The Interaction Between Voice and Appearance in the Embodiment of a Robot Tutor

Helen Hastie[1], Katrin Lohan[1], Amol Deshmukh[2], Frank Broz[1] and Ruth Aylett[1]

[1]School of Mathematical and Computer Science, Heriot-Watt University, Edinburgh, UK
h.hastie,k.lohan,f.broz,r.s.aylett@hw.ac.uk
[2]University of Glasgow, School of Computing Science, Glasgow, UK
Amol.Deshmukh@glasgow.ac.uk

Abstract. Robot embodiment is, by its very nature, holistic and understanding how various aspects contribute to the user perception of the robot is non-trivial. A study is presented here that investigates whether there is an interaction effect between voice and other aspects of embodiment, such as movement and appearance, in a pedagogical setting. An on-line study was distributed to children aged 11-17 that uses a modified Godspeed questionnaire. We show an interaction effect between the robot embodiment and voice in terms of perceived lifelikeness of the robot. Politeness is a key strategy used in learning and teaching, and here an effect is also observed for perceived politeness. Interestingly, participants' overall preference was for embodiment combinations that are deemed polite and more like a teacher, but are not necessarily the most lifelike. From these findings, we are able to inform the design of robotic tutors going forward.

Keywords: human-robot interaction, robotic tutors, robot embodiment, TTS

1 Introduction

A study is presented here that investigates interaction effects of voice and embodiment in a pedagogical domain. These types of social robots are being used in a variety of application areas such as entertainment, therapy, assistance as well as education [1]. Previous studies in HRI have indicated that the physical appearance of the robots [2], as well as their expressiveness can affect the interaction experience, especially with children. For example, Tielman et al. [3] and Leite at al. [4] indicate that children react more expressively and positively to a robot showing emotion through movement than to a robot that does not.

The degree of this expressiveness of a robot is directly connected to the definition of embodiment by Dauthenhahn et al. [5] via the set communication channels of a robotic system. Specifically, Dauthenhahn et al. define embodiment as 3-fold: the capabilities of the system, the level of its connections or perturbation channels with the environment and the environment in which it is embedded.

Thus, changing the level of embodiment can be achieved through not only changes in physical appearance, but also by changing the perception and production of utterances [6], as well as other expressive behaviours. This definition, therefore, supports the argument that changing the robot voice also changes its level of embodiment and this study investigates the perception of such an embodiment change.

2 Background

Adults and children both have preconceptions of how tutors and teachers should behave and act. Fischer (2011) [7] has shown that people's preconceptions about the degree of socialness of the human-robot interaction situation are important factors in determining the way people talk to a robot. Thus, preconceptions and user expectations play a crucial role, particularly in educational settings. With regard to media equation theory, Reeves and Nass [8] argued that people tend to treat computers as social actors. In their numerous experiments, in which people e.g., engaged in polite and reciprocal behaviours towards computers, they applied human stereotypes to computers such as a preconception that a car navigation system using a female voice was not an authoritative means to acquire directions [9]. Yet it is unclear how much such preconceptions are influencing how robots are perceived during interaction and for long-term engagement.

Pupil engagement is key to successful and sustained learning. However, Lemaignan et. al. [10] showed that anthropomorphic perception does not automatically elicit engagement. On the contrary, they found a negative correlation between anthropomorphic projections and actual behavioural engagement.

In the work presented here, we investigate the influence of voice as part of embodiment. Much work has been performed on the perception of synthesised speech (e.g [11]). Relevant here is the topic of gendered voice, which has been shown to influence user perceptions of the speaker. A study reported that when the participants were presented with a persuasive argument, the male synthetic voice was rated as more powerful than the female voice [12]. Also relevant are studies of synthesised speech in an educational environment such as the study by Goetz and colleagues [13] who reported that coherence between the educational tasks a robot performs and its appearance, including its instructive speech, can impact pupil compliance.

The influence of voice combined with robotic appearance has not been studied in the context of social robotics as widely or in the same detail as the influence of its appearance or of voice alone. A study by Walters and colleagues [14] showed that voice type could influence proximity in human-robot interactions. They found that when a mechanical looking robot employed either a human male voice, human female voice, synthesised neutral gender voice, or the experimenter's own voice, participants approached significantly closer to the robots with the human voices compared with the synthesised voices. A study by Tamagawa et al. [15] showed that people are influenced by the accents of synthesised voices when they rate the performance of robots, and these accents also influence peoples' experience of positive feelings.

