Speaking Properly with Robots

Julia Rosén
University of Skövde
Skövde, Sweden
julia.rosen@his.se

Erik Lagerstedt
University of Skövde
Skövde, Sweden
erik.lagerstedt@his.se

ABSTRACT

There is a risk of genuine but norm-breaking phenomena related to human-robot interaction remaining invisible, since their rarity make observed instances dismissed as anecdotes. In this extended abstract we present observations related to bias in who is understood in vocal interactions with robots. Noting the fundamentally political and intersectional nature of the problem, we call for a strategy for documenting such comparatively rare or messy events to break the invisibility and facilitate accumulation of evidence.

KEYWORDS

Robots, HRI, Speech Recognition, Vocal Interaction, Sociolinguistics

1 INTRODUCTION

There are many ways to interact with robots and vocal interaction is an important mode, not least for sociable robots [2]. Vocal interaction can be used for instructing the robot (when used as a tool or servant) or conversing with it (when used as a social companion). Although vocal interaction is much more than the process of sending and receiving distinct messages [12], sending and receiving messages are important aspects for both agents. In this abstract, we will mainly focus on the robot's (in)ability to recognize human utterances, based on observations from a recent experiment where participants had open-ended conversations with the humanoid robot Pepper. Although the interactions went generally well and participants were positive toward the robot, we noted communication breakdowns based on failures to recognize some human speech. The experiment was run in English with mostly non-native English speakers and we observed that the speech recognition failed more for some accents, leading to a bias in accessibility of the technology.

These observations were secondary findings in the study, and were too few and unique to allow for statistical analysis. Qualitative methods are generally better suited for such analysis, either deductively by understanding these special cases through established theories or categories, or inductively by examining what themes or patterns emerge from the data [18]. However, since the observations are secondary findings based on analysis of statistical outliers or other kinds of rare cases, such analysis might not be prioritized or dismissed as too anecdotal, instead favoring results directly related to the research questions. It is, however, important to document such findings (even if the intended purpose of the study is something else) for the literature to better reflect the messy nature of communication and interaction, and in turn allow for patterns related to rarer events to be visible. By highlighting these observations, we hope to provide examples of otherwise overlooked data. It is not intended as a criticism against the quality of state of the art technology, but rather highlight some of the problematic assumptions on which some of these technologies implicitly rely, and how that in turn might enforce marginalization and exclusion of some people. The brief background is followed with a short description of the experiment, and the abstract is concluded with a short discussion of the observations and our position.

2 SPEAKING WITH ROBOTS

There are many variations and nuances in spoken language and it can be difficult to distinguish between the categories language, dialect, and accent [14]. In some languages there are standard accents (such as the SSBE in Brittish English), however, such standards are idealized abstractions of the languages and therefore artificial [14]. A drawback of such standards is that they perpetuate the myth that some people speak the standard language and therefore have no accent. "Non-accent" speech is thus considered more appropriate, and people who most noticeably use "non-standard" accents are generally perceived more negatively [13]. Variations of a language can be based on geographical location and social status such as class, age, gender, and education, and will thus signal what social groups the speaker belongs to, and can in turn educe in-group biases [4, 11]. Attempts to identify and classify accents as more or less proper or correct is therefore fundamentally political [14] and intersectional [1, 16].

When humans are speaking, the understanding and production of accents is closely related; however, these are typically separate issues for robots. The understanding is related to speech recognition, such as speech-to-text modules, to transform input to a format that fits the processing. The part of the output intended as a vocal reply is also in the form of text, and a separate text-to-speech module is used to generate the actual sounds played by the robot. It is therefore quite possible that input is generated by a model trained to fit a different accent than the model generating the output.

