Artificial Voice on Humanoid Robot Ameca

Lala Jafarova
Supervised by Prof. Dr. Karsten Berns, MSc. Ashita Ashok

***Abstract*—The transdisciplinary field of human-robot interaction explores how people interact with robots. Users may find an expressive and natural-sounding artificial voice to be more comfortable and intuitive during interactions. Using various Text-to-Speech models, this study explores the voice profiles that the humanoid robot Ameca prefers. Twenty-four college students participated in the study and heard the several voice profiles offered by Ameca. Surveys assessing voice impression and quality as well as standardised monologues were used to gauge participants' opinions on the robot's voice. Significant changes in user perceptions based on speech profiles were found using repeated measures ANOVA results, underscoring the importance of voice in improving human-robot interactions. These results should help designers create more engaging and intuitive humanoid robots, which will make it easier for them to blend in with daily life and increase user acceptability and engagement. This research makes a vital contribution to the larger objective of making humanoid robots trustworthy and valued companions in a variety of social circumstances by studying and optimising voice features.**

***Index Terms*—Human-robot interaction (HRI), humanoid robots, voice characteristics, user perception, and Text-to-Speech (TTS)**

I. Introduction

HUMAN-Robot Interaction (HRI) has advanced significantly as a result of the convergence of science, engineering, and technology, changing our perception of and interactions with robotic objects. The multidisciplinary field of human-robot interaction (HRI) examines human-robot interaction. Robots are becoming more and more integrated into human-centered environments, interacting with people and taking part in a variety of activities as a result of ongoing advancements in robotic technology. This paradigm change highlights the importance of HRI research, which aims to promote social and engaging interactions between humans and robots, especially in vital areas like daily personal support, healthcare, and education.[1].

HRI is essential to establishing effective channels of communication between humans and humanoid robots. The key to this relationship is enabling expressive and genuine conversation between the user and the robot. An artificial voice that closely mimics human speech patterns and intonations can significantly improve the intuitiveness and comfort of these interactions. Studies have indicated that speech traits are critical in determining how comfortable and trusting individuals feel while interacting with robots. Extroverted, high-pitched voices are generally preferred by users since they improve the user experience overall and are seen as trustworthy.[2].

Not only can a realistic and appealing voice be useful, but it can also greatly improve the user experience by fostering communication and mutual trust between the human and the robot. Improving the user experience is crucial to encouraging people to embrace and incorporate humanoid robots into daily life. Furthermore, the creation of sophisticated artificial voices facilitates the smooth integration of humanoid robots into human societies, which may change societal norms and expectations around the use of technology.

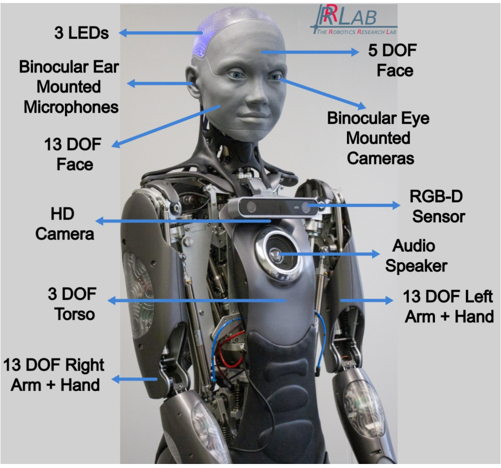


Fig. 1: This is the humanoid social robot, Ameca used in the presented HRI.

In this regard, our study leverages the various Text-to-Speech (TTS) systems to investigate the ideal voice profiles for humanoid robots. Ameca and other humanoid robots are trying to communicate with humans in a natural way, therefore it's important to give them voices that sound trustworthy and familiar in addition to providing information. These robots are capable of engaging in interactive behaviors that are similar to human interactions because of their humanoid form factor and an array of sensors, which include vision, touch, and audition. [3].

Ameca, the social robot employed in this research, is a ground-breaking development in social robotics because of its advanced capacity to recognize and predict human emotions and intentions through non-verbal clues such head positions, hand gestures, and facial expressions. Ameca also has a speaker on the chest, binocular eye-mounted cameras, eye-tracking based on face identification, and a ZED2 camera located on the chest in addition to binaural ear-attached microphones. These characteristics open the door for more organic and efficient HRI by enabling Ameca to interact in rich, human-like ways. (Figure 1.)