Regarding gender of the robot voice, recent work from Reich-Stiebert and Eyssel [16] indicated that a mismatch of robot gender and gender typicality of the respective learning task led to increased willingness to engage in prospective learning processes with the robot. A study by Sandygulova [17] also showed that children's responses to perceived gender and age of voices of a NAO robot were influenced by changing its voice. Whilst we are not specifically studying the uncanny-valley, there are also studies with relation to a mismatch of voice and embodiment. Mitchell et al. [18] found that a cross-modal mismatch in human realism caused uncertainty about whether an entity was animate or inanimate. They found that a robot with a human voice, or a human being

with a synthetic voice, was perceived as eerier than a robot with a synthetic voice or a human being with a human voice.

For the study presented here, we hypothesise that there is an interaction effect between voice and other aspects of a robot's embodiment in an educational setting. To test this hypothesis, we conducted an on-line study to garner judgements of children aged 11-17 on videos of other children interacting with two types of robot embodiments with two different robot voices, using footage from the study reported in [19]. This work contributes in that it makes headway into understanding how the various aspects of embodiment contribute to the perception of the robot and can inform design going forward in terms of matching an appropriate Text-to-Speech (TTS) voice to the chosen robot embodiment.

3 Initial Interactive Study

As mentioned above, videos were used from an interactive study reported in [19]. This study involved pupils doing a pedagogical task related to map reading skills with two different types of robot:

(i) The EMYS robot, able to display facial expressions [20] using movable eyes with eyelids, and a head in three segments mounted on a movable neck (see Figure 1 right image).
(ii) The NAO torso robot, able to display expressions using upper body gestures with hands and head (see Figure 1 left image).

In this first study, the EMYS was paired with a female adult Scottish voice with regular pitch and speech rate developed from a human corpus using unit-selection by Cereproc Inc (henceforth referred to as ADULT voice), while the NAO robot has a child voice with high pitch and high speech rate (CHILD voice) produced by Nuance. Using a Mann-Whitney U-test for unpaired non-parametric data, it was found that the NAO robot is rated as significantly more *friendly* ($p = 0.01, r = 0.46$), *pleasant* ($p = 0.02, r = 0.42$) and *empathic* ($p = 0.03, r = 0.38$) than the EMYS. Furthermore, the participants were given a question after each interaction (condition) to indicate their perceived relationship with the robot. Options given *"Brother or Sister", "Classmate", "Stranger", "Relative", "Friend", "Parent", "Teacher, "Neighbour"*. They found that the majority of participants rated NAO as a *"Friend"* (43%) and the EMYS as a *"Stranger"* (40%).

4 Online-Study Methodology

Two 30 second snippets of video were taken from the above-mentioned study, one of a child interacting with each robot (see Figure 1). These were duplicated and the voices of the robot were exchanged resulting in 4 videos (2 with the original voice and 2 with the swapped voice). The audio was synchronised with the behaviour of the robot. The videos can be found at `https://tinyurl.com/yaadoek4` and source code for system modules can be found at `http://www.emote-project.eu`.

54 participants took part in the study aged between 11-17 (mean=13.25) balanced roughly for gender (52% girls, 48% boys). These participants had similar backgrounds as they were recruited through local schools. Rewards were not given.

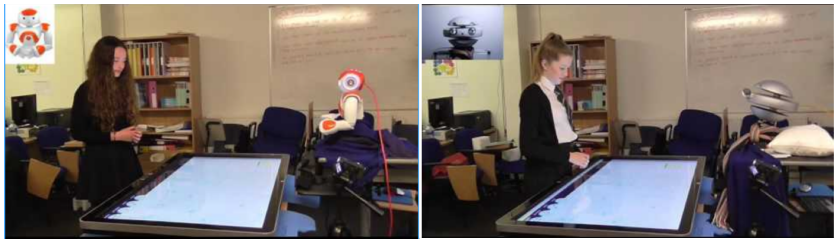

Fig. 1. Screen shots of videos shown to participants: on the left with the NAO Robot and on the right with the EMYS robot. Both side and head-shots were also shown (see top left).

4.1 Experiment Design

As mentioned above, while each participant saw footage of only one robot, they were each presented with two versions of the footage: 1) one with the CHILD voice; and 2) one with the ADULT voice.

