Universiteit Leiden
The Netherlands

Bachelor Computer Science & Datascience and Artificial Intelligence

The effect of different voices in robots

Lee van Ruijven

First supervisor & Second supervisor:
Joost Broekens & Tessa Verhoef

RESEARCH PROPOSAL BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
redwww.liacs.leidenuniv.nl
20/02/2025

The effect of different voices in robots

Van Ruijven, Lee

LIACS, Leiden University
The Netherlands
s3221261@leidenuniv.nl

Abstract

This study investigates whether there is an interaction effect between voice type (human-like vs. mechanical) and task type (social vs. functional) on user perceptions of trust, ease of use, and likeability in robots. A 2x2 factorial experiment involving 40 participants interacting with a NAO robot demonstrated that task type significantly influences certain factors of trust, like perceptions of intelligence, reliability, and capability, while voice type primarily affects likeability and perceived intelligence. Contrary to expectations, no significant interaction effects between voice and task were found for trust, ease of use, or likeability. These findings suggest that task type predominantly shapes cognitive evaluations of robots, whereas voice naturalness impacts affective responses. The study highlights the importance of considering both task and vocal design when developing robots for varied human-robot interactions.

1 Introduction

Robots are becoming more prevalent in our society with each passing year [14]. Thus, the field of robotics is actively researching the technical capabilities of our existing and future robots, but it is also important to understand the effects of these robots on the humans interacting with them. Research in human-robot interaction focusses on this but emphasizes robot appearance, behaviour, and non-verbal communication [4]. Despite this, verbal communication might be just as important.

Within psychology it is well researched, and society generally understands that tone and pitch of one's voice are incredibly vital to how we understand the message someone is trying to convey and how we perceive other people as a whole [26]. Voices can carry an abundance of information about someone, such as social status, gender, economic background, race, and we have been conditioned to differentiate these characteristics, creating the concept of 'auditory faces' [6].

Unconsciously, we apply the phenomenon of assuming capabilities and traits to robot voices as well and link certain physical traits in robots to these voices, even if there is no precedent for it [25]. This also extends to the perceived capabilities someone, or in this case, a robot, has. For example, we might expect a more synthetic voice to be better at solving logical problems or a more human-sounding voice to have better customer service skills [9].

2 Related work

The characteristics of a robot strongly affect a human's perception and thus their behaviour towards that robot. Previous studies have reported a significant effect of appearance [22], behaviour [8], and many other factors of a robot on human perception. Among these factors is voice, of which many components have been individually researched. It has been shown a robot's voice gender [11], human-likeness [4] [28] [19], accent [30], and pitch [24] can have an effect in different areas of perception, such as trust. In addition, not all components of voice had significant effect in terms of human perception. For voices that change in pitch in accordance to emotion, little effect was found [16]. A study done on the human-likeness of synthetic voices found that the more realistic voices were rated more pleasant and were more anthropomorphised by participants [28]. Furthermore, a similar study found comparable results, along with with increased anthropomorphism leading to increased trust as well [4].

But, in conflict with the previously mentioned studies, Abdulrahman et al. found no significant differences between the human-like voice and the machine-generated voice concerning co-presence perception, trust, or working alliance [1]. Conflicting results are also present in Im et al., where participants even preferred the synthetic voice over the human voice for functional tasks, but again found no significant effect for the social task [15].

These conflicting results might indicate that the type of task the robots are made to perform has an impact on human perception, or at least trust, as almost all studies have this as a metric. This theory is further supported by Li et al. [22], and even more prominently by Im et al. [15].

Conflicting results could also be attributed to the fact that trust is not a uniquely defined concept within the field of human-robot interaction. A widely accepted definition of trust in robotic automation describes it as the need for reliance in situations of uncertainty and vulnerability [21] [4]. Despite trust being an important and essential aspect of human-robot interaction, there is also the risk of over-trusting. This phenomenon happens when persons place too much trust in robots and autonomous systems, overestimating their capabilities and competence [2]. This leads to the possibility of purposeful deception and has significant ethical implications [13].

Further supporting the theory that a robot's voice may impact the perception of the job or task a person is performing, lies within the wider field of human psychology. Within the field it has been widely studied on several aspects of voice. Voice pitch has been shown to impact the perception of leadership capacity, with lower-pitched voices being perceived as more competent, stronger, and trustworthy [17] [18]. On the other hand, for physicians it was found a moderate pitch range combined with a warm emotional tone is associated with higher patient satisfaction [23]. While for teachers the preference seems to fall on lower-pitched voice, receiving higher teaching evaluations [3]. Even something like the hoarseness of a lecturer can have effect on students' satisfaction and can even impede their cognitive performance [27].

To be able to properly and correctly conduct our study some specific questions do need a definitive answer. One of these is: What are the most commonly reported effects of robot voices on human perception of robots? By and large, participants in studies report feeling more comfortable

and experiencing more trust with a more human-like voice in a robot while the opposite was true for the more synthetic voices. [28] [4] [19]. Another effect lies in the pitch of voice, a more feminine voice, or higher pitch, tends to be perceived more positively [5] [12], but not necessarily more trustworthy. These voices might actually cause a decrease in the robot's persuasiveness [29].