Our study aims to improve the conversational capacities of humanoid robots by utilising state-of-the-art technologies and incorporating prior research findings. This will help to close the gap between humans and robots and facilitate their smooth integration into human communities.

A. The Voices in different situations

The human voice can adjust its tone, pitch, and intensity to suit different contexts. This is due to a variety of factors, including emotional state, social dynamics, and cultural conventions. This flexibility makes it possible to communicate effectively in a variety of settings. The voice is a versatile and lively social tool with useful uses in self-presentation. Talkers have the ability to alter their voices in response to social or environmental cues. A study by Guldner et al. found that talkers adjust their voices in a shared trait space to fit social contexts. Beyond reactionary voice changes, talkers can also purposefully use their vocal behavior to reveal minor aspects of themselves to listeners. [4].

Moreover, vocal behaviour is shaped in part by cultural standards. Speech patterns, intonations, and loudness vary among cultures, which affects how people verbally express themselves in different contexts. In Kühne et al.'s study, for example [5], participants' self-reported proficiency in the Berlin dialect was positively correlated with the dependability they assigned to a dialect-speaking robot. Speaking Berlin accents, people found engaging with a robot more trustworthy and pleasant. The age, gender, amount of time the participants had lived in Berlin, and response device all had an impact on the evaluations. In the end, there was a positive correlation between the robot's perceived skill and reliability.

B. Gender effect

Despite the progress, little is understood about how a robot's gender depiction impacts users in real-world situations. "How will gendered representations affect everyday use?" is still the central question [6]. It's a well-known fact that robots are machines without a gender, but they also play social roles in our lives, such as friends, foes, relatives, and so forth. Three broad categories of humanoid robots, or androids, have been developed and marketed: entertainment robots (like toys), service robots (like receptionists or security guards), and companion robots (like teachers or house attendants). According to Carpenter et al.'s 2009 study [6], women felt less comfortable with robots in the home than men, with the degree of comfort with them approaching significance with gender as the between-subjects variable. They used videos of a non-human robot (Robovie) and a female robot (Repliee) that looked like humans. Nevertheless, some research indicates that women score lower than men on surveys measuring hostility towards robots [7].

Additionally, some research demonstrates the effective use of a voice-controlled robotic system that streamlines the learning process by enabling kids as young as four to communicate with robots in natural language [8]. Robust AI technologies, such as object detection and huge language models, have been used to improve the intuitiveness of robotic engagement for younger students. The voice-activated system allows young learners to engage with robotics and technology at an early age, creating new educational opportunities that were not available to them before. The report makes recommendations for possible advantages in several academic areas, including computational thinking, teamwork, and foundational courses like physics and maths.

C. Voice differences

The discussion over how people view social robots becomes more significant as humanoid robots proliferate. The field of how humans perceive synthetic voices is continually developing. According to the University of Potsdam study[9], women rated synthetic voices higher than men did on the majority of the measures. Surprisingly, voice evaluation was essentially unaffected by opinions about robots and interest in social robots. Male participants rated embodied male voices far higher than disembodied male voices, according to a different study by Crowell et al. Compared to the disembodied vocal entity, the subjects thought the robot was generally friendlier. Remarkably, regardless of the type of entity, judgements of female voices by male and female participants did not differ significantly [10].

D. Latest Text-to-Speech (TTS) Models

Advancements in text-to-speech (TTS) technologies have been remarkable in recent years, resulting in models that produce increasingly natural, expressive, and high-quality speech. This section highlights some of the latest TTS models, both open-source and commercially available, which showcase state-of-the-art techniques and features.

