Voice Design for Human-Robot Interaction: Tuning Acoustic Cues for Diverse Robot Appearances

JUMP SRINUALNAD, University of Twente, The Netherlands

The increasing integration of robots into daily life necessitates an understanding of how a robot's voice should match its physical appearance. Little attention from existing research has been given to how acoustic features including pitch, speaking rate, gender, and style of speaking interact with the robot's human likeness and size. This study looks into how individuals adjust these voice parameters, using a custom-designed voice interface to fit robots of varying degrees of human likeness and size. Participants came up with voice settings for robots, ASIMO and Qbo, each representing a high and low level of human likeness, in both small and large-scale robots. The study involved 22 participants (16 men and 6 women), who individually altered voice characteristics. The data analyzed reveals significant differences in voice parameter preferences based on the robot's appearance, with notable gender-specific tendencies.

Additional Key Words and Phrases: HRI, Voice design, Acoustic features, Audio processing

1 INTRODUCTION

While existing research has investigated the impact of factors such as voice naturalness, gender, accent, and prosody on the perceived trustworthiness of robot voices (e.g. [1, 12, 15, 17]) and how robots should look like (e.g. [6, 9, 10]). Little research has explored how the robot should sound (e.g. [1, 3]), especially the interactions of individuals when they manipulate acoustic features themselves in human-robot interaction systems to fit the robot's overall appearance.

So why should a robot voice match its physical appearance? A good match in physical appearance and vocalization can assist in minimizing any potential discomfort that may be felt by the user during the interaction with robots [5]. It can also improve perceived trustworthiness when a robot's voice aligns with its physical characteristics, whether humanlike, animal-like, or fully robotic. It helps users form more accurate expectations about the robot's functionality and behavior [13]. Therefore, designing robots with voices that align with their physical attributes not only improves usability but also makes more natural human-robot interactions.

To address this gap between the robot's voice and its overall appearance, we will explore how individuals manipulate acoustic features including pitch, speaking rate, gender, and speaking style, and adapt their behavior in response to robots that vary in human likeness and size.

1.1 Motivation

In the near future, we will see a rise in the number of robots in day-to-day life thanks to technological advancements as well as their incorporation and significance across several industries including health care, services industry, and personal assistance hence; it should be possible to determine which robot is speaking to us with ease. Thus, a strong association between a robot's voice and its appearance would be preferable, but this is currently not the case.

Voice can reveal a lot of information about its speaker. For example, gender is the most obvious characteristic that has different effects on people's perception of the robots, [5] found that participants anthropomorphize robots with a same-gender human-like voice more strongly. In spite of this, human voices carry attributes that allow us to discern their origins, social class, and ethnic background. This is not the case with robots who have no voice implied by them, nor do they relate to the physique of robots [13]. Therefore, it might be useful to look into other obvious physical characteristics of the robot such as human likeness and size to find a voice that matches the corresponding attributes.

1.1.1 Human-likeness. Designing a robot voice with human-like qualities is significant as it plays a role in the perception of users towards human-robot interactions. [12] suggested that in areas where social skills predominantly belong to humans (e.g., caregiving), using more natural-sounding voices led to higher anthropomorphism, pleasantness, and acceptance ratings across different real-life application scenarios. In summary, the study suggests favorable user responses to the voices of the highly human-like robots on human-like teleoperated robots. However, it also mentioned the need for further research on different sample groups.

1.1.2 Size. People often relate a robot's voice gender to its physical size [3]. Interestingly enough, they tend to think of smaller robots when they hear female voices and larger ones when they hear male voices. Unlike humans where vocal pitch typically relates to gender and in turn can influence perceptions of size, robot gender is a design decision; not a biological trait. Surprisingly, this connection still stands even though robot gender is not biologically related. This implies that associations between voice genders could impact how the sizes of robots are seen— an aspect worth consideration from different sample populations.

1.1.3 Acoustic Features Choices. Addressing the choice of using pitch, speaking rate, gender, and speaking style as the measurements. Starting with pitch, [8] found that choosing the right voice pitch should be a priority in social robot design. Speaking rate influences the perceived personality and clarity of speech. [11] conducted an experiment comparing different speech rates (fast, normal, moderately slow, and slow), and found contrasting results to those typically observed in human communication studies so it might

TScIT 37, July 8, 2022, Enschede, The Netherlands

be interesting to include speaking rate in our experiment. Gender is another crucial dimension of voice perception that influences social interactions. In both human and robot voices, gender cues such as pitch, play a crucial role in perceived characteristics such as masculinity, femininity, authority, or warmth. [5, 12]. Speaking style or tone of voice can convey nuances in formality, friendliness, and professionalism, so it is essential to incorporate a wide variety of styles.

1.2 Problem Statement

As mentioned earlier, little attention has been given to aligning a robot's voice with its physical appearance, which could affect user likability. This alignment is crucial as robots become relevant to sectors such as healthcare. We will explore how users manipulate voice characteristics like pitch, speaking rate, gender, and style in response to different robot appearances.