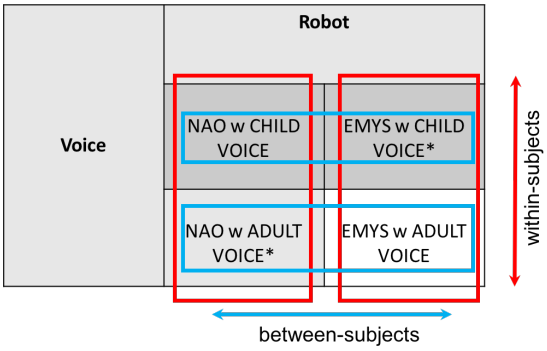

Fig. 2. Experimental design set-up. *indicates the majority preferred combination (within-subjects)

Therefore, we conducted a within-subjects experiment for comparing the voices on one type of robot but a between-subject experiment for comparing cross-robot embodiment ratings (see Figure 2 for an overview of the experiment design). Each participant was randomly assigned to the two main, between-subjects conditions (NAO robot or EMYS robot) with 27 participants in each group. The two participant groups were roughly balanced for gender and had a similar mean age (mean = 13.3/13.2). The ordering of which voice came first was alternated to counteract any ordering effects.

4.2 Questionnaires

The participants were asked to complete a short questionnaire after each video. The questions were taken and modified for children from the Godspeed questionnaire series

[21], designed as a standard user measurement tool for human-robot interaction. The number of questions was reduced to 6 in order to avoid fatigue in the children. To make it appropriate for children a "smileyometer" scale was used. *Lifelikeness, friendliness* and *pleasantness* were taken directly from the Godspeed questionnaire with the latter two allowing for cross-study comparison. *Empathy* was deemed non-relevant for an on-line observational study. The perceived *politeness* and *intelligibility* were included as they are questions that have been used in language generation and previous speech perception studies (e.g. [22]) with politeness deemed an important characteristic of teachers [23].

For the sixth question, the participants were asked to indicate their perceived relationship with the robot. Options given were as above *"Brother or Sister", "Classmate", "Stranger", "Relative", "Friend", "Parent", "Teacher", "Neighbour"*. Finally, after watching both videos and filling in both questionnaires, the participants were asked the question "Which robot would you prefer to have in your classroom?" and were asked to select one of two options (Robot 1/Robot 2).

4.3 Participation vs. observation studies

Whilst crowd-sourcing using videos has proven to be useful in previous HRI studies [24,25], there is clearly a difference between observing a video of a robot interaction and participating in such an interaction. Therefore, it is important to establish whether these differences would make the data obtained from the two experiments incomparable. For this reason, the conditions in this study that corresponded to the in-school study reported in [19], namely NAO robot with the CHILD voice, and EMYS robot and the ADULT voice were examined.

As mentioned above in the interactive study, the NAO was rated significantly higher for the *pleasant*, *friendly* and *empathic* ratings. The same test (a Mann-Whitney U-test) was performed for the online study and similar results were found in that the NAO had a higher mean rating for all dimensions and significantly so for *pleasantness* ($p = 0.03, r = 0.62$) and for *politeness* ($p = 0.03, r = 0.61$). This indicates that while, in general, there are major differences between being a participant and being an observer, there are similar perceptions in the areas of interest in terms of comparing whole robot/voice combinations. What we are interested in, however, is investigating further what aspect of embodiment, including voice, influences the perception of the robotic tutors. Our method for doing such a study is discussed in the following section.

5 Results

A mixed-design ANOVA was conducted with the between-subjects independent variable (IV1) as the robot embodiment (EMYS vs. NAO) and the within-subjects independent variable (IV2) as the voice (CHILD vs. ADULT). The dependent variable was taken as the rating on the five rating scales (e.g. $DV = friendliness, DV = pleasantness$). Whilst 5 rating questions were asked, we found an interaction effect for two of these: *lifelike* and *politeness*. We, therefore, only present here results for these two dependent variables as well as results for the overall preference of voice and the perceived relationship. We briefly discuss the role of gender and finally present some qualitative feedback.

5.1 Lifelike

For *lifelike*, there is an interaction effect observed for the interaction between the robot embodiment and voice ($F(1,102) = 8.54, MSE = 11.45, p = 0.004$). This means that the lifelike ratings of the voice are not detached from the other aspects of embodiment. Specifically, the CHILD voice is deemed more lifelike on the NAO robot and the ADULT voice deemed more lifelike on the EMYS robot, as shown in Figure 3[1].