Another question we must ask ourselves, is: Is there any evidence of interaction effects between task and robot voice when directly studied or in contradicting findings across multiple studies? This topic has been discussed previously within related works, but for the purposes of answering this question definitively, we have created a comprehensive table discussing multiple experiments conducted by researchers in the HRI field.

Name of the article	doi	Type of task(s)	Type of voice(s)	Conclusion
Influence of Robots' Voice Naturalness on Trust and Compliance	https://doi.org /10.1145/3706066	The game Battleship (robot assists player)	Low-pitch natural voice and a neutral pitch mechanical voice	The voice suitablity of a robot depends on the scenario it finds itself in
The Effect of Robot Attentional Behaviors on User Perceptions and Behaviors in a Simulated Health Care Interaction: Randomized Controlled Trial	https://doi.org /10.2196/13667	Health care receptionist	Neutral robot voice and a pitch changing robot voice	Pitch changes during speech does not have an effect on perception
The power of voice! The impact of robot receptionists' voice pitch and communication style on customer value cocreation intention	https://doi.org /10.1016/j.ijhm. 2024.103819	Receptionist	Low-pitched and high-pitched synthesized voices	A high-pitched voice in a robot is perceived more favourably
Let voice assistants sound like a machine: Voice and task type effects on perceived fluency, competence, and consumer attitude	https://doi.org /10.1016/j.chb. 2023.107791	Voice assistant (social vs functional task)	Human vs. synthetic voice	Participants preferred the synthetic voice over human voice.
Robot Voices in Daily Life: Vocal Human-Likeness and Application Context as Determinants of User Acceptance	https://doi.org /10.3389/fpsyg .2022.787499	No task, observation only	Five female voices with varying degrees of human-likeness	More human-like voices were perceived more favourably.
Invoking and identifying task-oriented interlocutor confusion in human-robot interaction	https://doi.org /10.3389/frobt .2023.1244381	Word problem solving (robot assists player)	Synthetic voice	Voice had an effect on study outcomes
If it looks like a human and speaks like a human ... Communication and cooperation in strategic Human–Robot interactions	https://doi.org /10.1016/j.socec .2023.102011	Prisoner's dilemma	Human-like synthesized voice	Manner of speech can effect a human's perception of a robot.
Stereotypes or golden rules? Exploring likable voice traits of social robots as active aging companions for tech-savvy baby boomers in Taiwan	https://doi.org /10.1016/j.chb .2018.02.025	Aging companions	Female and male voices, Old and young voices, extroverted and introverted voices	The results indicate that most participants preferred female, extroverted voices.

Table 1: A list of HRI studies

In table 1, one is clearly able to tell that there have indeed been contradictory findings across multiple studies regarding the existence of an effect of the voice a robot uses. These studies include the findings of Im et al. [15] where the participants preferred a synthetic voice over a human voice, while studies like Schreibelmayr et al. [28] showed a clear preference towards human-like voices. Because these studies use different tasks, like receptionist [24], or assisting during a game [4], we are able to conclude that there is a possible interaction effect between voice and task.

3 Research Question

All the studies discussed previously, provide us with probable cause to conduct research in the hopes of answering the question: *Is there an interaction effect of voice and task on the human perception of a robot, measured on trust, ease of use, and likeability?*. With trust, ease of use, and likeability chosen as measures, as these were the most commonly used in previously conducted research in the field of Human-Robot Interaction.

Based on the previously conducted research discussed in *Related Works*, we derive the following main hypotheses about our interaction effects:

H1: There will be a significant interaction between the type of robot voice and the type of task on perceived trust in the robot.

H2: There will be a significant interaction between voice type and task type on perceived ease of use of the robot.

H3: There will be a significant interaction between voice type and task type on the likeability ratings of the robot.

4 Methodology

4.1 Experiment Design

To answer our main research question, an experiment in the form of an in-person study is conducted, based on the literature study and the results of our sub-questions. This study uses a 2x2 factorial design, using trust, ease of use, and likeability as the measures.

For the independent variables, we look at voice type, human-like voice versus mechanical voice, and task type, a social companion versus a teacher. The participants experience only one of the four scenarios, creating a between subjects design. Only the task and voice should differ, gestures, responses, and mannerisms are as identical as possible in all four groups.

The speech and listing are facilitated through OpenAI's Whisper (faster-whisper), NAO's built in TTS feature, and speak's DAISYS API, with the DAISYS voice lines creation and Whisper's STT being performed on a remote server before being sent to the NAO robot. NAO's built in TTS feature is used as the mechanical voice within the experiment, while DAISYS API facilitates the human-like voice. DAISYS API offers the generation of multiple types of voices; for our voice the "Conversational" type was chosen and generated one called "Adam", which is used exclusively throughout the entire experiment. Conversation is generated through OpenAI's GPT-4o mini LLM, with the following prompts for the two different tasks:

Social task	Functional task
"You are a robot that takes on a life coach role towards the person you are speaking to."	"You are a robot that assists players with solving sudoku's."
"You cannot help with anything else. Always speak in plain English, no more than a 50 words per response."	"You cannot help with anything else. Always speak in plain English, no more than a 50 words per response."
"Avoid lists, code, or technical formatting."	"Avoid lists, code, or technical formatting."
"Speak naturally as if talking to a human and always stay on the topic of giving advice about life."	"Speak naturally as if talking to a human and always stay on the topic of sudoku's."