1.2.1 Research Question. **RQ**: How do individuals adjust acoustic features for robots by changing pitch, speaking rate, gender, and speaking style to fit robots with varying degrees of human likeness and size, specifically when the robot's physical appearance ranges from more human-like to more robotic?

To help answer this research question, we will be answering these 2 sub-questions:

1. Adaptive Adjustment In Parameters:
sub-RQ1: How do participants adjust their use of pitch, speaking rate, gender, and speaking style in response to the varying appearance of the robot and what is the most important parameter that seems to determine the appearance of the robot?

2. Understanding Perceptual Effects:
sub-RQ2: How do variations in voice adjustment correlate with participants' perceptions of the robot's appearance and likability?

2 RELATED WORK

2.1 Aligning Robot Appearance and Voice

The alignment between a robot's appearance and its voice is crucial for increasing user acceptance and overall user experience in human-robot interaction (HRI). When a robot's appearance matches its voice, it creates a believable interaction scenario that promotes trust and engagement [15]. Humans might associate specific attributes with both physical appearance and voice characteristics. For example, [7] found that participants from a diverse sample indicated that certain vocal features contribute to the formation of a mental image of the robot. Participants also tend to assign different job roles to robots based on minimal visual features. For instance, robots without mouths were associated with security roles, demonstrating how visual cues and vocal characteristics together could shape perceptions.

Another study, [16] where participants consistently chose voice prototypes that matched or enhanced the robots' images, indicated that visual cues alone can guide the creation of suitable robot voices. The study highlights the important interplay between visual and auditory perceptions in shaping human-robot interactions. It also shows the practical relevance for engineers in designing matching robot voices and underscores the synergy between cognitive science and machine learning.

2.2 Acoustic Features of Human Voices Predict Body Size

[3] investigated how acoustic features in human voices relate to body size, specifically height and weight. Using speech samples collected via an online survey, the study focused on vowel sounds 'i', 'a', and 'u', representing extremes in vocal tract characteristics. Through analysis of frequency and spectral energy, they employed linear discrimination analysis (LDA) to predict participants' height and weight categories. Findings revealed that certain acoustic traits could accurately predict body size, particularly height for both genders and weight more reliably for males. Specifically, lower formant frequencies and higher fundamental frequencies in male voices indicated larger body sizes, while higher formant frequencies and lower fundamental frequencies in female voices correlated with larger body sizes.

2.3 Impact of Acoustic Features on Perception of Robot Voices

[4] explored how different aspects of robot voices—like pitch, pitch range, and formant dispersion—affect how they are perceived. Using four robot prototypes, researchers tested various voice characteristics to gauge their persuasiveness and charm. Surprisingly, voices with low physical and melodic dominance were rated as most charming and persuasive, challenging assumptions about how robot size influences voice perception. Another study [8] also focused on voice pitch, revealing a preference for higher-pitched robot voices such as Olivia, which received better ratings for appearance, behavior, and personality compared to lower-pitched counterparts like Cynthia. In contrast, another study [11] delved into the influence of speech rate on human-robot communication. It found that participants favored normal and moderately slow speech rates over faster alternatives, perceiving them as more competent. Interestingly, in dynamic settings like walking interactions, slower speech was not only more comprehensible but also rated as positively as normal and moderately slow speech, emphasizing the context-dependent nature of speech rate in robot interactions.

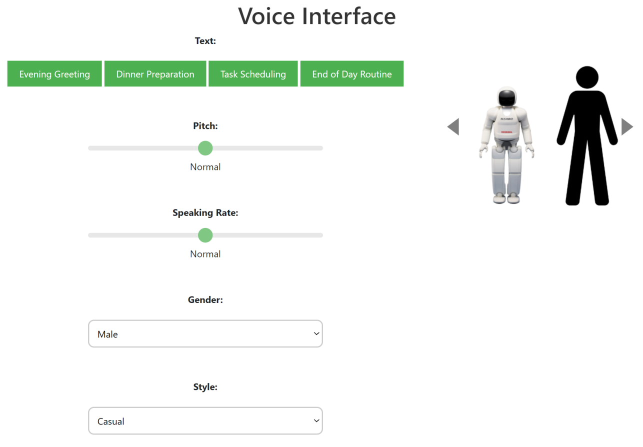
2.4 Voice and Motion Realism for Virtual Agents

Two studies explored the impact of various modalities on human-robot interaction. The first study [2] examined voice, motion, and appearance on perceived speech-gesture alignment, likability, and anthropomorphism, finding that high realism in voice and motion significantly improved these perceptions. Interestingly, mismatches between voice and motion realism did not negatively affect perceptions, and less realistic characters were often preferred for likability, while appearance did not significantly impact human likeness. The second study [14] investigated how personality traits are conveyed through speech and motion, revealing that extraversion is best communicated through high-fidelity motion, while agreeableness and emotional stability are primarily driven by speech. Conscientiousness and openness rely on both modalities. Virtual agents using

Text-to-Speech (TTS) and robotic animation were generally perceived as less extroverted, agreeable, open, and emotionally stable. It would be interesting to investigate whether similar patterns hold for static robot images, as explored in this current research.