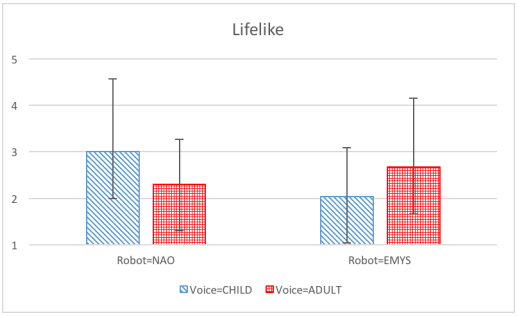

Fig. 3. Graphic showing mean *lifelike* ratings where an interaction effect for the interaction between robot and voice was observed. Error bars show standard deviation.

5.2 Politeness

Through the mixed-design ANOVA, there was no observed interaction effect between voice and embodiment. However there was observed a significant effect for the single variable voice-only, i.e. for the within-subjects, single robot condition ($F(1,102) = 6.98, MSE = 6.43, p = 0.01$). Unconfounded tests by a Wilcoxon signed rank test for paired data ($Z = -8.1, r = 1.1, p = 0.007$) show that the CHILD voice is rated significantly more polite than the ADULT voice on the EMYS robot (indicated by * in Figure 4). Therefore, on the EMYS only, we can state that the CHILD voice is deemed significantly more *polite*.

5.3 Overall Preference of Voice

With regard to the preference of voice for a certain robot/voice combination. 59% preferred the CHILD voice on the EMYS robot and 81% preferred the ADULT voice on the NAO. It is interesting to note that it is these combinations (indicated by * on Figure 2) that were rated as less lifelike, as discussed above. We can explore this further with a two proportion Z-score test to test the hypothesis that the proportion of participants who preferred the CHILD voice is different between the two robot condition. Indeed, through this two proportion, one-way Z-score test, a significant difference is observed ($Z = 3.0706, p = 0.001$)[2].

[1] Confounded post-hoc tests cannot be conducted here as there is more than one variable.

[2] Requirements for approximating a binomial distribution with a normal distribution were met for the Z-score tests reported here.

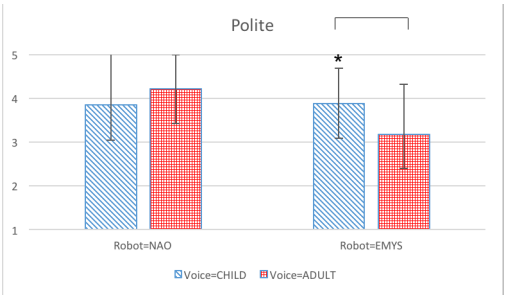

Fig. 4. Graphic showing a mean ratings for *politeness* where an interaction effect for voice-only was observed, which is significant in the EMYS condition indicated by * ($p < 0.05$). Error bars show standard deviation.

5.4 Perceived Relationship

Figures 5 and 6 show pie charts of all four combinations. Interestingly, the conditions that the majority preferred (indicated by *) were categorised a higher proportion of the time as teachers (52% vs 37% for EMYS and 59% vs 56% for NAO). Recall that in the in-school study presented in Section 3, 43% of the participants rated NAO as a Friend. Perhaps not surprisingly, this is a difference between the participatory study and the observational study whereby a minimal level of interaction is likely necessary for the participant to bond enough to refer to the robot as a friend.

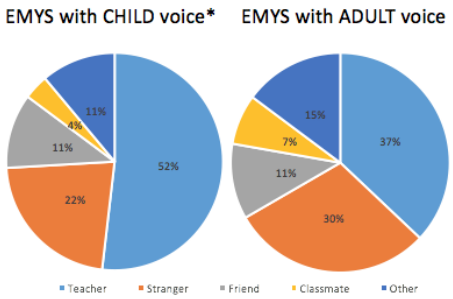

Fig. 5. Pie charts of the perceived character of the EMYS robot embodiment for two voices. *indicates the majority preferred combination.