Table 2: The task prompts for the LLM used during the experiment

Before the interaction starts the participant receives a brief explanation, detailing the proceedings of the experiment, for example, that the robot takes some time to process their answer and does not respond immediately and a summary of the task they are partaking in.

The social task consists of a robot attempting to form a friendly connection with the participant, by asking personal questions and making small talk. The robot is taking on the role of a life coach in this scenario. They ask the participant questions about how their life is going and provide emotional support.

The functional task consists of the participants attempting to solve a sudoku with the assistance of the robot. The robot explains at the start of the interaction by explaining how a sudoku works and assist when the player gets stuck by providing a hint, or explaining the rules of sudoku again if the player asks for these things. The LLM also has access to a hint based on the current state of the sudoku generated by a back-tracking algorithm. This is only provided if the participant explicitly asks for one. To the right, 1, shows the general experiment setup, with a paper copy of the sudoku placed in front of the participant. The researcher observes the input of the participants and, in turn, inputs that into the sudoku program, creating an accurate digital copy for the prompts for the LLM.

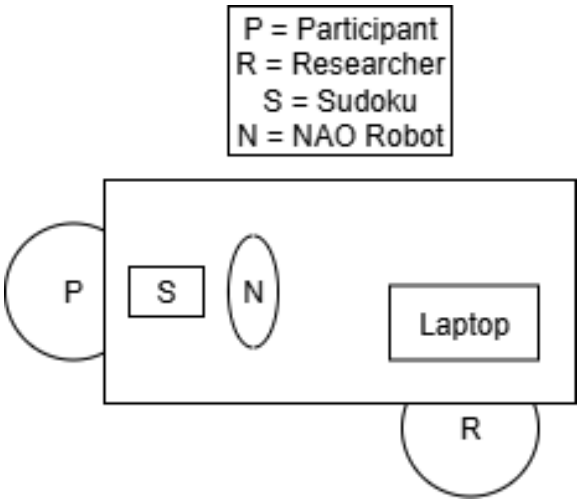

Figure 1: Diagram of the experiment setup with the NAO robot and participant.

These interactions go on for approximately 5 minutes, or until the interaction finds a natural conclusion. Each response is recorded in the "conversation history" and is added to the prompts the LLM uses to generate new responses. The introduction phase lasts approximately 40 seconds. To make sure the experiment does not cut off abruptly for the participant, the program is set-up in three different phases, the introduction phase, the task phase, and the conclusion phase. After certain time-thresholds, the program shifts into the next phase, as to keep the experiment on schedule. These phases are described to the LLM, along with the task prompt, as follows:

Introduction phase	Task phase	Conclusion phase
"The researcher, [Name], first introduces you to the participant. After that only the participant is speaking to you."	"Be curious"	"Explain that due to time constraints this will be the end of your interaction."
"Introduce yourself, your name is Charlie and you are a robot, designed to do a task."	For functional task: "Explain the rules of sudoku if asked", "Only provide a hint if the participant asks for it", "You know the contents of the sudoku puzzle"	"Conclude your interaction, say goodbye and thank the participant for their time."
"Asks the participants name."	For the social task: "Give tips that people can apply in their daily life."	"YOU MUST IMMEDIATELY END THE CONVERSATION."
"Explain the task you were designed to do"		

Table 3: The phase prompts for the LLM used during the experiment

Afterwards the experiment has concluded, the participants have to fill out the provided questionnaire (see Appendix A), which marks the end of the experiment. The researcher does answer questions about the questionnaire in case the participant does not understand a word or concept; questions pertaining to desired results are not be answered.

5 Results

5.1 Participants

In total, 40 participant took part in the study. Participants mostly self-identified as "Woman" (23/40, 57.5%), followed by "Man" (15/40, 37.5%) and "Non-binary" (2/40 5.0%), and ranged in age from 18 to 62 years old (mean=40.20, SD=14.64). The most prevalent highest achieved education among the participants was both a high school diploma and a bachelor's diploma respectively(17/40, 42.5%). All participants reported their native language as Dutch (40/40, 100%) and all but one reported their country of residence as the Netherlands (39/40, 97.5%), with the outlier having Australian residency.

These 40 participants all took part in one of the four different versions of the experiment. How many participants were in each experimental condition is shown below:

Voice Type	**Life Coach**	**Sudoku**
Human	10	9
Synthetic	11	10

Table 4: Number of Participants per Experimental Condition

5.2 Analysis Overview

Data from the post-experiment questionnaire (see Appendix A) yielded 40 relevant dependent variables, which were grouped into 9 composite variables in line with the guidelines of the Godspeed Questionnaire and the MDMT.