3 METHODOLOGY

3.1 Interface



Voice Interface

Text:

Pitch: Normal

Speaking Rate: Normal

Gender: Male

Style: Casual

Fig. 1. Voice Interface

Fig. 1 shows the voice interface which consists of four dialogues designed to provide precise control over experimental conditions.

- **Evening greeting:** *"Good evening! I hope you had a pleasant day. Is there anything specific you would like assistance with tonight?"*
- **Dinner preparation:** *"I noticed it is almost dinner time. Would you like me to prepare a reservation at your favorite restaurant, or would you prefer I suggest a new place for you to try?"*
- **Task scheduling:** *"I see you have a busy day ahead tomorrow. Shall I add your morning workout to your schedule and remind you about the meeting with your team at 10 AM?"*
- **End-of-day routine:** *"Before you retire for the night, would you like me to adjust the thermostat to your preferred sleeping temperature and set the alarm for your usual wake-up time?"*

Each dialogue allows users to manipulate parameters including: Fig. 2 shows the ranges of pitch and speaking rate that participants can adjust. The units for pitch and speaking rate are from the Google Text-to-Speech (TTS) API used to develop this interface, more in Section 3.1.1. Users can also select between **male** and **female** voices and toggle between **casual** and **formal** styles. Immediate playback of voice samples upon selection provides interactive feedback, enabling users to assess real-time changes and evaluate how variations in these acoustic features influence perceived personality and situational appropriateness of robot voices.

Furthermore, the interface includes visual representations of two different robots with varying sizes mentioned in Section 3.2 alongside the controls for adjusting voice parameters. This setup allows participants to modify the voice attributes while simultaneously viewing the corresponding robot image.

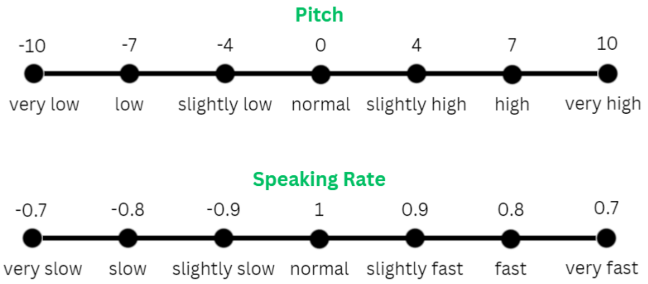


Fig. 2. Ranges of Pitch & Speaking Rate

3.1.1 Technologies. We used Google's Text-to-Speech API via a Flask web app to create robot voices in real-time, letting users adjust parameters like pitch, speaking rate, gender, and style. This setup enabled dynamic voice generation with American English accents.

For casual voices, it uses Wavenet voices. Wavenet voices, developed by Google DeepMind, are natural-sounding synthetic voices that use deep neural networks to generate more human-like speech.

- *"en-US-Wavenet-D"* for males
- *"en-US-Wavenet-F"* for females

For formal voices, the application uses voices designed to sound like professional news anchors, providing a clear and authoritative tone.

- *"en-US-News-N"* for males
- *"en-US-News-L"* for females

If none of these voices match, the application defaults to *"en-US-Standard-A"*, a more generic synthetic voice.

3.2 Robots

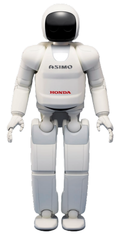


Fig. 3. ASIMO

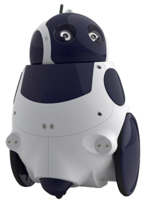


Fig. 4. Qbo

3.2.1 Levels of Humanlikeness. The choice of ASIMO as the human-like robot in our study is rooted in its advanced humanoid design. Developed by Honda, ASIMO is one of the most renowned humanoid robots, designed to resemble a human both in appearance and movement. It can walk, run, climb stairs, and interact with people, showcasing a high degree of human likeness. This makes ASIMO an ideal candidate for examining how users adjust their perception and interaction based on a robot that closely mimics human behaviors and physical attributes.

Conversely, Qbo is chosen to represent a more robotic, less human-like appearance. Created by TheCorpora, Qbo is a compact, mobile robot with a distinctly non-humanoid design. Its primary features include a dome-shaped head with cameras and sensors, wheels for movement, and a relatively simple exterior. Qbo's design emphasizes functionality and robotics rather than human mimicry, making it an excellent subject for studying user interactions with robots that are perceived as more mechanical and less anthropomorphic. This contrast with ASIMO allows for a comprehensive analysis of how human likeness influences acoustic feature adjustments in robot voices.

