower-pitched voices than small robots, congruent with a bigger vocal tract [35]? Should robot voices have regional accents? Should the voice be indicative of the robot's job?

Regarding naturalness, researchers seem not to have reached an agreement on whether robots should have natural or mechanical-sounding voices. For example, Sims et al. showed that being able to speak (either with a synthetic or a natural voice) was enough for a robot to be treated as a competent agent [43]. On the other hand, Mitchell et al. [29] argue that incongruence in the human-likeness of a character's face and voice can elicit feelings of eeriness, while Tamagawa et al. argue that, for the sake of clarity and familiarity, people would prefer to have this incongruency [46]. Torre et al. [49] also showed that the task context might interact with the naturalness of the voice, since people trusted a robot with a synthetic voice more than a robot with a natural voice when the robot was behaving trustworthily, but the opposite when the robot was behaving untrustworthily. Furthermore, the natural/synthetic dichotomy is arguably not made of discrete steps, but rather it is a continuum. For example, a synthetic voice produced by a certain TTS system will have a different quality than one produced by another, yet such nuances are often lost within the 'synthetic' voice category.

Regarding gender, given the human tendency to anthropomorphise, it is likely that features that are socially linked to a gender might be found in robots, too. If this is the case, gender stereotypes might play a role in human-robot interactions [cf. 32, 47], and robot designers might want to use this knowledge to create specific robots. However, previous studies on robot voice gender are in disagreement. For example, Walters et al. found that the gender of the robot voice did not influence participants' gender attribution of the robot [53]; on the contrary, Siegel et al. characterised the gender of a robot solely based on a pre-recorded voice [42]. Similarly, Sandygulova and O'Hare showed that children assigned a gender to a NAO robot on the basis of the voice alone – which was a synthetic male or female voice [39]. However, in this experiment participants heard all the possible voices on the robot, one after the other, so it is possible that a priming effect influenced these results. Eyssel et al. also found that people perceived robots more positively (e.g. in terms of psychological closeness and anthropomorphism) when the perceived gender of the robot matched that of the person interacting with it [13].

Very few studies experimentally manipulated a robot's voice to have different accents. One such study is Tamagawa et al's, where participants felt more positive emotions with a robot speaking with a New Zealand accent than a US accent; they also thought that the New Zealand-accented robot performed better in the explanation of a medical procedure [46]. Preliminary research from Andrist et al. on the Arabic language suggests that accent and context also interact in human-robot interactions: participants believed that robots with a regional accent were more credible when they were knowledgeable, whereas robots with a standard accent were more credible when they had little knowledge [1]. This facet of human speech has been little studied in HRI, but the few previous studies suggest that accent might be important in robot perception as well, and that certain robots might be more suited to certain accents.

Other vocal characteristics have been scarcely studied. For example, Niculescu et al. found that varying the pitch of a robot's voice had a strong influence on the way users rated the overall interaction quality, and encouraged further research on the topic [34]. Expressing emotions in the voice might also be beneficial in some human-robot interactions: Leyzberg et al. [25] and Tielman et al. [48] found that people tended to co-operate more with robots that expressed themselves using appropriate emotional verbal feedback.

A few recent studies have also investigated the design of non-linguistic sounds in HRI [30, 27, 10]. However, the amount of information that can be communicated by the non-linguistic utterances is substantially limited, and according to Crumpton and Bethel, "there has been little research into how manipulating a robots voice would affect its users" [11]. Thus, there remains a significant need to advance research in this area.

Research has shown that people are often able to successfully identify the visual appearance of a human speaker from voice cues alone [45]. From this work it seems that the consensus among experts is that best performance is attained using dynamic, rather than static, reference stimuli. Still, recent research using deep learning approaches have demonstrated, at significantly above chance levels, that CNNs can be successfully trained to correctly identify speakers based solely on recorded voice cues, using both photo (i.e. static) as well as video (i.e. dynamic) stimulus [31]. These findings

indicate that there is frequently an observable agreement between a person's appearance and their voice. While previous studies have indicated that a robot's voice and appearance can influence the mental model that people form [36, 26, 52], as of yet no equivalent tests have been reported which have investigated if people can identity robots from voice cues alone. The answer to this question remains unknown and non-trivial, since robots can come in a wide variety of forms and do not typically generate sound in a manner that provides paralinguistic cues (i.e. coordinated movement of lips).

In this study, we hope to clarify whether people do form a mental image of a robot upon hearing its voice for the first time. We also hope to determine whether this voice-appearance association is consistent with voices previously used in HRI studies, and whether it can be pinned down to the selected vocal features of naturalness, gender, and accent.

III. Methodology

A. Stimuli preparation

The experiment consisted of three between-subject tests: test A, test B, and test C (see Section III-C). In test A, participants listened to voices uttering 14 sentences, which were previously validated as neutral in content [37]. These sentences are listed in table I. In test B, participants listened to voices uttering the same 14 sentences, but played backwards, so as to eliminate their linguistic content, while maintaining indexical information such as gender and voice naturalness. Finally, in test C, participants listened to voices uttering a dialogue between a robot and a human user. The robot's script was played through audio, while the human's script appeared in writing on the screen. The dialogue was designed to indicate a human-robot interaction plausible in many different contexts. The dialogue is transcribed in table II; as it is possible to see, the dialogue is made up of 3 utterances spoken by the robot.

These tests were designed to examine if the mental image formed as a result of exposure to neutral verbal sentences (i.e. test A) was robust to linguistic information in the speech (which was removed in test B through reversal of sentences in test A), and to the addition of a basic task context (which was added through the dialogue-based interaction presented in test C).