Since the recognition aspect of the problem is mainly related to speech recognition (both in terms of vocabulary and pronunciation), there is much work in human-computer interaction (HCI) to rely on. For instance, problems related to the over-representation of a fairly specific set of accents (emulating female and American accents) for default voices among digital assistants [29] and the lack of sense-making when semantic information is derived automatically from language corpora [3, 26] are relevant for both HCI and HRI. Another example is the identification and mitigation of bias in speech recognition technology, which has been studied in HCI with respect to aspects such as gender, race, and disability [17]. It is necessary to also study vocal interaction and speech recognition systems in HRI, since the embodied nature of the robots change the context of such interactions. Issues related to accents in HRI have been acknowledged previously, both for recognition [8, 25], semantic processing [7], and production [19, 20, 25, 27], but work on recognition is still particularly rare. Irfan et al. [8] report on robots failing to recognize speech disproportionately from non-native English speakers, and the proposed potential solutions were to improve the robot's hardware, increase the number of languages the robot can interact in, while constraining the grammar in each respective language (which is also recommended for vocal child-robot interaction [10]). Albeit reasonable, they might not be sufficient from an inclusion perspective, since identifying what languages and accents to be included can be a way to define and enforce a norm for what vocalizations are valid (cf. [14]). Further, deploying robots not able to handle the relevant accents can be counter-productive, or even harmful [25].

3 THE STUDY AND THE OBSERVATIONS

The observations reported in this abstract are from a within-subject experiment conducted to study how participants behave and react in interaction with a social robot. The participants were told that (1) they could freely explore what is possible in the interaction, and (2) the purpose of the study was to evaluate how well people interact with a robot intended as a companion in the home. Each participant had two interaction sessions with the robot (2.5 min. each), with a short break to fill in questionnaires. The entire interaction was in English. A customized dialogue system using Google's speech-to-text cloud service, OpenAI GPT-3 language model for responses, and NaoQi ALAnimatedSpeech service for the robot speech was used. The technology used in the experiment is state of the art, which we highlight since we argue that the observed problems are not mainly due to old technology, but more fundamental in nature.

The participants' (N=31, each rewarded with a movie ticket) ages ranged between 20–54 years old (M=29) and were recruited by fliers on campus. The participants consisted mainly of students and teachers from various faculties, with 45% identifying as male and 55% identifying as female (no-one self-described or chose non-binary). The experiment was conducted in Skövde, Sweden, with participants having a wide range of first languages: there were native speakers of English (7%), Swedish (56%), Spanish (16%), and Arabic (10%), and the remaining 11% had one participant per language, namely native German, Portuguese, Turkish, and bilingual Spanish/Arabic speakers. All participants were easily understood by the facilitator when speaking English. We collected data on the first languages as we observed during piloting that people with stronger non-American accents were in general understood less. We did also observe indications of bias in speech recognition based on gender, with more issues for female than male participants.

While analyzing the data (a still ongoing process) from the experiment, some secondary observations were made regarding instances of inability of the robot. Despite a similar theme, each observation was an outlier in some way, meaning they would typically be discarded as anecdotes, in favor of clarity when compiling the study's results. For this abstract, we mainly rely on six participants who could be identified as outliers in the quantitative data, since the robot made practically no utterances during at least one of the two interactions of those participants (the robot did not recognize any utterances from the participant). None of these participants had American accents, but one was a native speaker of a British accent. Overall, we identified two main issues during the vocal interactions: (1), the speech is not recognized at all and there is no response, and (2), the speech is misheard and it prompts an incorrect response.