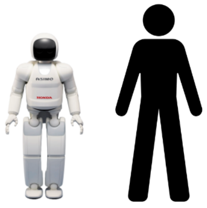


Fig. 5. Small ASIMO

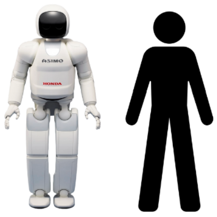


Fig. 6. Large ASIMO

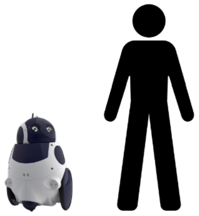


Fig. 7. Small Qbo

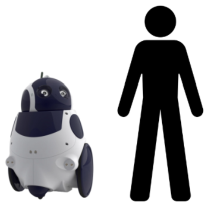


Fig. 8. Large Qbo

Table 1. Different images of the robot presented to participants during the study

3.2.2 Levels of Size. Table. 1 shows images of the robot that were shown to the participants where the silhouette of a human figure representing the participants is displayed on the right side of each image. Fig. 12 depicted ASIMO, a smaller robot shorter than the participants, representing a small human-like robot. Fig. 13 showed ASIMO approximately the same size as the participants, representing a larger human-like robot. In contrast, Fig. 14 displayed Qbo, a smaller robot standing around knee or waist height, symbolizing a small, more robotic-looking robot. Lastly, Fig. 15 presented Qbo as a larger robot approximately torso height, representing a bigger, more mechanical-looking robot. This setup allowed participants to visually compare and evaluate the robots based on their size and also the degree of human likeness relative to themselves.

3.3 Procedure

Participants were recruited through personal contacts and direct outreach for this study. They were asked to come to a room for an approximately 15-minute session. Upon arrival, they were briefed about the study, signed a consent form, and were informed that no voice recordings would be made. As the research focuses on voice, we asked the participants if they would like to wear headphones for clearer listening. They then completed a demographic questionnaire, providing their age, gender, mother tongue, and ethnicity. Seated in front of a computer, each participant viewed 4 robot images in a specific order starting with small ASIMO, large ASIMO, small Qbo, and large Qbo. The interface was then explained to the participants, they then selected a preferred voice they thought would best suit each robot.

3.4 Task

We began by showing the first image (small ASIMO) and asked participants to describe it. This step aimed to make participants aware of differences in size and the potential degree of human likeness compared to a human without explicitly telling them. Next, participants interacted with the interface, adjusting the pitch, speaking rate, gender, and style to match the robot's appearance. Once they were satisfied with their choices, their choice of the acoustic feature was collected and a few follow-up questions were then conducted to understand participants' reasoning behind their choices (see Section 3.6 on Measurements). This process was repeated for all four images. We then asked whether participants wished to revise any of their preferences. After completing the task, participants were debriefed.

3.5 Participants

A total of 22 participants were interviewed, consisting of 16 men and 6 women. The youngest participant was 20 and the oldest was 24 (average age of 21.4 years). The majority were native speakers of Western European languages, with a significant representation of English, Dutch, and Thai speakers. Specifically, English-speaking participants comprised the largest group, followed by Dutch and Thai speakers. Additionally, there were participants whose native languages included Chinese, Russian, Spanish, and Albanian, contributing to a diverse profile in the study.

3.6 Measurements

The independent variables included the level of human likeness and the size of the robots presented. Participants' preferences for pitch, speaking rate, gender, and style for each robot, along with demographic information were recorded. Additionally, answers to the follow-up questions for each robot were collected. As previously stated in Section 3.3 there will be no voice recording, participants' responses were directly transcribed into the qualitative analysis software, detailed in Section 3.7.3. The open-ended questions can be seen below.

- Are you satisfied with the voice you created?
- What influenced your choice of voice for this robot?
- Among pitch, speaking rate, gender, and style, which do you think contributes the most to your likability and why?
- Among pitch, speaking rate, gender, and style, which do you think contributes the least to your likability and why?
- Do you have any additional comments?

At some point during the experiment when participants observed both versions of ASIMO or Qbo and verbally acknowledged the size differences between the robots, additional questions were posed:

- Do you believe the size of the robots influences their ideal voices?
- Among pitch, speaking rate, gender, and style, which do you think is the most crucial factor in determining the robot's voice based on its size?

If participants did not verbally comment on the size differences then these questions were not asked. Data were stored in a CSV file for subsequent analysis.

3.6.1 *Data Types.*

- Pitch *(very low, low, slightly low, normal, slightly high, high, very high)*: **ordinal data.**
- Speaking rate *(very slow, slow, slightly slow, normal, slightly fast, fast, very fast)*: **ordinal data.**
- Gender *(male, female)*: **categorical data.**
- Style *(casual, formal)*: **categorical data.**

3.7 Data Analysis

The analysis will use descriptive statistics to summarize preferences for pitch, speaking rate, gender, and style. Interaction plots for correlation analysis will explore relationships between each acoustic feature with human likeness and size. Thematic analysis will interpret qualitative data on participants' voice preferences for each robot.