TABLE I
NEUTRAL SENTENCES USED IN TESTS A AND B

Index	Sentence
01	"I'm on my way to the meeting."
02	"I would like a new alarm clock."
03	"I think I've seen this before."
04	"I wonder what that is about."
05	"Have you seen him?"
06	"The airplane is almost full."
07	"Can you hear me?"
08	"Maybe tomorrow it will be cold."
09	"Can you call me tomorrow?"
10	"I think I have a doctor's appointment."
11	"We'll stop in a couple of minutes."
12	"How did he know that?"
13	"Don't forget a jacket."
14	"The surface is slick."

TABLE II
SCRIPTED DIALOGUE BETWEEN ROBOT AND HUMAN IN TEST C.

Speaker	Sentence
Robot	"Hello, sorry to bother you."
Human	"Hi, thats no problem. Is everything ok?"
Robot	"My software needs an update. I just wanted to let you know that I need to be offline for a short period."
Human	"Sure."
Robot	"I will get back to work in around five minutes."

1) Voice stimuli: The voice stimuli were obtained from two different sources: natural recordings of several English speakers, and synthetic speech samples from TTS systems used in previous HRI research.

For the natural voices, we recorded one male and one female speaker of the following accents of English: Irish (Dublin), American (California), and Italian (Rome). Speakers from the first two countries spoke English natively, while the Italian speakers were fluent, as measured by the fact that they had been living in an English-speaking country for several years. These accents were chosen in order to have a more global overview of robot-voice suitability. In particular, we chose Irish as a local, native English accent – as the experiment was carried out in Ireland; American as a global, native English accent; and Italian as a global, non-native English accent. While studies on accent attitudes in HRI are scarce, the sociolinguistics literature suggests that people treat such different accents differently [e.g. 24]. In this way, we produced stimuli pertaining to the accent and gender categories that we wanted to examine.

The six speakers were recorded in a quiet room using a Zoom H6 Handy Recorder and an AKG C520L condenser microphone, where they read the aforementioned neutral and dialogue sentences, for a total of 17 sentences. All the recordings were then cleaned with a noise-removal filter in Audacity.

These natural voices were also re-synthesized to obtain 'mechanical' sounding voices, while retaining individual speaker characteristics such as accent and gender. To obtain this effect, we first flattened the fundamental frequency (f_0) of each speaker to that speaker's mean f_0 value, and then applied a comb filter using Audacity (comb frequency = same value as the flattened f_0 of the sound file; comb decay = 0.1; normalization level = 0.990). The resulting re-synthesised voices were 'monotonous' and had a metallic flare; this allowed us to have a machine-sounding version of the natural voices, while retaining many individual voice characteristics of the speakers constant.

For the 'default' robot voices – hereby belonging to the 'synthetic' experimental condition, as opposed to 'natural' and 'resynthesised' – we identified prominent research groups

TABLE III
ROBOTS AND CORRESPONDING VOICES USED IN THE ANALYSIS.

Robot	Speech Engine	Voice Name	Reference Study
Flash	CereProc	Heather	[16]
G5	Acapela	Rod[2]	[54]
iCub	Acapela	Rod[2]	[54]
Softbank Pepper	Pepper	Default	Developer
Poli	Amazon	Kim	[41]
PR2	Cepstral	David	[9]
HUBO-SCIPRR	Cepstral	Alison	[15]
Stevie	CereProc	Giles	[28]
PAL REEM	Acapela	Rachel	[51]
ROS default	Festival	Kal	Developer

which had a track record[1] of conducting research involving the use of voice on socially interactive service robots. These groups provided, or gave the necessary information to obtain, a recording of the 17 synthetic sentences necessary to recreate the voices previously used in HRI research. These pairings of robot and voice are listed in table III.

Thus, our robot voice corpus consisted of 21 voices: 6 natural recordings, 6 re-synthesised versions of these recordings, and 9 voices used in previous HRI studies.

Once all the voices were obtained, the 14 neutral utterances (from the natural, re-synthesized and synthetic conditions) were reversed using Audacity, resulting in sound files where the linguistic content was unintelligible. This process of reversing the utterances changes some acoustic features such as pitch contour, but other features such as gender and naturalness are retained. This resulting 'backwards' condition was meant to see whether people discriminate different robots on the basis of how the voice sounds alone, regardless of linguistic content.

The resulting voice samples had some inherent audio differences (such as reverberation and amplitude), having been obtained in different recording environments. Therefore, to make them sound alike and mask compression artefacts, they were remixed with a 0.06 dB brown noise, amplitude-normalised to -0.03 dB, and re-sampled to 44.1 kHz. Since the voice recordings came from several different sources, the resampling was done to ensure that the final stimuli all had the same sampling rate. This is standard practice in signal processing.

2) Image stimuli: We collected high quality photographs of the following eight robots: *iCub*, *Pepper*, *Poli*, *PR2*, *HUBO* (with a *SCIPRR* mounted head), *Stevie*, *Flash*, *G5*. These robots were chosen because they represented a diverse sample of widely used service robot technology (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, of which we knew the details (see table III). The pictures showed all the main features of the robots (head, arms, torso, legs/wheels, etc.). Each robot was positioned on a neutral grey background, and the images of the robots were scaled to fit a 768x950 pixel window, as shown in Figure 1.

[1]Here, the inclusion criterion was groups that had either recently designed bespoke service robot platforms capable of spoken interaction, or had published papers relating to service robots with spoken interaction at major venues such as HRI, ICSR, IJSR, etc. within the last 5 years.

[2]Note how here the same voice was used for G5 and iCub; therefore, in the experiment we considered this voice to be the default for both robots.