4 DISCUSSION AND CONCLUSIONS

Breakdown in speech recognition can be due to many reasons, including speech-to-text, the language model, and the robot's hardware (e.g., the built-in microphone). In the experiment we observed, for instance, that participants speaking in soft voices were often not registered by the robot. This phenomenon intersects potentially with gender and other aspects impacting the accent. Such aspects are important, but it is also necessary to consider the context in which interactions take place. The robot might, for instance, encourage an anthropomorphic expectation—where past experiences with humans are used as a template for understanding the interaction with the robot—through its particular embodiment [6, 15, 23]. Creating robots that signal social capabilities they do not possess could therefore constitute a kind of harmful deception, where humans expect to be able to communicate in a way that is not possible [9, 22, 24]. If this illusion remains despite breakdown, the robot's response can be perceived as intentional rudeness. In our study we, for instance, had a participant who's question about singing was misconstrued as a statement regarding pregnancy, leading to a response from the robot that prompted a flustered *"do I look fat?"* from the participant. Another way of handling communication breakdowns appears when the illusion is stronger than the user's confidence, and the user attributed the blame to their own inability [25]. The participant who got the fewest responses from the robot (4 in total) fell into this category. The participant was very interested in technology and stated after the interactions that *"I don't think the robot understand me. It happens a lot when I try to talk to Google. I don't know if it my voice or if it's the robot."* Some participants did simply give up if the robot did not perform sufficiently well, ascribing the failure to the robot. Blame attribution is important for maintaining cooperation [28], making bias in this regard problematic.

Similarly, several of the participants with major issues with being understood expressed a positive interaction with the robot. For example, one participant said *"It was a great experience [...] it was a first time for me to see a robot. It's easy to contact with it."*, despite almost no utterances from the robot. Expressed positive user experience is thus possible even after poor interactions. It is possible that people who are used to not be understood by speech recognition might settle for inferior levels of interaction and therefore accept what is offered to them. This mirrors results from HCI where non-native English speakers had equal satisfaction but reduced usability of a smart home compared to native English speakers [21]. We do not, however, consider it acceptable with such inequity in terms of access to technology. We believe that this sentiment is shared with most roboticists, and that increased positive user experience for all through inclusive design is generally a desired goal.

We do not have a clear solution for how to address the problems of bias in who is being understood. The issue is complex, and there are risks associated with both the strategy of ignoring and of recognizing different accents [30]. The problem should, however, remain a difficult problem to wrestle with for the roboticists, but not for the users. As such, it is necessary to acknowledging that it is a problem and keep the issue open until it is properly resolved (if that will ever be the case). For designers of robots, it is important to be honest and transparent when relying on (and therefore perpetuating) a model of language that is fundamentally and problematically flawed. A step toward addressing and study the issue among researchers would be to identify ways to systematically document rare events (including what demographic data to collect and how to report it) that together would paint a clearer picture, but would not be sufficient on their own to provide clear results. The intersectional nature of these phenomena mean that it might be necessary to collect information related to e.g., gender and first language, but it might not be sufficient to have the variables aggregated and in isolation [5].

ETHICAL REVIEW

This project was submitted for ethical review to the The Swedish Ethical Review Authority (#2022-02582-01, Linköping) and was found to not require ethical review under Swedish legislation (2003:615).

REFERENCES

[1] Megan E Birney, Anna Rabinovich, and Thomas A Morton. 2020. Where are you from? An investigation into the intersectionality of accent strength and nationality status on perceptions of nonnative speakers in Britain. *Journal of Language and Social Psychology* 39, 4 (2020), 495–515.

[2] Cynthia Breazeal. 2003. Toward sociable robots. *Robotics and autonomous systems* 42, 3-4 (2003), 167–175.

[3] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[4] Jack K. Chambers. 2009. *Sociolinguistic Theory: Linguistic Variation and its Social Significance* (revised ed.). Wiley-Blackwell.

[5] Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism.* MIT press.

[6] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review* 114, 4 (2007), 864–886.

[7] Charles Hannon. 2018. Avoiding bias in robot speech. *Interactions* 25, 5 (2018), 34–37.

[8] Bahar Irfan, Mehdi Hellou, Alexandre Mazel, and Tony Belpaeme. 2020. Challenges of a real-world HRI study with non-native english speakers: Can personalisation save the day?. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 272–274.

[9] Ryan Blake Jackson and Tom Williams. 2019. On perceived social and moral agency in natural language capable robots. In *2019 HRI workshop on the dark side of human-robot interaction*. 401–410.

[10] James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child speech recognition in human-robot interaction: evaluations and recommendations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 82–90.

[11] Katherine D Kinzler, Kathleen H Corriveau, and Paul L Harris. 2011. Children's selective trust in native-accented speakers. *Developmental science* 14, 1 (2011), 106–111.