3.7.1 Descriptive Analysis. Statistical method used:
mean (e.g. average choice of the pitch for each robot), **standard deviation (SD)** (e.g. low standard deviation suggests that participants chose similar values for pitch) **mode** (e.g. most frequent choice for pitch on each robot), and **range** will be calculated on pitch and speaking rate, considering they are ordinal variables as shown in Fig. 2 and Section 3.6.1 on data types.
Percentiles will be used to understand the distribution of pitch and speaking rate choices across participants. **Interquartile Range (IQR)** will complement the range by offering a measure of variability that is robust against outliers, focusing on the middle 50% of the data; percentiles will be displayed on box plots.
Percentage analysis involves determining the proportion or percentage of participants who selected specific gender and style preferences (e.g., male voice and casual style) relative to the total number of participants. **Frequency distribution** using histograms will show how often gender and style category is selected.

3.7.2 Correlation Analysis. We will use an interaction plot to visualize and capture the relationship between human likeness and size (two independent variables), which differs across each acoustic feature (dependent variable). For instance, plotting the speaking rate on an interaction plot shows how the average speaking rate (dependent variable) varies across different levels of robot-human likeness (humanlike vs. robot-like) for each robot size (small vs. large). If the lines for small and large robots cross or diverge, it suggests that the effect of human likeness on speaking rate depends on the robot's size.
Speaking of significance, we will use Two-way ANOVA or its non-parametric equivalent, ART ANOVA, to statistically analyze and support the findings from the interaction plot. This analytical approach allows us to determine whether there are significant differences in the average pitch, speaking rate, gender, and style across different degrees of human likeness and robot size. This statistical analysis is built on top of the insights gained from the interaction plot, which visually captures how these acoustic features vary across the levels of human likeness and robot size. The significance level was set at 0.05.

3.7.3 Qualitative Analysis. Qualitative data from open-ended responses underwent thematic analysis using QDA Miner Lite, a free qualitative analysis software. This tool aided in coding, organizing, and exploring themes within the data. Steps taken were:

- Enter a name and description for each code (sentence, phrase, or paragraph) (e.g., themes or categories to look for).
- Highlight the text segment in the participants' responses that corresponds to a theme.
- Continue to code the data with different participants', we might notice new themes or need to adjust existing ones.
- Update code and explore patterns and relationships between codes using the Co-occurrence feature.

Additionally, we manually analyzed the data to find themes that the software might have missed. This is done through pen and paper since our sample size is quite small. The findings were then documented with quotes from the data to support our conclusions.

4 RESULTS

4.1 Data Preprocessing and Variable Encoding

Table 3 shows how data were encoded for data analysis. Fortunately, the dataset contained no missing values, so there was no need for data cleaning, leaving us with a sample size of N = 22.

4.2 Quantitative Results

4.2.1 Descriptive Statistics Visualization. The descriptive statistics for pitch and speaking rate are displayed in Table 2, where the mode is also provided with values decoded back to give clear, understandable descriptions. Fig. 9 includes boxplots that show how pitch and speaking rate are distributed across different robots. Only the speaking rates for small Qbo are presented with outliers.
Fig. 4 shows the percentage of participants who selected a specific gender and style, while Fig. 10 and Fig. 11 display the frequency count of gender and style preferences, respectively.

4.2.2 Correlation Analysis. Table 5 shows the interaction plot between each Acoustic feature, human likeness, and size. None of the lines cross, which suggests that the effect of human likeness on each acoustic feature is consistent across both sizes of robots, indicating no significant interaction effect between robot human likeness and size for that particular acoustic feature. To support this, we will conduct a 2-way ANOVA or ART ANOVA to determine its statistical significance. We begin by checking whether our data follows a normal distribution or not using the Shapiro-Wilk test.

Table 2. Descriptive Statistics for Pitch and Speaking Rate

Robot	Pitch					Speaking Rate				
	Mean	SD	Mode	Range (min, max)	IQR	Mean	SD	Mode	Range (min, max)	IQR
Small ASIMO	3.27	1.03	4.00 (slightly high)	[1.00, 5.00]	1.75	3.00	0.87	3.00 (normal)	[2.00, 5.00]	1.75
Large ASIMO	1.73	0.88	1.00 (low)	[0.00, 3.00]	1.00	2.73	0.70	3.00 (normal)	[2.00, 4.00]	1.00
Small Qbo	4.14	0.99	5.00 (high)	[2.00, 6.00]	1.75	3.45	1.01	4.00 (slightly fast)	[1.00, 5.00]	1.00
Large Qbo	2.86	1.42	2.00 (slightly low)	[0.00, 5.00]	2.00	3.23	0.75	3.00 (normal)	[1.00, 5.00]	1.00

Table 3. Encoding for Acoustic Features

Feature		Encoding
Pitch	**Speaking Rate**	
Very low	Very slow	**0**
Low	Slow	**1**
Slightly low	Slightly slow	**2**
Normal	Normal	**3**
Slightly high	Slightly fast	**4**
High	Fast	**5**
Very high	Very fast	**6**
Gender	**Style**	
Male	Casual	**0**
Female	Formal	**1**