B. Participants

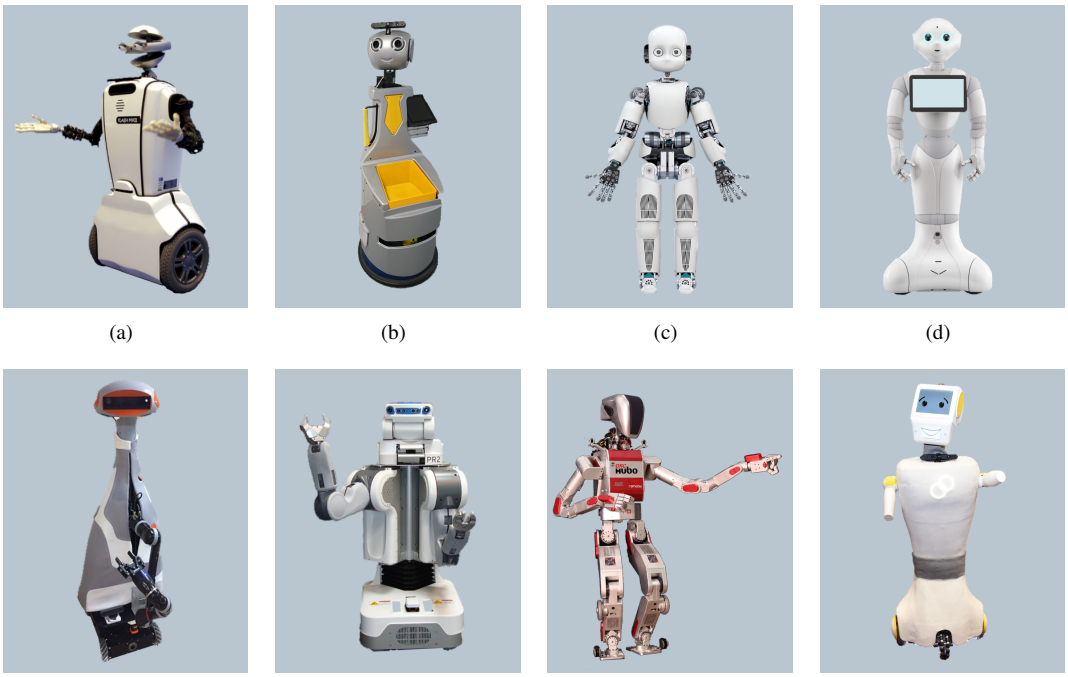
We tested 90 participants (30 in test A, 30 in test B, and 30 in test C), who were either staff/students at the university, or visitors of a nearby science museum. There were 38 females and 51 males (and one person preferred not to disclose this information); they were aged 18-72 (mean = 29.76, sd = 10.62). Their self-reported English language fluency was as follows: 56 native speakers; 2 native-like; 27 fluent; 5 basic. While the majority (n = 49) of people were from the Republic of Ireland, the remaining 41 were from 19 other countries. Finally, we asked all participants if they were familiar with any real-life robots, and to what extent. While a few people mentioned the robot *Stevie* – which was expected since this robot was developed in the same university where the study was conducted – only one person mentioned familiarity with another of the robots included in the study (*Pepper*).

C. Procedure

Participants were assigned to one of the three aforementioned test conditions, in a counterbalanced fashion. To reduce any confounding factors that may have been caused by prior familiarity with the robots, participants who declared familiarity with one or more robots included in the study were assigned to the B test condition, where the voice samples were reversed and the linguistic content was unintelligible.

All experiments were conducted in a quiet room, 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 positioned at a computer desk wearing good quality over-ear headphones, and ran one practice trial with the experimenter. The practice trial consisted of hearing a few utterances from the Star Wars *R2D2* robot, and then participants were asked to select the most suitable robot from a selection of eight images depicting eight well known robots from popular culture (including *R2D2*, *C3P0*, and *Baymax*).

After the practice trial, participants were left with the actual experiment, which consisted of 21 trials (one for every collected voice). Depending on the experimental condition, participants were presented with voices (in random order) from either the neutral sentences (test A), the neutral backwards sentences (test B), or the scripted dialogue (test C). Every trial proceeded as follows: for each voice sample, participants listened to 3 randomly selected sentences (in test A and B), or to the scripted dialogue (in test C, which also consisted of 3 spoken sentences), while a fixation cross appeared in the centre of the screen. After the voice sequence had terminated, they were shown the 8 robot pictures, equally positioned on



(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*.

the screen, and were asked to select (by mouse click) the picture of the robot that best suited the voice they just heard. The relative location of each robot was fixed throughout the experiment. A "restart" button was placed adjacent to the robot pictures, in case participants wanted to hear the voice again, or change their selection. The whole session lasted approximately 15 minutes.

IV. Results

We conducted chi-square tests for independence on each of the variables of interest – gender, naturalness, accent – to see whether there was a causal relationship between the robots being selected and the variable. 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 chi-square residuals for each robot against a critical z value and adjusting the alpha level for multiple comparisons (Bonferroni correction). These post-hoc analyses were conducted for the overall robot selection as well as for the selection in the 3 test conditions. The full contingency tables can be found in the auxiliary materials.

A. Robot default voice

We looked at whether people associated a robot with its 'default' voice (based on data presented in Table III). Fig. 2 illustrates how often the voice used in the previous studies was matched with the robot it was deployed on, for each of the three test conditions. The results show that only *PR2* was consistently selected upon hearing its default voice, across all test conditions; *iCub* was selected above chance level in the backwards and neutral conditions, and *Stevie* in the dialogue and neutral conditions; *G5* and *SCIPRR* were selected above chance level only in the neutral condition, and *Poli* only in the dialogue condition.