Open-source TTS models have democratized access to advanced speech synthesis technology, allowing researchers, developers, and enthusiasts to experiment and innovate. Here are some notable recent developments:

- **Matcha-TTS** (https://github.com/shivammehta25/Matcha-TTS): Matcha-TTS introduces a new encoder-decoder architecture designed for efficient TTS acoustic modeling. By utilizing optimal-transport conditional flow matching (OT-CFM), Matcha-TTS achieves a significant improvement in synthesis speed and quality. The model features an ODE-based (Ordinary Differential Equation) decoder that requires fewer synthesis steps compared to traditional models trained with score matching. Additionally, the careful design of the synthesis process ensures each step runs quickly, making Matcha-TTS a powerful option for real-time applications[11].
- **Parler-TTS** (https://github.com/huggingface/parler-tts/tree/main/parler_ tts): Parler-TTS is an open-source TTS model that focuses on generating high-quality, natural-sounding speech. One of its standout features is the ability to control various aspects of the synthesized speech through simple text prompts. Users can adjust parameters such as gender, background noise, speaking rate, pitch, and reverberation. This flexibility makes Parler-TTS a versatile tool for various applications. All datasets, pre-processing scripts, training code, and model weights are released under a permissive license, encouraging the community to build upon and enhance the model [12]
- **Style-TTS2**(https://github.com/yl4579/StyleTTS2): Style-TTS2 represents a significant evolution in the field

of TTS by leveraging style diffusion and adversarial training with large speech language models (SLMs). Unlike its predecessor, Style-TTS2 models styles as a latent random variable through diffusion models, allowing it to generate the most appropriate style for the text without requiring reference speech. The model benefits from efficient latent diffusion and the diverse synthesis capabilities of diffusion models. Additionally, large pre-trained SLMs, such as WavLM, are used as discriminators in the training process, along with novel differentiable duration modeling. This approach results in improved speech naturalness and expressiveness [13]

In addition to open-source models, several commercially available TTS solutions offer robust features and high-quality speech synthesis. These solutions are widely used in various industries, from virtual assistants to content creation.

- **Acapela**: Acapela Group provides a range of voices and TTS solutions, including the exciting Ameca voices, which are designed to deliver natural and expressive speech. Acapela's technology is used in numerous applications, from assistive devices to customer service platforms, offering voices that can be customized to meet specific needs and preferences.
- **Plus Q**: Plus Q offers a unique TTS solution with a gender-neutral voice, known as Q. Unlike traditional voice assistants like Alexa, Siri, or Google Assistant, Q is designed to reflect the diversity of human voices and reduce gender bias. The gender-neutral voice of Q is an important step towards inclusivity, providing a voice that can appeal to a broader audience and align with modern values of diversity and equality [14].
- **Amazon Polly**: mazon Polly is a well-known TTS service that uses advanced deep learning technologies to convert text into lifelike speech. With a broad set of languages and dozens of natural-sounding voices, Polly enables developers to create speech-activated applications and convert written content into spoken word. The versatility and scalability of Amazon Polly make it a popular choice for various applications, including news reading, e-learning, and interactive voice response (IVR) systems [15].

These models and solutions highlight the rapid advancements and diverse applications of TTS technology, enabling more natural, expressive, and accessible speech synthesis across different platforms and use cases.

E. Current study:

This study focuses on the quest for the optimal voice profiles for Ameca, the humanoid robot. As we delve into the intricacies of HRI, it becomes evident that the nuances of voice play a critical role in shaping users' perceptions and experiences. By understanding participants' preferences regarding these voice attributes, we aim to unlock insights into how Ameca's voice can affect humans's perception. We anticipate uncovering valuable insights that will pave the way for more intuitive and meaningful interactions between humans and Ameca, ultimately enhancing the integration of humanoid robots into our daily lives.

By integrating insights from previous research and leveraging cutting-edge technologies, our study endeavors to enhance the communicative capabilities of humanoid robots, thereby bridging the gap between humans and robots and paving the way for their seamless integration into human communities.

II. Methods

The main goal of this study is to investigate how different voice profiles of Ameca (the humanoid robot) affect human perception. Our null hypothesis is that there is no difference in human perception based on the different voice profiles of Ameca, while the alternative hypothesis posits that there is a difference in human perception based on the different voice profiles of Ameca.