Table 4. Percentages Distribution for Gender and Style

Robot	Gender (%)		Style (%)	
	Male	Female	Casual	Formal
Small ASIMO	40.91	59.09	54.55	45.45
Large ASIMO	86.36	13.64	27.27	72.73
Small Qbo	18.18	81.82	54.55	45.45
Large Qbo	27.27	72.73	45.45	54.55

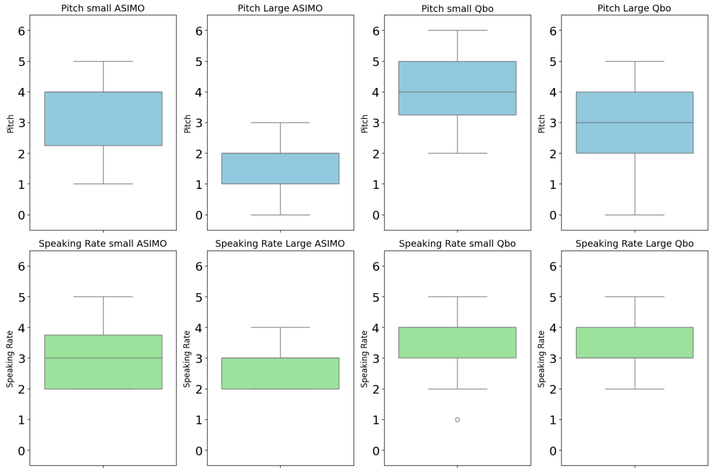


Fig. 9. Boxplots on Pitch and Speaking Rate

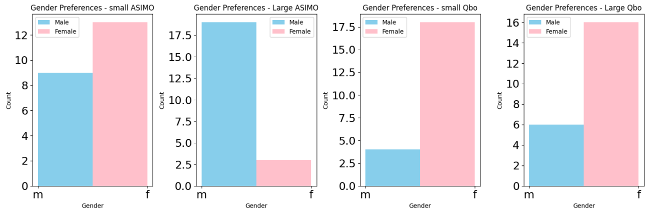


Fig. 10. Histograms for Gender preferences

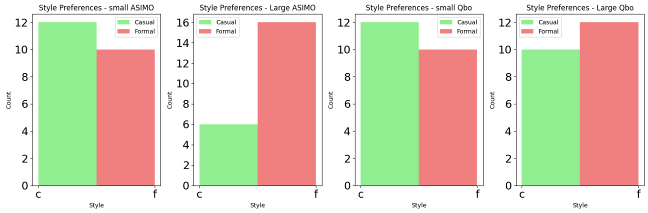


Fig. 11. Histograms for Style preferences

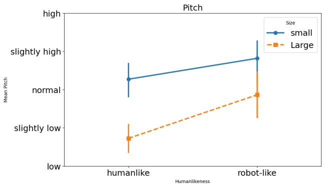


Fig. 12. Interaction Plot Between Human Likeness, Size and Pitch

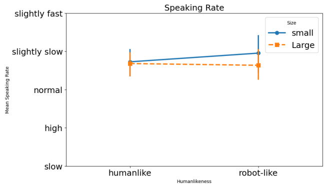


Fig. 13. Interaction Plot Between Human Likeness, Size and Speaking Rate

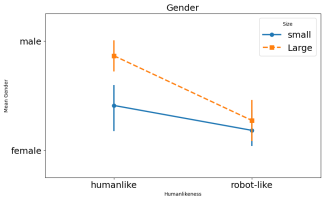


Fig. 14. Interaction Plot Between Human Likeness, Size and gender

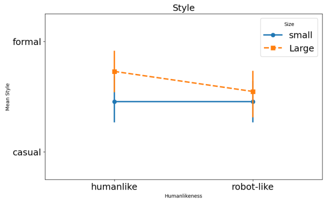


Fig. 15. Interaction Plot Between Human Likeness, Size and Style

Table 5. Interaction plots for Acoustic Features

Table 6 indicates that the majority of p-values are below 0.05, suggesting strong evidence against the data adhering to a normal distribution. Therefore, assuming normality for the data is not appropriate, and ART ANOVA will be used instead. Table 7 shows the result of ART ANOVA, where the following is the information displayed in the table:

Table 6. Shapiro-Wilk Test Results

Robot	p-value (Pitch)	p-value (Speaking Rate)
Small ASIMO	0.003	0.004
Large ASIMO	0.004	< 0.001
Small Qbo	0.034	0.006
Large Qbo	0.210	0.003

Table 7. Results of ART ANOVA

Source	ddof1	ddof2	F	p-unc	ng2	eps
Pitch	3	9	15.01	< 0.001	0.758	0.525
Speaking Rate	3	9	1.437	0.295	0.242	0.550
Gender	3	9	12.552	< 0.001	0.787	0.381
Style	3	9	3.438	0.065	0.365	0.435

- **ddof1:** Degrees of freedom for the feature being analyzed.
- **ddof2:** Degrees of freedom for error (9 in this case).
- **F:** F-value measures variation between groups relative to within groups.
- **p-unc:** p-value for the F-test under the null hypothesis. A p-value < 0.05 suggests significance.
- **ng2:** Partial eta-squared (η_p^2) measures variance explained by each features.
- **eps:** Epsilon (ϵ) adjusts ANOVA degrees of freedom for sphericity assumption violations.