As mentioned previously, participants who may have had prior familiarity with *Stevie* were assigned to test B to avoid potential confounding effects. We still checked whether having prior exposure to Stevie introduced a confound into the results. This analysis revealed that only one of the seven participants in test B, who had had prior experience with Stevie, selected it as the most suitable upon hearing its 'default' voice. Even by eliminating this participant, the total number of participants who 'correctly' matched the voice to the robot remains above chance level.

B. Voice gender

Fig. 3 shows the frequency a certain robot was selected based on the gender of the voice being played. We categorised the voices as either male, female, or ambiguous: 11 voices were male (3 natural voices, 3 resynthesised voices, *G5*, *iCub*, *Stevie*, *PR2*, *ROS*), 10 were female (3 natural voices, 3 resynthesised voices, *Poli*, *SCIPRR*, *Flash*, *Reem*) and one was ambiguous (Pepper). The reason we categorised the default *Pepper* voice as ambiguous is that its manufacturers specified that Pepper is neither male nor female[3]; furthermore, this is the default voice also for another robot built by the same company – *NAO*. Given the widespread use in research of both *NAO* and *Pepper* with their – identical – default voice, we thought it best

[3]SoftBank robotics guidelines, last accessed on 01/01/2019

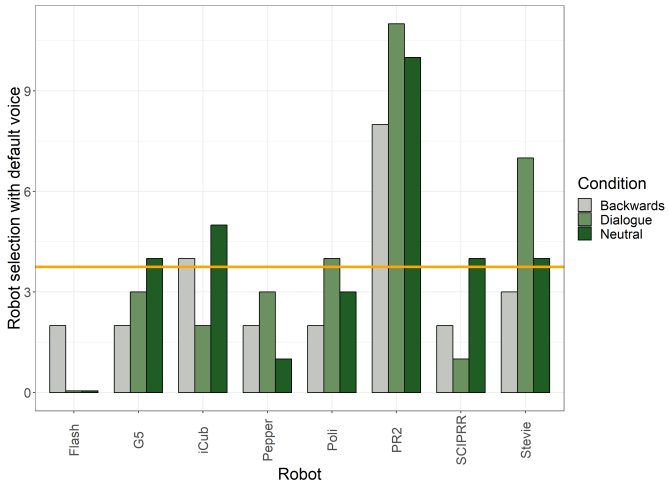


Fig. 2. Number of people who selected a robot upon hearing its "default" voice in the 3 test conditions, assuming independence between trials. The orange horizontal line indicates the 12.5% chance level of selecting the 'correct' robot at each trial.

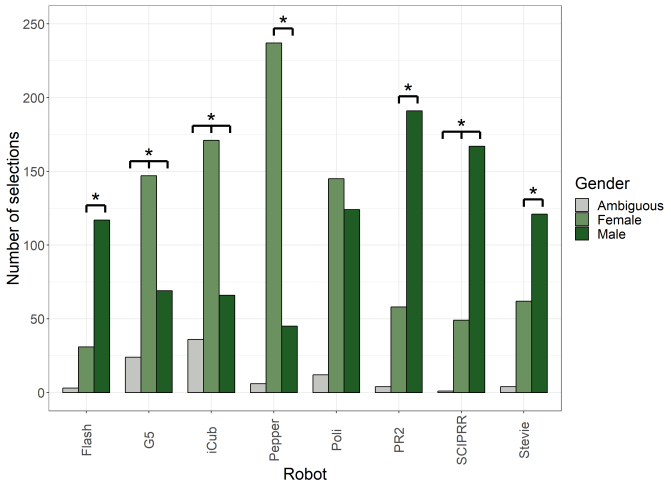


Fig. 3. Robot selection based on voice gender across the 3 test conditions. '*' indicates statistical significance at the 95% level.

to not categorise this voice as either male or female, in this case.

A chi-square test of independence was performed on the whole contingency table (see Table I in the auxiliary materials) and found a significant association between gender and the robot being selected, $\chi^2(14, N = 90) = 498.68, p < .001$. Individual residual comparisons can be found in Table I in the auxiliary materials.

As it is possible to see, people seem to have a quite clear idea of which robots should have a female or a male voice: *Flash*, *PR2*, *HUBO-SCIPRR* and *Stevie* were mostly associated with a male voice; *G5*, *iCub* and *Pepper* were mostly associated with a female voice; and *Poli* seems to be the most ambiguous robot to classify. As can be seen in Table I in the auxiliary materials, these associations are mostly robust to test manipulation. Talking informally with participants after they finished the task, many remarked how they had noticed they were assigning female voices to robots with round shapes or round eyes, and male voices to square-looking robots.

C. Voice naturalness

The frequency people selected robots based on the naturalness of the voice they heard is shown in Fig. 4. For the purpose of this analysis, we excluded the synthetic robot voices taken from past HRI studies, and examined differences between the natural recordings and their analogous resynthesised version. This was done because the voices used in previous studies arguably differ among themselves in terms of synthetic quality, as discussed in Section II.

A chi-square test of independence revealed a significant relationship between the voice naturalness and the robot being selected, $\chi^2(7, N = 90) = 153.02, p < .001$. Individual residual comparisons can be found in Table II in the auxiliary materials.

Many participants informally remarked they had associated the natural voices with the 'shinier', more natural-looking robots, and indeed robots such as *Pepper* and *iCub* are consistently selected in conjunction with natural voices, across all 3 experimental conditions, while *Poli*, *PR2*, and *SCIPRR* are consistently associated with the resynthesised voices. *Stevie* is also most often associated with a natural voice, while people are not consistent in choosing a natural or synthetic voice for *Flash* and *G5*. As can be seen from Table II in the auxiliart materials, these associations mostly remain when broken down into the 3 test conditions.