A. Participants

The pilot study included 24 participants, all of whom were university students from the University of Kaiserslautern. The participants' ages ranged from 22 to 38 years, with a mean age of 27.083 years (SD = 3.752). The gender distribution was 16 females and 8 males. Inclusion criteria included having an intermediate level English proficiency level. Informed consent was obtained from all participants before their involvement in the study. The study was conducted within a controlled laboratory environment at the Robotics Research Lab at RPTU Kaiserslautern.

B. Materials

The materials utilized in this study included:

- **Ameca**: The humanoid robot used for conducting the experiments. Ameca's capabilities include advanced facial expressions and speech synthesis, making it suitable for human-robot interaction studies.
- **Acapela Lucy and Acapela Graham**: Two default synthetic voices representing a female and a male voice, respectively.
- **Plus Q**: The first genderless voice, which was created by using a wav file of the Plus Q voice in StyleTTS2, a state-of-the-art text-to-speech synthesis tool.
- **2 Amazon Polly - Joanna and Mathew** : These voices were generated using the trial option of Amazon Polly, an advanced text-to-speech service by Amazon.
- **Standardized Monologues for Each Voice** (See Appendix G): consistent scripts used for each voice condition to ensure uniformity across experimental trials. These monologues were designed to be neutral in content and length to prevent any bias that might arise from the script itself [16].
- **Post -Questionnaire**: This included two components: the Godspeed questionnaire to assess voice impression [17], and the Mean Opinion Score (MOS) to evaluate voice quality [18]. These instruments are well-validated measures in the field of human-robot interaction.
- **Pre -Questionnaire**: This included demographic information, a consent form, and questions about previous

human-robot interaction (HRI) experience [19]. This was used to control for any confounding variables related to participants' backgrounds and familiarity with robotic systems.

C. Experimental Design and Procedure

The study employed a within-subjects design to assess participants' preferences for specific voice profiles exhibited by Ameca. Each participant first completed the pre-questionnaire (Appendices A and B), which collected demographic information and prior experience with HRI.

The experimental procedure was as follows:

1) Participants were brought into the laboratory in groups of five.
2) Ameca presented each standardized monologue (Appendix E) in a randomized order, ensuring that each participant heard all five voice profiles (Graham, Joanna, Lucy, Matthew, and Plus Q) without any fixed sequence that could bias their responses.
3) After listening to each monologue, participants completed the related post-questionnaire (Appendices C and D), rating their impressions and the quality of the voice they had just heard.
4) This process was repeated for each voice profile, with participants providing feedback immediately after each presentation to ensure their impressions were accurately recorded.

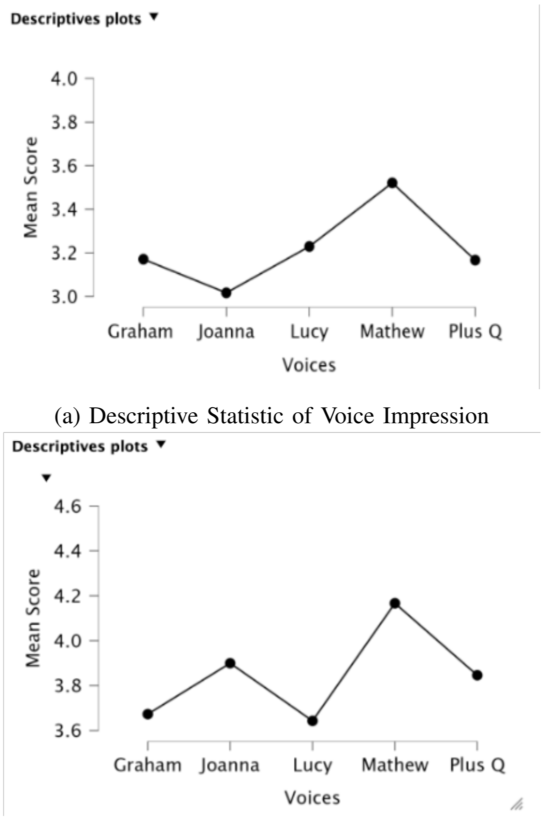
The experiment was conducted in the Robotics Research Lab at RPTU Kaiserslautern, where Ameca is located. The controlled environment ensured that external variables were minimized, and the group setting (five participants per group) facilitated an efficient data collection process while maintaining individual response integrity.