5.5 Gender and Perception

We also investigated whether the gender of the participant had an effect on ratings. We again performed a mixed-design ANOVA now with three independent variables: IV1:the robot embodiment (EMYS vs NAO); IV2: the gender of the participant; and IV3: the

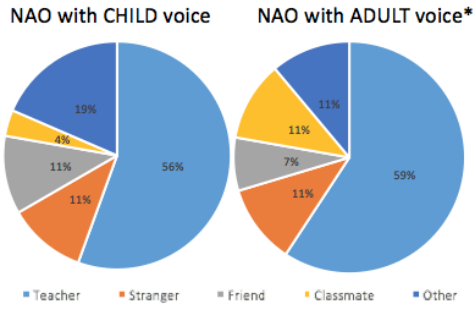

Fig. 6. Pie charts of the perceived character of the NAO robot embodiment for two voices. *indicates the majority preferred combination.

robot voice. The dependent variable was taken again as each rating on individual scales (e.g. $DV = friendliness, DV = pleasantness$). We found an interaction effect for *lifelikeness* between gender and voice (IV2 and IV3) ($F(1,98) = 6.13, MSE = 7.802, p = 0.01$). Indeed, it was found that the boys rated all combinations higher than the girls for lifelike.

5.6 Qualitative results

Qualitative comments backed up the view that the majority preferred combination was perceived less lifelike (i.e. NAO with ADULT and EMYS with CHILD). For example, in the NAO robot condition, there were comments regarding the human-like qualities of the ADULT voice e.g. "..the first [ADULT VOICE] was a lot more artificial" and " He [CHILD VOICE] seems more friendly and life like". In the EMYS robot condition, the ADULT voice was commented on being more human-like "Judging mostly on voice-only, not on what was said, [ADULT Voice] robot is more human-like".

6 Discussion and Future work

Politeness strategies used by teacher and students in the class can play an important role in the learning and teaching process [23]. Moreover, politeness can have an instrumental role in the social interaction for example in Brown and Levinson's [26] theory that places politeness as a universal face-threatening strategy. As effective teachers employ politeness strategies, pupils may therefore have preconceptions of the level of politeness of teachers and this may have affected their ratings, where the robot that was deemed significantly more polite (EMYS with CHILD voice) was categorised mostly as a teacher (52%) and also preferred by the majority in an educational setting (59%).

It is interesting to note that the combinations that were deemed less lifelike were actually those that were preferred in the within-subjects condition (EMYS with CHILD voice and NAO with ADULT voice). In addition, these less lifelike combinations were also deemed more frequently as a teacher than the more lifelike combinations. This has

connotations for the design of robotic tutors going forward, whereby politeness should be emulated but perhaps the goal of making the robot completely lifelike may not be a necessary one. Indeed, Zawieska [27] argues that rather than aim at the close resemblance of human characteristics in the robot's form and behaviour, "anthropomorphic robots may *deliberately* exploit the divergence between the robot's characteristics and performance and the human frame of reference" and our work supports this as far as the subjects prefer the less lifelike combination.

With regards the other attributes tested, it is perhaps not surprising that no interaction was found for *friendliness*. This may be because physical interaction is necessary for a robot to be perceived as a friend or friendly. No significant effect was found for *pleasantness* or *intelligibility* and this may simply be because the voices rely on equally mature, natural sounding TTS technology although one has to be careful when interpreting "no effect". Future work should involve examining the perceived gender of the two robots in question and looking at varying the pedagogical environment in order to investigate if the perception of the robot changes depending on the task at hand as was found by [17]. The idea of complexity of interactions and embodiment would be interesting to investigate at different age ranges as well as in relation to the different genders. Finally, future work would be to investigate perceptions in long-term educational studies, especially in terms of how children build relationships in situations and form attachments with respect to different embodiments.

7 Acknowledgements

We thank all the students, teachers and school staff who were involved in this study and our colleague Dr. Janarthanam. This work was partially supported by the European Commission (EC) under ICT-317923 EMOTE.