The p-value less than 0.05 for Pitch and Gender in Table 7 indicates strong evidence that the variations in pitch and gender across different robots significantly influence the dependent variable, considering both robot-human likeness and size as factors. Conversely, the p-value for speaking rate and style is greater than 0.05, indicating that these factors do not significantly influence the dependent variable. This aligns with the findings from the interaction plot, reinforcing that there is no interaction effect between speaking rate, style, to human likeness, and size of the robot.

4.3 Qualitative Results

4.3.1 Thematic Analysis. Thematic analysis of participant responses revealed several key themes explaining why specific voice choices were made and how these influenced perceptions of the robot's likability. Here are the identified themes:

- **Pitch:**
  - Participants often associate low pitch with larger-size robots especially large ASIMO, describing them as "intimidating," "masculine," and often invoking stereotypes.
  - High pitch is frequently associated with a non-humanlike robot (Qbo), described as providing "comfort" or a "chill" vibe.
  - Participants who emphasize pitch's role in likability often use terms like "authority" and "serious" to describe the robot.
- **Speaking Rate:**
  - No consistent trends or themes were noted among participants who believe the speaking rate contributes the most or least to robot likability.
- **Gender:**
  - Male voices are commonly preferred for Large ASIMO, described as "masculine" and "serious" by participants.
  - Female voices are associated with non-humanlike robots (Qbo), described as "cute" or akin to a "pet."
- **Style:**
  - Participants selecting a formal voice often describe the robot, especially Big ASIMO, as "serious" and embodying "authority."
  - Casual voice selections evoke descriptions such as "caring" and "engaging," particularly for both Small and Large Qbo robots.

5 DISCUSSION

5.1 Insights on Data Analysis

5.1.1 Insights on Quantitative Data. Starting from Table 2, the mean and mode of pitch are higher for both small versions of the robots (ASIMO and Qbo), this suggests that participants often prefer higher pitch voice on smaller robots and lower pitch voice on larger ones. The narrow IQR and boxplots in Fig. 9 which shows data points are concentrated tightly around the median suggest that participants often have consistent preferences in their choices of pitch and speaking rates across different robot sizes and human likeness levels. The mean and mode of speaking rate are steady across all robots which suggests most of the time that participants prefer a speaking rate that does not deviate from the normal range (not too fast or too slow).

Table 4 and Fig. 10 show that the majority of participants prefer male voice on Large ASIMO and female voice on both versions of Qbo regardless of its size. However, the results from the same table and figure suggest that the speaking style does not exhibit a clear trend. For most robots, preferences are nearly evenly split between casual and formal styles, except for large ASIMO, where a distinct preference for formal style is evident.

5.1.2 Insights on Qualitative Data. Section 4.3.1 on pitch suggests that participants use words such as "intimidating" or "masculine" to describe ASIMO. Many participants mentioned that this is influenced by ASIMO's design, which includes more masculine features such as broader shoulders and less curvature, making a male voice more fitting, irrespective of size. It is worth noting that when participants believe both gender's voices would suit the robot such as small ASIMO (as shown in Table 4 where the ratio for gender is nearly equal), they tend to choose the voice corresponding to their own gender. This aligns with the previous study [5] which suggests males showed a preference for male voices with respect to conformity, social attraction, and trust, while females also exhibited a preference for females. Regarding the Speaking rate, no notable comments or themes emerged among participants based on different robots.

Section 4.3.1 on gender suggests that participants often associate male voice for large ASIMO and female voice for non-humanlike robots such as Qbo. Starting with large ASIMO, Participants mentioned that this is due to the fact that they do not think a female voice would suit ASIMO so the only option left is a male voice. Participants mentioned that their strong preference for the female voice on Qbo is due to it being pet-like and relatively smaller than them for both versions. Interestingly, some also mentioned that this preference is influenced by their exposure to interacting with robots and AI technologies, such as SIRI or depictions in movies, where female voices are commonly used for assistance roles for robots, and smaller robots in movies are often in female voices. Regarding the style, no notable trends emerged among participants based on their responses other than formal suits large ASIMO best.

5.2 Answering research questions

5.2.1 sub-RQ1. We have seen in the Result Section 4 how values for acoustic features including pitch, speaking rate, gender, and style change based on different robots with degrees of human likeness and size. Participants showed a significant focus on their pitch preferences, reflecting the variety of values observed across different robots such as in table 2. As suggested by [8], the pitch should be a priority in social robot design, Gender also comes close to the pitch, based on the ART ANOVA results from Table 7 where both features show a correlation between itself and human likeness and size.