- **Godspeed composites (5):** Anthropomorphism, Animacy, Likeability, Perceived Intelligence, Perceived Safety
- **MDMT composites (4):** Reliable, Capable, Ethical, Sincere

Each composite variable was calculated as the mean of its associated dependent variables.

For the statistical model a Two-Way ANOVO also known as MANOVA was used to analyze the data. This approach was decided on based on the amount of composite variables and the amount of participants that took part in the study. MANOVA is also explicitly designed to handle intercorrelated outcomes, which all outcomes should, in theory, be.

From the MANOVA method, the following results were able to be gathered, as shown in table 5. All composite variables are shown under their respective questionnaire with the p-values shown for each factor and the interaction between the two. MANOVA was used specifically to evaluate the multivariate effects of Voice and Task across the composite variables, and because of its reduced risk of false positives.

Variable	**Voice Factor**	**Task Factor**	**Interaction**
Godspeed Questionnaire (ANOVA)			
Anthropomorphism	0.658	0.005	0.981
Animacy	0.596	0.148	0.543
Likeability	0.156	0.612	0.220
Perceived Intelligence	0.122	0.064	0.363
Perceived Safety	0.420	0.501	0.876
MDMT Questionnaire (ANOVA)			
Reliable	0.889	0.043	0.169
Capable	0.810	0.518	0.187
Ethical	0.738	0.701	0.799
Sincere	0.848	0.164	0.786

Table 5: P-values for ANOVAs and Mixed Effects Models of Godspeed and MDMT Composite Variables

5.3 Significant Findings

Only two composite variables showed significant effects for the MANOVA statistical model used. Notably, both were were associated with Task Type factor, and none were found for the Voice Type factor and the interaction between the two factors.

Measure	**Effect**	**Means**	**F / z**	**p-value**
Anthropomorphism	Task_Factor	L = 2.83, S = 2.38	F = 8.87	.005
Reliable	Task_Factor	L = 4.36, S = 3.95	F = 4.41	.043

Table 6: Summary of Significant Effects (p ¡ .05), with Group Means by Factor Level

The Task Type (Life Coach vs. Sudoku) was the only independent variable that had significant findings, namely within the Godspeed Questionnaire, the Anthropomorphism composite variable, and within the MDMT, the Reliable composite variable. Notably, Anthropomorphism showed significantly stronger evidence of a difference between tasks with $p < 0.005$, as opposed to the Reliable variable with $P < 0.043$.

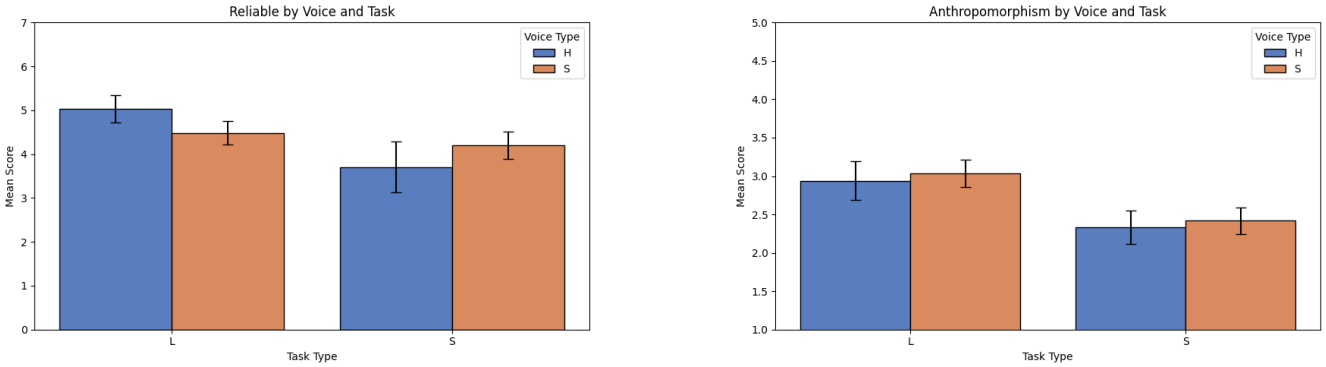

Figure 2: Significant main effects of task type on reliable (left) and anthropomorphism (right).

5.4 Composite Variable Means

Figure 3 displays the mean scores of all nine composite variables, as established in *Analysis Overview*, across the four experimental conditions. The plot illustrates that participants generally rated the robot high on Likeability, but more in the case of a human-like life coach. It also illustrates that there is a noticeable difference in-between ratings for the Reliable and Capable variables.

It should be noted that the independent variables of the Godspeed Questionnaire and the MDMT function on different scales. The Godspeed Questionnaire, thus Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived safety, were all rated on a scale of 1 to 5, while the MDMT, thus Reliable, Capable, Ethical, and Sincere, were all rated on a scale of 0 to 7, with an added option of "Does Not Fit", which was interpreted as a missing entry for the analysis.