D. Data Preparation and Analysis

After the experiment, the collected data was organized in an Excel file. Each file was categorized by voice: V1 (Graham), V2 (Joanna), V3 (Lucy), V4 (Matthew), and V5 (Plus Q). These files were then converted into CSV format for analysis using JASP 0.18.3, a comprehensive statistical software package.

Data analysis involved the following steps:

1) **Descriptive Statistics**: Initial analysis included calculating the mean and standard deviation for each voice profile across all participants.
2) **Normality Check**: Tests such as the Shapiro-Wilk test were used to ensure that the data met the assumptions for parametric testing.
3) **Repeated Measures ANOVA**: This was used to determine if there were statistically significant differences in participants' impressions and perceived quality of the different voice profiles. The within-subjects design allowed us to control for individual differences by comparing each participant's responses across the different conditions.
4) **Post-hoc Analysis**: If the ANOVA indicated significant differences, post-hoc tests were conducted to identify which voice profiles differed significantly from each other.



(a) Descriptive Statistic of Voice Impression

(b) Descriptive Statistic of Voice Quality

Fig. 2: Comparison of Voice Impression and Voice Quality

The experiment aims to provide insights into preferred voice profiles for Ameca, contributing to the optimization of human-robot interactions. By analyzing participant responses and reverberation measurements, we can understand how these voices impact user experience and inform future developments in humanoid robot design.

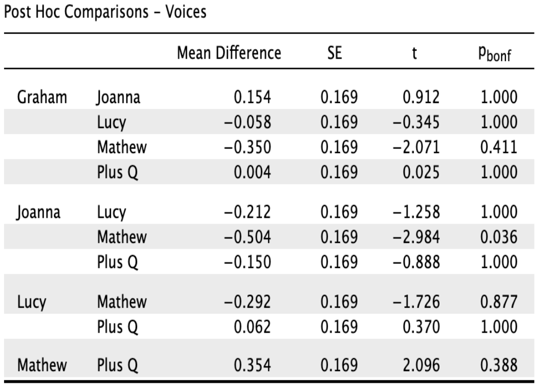
III. Results

A. Voice Impression

Voice impression was assessed using the Godspeed questionnaire [17], administered after participants listened to each voice profile. The questionnaire included measures of likeability, perceived intelligence, and perceived naturalness of the voice.

Descriptive statistics for the voice impression scores are shown in Figure 3b. Each voice profile (Graham, Joanna, Lucy, Matthew, Plus Q) was rated by participants, and the mean scores with standard deviations were presented.

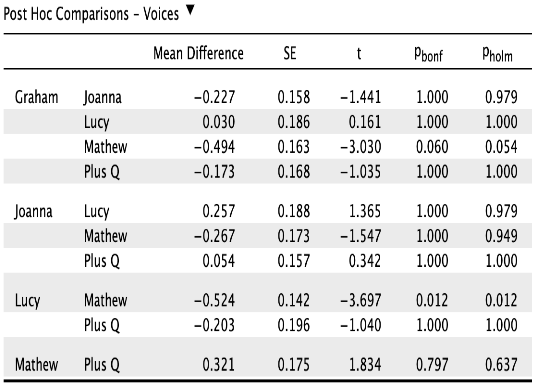
A repeated measures ANOVA was conducted to evaluate the differences in voice impression scores across the five voice profiles. The results indicated a trend toward significance, $p = 0.055$. While this p-value is slightly above the conventional threshold of 0.05, it suggests that there may be meaningful differences worth exploring further.