[12] Klaus Krippendorff. 1993. Major metaphors of communication and some constructivist reflections on their use. *Cybernetics & human knowing* 2, 1 (1993), 3.

[13] Shiri Lev-Ari and Boaz Keysar. 2010. Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of experimental social psychology* 46, 6 (2010), 1093–1096.

[14] Rosina Lippi-Green. 2012. *English with an accent: Language, ideology, and discrimination in the United States.* Routledge.

[15] Manja Lohse. 2009. The role of expectations in HRI. *New Frontiers in Human-Robot Interaction* (2009), 35–56.

[16] Larry R Nelson Jr, Margaret L Signorella, and Karin G Botti. 2016. Accent, gender, and perceived competence. *Hispanic Journal of Behavioral Sciences* 38, 2 (2016), 166–185.

[17] Mikel K Ngueajio and Gloria Washington. 2022. Hey ASR System! Why Aren't You More Inclusive? Automatic Speech Recognition Systems' Bias and Proposed Bias Mitigation Techniques. A Literature Review. In *HCI International 2022–Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings*. Springer, 421–440.

[18] Michael Quinn Patton. 2014. *Qualitative research & evaluation methods: Integrating theory and practice.* Sage publications.

[19] Michael Pucher. 2022. On the sociolects of robots. In *Workshop Robo-Identity 2: Exploring Artificial Identity and Emotion via Speech Interactions*.

[20] Michael Pucher, Gudrun Schuchmann, and Peter Fröhlich. 2009. Regionalized text-to-speech systems: Persona design and application scenarios. In *Multimodal Signals: Cognitive and Algorithmic Issues: COST Action 2102 and euCognition International School Vietri sul Mare, Italy, April 21-26, 2008 Revised Selected and Invited Papers*. Springer, 216–222.

[21] Aung Pyae and Paul Scifleet. 2018. Investigating differences between native English and non-native English speakers in interacting with a voice user interface: A case of Google Home. In *Proceedings of the 30th Australian conference on computer-human interaction*. 548–553.

[22] Julia Rosén, Erik Lagerstedt, and Maurice Lamb. 2022. Is human-like speech in robots deception?. In *HRI'22 Workshop—Robo-Identity 2, March 6, 2022, online*. 1–3.

[23] Julia Rosén, Jessica Lindblom, and Erik Billing. 2022. The Social Robot Expectation Gap Evaluation Framework. In *Human-Computer Interaction. Technological Innovation: Thematic Area, HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part II*. Springer, 590–610.

[24] Julia Rosén, Kathleen Richardson, Jessica Lindblom, and Erik Billing. 2018. The robot illusion: facts and fiction. In *Workshop in Explainable Robotics System, workshop in conjunction with HRI Human Robot Interaction 2018, Chicago, USA, March, 5-8, 2018*.

[25] Divyanshu Kumar Singh, Manohar Kumar, Eduard Fosch-Villaronga, Deepa Singh, and Jainendra Shukla. 2022. Ethical Considerations from Child-Robot Interactions in Under-Resourced Communities. *International Journal of Social Robotics* (2022), 1–17.

[26] Dias Oliva Thiago, Antonialli Dennys Marcelo, and Alessandra Gomes. 2021. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online. *Sexuality & culture* 25, 2 (2021), 700–732.

[27] Ilaria Torre and Sébastien Le Maguer. 2020. Should robots have accents?. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 208–214.

[28] Diede PM Van der Hoorn, Anouk Neerincx, and Maartje MA de Graaf. 2021. "I think you are doing a bad job!" The Effect of Blame Attribution by a Robot in Human-Robot Collaboration. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*. 140–148.

[29] Mark West, Rebecca Kraut, and Han Ei Chew. 2019. I'd blush if I could: closing gender divides in digital skills through education. (2019).

[30] Tom Williams. 2023. The Eye of the Robot Beholder: Ethical Risks of Representation, Recognition, and Reasoning over Identity Characteristics in Human-Robot Interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*.