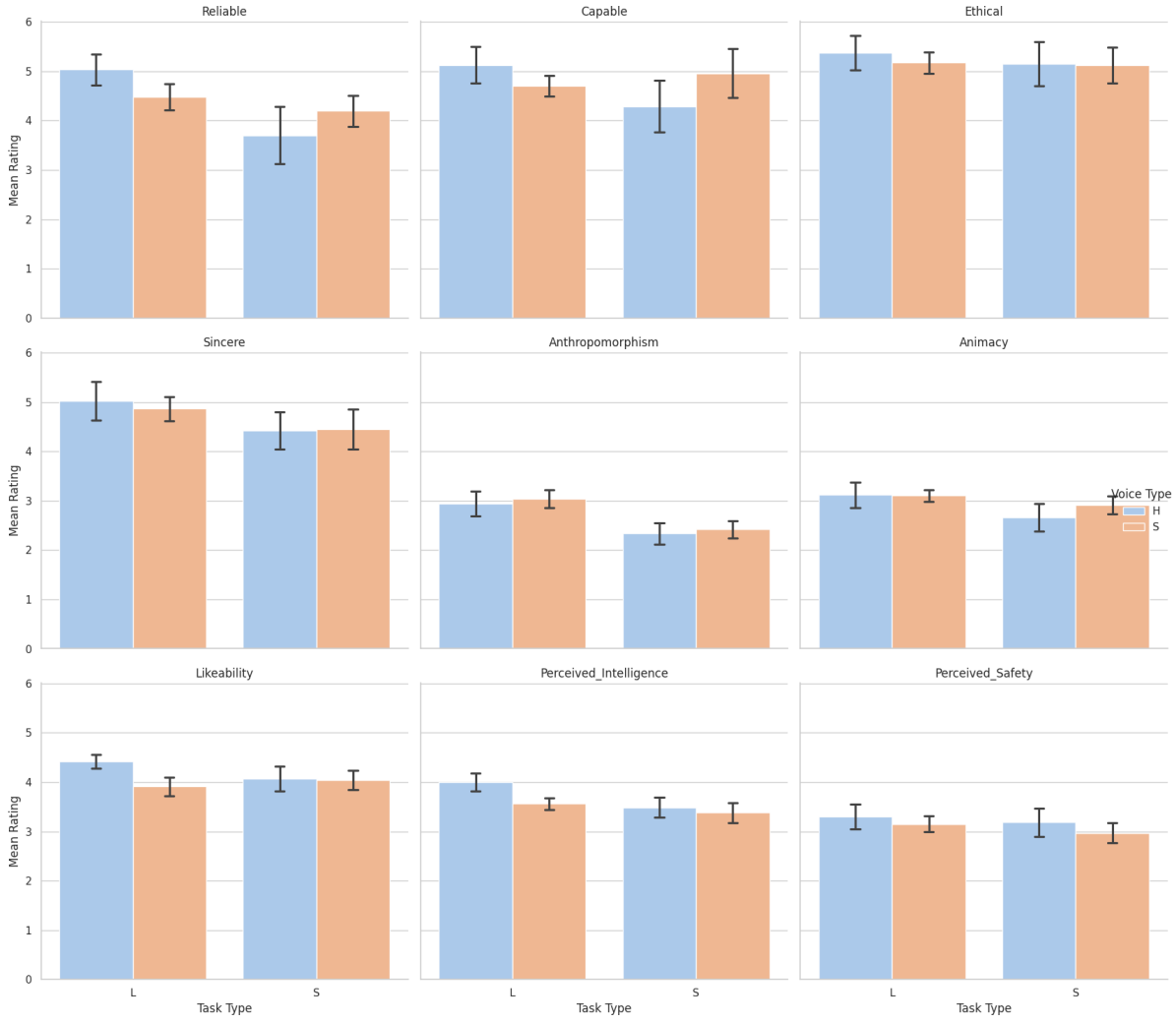

Figure 3: Mean scores for the nine composite variables by experimental condition. Error bars represent standard error of the mean.

6 Discussion

This study examined how the interaction between robot voice type (Human vs. Synthetic) and task type (Life Coach vs. Sudoku) influences key perceptions of trust, ease of use, and likeability in a human-robot interaction setting.

Hypothesis 1 predicted a significant interaction effect on perceived trust. However, the results did not support this assumption. While the analysis did reveal a significant main effect of task type on the Reliable dimension of trust, indicating that users found the robot more reliable in one task context over another (with functional tasks likely reducing perceived reliability), there was no significant interaction between voice and task type. This suggests that the combination of a human-like voice with a social task, or a mechanical voice with a functional task, did not produce any additive or synergistic effect on trust perceptions. Instead, it appears that task framing alone plays a more dominant role in shaping cognitive trust judgments such as reliability. The absence of interaction effects implies that users may assess trustworthiness based primarily on the robot's perceived role or function, rather than how it sounds [22]. This suggests that the nature of the task may play a more critical role in shaping trust than vocal characteristics, aligning with previous findings emphasising the influence of task type on trust in robots [15].