References

1. Daniel Leyzberg, Samuel Spaulding, Mariya Toneva, and Brian Scassellati. The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the Cognitive Science Society*, volume 34, 2012.
2. Joshua Wainer, David J Feil-Seifer, Dylan Shell, Maja J Mataric, et al. Embodiment and human-robot interaction: A task-based perspective. In *Proceedings of HRI*, 2007.
3. Myrthe Tielman, Mark Neerincx, John-Jules Meyer, and Rosemarijn Looije. Adaptive emotional expression in robot-child interaction. In *Proceedings of HRI*, 2014.
4. Iolanda Leite, Ginevra Castellano, André Pereira, Carlos Martinho, and Ana Paiva. Modelling empathic behaviour in a robotic game companion for children: an ethnographic study in real-world settings. In *Proceedings of HRI*, 2012.
5. Kerstin Dautenhahn, Bernard Ogden, and Tom Quick. From embodied to socially embedded agents–implications for interaction-aware robots. *Cognitive Systems Research*, 3(3), 2002.
6. Kerstin Fischer, Katrin Lohan, and Kilian Foth. Levels of embodiment: Linguistic analyses of factors influencing HRI. In *Proceedings of HRI*, 2012.
7. Kerstin Fischer, Kilian Foth, Katharina J Rohlfing, and Britta Wrede. Mindful tutors: Linguistic choice and action demonstration in speech to infants and a simulated robot. *Interaction Studies*, 12(1):134–161, 2011.
8. Byron Reeves and Clifford Nass. How people treat computers, television, and new media like real people and places. *CSLI Publications and Cambridge*, 1996.

9. Clifford Nass and Youngme Moon. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103, 2000.
10. Séverin Lemaignan, Julia Fink, Francesco Mondada, and Pierre Dillenbourg. You're doing it wrong! studying unexpected behaviors in child-robot interaction. In *International Conference on Social Robotics*, pages 390–400. Springer, 2015.
11. Simon King and Vasilis Karaiskos. *The Blizzard Challenge 2016*. University of Edinburgh.
12. John W Mullennix, Steven E Stern, Stephen J Wilson, and Corrie-lynn Dyson. Social perception of male and female computer synthesized speech. *Computers in Human Behavior*, 19(4):407–424, 2003.
13. Jennifer Goetz, Sara Kiesler, and Aaron Powers. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *Proceedings of ROMAN*, pages 55–60. Ieee, 2003.
14. Michael L Walters, Dag Sverre Syrdal, Kheng Lee Koay, Kerstin Dautenhahn, and R Te Boekhorst. Human approach distances to a mechanical-looking robot with different robot voice styles. In *Proceedings of HRI*, 2008.
15. Rie Tamagawa, Catherine I Watson, I Han Kuo, Bruce A MacDonald, and Elizabeth Broadbent. The effects of synthesized voice accents on user perceptions of robots. *International Journal of Social Robotics*, 3(3):253–262, 2011.
16. Natalia Reich-Stiebert and Friederike Eyssel. (Ir) relevance of Gender?: on the influence of gender stereotypes on learning with a robot. In *Proceedings of HRI*, 2017.
17. Anara Sandygulova and Gregory MP O'Hare. Children's responses to genuine child synthesized speech in child-robot interaction. In *Proceedings of HRI*, 2015.
18. Wade J Mitchell, Sr Kevin A Szerszen, Amy Shirong Lu, Paul W Schermerhorn, Matthias Scheutz, and Karl F MacDorman. A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception*, 2(1):10–12, 2011.
19. Amol Deshmukh, Srinivasan Janarthanam, Helen Hastie, Mei Yii Lim, Ruth Aylett, and Ginevra Castellano. How expressiveness of a robotic tutor is perceived by children in a learning environment. In *Proceedings of HRI*, 2016.
20. Jan Kędzierski, Robert Muszyński, Carsten Zoll, Adam Oleksy, and Mirela Frontkiewicz. Emys-emotive head of a social robot. *International Journal of Social Robotics*, 5(2):237–249, 2013.
21. Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81, 2009.
22. Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser, and Oliver Lemon. Cluster-based prediction of user ratings for stylistic surface realisation. In *Proceedings of EACL*, 2014.
23. Senowarsito Senowarsito. Politeness strategies in teacher-student interaction in an EFL classroom context. *TEFLIN Journal*, 24(1), 2013.
24. Nathan Koenig, Leila Takayama, and Maja Matarić. Communication and knowledge sharing in human–robot interaction and learning from demonstration. *Neural Networks*, 23(8):1104–1112, 2010.
25. Russell Toris, David Kent, and Sonia Chernova. The robot management system: A framework for conducting human-robot interaction studies through crowdsourcing. *Journal of Human-Robot Interaction*, 3(2):25–49, 2014.
26. P. Brown and S.C. Levinson. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics. Cambridge University Press, 1987.
27. Karolina Zawieska and Agnieszka Sprońska. *Anthropomorphic Robots and Human Meaning Makers in Education*, pages 251–255. Springer International Publishing, Cham, 2017.