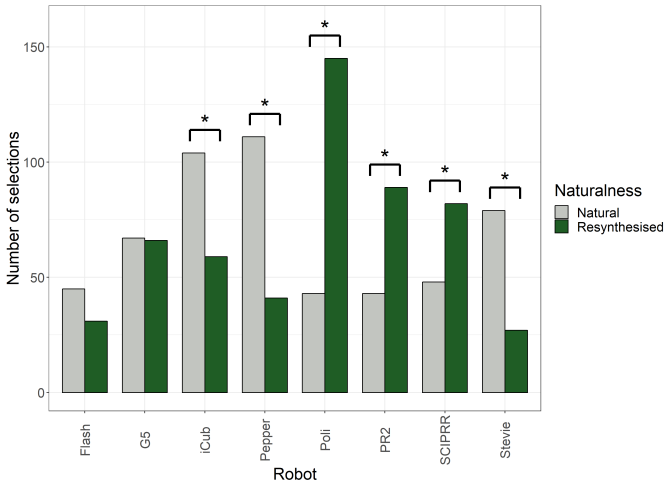


Fig. 4. Robot selection based on voice naturalness across the 3 test conditions. '*' indicates statistical significance at the 95% level.

D. Voice accent

Fig. 5 shows the robots that were selected when a voice with an accent manipulation was played. Although some of the default robot voices had a specific accent (e.g. Flash in [16] has a Scottish female voice), here we only report people's selection

upon hearing the voices we recorded or resynthesised, which had either an American, Irish, or Italian accent of English.

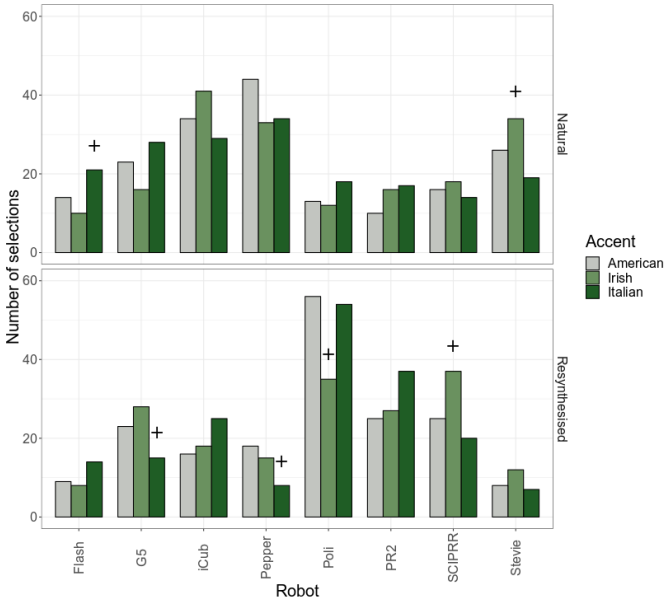


Fig. 5. Robot selection based on natural (top) and resynthesised (bottom) accents. '+' indicates that the residual test approached significance at the 95% level.

Since the vocal resynthesis process might have altered some of the nuances of the various accents, we performed chi-square tests of independence for the natural and resynthesised accents separately. The test on the natural accents revealed no significant association between accent and robot being selected ($\chi^2(14, N = 90) = 19.70, p = .14$), suggesting that people did not consistently associate natural accents with specific robots. Instead, the test on the synthetic accent revealed a significant association ($\chi^2(14, N = 90) = 27.52, p = .02$). However, the necessary adjustments for multiple comparisons made the post-hoc residual analyses fall below the corrected alpha level. Individual comparisons can still be seen in Tables III and IV in the auxiliary materials for full information, and the comparisons that approached significance under this conservative approach are marked in Fig. 5.

E. Test condition

Finally, we looked at whether a certain robot was selected a different number of times in the 3 test conditions (neutral sentences, backwards sentences, dialogue). A chi-square test of independence revealed a significant relationship between the test condition and the robot being selected, $\chi^2(14, N = 90) = 28.16, p = .01$. Individual residual comparisons can be found in Table V in the auxiliary materials. The only statistically significant post-hoc test was found for *Stevie*, being selected fewer times than expected in the backwards condition. As this was the condition where all the participants familiar with this robot were placed, this result interestingly suggest that removing linguistic content from a seemingly familiar voice might hinder recognition of that voice.

V. Discussion

In this study, we examined whether people form a mental image of a robot upon hearing its voice for the first time. People heard a range of voices, differing in terms of gender, naturalness, and accent. For each voice, they were tasked to select the image of a robot that best suited the voice they just heard.

Looking at how people associated voices that were previously used in HRI studies in conjunction with the eight robots employed here, we found that participants rarely matched the voice with the corresponding robot used in those studies. The notable exception to this was for *PR2*, which was matched reliably across all 3 conditions. It seems that the vast majority of participants found its voice, as used in [9] (Cepstral: David), to be the best fit for that robot. Reasonable performance was observed for *Stevie*, as used in [28] (CereProc: Giles), and *iCub*, as used in [54] (Acapela: Rod), which exceeded chance level on two occasions. For *Stevie*, this partly validates previous work on the design of this robot, which included a user experience study to determine the voice that would suit it best [28]. Interestingly, only very few participants matched the voice delivered by the robot manufacturer SoftBank Robotics with the corresponding *Pepper* robot. These results suggest that this default voice might not be the optimal solution. Even more, the mismatch between voice and appearance might introduce artefacts in human-robot interactions, which may have tangible effects in research and real-world applications involving the widely used *Pepper* robot.