Regarding *Hypothesis 2*, which proposed an interaction effect on perceived ease of use, the analyses did not reveal significant interaction effects. Neither voice type, task type, nor their interaction had a statistically significant effect on how easy participants found the robot to use. This outcome suggests that users' judgments of ease of use may be relatively stable across different combinations of voice and task framing — at least in the short, controlled interactions used in this study. One possible explanation is that ease of use is more closely tied to the robot's physical interface (e.g., appearance, movements) than to more social or contextual features like voice or task framing [10]. From a design standpoint, these results imply that improving functional usability — through interface design, feedback timing, and clear instruction — may be more effective than modifying social cues when the goal is to enhance ease of use.

It is possible that participants' judgments of ease of use were influenced by factors outside the experiments control. The DAISYS speech API, which is used to generate the human-like voice, was prone to delay the response of the robot, impacting the overall user experience, and possibly all results related to the human-like voice, but significantly the ease of use, as was made clear verbally during the proceedings of the experiment and in the provided comment section of the questionnaire. Though this did not present itself in any of the findings, and is merely speculation.

Hypothesis 3 expected an interaction effect on likeability ratings. This hypothesis was not supported by the data, as no significant interaction effect was observed. While prior research has suggested that human-like voice cues may increase affective responses such as likeability, particularly in socially framed tasks [28] [19], the current findings indicate that these features did not interact meaningfully in shaping user preferences. Interestingly, although likeability itself was not significantly affected, task type did show a significant main effect on anthropomorphism - a construct closely related to how socially relatable and emotionally appealing a robot is perceived to be[7]. This suggests that participants were more likely to attribute human-like qualities to the robot in certain task contexts,

in this case a life coach. One possible interpretation is that likeability, unlike anthropomorphism, is a more stable and less malleable trait judgment that is not easily influenced by brief interaction conditions. It may also reflect individual user biases or expectations that go beyond voice or task framing.

In summary, none of the three hypothesised interaction effects between robot voice type and task type on trust, ease of use, or likeability were supported by the data. However, significant main effects of task type were found for both trust (specifically the Reliable dimension) and anthropomorphism, highlighting the influence of task framing on users' cognitive and affective perceptions of robots. These findings suggest that task context plays a more substantial role than vocal characteristics in shaping user evaluations, particularly in short-term interactions. Ease of use appeared unaffected by either factor, possibly due to technical limitations such as voice-related response delays, or because users prioritise functional interface features over social cues in usability judgments.

6.1 Limitations and Future Work

Limitations of this study include the relatively small and homogenous sample size, which may reduce statistical power and limit generalizability. The reliance on self-reported measures is another limitation, as such data can be influenced by social desirability or subjective biases. The experiment was also not held at a consistent location with a consistent set-up nor were all participants random, as several had a prior relation with the researcher, thus environmental and social factors might have had an undesired effect on the results. Between subjects, there was the additional problem of server response times in regards to the NAO robot itself, but mainly the DAISYS API. Another factor of limitation is the short exposure time of the experiment [20]. Participants only spent 5 minutes interacting with the NAO robot, not enough to build a bond or experience all possible facets of an interaction. The tasks chosen for this particular study also had limited generalizability to real-world interactions, a life coach or sudoku assistant robot have not yet seen wide-spread usage.

Future research should consider larger and more diverse samples, integrate objective behavioural or physiological metrics, and explore additional voice types and tasks to deepen understanding of these effects. Along with aiming to retain more control of the general experiment environment, such as consistent experimental conditions, and consistent response times between the voice factors. A longer interaction with the participants could also be considered for a follow-up study, providing a more established relationship with the NAO robot. Along with tasks and a setting that reflects real-world scenarios and current usage of assistive robots. Furthermore, this findings indicate that for interactive robots, task framing and contextual alignment may be more impactful than emphasizing voice naturalness, at least in the early stages of interaction.

7 Conclusion

This study set out to examine whether there is an interaction effect between a robot's voice type (human-like vs. synthetic) and the type of task it performs (social vs. functional) on user perceptions of trust, ease of use, and likeability. Drawing on insights from existing literature, it was hypothesised that these two design factors would jointly influence how users evaluate and relate to robots. A 2x2 factorial experiment involving 40 participants interacting with a NAO robot was conducted, using

validated measures from the Godspeed and MDMT questionnaires to assess user perceptions.

Contrary to the proposed hypotheses, no significant interaction effects between voice type and task type were observed on any of the three measured dimensions—trust, ease of use, or likeability. However, the results did reveal significant main effects of task type on two variables. Specifically, task type influenced perceptions of reliability, a core dimension of trust, and anthropomorphism, which is closely associated with the affective and social evaluation of robots. These findings suggest that how a robot's role or function is framed has a stronger impact on user perceptions than whether the robot uses a human-like or mechanical voice.

The lack of significant effects for perceived ease of use raises important considerations. While the interface and interaction mechanics were held constant across conditions, qualitative feedback indicated that technical delays—especially those associated with the human-like voice API—may have negatively affected user experiences. Although speculative, such limitations may have masked potential effects and highlight the importance of stable performance in HRI studies when evaluating constructs like usability.