Post Hoc Comparisons – Voices

		Mean Difference	SE	t	p_{bonf}
Graham	Joanna	0.154	0.169	0.912	1.000
	Lucy	−0.058	0.169	−0.345	1.000
	Mathew	−0.350	0.169	−2.071	0.411
	Plus Q	0.004	0.169	0.025	1.000
Joanna	Lucy	−0.212	0.169	−1.258	1.000
	Mathew	−0.504	0.169	−2.984	0.036
	Plus Q	−0.150	0.169	−0.888	1.000
Lucy	Mathew	−0.292	0.169	−1.726	0.877
	Plus Q	0.062	0.169	0.370	1.000
Mathew	Plus Q	0.354	0.169	2.096	0.388

(a) Descriptive Statistic of Voice Impression



Post Hoc Comparisons – Voices ▼

		Mean Difference	SE	t	p_{bonf}	p_{holm}
Graham	Joanna	−0.227	0.158	−1.441	1.000	0.979
	Lucy	0.030	0.186	0.161	1.000	1.000
	Mathew	−0.494	0.163	−3.030	0.060	0.054
	Plus Q	−0.173	0.168	−1.035	1.000	1.000
Joanna	Lucy	0.257	0.188	1.365	1.000	0.979
	Mathew	−0.267	0.173	−1.547	1.000	0.949
	Plus Q	0.054	0.157	0.342	1.000	1.000
Lucy	Mathew	−0.524	0.142	−3.697	0.012	0.012
	Plus Q	−0.203	0.196	−1.040	1.000	1.000
Mathew	Plus Q	0.321	0.175	1.834	0.797	0.637

(b) Post-hoc comparisons for Voice Impression Scores

Fig. 3: Post-Hoc Comparisons for Voice Quality Score

Given the marginally non-significant result, post-hoc comparisons using the Bonferroni correction were conducted to identify potential differences between specific voice profiles. The posthoc results are summarized in Figure 3a.

These results indicate that the voice profiles of Lucy and Plus Q were significantly different from Graham's in terms of voice impression.

B. Voice Quality

Voice quality was assessed using the Mean Opinion Score (MOS) [18], administered after each voice presentation. The MOS provides a numerical indication of the perceived quality of the voice.

Descriptive statistics for the voice quality scores are presented in Figure 3b. Each voice profile was evaluated by participants, and the mean scores with standard deviations were displayed.

A repeated measures ANOVA was conducted to evaluate differences in voice quality scores across the five voice profiles. The results indicated a significant effect, $p = 0.02$, suggesting that participants perceived differences in voice quality among the profiles. Post-hoc comparisons using the Bonferroni correction were conducted to identify specific differences between voice profiles. The posthoc results for voice quality are summarized in Figure 3b

These results indicate that the voice profiles of Joanna and Matthew were significantly different from Graham in terms of voice quality.

C. Summary of Findings

The analysis revealed that there were significant differences in both voice impression and voice quality among the different voice profiles of Ameca. Specifically: Voice impression scores indicated a trend toward significance, with post-hoc tests showing that Lucy and Plus Q were significantly different from Graham. Voice quality scores showed a significant effect, with Joanna and Matthew being rated significantly different from Graham. These findings suggest that different voice profiles can indeed affect human perception, highlighting the importance of selecting appropriate voice characteristics for humanoid robots like Ameca to optimize human-robot interaction.

IV. Discussion

This study aimed to investigate how different voice profiles of Ameca, a humanoid robot, affect human perception. Our findings provide valuable insights into the significance of voice characteristics in human-robot interaction. Specifically, we found:

- A trend towards significance in voice impression, with post-hoc analyses revealing significant differences between the voice profiles of Lucy and Plus Q compared to Graham.
- A significant effect in voice quality, with Joanna and Matthew being rated significantly different from Graham.

A. Interpretation of Results

The trend towards significance in voice impression suggests that participants perceived the voice profiles differently, though not all comparisons reached conventional levels of significance. The significant differences found in post-hoc tests imply that certain voice profiles, such as Lucy and Plus Q, may be more or less appealing to users than others.

The significant differences in voice quality indicate that participants clearly differentiated between the voice profiles in terms of perceived quality. The higher ratings for Joanna and Matthew suggest these voices were perceived as more natural or pleasant, which may be due to their synthesis quality or other acoustic characteristics.

B. Implications for Humanoid Robot Design

These results have important implications for the design of humanoid robots:

- **Customization of Voice Profiles:** Designers should consider offering customizable voice profiles to cater to user preferences. Given that different voices significantly affect perception, allowing users to select or personalize the robot's voice could enhance user satisfaction and engagement.
- **Voice Selection for Different Contexts:** Different contexts may require different voice profiles. For instance, a gender-neutral voice like Plus Q might be preferred in environments seeking to promote inclusively and avoid gender biases.