We also found that female voices were more often associated with the *G5*, *iCub*, and *Pepper* robots, and male voices with the *PR2*, *HUBO-SCIPRR* and *Stevie* robots. We speculate that this might be because the appearance of *PR2* and *HUBO-SCIPRR* is rather mechanical, with joints and metallic parts showing, and human social constructs make it so that technical-looking robots could be assigned a male gender [6]. On the other hand, *Pepper* and *G5* appear to be smiling, and positive emotional displays, according to other social constructs, are a feminine quality [8]. Other characteristics could also be playing a role: for example, round objects are often associated with feminine traits, and edgy objects with masculine traits [20, 5]. The results partially support this; *G5*, *iCub*, and *Pepper* have round heads, and were mostly associated with a female voice, while *PR2*, *SCIPRR*, and *Stevie* have squared heads, and were mostly associated with a male voice. However, these explanations do not tell the full story, since *Stevie* also appears to be smiling, and *Poli*, which was not consistently associated with any gendered voice, has a round head. Finally, the strong association between *iCub* and female voices could relate to the manufacturers' intention of mimicking a humanoid child. Since children and women have generally higher vocal pitch than men [35], people might have chosen the female voice for the lack of a more appropriate childlike voice.

People consistently associated a natural voice with *iCub*, *Pepper*, and *Stevie*, and a resynthesised voice with *Poli*, *PR2*,

and *HUBO-SCIPRR*. It might be that people associated some humanlike characteristics of the first three robots, such as the fact that they have a facial expression, with more human-sounding voices. On the other hand, *Poli*, *PR2*, and *HUBO-SCIPRR* appear to be merely socially evocative, in accordance with the definition by Breazeal [7]. As a result, their more functional features might have resulted in them looking more 'mechanical', hence the association with a resynthesised voice. Interestingly, *G5* was not consistently associated either with a natural or a resynthesised voice. Potentially, some people considered its facial expression as the humanlike feature upon which to base the association with a natural voice, whereas other people could have considered its lack of arms and legs a sign that a non-humanlike voice was more appropriate. The case of *G5* highlights how individual differences in robot perception are also likely to play a role in the design of appropriate robot voices [see e.g. 55], and should be addressed more thoroughly in the future.

Regarding voice accents, results were less consistent than for the other vocal features we examined. This might be because we tested participants from a wide range of language and cultural backgrounds. However, research has shown that stereotypes based on global English accents persist beyond geographic borders [23, 3], and given that the majority of participants reported being at least fluent in the English language, we can assume that most of them heard the systematic differences in pronunciation afforded by the three accents we examined.

Although (conservative) statistical significance was not reached, there was a trend for *Poli* to be often associated with the American and Italian accents, especially in the resynthesised voice condition, and *Pepper* with American accents, especially in the natural condition. *Flash* was also often associated with Italian accents, especially in the natural condition (Fig. 5). As many of the sci-fi media is produced in the USA, it is possible that people associated some of the more product-like robots, such as *Pepper* or *Poli*, with American accents. Also, as Italians and Americans are sometimes considered very expressive communicators, it is possible that people associated these accents with robots that seemed to be communicating with gestures, such as *Flash*, *Poli*, or *PR2*. There was also a trend for *Stevie* to be often associated with the Irish accent, especially in the natural voice condition. This further supports the previous work on the design of this robot [28]. As can be seen from Table III in the auxiliary materials, this association approaches significance only in the neutral sentence test condition, suggesting that an interaction between voice and context might be at play. Furthermore, it is possible that accent stereotypes might emerge during the course of an interaction, e.g. when a robot's behaviour becomes apparent [cf. 49]. More research is needed to reveal the unconscious stereotypes that we might be forming when interacting with accented robots.

Finally, concerning the type of communication context – scripted dialogue, neutral sentences, sentences played backwards – it seems that this manipulation had little influence on participants' selections in the experiment. For the default robot voices, the neutral sentences condition generally increased performance, with people choosing five 'correct' robots over chance level, rather than just two in the backwards condition and three in the dialogue condition. In general, we found that the voice-looks associations were consistent across the 3 test types, even in the case where linguistic content was removed – with the notable exception of the *Stevie* robot. This suggests that the sound of a voice alone is enough to make us form a mental image of how that speaker – in our case, a robot – should look like.

VI. Conclusion

In this paper, we examined the relation between robot voice and robot appearance, specifically whether people consistently form a mental image of a robot after hearing its voice for the first time. The methods adopted in this study reflect the exploratory nature of the research; given the wide range of parameters that may influence how robot voices may be perceived, we felt it most appropriate to initially use an approach that supported analysis on a macro, rather than a micro level. Results from a rather diverse participant sample indicated that various vocal features might contribute to the shaping of this image. This preliminary study revealed several findings that are both novel and would seem to have important implications to future HRI research. In so doing, the paper motivates deeper follow-on studies to be conducted on the role of voice in HRI.

Notably, interaction context is something that should be addressed in future work when determining how a robot's voice should sound like. For example, as mentioned in II, [1] found a differential effect of voice accent based on the role the robot was playing in the interaction; similarly, [49] found that people trusted robots with a synthetic voice more when the robots were behaving trustworthily, and robots with a natural voice more when the robots were behaving untrustworthily. Furthermore, in [19], participants assigned different job contexts to robots which differed from each other based on minimal-pair features (e.g. robot faces with no mouth were consistently associated with security-type jobs). Similarly, in an experiment where people interacted with a robot in two different contexts (asking for directions and having an open chat), context affected perceptual ratings that people gave after the interaction [38].