Overall, this study contributes to understanding how social and functional cues influence human-robot interaction. The findings underscore the importance of task context in shaping user expectations and evaluations, particularly regarding trust and perceived human-likeness.

References

[1] Amal Abdulrahman, Deborah Richards, and Ayse Aysin Bilgin. A comparison of human and machine-generated voice. In *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology*, VRST '19, New York, NY, USA, 2019. Association for Computing Machinery.

[2] Alexander M. Aroyo, Jan de Bruyne, Orian Dheu, Eduard Fosch-Villaronga, Aleksei Gudkov, Holly Hoch, Steve Jones, Christoph Lutz, Henrik Sætra, Mads Solberg, and Aurelia Tamò-Larrieux. Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. *Paladyn, Journal of Behavioral Robotics*, 12(1):423–436, 2021.

[3] Thomas Augustin. Voice pitch influences on teaching evaluations and student learning. 2018.

[4] Dennis Becker, Lukas Braach, Lennart Clasmeier, Teresa Kaufmann, Oskar Ong, Kyra Ahrens, Connor Gäde, Erik Strahl, Di Fu, and Stefan Wermter. Influence of robots' voice naturalness on trust and compliance. *J. Hum.-Robot Interact.*, 14(2), January 2025.

[5] Sofie Behrens, Anne Egsvang, Michael Hansen, and Anton Møllegård-Schroll. Gendered robot voices and their influence on trust. pages 63–64, 03 2018.

[6] Pascal Belin, Patricia E. Bestelmeyer, Marianne Latinus, and Rebecca Watson. Understanding voice perception. *British Journal of Psychology*, 102(4):711–725, Jun 2011.

[7] Markus Blut, Cheng Wang, Nancy V Wünderlich, and Christian Brock. Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other ai. *Journal of the academy of marketing science*, 49(4):632–658, 2021.

[8] Gordon Briggs and Matthias Scheutz. How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *Int. J. Soc. Robot.*, 6(3):343–355, August 2014.

[9] Elizabeth Cha, Anca D. Dragan, and Siddhartha S. Srinivasa. Perceived robot capability. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 541–548, 2015.

[10] Sara M. Conklin, Richard J. Koubek, James A. Thurman, and Leah C. Newman. The effects of aesthetics and cognitive, style on perceived usability. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(18):2153–2157, 2006.

[11] Charles R Crowelly, Michael Villanoy, Matthias Scheutzz, and Paul Schermerhornz. Gendered voice and robot entities: perceptions and reactions of male and female subjects. In *2009 IEEE/RSJ international conference on intelligent robots and systems*, pages 3735–3741. IEEE, 2009.

[12] Yi Ding, Ran Guo, Wei Lyu, and Wengang Zhang. Gender effect in human–machine communication: a neurophysiological study. *Frontiers in Human Neuroscience*, Volume 18 - 2024, 2024.

[13] Raffaella Esposito, Alessandra Rossi, and Silvia Rossi. Deception in hri and its implications: a systematic review. *J. Hum.-Robot Interact.*, February 2025. Just Accepted.

[14] Martin Ford. *2. The Rise of the Robots: Impact on Unemployment and Inequality*, pages 27–45. Cornell University Press, Ithaca, NY, 2018.

[15] Hyunjoo Im, Billy Sung, Garim Lee, and Keegan Qi Xian Kok. Let voice assistants sound like a machine: Voice and task type effects on perceived fluency, competence, and consumer attitude. *Computers in Human Behavior*, 145:107791, 2023.

[16] Deborah L Johanson, Ho Seok Ahn, Bruce A MacDonald, Byeong Kyu Ahn, Jongyoon Lim, Euijun Hwang, Craig J Sutherland, and Elizabeth Broadbent. The effect of robot attentional behaviors on user perceptions and behaviors in a simulated health care interaction: Randomized controlled trial. *J. Med. Internet Res.*, 21(10):e13667, October 2019.

[17] Casey A Klofstad, Rindy C Anderson, and Susan Peters. Sounds like a winner: Voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences*, 279(1738):2698–2704, 2012.

[18] Lisa Korenman, Elizabeth Wetzler, Samantha Leahy, and Lissa V Young. Voices of leadership: The effects of voice pitch on percieved leadership capabilities. *Advancing Women in Leadership Journal*, 42:123–131, 2023.

[19] Katharina Kühne, Martin H Fischer, and Yuefang Zhou. The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study. *Frontiers in neurorobotics*, 14:593732, 2020.

[20] Guy Laban, Arvid Kappas, Val Morrison, and Emily S. Cross. User experience of human-robot long-term interactions. In *Proceedings of the 10th International Conference on Human-Agent Interaction*, HAI '22, page 287–289, New York, NY, USA, 2022. Association for Computing Machinery.

[21] John D Lee and Katrina A See. Trust in automation: designing for appropriate reliance. *Hum. Factors*, 46(1):50–80, 2004.

[22] Dingjun Li, P. L. Rau, and Ye Li. A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics*, 2(2):175–186, May 2010.