- **Improving Voice Quality:** Emphasis should be placed on using high-quality voice synthesis tools, as seen with the positive reception of Amazon Polly voices (Joanna and Matthew). This can enhance the perceived intelligence and naturalness of the robot, leading to better user experiences.

C. Limitations and Future Research

While the study provides valuable insights, it also has limitations:

- **Sample Size and Demographics:** The sample size was relatively small and consisted solely of university students, which may limit the generalizability of the findings. Future studies should include a more diverse sample to validate these results across different demographics.
- **Laboratory Setting:** The controlled laboratory environment, while useful for eliminating external variables, may not perfectly replicate real-world interactions. Future research should explore these effects in more naturalistic settings.
- **Limited Voice Profiles:** Only five voice profiles were tested. Including a broader range of voices in future studies could provide a more comprehensive understanding of how various vocal characteristics influence human perception.

D. Conclusion

This study demonstrates that the voice profiles of a humanoid robot like Ameca significantly impact human perception in terms of both voice impression and quality. The findings highlight the need for careful consideration of voice selection in the design of humanoid robots to optimize human-robot interaction. Future research should continue to explore this area, incorporating a wider range of voices, diverse participant samples, and real-world settings to further refine our understanding of the role of voice in human-robot interaction.

V. Appendix

A. Demographic Information:

- Gender
- Age
- Residence History 1
- Residence History 2
- Languages known
- Field of study

B. Robot Contact:

I have contact with social robots at university. I own and use an assistant robot (Alexa, Siri, etc)
[ ] Yes
[ ] No

I have high expectations from social robots.
[ ] Yes
[ ] No

Have you met Ameca/Emah in person before?
[ ] Yes
[ ] No

Have you watched videos or viewed images of Ameca/Emah before?
[ ] Yes
[ ] No

Which gender do you prefer Ameca/Emah to be assigned?
[ ] Male
[ ] Female
[ ] It (Neutral)

C. MOS (Mean Opinion Score) [18]

How do you rate the sound quality of the voice you have heard?
[ ] Excellent
[ ] Good
[ ] Fair
[ ] Poor
[ ] Bad

How do you rate the degree of effort you had to make to understand the message?
[ ] No effort required
[ ] Slight effort required
[ ]Effort required
[ ] Major effort required
[ ] Message not understood with any feasible effort

Did you find single words hard to understand?
[ ] None
[ ] Few
[ ] Some
[ ] Many
[ ] Every word

Did you distinguish the speech sounds clearly?
[ ] Yes, very clearly
[ ] Yes, clearly enough
[ ]Fairly clear
[ ] No, not very clear
[ ] No, not at all

Did you notice any anomalies in the naturalness of sentence pronunciation?
[ ] No
[ ] Yes, but not annoying
[ ] Yes, slightly annoying
[ ] Yes, annoying
[ ] Yes, very annoying

Did you find the speed of delivery of the message appropriate?
[ ] Yes
[ ] Yes, but slower than preferred
[ ] Yes, but faster than preferred
[ ] No, too slow
[ ] No, too fast

Did you find the voice you have heard pleasant?
[ ] Very pleasant
[ ] Pleasant
[ ] Fair
[ ]Unpleasant
[ ] Very unpleasant

D. Godspeed [17]

Please rate your impression of Ameca on a 1-5 scale:
How human-like did the voice sound to you?
Fake 1-2-3-4-5 Natural
Machinelike 1-2-3-4-5 Humanlike
How lively did the voice sound to you?
Dead 1-2-3-4-5 Alive
Stagnant 1-2-3-4-5 Lively
How much did you like the voice you heard?
Dislike 1-2-3-4-5 Like
Annoying 1-2-3-4-5 Pleasing
How intelligent did the voice sound to you?
Ignorant 1-2-3-4-5 Knowledgeable
Unintelligent 1-2-3-4-5 Intelligent
How safe did you feel listening to the voice?
Anxious 1-2-3-4-5 Relaxed
Agitated 1-2-3-4-5 Calm