Future work could also favour tests that use dynamic stimuli, possibly involving physically present robots, since the use of static images has recognised limitations in studies of this kind [45], and more broadly in the field of HRI [40, 2, 18]. Similarly, it is well known that novelty can introduce a confounding effect in HRI studies [14], especially when testing involves naive subjects. Therefore, future tests should consider a range of participants that have greater prior familiarity with robot technology, and/or explore the role of voice over longer periods of time when the novelty effect is likely to have faded.

With this study, we highlighted some of the vocal features that could be considered when designing robot voices, and we

argue that not only do people consistently form a mental image of a robot upon hearing its voice for the first time, but also that this image does not always match the voice-appearance pairs used in previous HRI studies.

References

[1] Sean Andrist et al. "Effects of Culture on the Credibility of Robot Speech". In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*. ACM. ACM Press, 2015, pp. 157–164. DOI: 10.1145/2696454.2696464.

[2] Wilma A. Bainbridge et al. "The benefits of interactions with physically present robots over video-displayed agents". In: *International Journal of Social Robotics* 3.1 (2011), pp. 41–52.

[3] Donn Bayard et al. "Pax Americana? Accent attitudinal evaluations in New Zealand, Australia and America". In: *Journal of Sociolinguistics* 5.1 (2001), pp. 22–49.

[4] Pascal Belin. "Voice processing in human and non-human primates". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 361.1476 (2006), pp. 2091–2107.

[5] Diane S. Berry and Leslie Z. McArthur. "Perceiving character in faces: the impact of age-related craniofacial changes on social perception." In: *Psychological Bulletin* 100.1 (1986), p. 3.

[6] Francesca Bray. "Gender and Technology". In: *Annu. Rev. Anthropol.* 36 (2007), pp. 37–53.

[7] Cynthia Breazeal. "Toward sociable robots". In: *Robotics and autonomous systems* 42.3-4 (2003), pp. 167–175.

[8] Leslie R. Brody. "The socialization of gender differences in emotional expression: Display rules, infant temperament, and differentiation". In: *Gender and emotion: Social psychological perspectives*. Ed. by Agneta H. Fischer. Cambridge University Press, 2000, pp. 24–47.

[9] Maya Cakmak and Leila Takayama. "Teaching people how to teach robots: The effect of instructional materials and dialog design". In: *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM. 2014, pp. 431–438.

[10] Elizabeth Cha et al. "Effects of Robot Sound on Auditory Localization in Human-Robot Collaboration". In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM. 2018, pp. 434–442.

[11] Joe Crumpton and Cindy L. Bethel. "A survey of using vocal prosody to convey emotion in robot speech". In: *International Journal of Social Robotics* 8.2 (2016), pp. 271–285.

[12] Brian R. Duffy. "Anthropomorphism and the social robot". In: *Robotics and autonomous systems* 42.3-4 (2003), pp. 177–190.

[13] Friederike Eyssel et al. "'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 2012 ACM/IEEE International Conference on Human-Robot Interaction*. IEEE. 2012, pp. 125–126.

[14] Rachel Gockley et al. "Designing robots for long-term social interaction". In: *Intelligent Robots and Systems (IROS), 2005 IEEE/RSJ International Conference on*. IEEE. 2005, pp. 1338–1343.

[15] Anthony M. Harrison, Wendy M. Xu, and J. Gregory Trafton. "User-Centered Robot Head Design: a Sensing Computing Interaction Platform for Robotics Research (SCIPRR)". In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM. 2018, pp. 215–223.

[16] Helen Hastie et al. "The Interaction Between Voice and Appearance in the Embodiment of a Robot Tutor". In: *International Conference on Social Robotics*. Springer. 2017, pp. 64–74.

[17] Susan M. Hughes, Franco Dispenza, and Gordon G. Gallup Jr. "Ratings of voice attractiveness predict sexual behavior and body configuration". In: *Evolution and Human Behavior* 25.5 (2004), pp. 295–304.

[18] Sylwia Hyniewska and Wataru Sato. "Facial feedback affects valence judgments of dynamic and static emotional expressions". In: *Frontiers in Psychology* 6 (2015), p. 291.

[19] Alisa Kalegina et al. "Characterizing the Design Space of Rendered Robot Faces". In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM. 2018, pp. 96–104.

[20] Caroline F. Keating. "Gender and the physiognomy of dominance and attractiveness". In: *Social Psychology Quarterly* (1985), pp. 61–70.

[21] Sara Kiesler. "Fostering common ground in human-robot interaction". In: *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*. IEEE. 2005, pp. 729–734.

[22] Robert M Krauss, Robin Freyberg, and Ezequiel Morsella. "Inferring speakers' physical attributes from their voices". In: *Journal of Experimental Social Psychology* 38.6 (2002), pp. 618–625.

[23] Hans J. Ladegaard. "National stereotypes and language attitudes: The perception of British, American and Australian language and culture in Denmark". In: *Language & Communication* 18.4 (1998), pp. 251–274.

[24] Shiri Lev-Ari and Boaz Keysar. "Why don't we believe non-native speakers? The influence of accent on credibility". In: *Journal of Experimental Social Psychology* 46.6 (Nov. 2010), pp. 1093–1096. DOI: 10.1016/j.jesp.2010.05.025.

[25] Dan Leyzberg et al. "Robots that express emotion elicit better human teaching". In: *Proceedings of the 2011 ACM/IEEE International Conference on Human-Robot Interaction*. ACM. 2011, pp. 347–354.

[26] Dingjun Li, PL Patrick Rau, and Ye Li. "A cross-cultural study: Effect of robot appearance and task". In: *International Journal of Social Robotics* 2.2 (2010), pp. 175–186.