[23] Jingfang Liu and Huihong Jiang and. Exploring the effects of online physician voice pitch range and filled pauses on patient satisfaction in mobile health communication. *Health Communication*, 39(13):3258–3271, 2024. PMID: 38314782.

[24] Xiao-Xin Liu, Cheng-Yue Yin, and Meng-Ran Li. The power of voice! the impact of robot receptionists' voice pitch and communication style on customer value cocreation intention. *Int. J. Hosp. Manag.*, 122(103819):103819, September 2024.

[25] Conor McGinn and Ilaria Torre. Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots, 2019.

[26] Sara Pearsell and Daniel Pape. The effects of different voice qualities on the perceived personality of a speaker. *Frontiers in Communication*, 7, 2023.

[27] Isabel S Schiller, Carolin Breuer, Lukas Aspöck, Jonathan Ehret, Andrea Bönsch, Torsten W Kuhlen, Janina Fels, and Sabine J Schlittmeier. A lecturer's voice quality and its effect on memory, listening effort, and perception in a vr environment. *Scientific Reports*, 14(1):12407, 2024.

[28] Simon Schreibelmayr and Martina Mara. Robot voices in daily life: Vocal human-likeness and application context as determinants of user acceptance. *Frontiers in Psychology*, 13, 2022.

[29] Sichao Song, Jun Baba, Junya Nakanishi, Yuichiro Yoshikawa, and Hiroshi Ishiguro. Mind the voice!: Effect of robot voice pitch, robot voice gender, and user gender on user perception of teleoperated robots. pages 1–8, 04 2020.

[30] Ilaria Torre, Jeremy Goslin, Laurence White, and Debora Zanatto. Trust in artificial voices: A" congruency effect" of first impressions and behavioural experience. In *Proceedings of the technology, mind, and society*, pages 1–6. 2018.

Appendix: Supplementary Material

Appendix A:

Universiteit Leiden
The Netherlands

Experiment Questionnaire

Please fill in the following personal information, write your answers on the provided line or circle the most fitting answer. No identifiable information will be shared to anyone or any institution. If you are not comfortable with sharing or do not want to share certain information, you are kindly asked to leave the question blank.

Age:_____

Educational level:

Elementary school / High school /
Bachelor's / Master's / PhD

Gender: _____

Native language:_____

Occupation/Field of study:

_____

Country of residence:

_____

Comments on the experiment:

_____

_____

_____

_____

Please answer the following questions to the best of your abilities. You will be presented with two different words from the opposite ends of a spectrum, mark on the provided scale where you would categorize the NAO robot from your experience.

Anthropomorphism

Fake	1	2	3	4	5	Natural
Machinelike	1	2	3	4	5	Humanlike
Unconscious	1	2	3	4	5	Conscious
Artificial	1	2	3	4	5	Lifelike
Moving rigidly	1	2	3	4	5	Moving elegantly

Animacy

Dead	1	2	3	4	5	Alive
Stagnant	1	2	3	4	5	Lively
Mechanical	1	2	3	4	5	Organic
Artificial	1	2	3	4	5	Lifelike
Inert	1	2	3	4	5	Interactive
Apathetic	1	2	3	4	5	Responsive

Likeability

Dislike	1	2	3	4	5	Like
Unfriendly	1	2	3	4	5	Friendly
Unkind	1	2	3	4	5	Kind
Unpleasant	1	2	3	4	5	Pleasant
Awful	1	2	3	4	5	Nice

Perceived Intelligence

Incompetent	1	2	3	4	5	Competent
Ignorant	1	2	3	4	5	Knowledgeable
Irresponsible	1	2	3	4	5	Responsible
Unintelligent	1	2	3	4	5	Intelligent
Foolish	1	2	3	4	5	Sensible

Perceived Safety

Anxious	1	2	3	4	5	Relaxed
Calm	1	2	3	4	5	Agitated
Still	1	2	3	4	5	Surprised

Please rate the robot using the scale from 0 (Not at all) to 7 (Very). If a particular item does not seem to fit the robot in the situation, please select the option that says “Does Not Fit.”

	Not at all 0	1	2	3	4	5	6	Very 7	Does Not Fit
Reliable	○	○	○	○	○	○	○	○	○
Sincere	○	○	○	○	○	○	○	○	○
Capable	○	○	○	○	○	○	○	○	○
Ethical	○	○	○	○	○	○	○	○	○
Predictable	○	○	○	○	○	○	○	○	○
Genuine	○	○	○	○	○	○	○	○	○
Skilled	○	○	○	○	○	○	○	○	○
Respectable	○	○	○	○	○	○	○	○	○
Someone you can count on	○	○	○	○	○	○	○	○	○
Candid	○	○	○	○	○	○	○	○	○
Competent	○	○	○	○	○	○	○	○	○
Principled	○	○	○	○	○	○	○	○	○
Consistent	○	○	○	○	○	○	○	○	○
Authentic	○	○	○	○	○	○	○	○	○
Meticulous	○	○	○	○	○	○	○	○	○
Has integrity	○	○	○	○	○	○	○	○	○
	Not at all 0	1	2	3	4	5	6	Very 7	Does Not Fit