E. Introduction monologue of Ameca

”Hello! I’m Emah, an advanced humanoid robot from Robotics Research Lab in Germany. I was developed at the company, Engineered Arts, in London. My main goal at this research group as a social robot, is to explore human speech, emotions, and social interaction. I am 187cm tall and I weigh about 49 kilograms.I can look for a specific person with my “Eye-mounted, binocular cameras” and “High-resolution chest-mounted, camera”. I also can, hear you using my Binaural, ear-mounted microphones. I have 51 degrees of freedom and can produce more than 50 gestures and facial expressions. And that is some technical information about me. Thanks for listening! Have a great day!”

References

[1] J. Jesin, I. W. Catherine, and M. Bruce, “Artificial empathy in social robots: An analysis of emotions in speech,” August 2018.

[2] N. Andreea, v. D. Betsy, N. Anton, L. Haizhou, and L. S. Swee, “Making social robots more attractive: The effects of voice pitch, humor and empathy,” 2013.

[3] I. Hiroshi, O. Tetsuo, T. Michita, Imai, K. Maeda Takayuki, and N. Ryohei, “Robovie: an interactive humanoid robot,” 2001.

[4] G. Stella, L. Nadine, L. Clare, W. Lisa, N. Frauke, F. Herta, and M. Carolyn, “Human talkers change their voices to elicit specific trait percepts,” 2024.

[5] K. Kühne, E. Herbold, O. Bendel, Y. Zhou, and M. H. Fischer, ““ick bin een berlina”: dialect proficiency impacts a robot’s trustworthiness and competence evaluation,” *Frontiers in Robotics and AI*, vol. 10, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2023.1241519

[6] C. Julie, M. D. Joan, E.-S. Norah, R. L. Tiffany, D. B. John, and V. Nancy, “Gender representation and humanoid robots designed for domestic use,” 2009.

[7] T. Nomura, T. Kanda, and T. Suzuki, “Experimental investigation into the influence of negative attitudes toward robots on human-robot interaction,” *AI Soc.*, vol. 20, pp. 138–150, 03 2006.

[8] C. A. Aguilera, A. Castro, C. Aguilera, and B. Raducanu, “Voice-controlled robotics in early education: Implementing and validating child-directed interactions using a collaborative robot and artificial intelligence,” *Applied Sciences*, vol. 14, no. 6, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/6/2408

[9] K. Katharina, H. F. Martin, and Z. Yuefang, “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*, vol. 14, pp. 3735–3741, 12 2020.

[10] R. C. Charles, S. Matthias, S. Paul, and V. Michael, “Gendered voice and robot entities: Perceptions and reactions of male and female subjects,” *RSJ International Conference on Intelligent Robots and Systems*, pp. 3735–3741, 10 2009.

[11] S. Mehta, R. Tu, J. Beskow, Éva Székely, and G. E. Henter, “Matcha-tts: A fast tts architecture with conditional flow matching,” 2024.

[12] Y. Lacombe, V. Srivastav, and S. Gandhi, “Parler-tts,” 2024.

[13] Y. A. Li, C. Han, V. S. Raghavan, G. Mischler, and N. Mesgarani, “Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models,” 2023.

[14] C. Pride, Virtue, E. AI, K. Interactive, and thirtysoundsgood. (2019) Genderless voice. [Online]. Available: https://genderlessvoice.com/

[15] Amazon. (2016) Amazon polly: Deploy high-quality, natural-sounding human voices in dozens of languages. [Online]. Available: https://aws.amazon.com/polly/

[16] S. Göde, “Modulating robot behavior during self-introduction scenario,” 2023.

[17] C. E. K. D. . Z. S. Bartneck, C., “Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots,” *International Journal of Social Robotics*, vol. 1(1), 2009.

[18] J. Lewis, “The revised mean opinion scale (mos-r): Preliminary psychometric evaluation,” 01 2001.

[19] F. L. . C. F. Sorrentino, A., “From the definition to the automatic assessment of engagement in human-robot interaction: A systematic review.” *International Journal of Social Robotics*, 2024.