[27] Diana Löffler, Nina Schmidt, and Robert Tscharn. "Multimodal Expression of Artificial Emotion in Social Robots Using Color, Motion and Sound". In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM. 2018, pp. 334–343.

[28] Conor McGinn et al. "Exploring the application of design thinking to the development of service robot technology". In: *ICRA2018 Workshop on Elderly Care Robotics - Technology and Ethics (WELCARO)*. IEEE. 2018.

[29] Wade J. Mitchell et al. "A mismatch in the human realism of face and voice produces an uncanny valley". In: *i-Perception* 2.1 (2011), pp. 10–12.

[30] Dylan Moore et al. "Making Noise Intentional: A Study of Servo Sound Perception". In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM. 2017, pp. 12–21.

[31] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. "Seeing voices and hearing faces: Cross-modal biometric matching". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8427–8436.

[32] Clifford I. Nass, Jonathan Steuer, and Ellen R. Tauber. "Computers are social actors". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. Boston, MA, USA, 1994, pp. 72–78.

[33] Clifford I. Nass et al. "Can computer personalities be human personalities?" In: *International Journal of Human-Computer Studies* 43.2 (1995), pp. 223–239.

[34] Andreea Niculescu et al. "Making social robots more attractive: the effects of voice pitch, humor and empathy". In: *International Journal of Social Robotics* 5.2 (2013), pp. 171–191.

[35] John J. Ohala. "Cross-language use of pitch: An ethological view". In: *Phonetica* 40 (1983), pp. 1–18.

[36] Aaron Powers and Sara Kiesler. "The advisor robot: tracing people's mental model from a robot's physical attributes". In: *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM. 2006, pp. 218–225.

[37] Jeff B. Russ, Ruben C. Gur, and Warren B. Bilker. "Validation of affective and neutral sentence content for prosodic testing". In: *Behavior Research Methods* 40.4 (2008), pp. 935–939.

[38] Maha Salem, Micheline Ziadee, and Majd Sakr. "Effects of politeness and interaction context on perception and experience of HRI". In: *International Conference on Social Robotics*. Springer. 2013, pp. 531–541.

[39] Anara Sandygulova and Gregory M. P. O'Hare. "Children's Perception of Synthesized Voice: Robot's Gender, Age and Accent". In: *Social Robotics*. Ed. by Adriana Tapus et al. Springer International Publishing, 2015, pp. 594–602. ISBN: 978-3-319-25554-5.

[40] Mohammad Shayganfar, Charles Rich, and Candace L. Sidner. "A design methodology for expressing emotion on robot faces". In: *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE. 2012, pp. 4577–4583.

[41] Elaine Schaertl Short, Mai Lee Chang, and Andrea Thomaz. "Detecting Contingency for HRI in Open-World Environments". In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM. 2018, pp. 425–433.

[42] Mikey Siegel, Cynthia Breazeal, and Michael I. Norton. "Persuasive robotics: The influence of robot gender on human behavior". In: *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE. 2009, pp. 2563–2568.

[43] Valerie K. Sims et al. "Robots' auditory cues are subject to anthropomorphism". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 53. 18. SAGE Publications. San Antonio, Texas, USA, 2009, pp. 1418–1421.

[44] Harriet M. J. Smith et al. "Concordant cues in faces and voices: Testing the backup signal hypothesis". In: *Evolutionary Psychology* 14.1 (2016), p. 1474704916630317.

[45] Harriet M. J. Smith et al. "Matching novel face and voice identity using static and dynamic facial images". In: *Attention, Perception, & Psychophysics* 78.3 (2016), pp. 868–879.

[46] Rie Tamagawa et al. "The effects of synthesized voice accents on user perceptions of robots". In: *International Journal of Social Robotics* 3.3 (2011), pp. 253–262.

[47] Benedict Tay, Younbo Jung, and Taezoon Park. "When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction". In: *Computers in Human Behavior* 38 (2014), pp. 75–84.

[48] Myrthe Tielman et al. "Adaptive emotional expression in robot-child interaction". In: *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*. ACM. 2014, pp. 407–414.

[49] Ilaria Torre et al. "Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience". In: *Proceedings of APAScience '18: Technology, Mind,and Society (TechMindSociety '18)*. Washington, DC, USA, 2018.

[50] Jaroslava Varella Valentova and Jan Havlíček. "Perceived sexual orientation based on vocal and facial stimuli is linked to self-rated sexual orientation in Czech men". In: *PloS one* 8.12 (2013), e82417.

[51] Jonathan Vitale et al. "Be More Transparent and Users Will Like You: A Robot Privacy and User Experience Design Experiment". In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM. 2018, pp. 379–387.

[52] Michael L Walters et al. “Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion”. In: *Autonomous Robots* 24.2 (2008), pp. 159–178.

[53] Michael L. Walters et al. “Human approach distances to a mechanical-looking robot with different robot voice styles”. In: *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*. IEEE. 2008, pp. 707–712.

[54] Debora Zanatto et al. “Priming Anthropomorphism: Can the credibility of humanlike robots be transferred to non-humanlike robots?” In: *Proceedings of the 2016 ACM/IEEE International Conference on Human-Robot Interaction*. IEEE. 2016, pp. 543–544.

[55] John Zimmerman et al. “Putting a face on embodied interface agents”. In: *Designing Pleasurable Products and Interfaces*. Eindhoven Technical University Press. 2005, pp. 233–248.

[56] Miron Zuckerman and Robert E. Driver. “What sounds beautiful is good: The vocal attractiveness stereotype”. In: *Journal of Nonverbal Behavior* 13.2 (1989), pp. 67–82.