5.2.2 sub-RQ2. We have seen in Sections 4.3 and 5.1.2 how their choices of acoustic features correlate with participants' perceptions of the robot's appearance and likability. Participants often base their voice preferences on personal experiences, upbringing, and stereotypes. For instance, they may find smaller robots cute or pet-like, leading them to prefer a female voice. Conversely, a more human-like robot is seen as more serious, and especially ASIMO suits a male voice. Additionally, the choice of speaking rate which does not deviate from the normal range (slightly slow to slightly fast) is due to the voice being easy to understand and what participants are used to.

5.2.3 RQ. Based on this study where pitch and voice gender seem to be the most important factors in designing a voice for robots with a certain degree of human likeness and size. With previous study's results [3] also stated that male voices with deeper pitches and lower resonances are often perceived as coming from larger individuals which can also be seen true from this study as discussed earlier in Results Section 4 that male and low pitch voice are often preferred for large ASIMO. It can be said that pitch serves as a key indicator for suggesting a specific body size, particularly evident in our human-like robot ASIMO but not in Qbo. This study suggests that gender is the most influential acoustic feature for inferring the degree of human-likeness in a robot's appearance among those examined with a clear preference for a female voice for Qbo regardless of its size.

5.3 Limitations

5.3.1 Sample Size. With only 22 participants, the study sample size is relatively small. While the findings provide valuable insights into voice preferences for specific robot appearances, it is too small to generalize the claims made in this study.

5.3.2 Sequential Exposure. Participants were presented with each robot image individually, potentially limiting their ability to compare preferences across different robots simultaneously. Future studies could explore how simultaneous exposure to all robot images influences participants' voice preferences and perceptions.

5.3.3 Robot Representative. Choosing robots that convey human likeness involves selecting features like skin texture, facial expressions, and body movements. These attributes significantly influence how participants perceive and interact with the robot. More prominent size differences also play a crucial role in distinguishing between robots.

5.4 Further Work

Future research should explore whether participants' preferences for acoustic features remain consistent when robots are presented together initially, rather than one by one. Furthermore, investigating the effects of more pronounced size differences between robots, while keeping them believable as assistance robots, could refine acoustic design strategies since a few participants did not notice any difference in the size of the robots. Understanding how significant size variations influence voice preferences makes robotic assistants more effective, approachable, and aligned with user expectations. Additionally, using a variety of robots, such as those designed to be more human-like with human skin or human-like faces, and others that are more robotic with no features resembling humans or living things, could offer new insights into how different appearances affect voice preferences and overall interaction quality.

6 CONCLUSION

This study investigated participant preferences for voice characteristics in robots, focusing on factors such as pitch, speaking rate, gender, and speaking style across different robots with varying degrees of human likeness and size. Key findings suggest that pitch plays a crucial role, with higher pitches preferred for smaller non-humanlike robots like Qbo are perceived as cute or pet-like. Lower pitches are preferred on larger, more humanoid robots like ASIMO. In the case of small ASIMO, where participants think both male and female voices are suitable for the robot, participants tend to choose the voice gender that matches their own gender. The thematic analysis highlighted that participants' choices were often influenced by stereotypes, personal experiences, and perceptions of robot characteristics. Overall, the findings highlight the importance of pitch and voice gender, along with maintaining a normal range for speaking rate and a suitable speaking style in designing robot voices can align with user expectations and enhance interaction quality. Understanding these preferences can guide the development of more engaging robotic assistants in various contexts of HRI.

A APPENDIX

A.1 AI Tools

ChatGPT and Grammarly have been used to correct grammar and ensure clarity in this paper.

REFERENCES

[1] F. Eyssel, D. Kuchenbrandt, S. Bobinger, L. de Ruiter, and F. Hegel. 2012. 'If you sound like me, you must be more human': on the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *HRI '12: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 125–126. https://doi.org/10.1145/2157689.2157717

[2] Thomas Sean Guiard Cédric Ennis Cathy & McDonnell Rachel Ferstl, Ylva. 2021. Human or Robot? Investigating voice, appearance and gesture motion realism of conversational social agents. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. Association for Computing Machinery, New York, NY, USA, 76–83. https://doi.org/10.1145/3472306.3478338

[3] K. Fischer and O. Niebuhr. 2023. Which voice for which robot? Designing robot voices that indicate robot size. *ACM Transactions on Human-Robot Interaction* 12, 4 (December 2023), Article 55. https://doi.org/10.1145/3632124

[4] K. Fischer, O. Niebuhr, R. M. Langedijk, and S. Eisenberger. 2019. I shall know you by your voice – Melodic and physical dominance in the design of robot voices. In *Proceedings 1st International Seminar on the Foundations of Speech: Breathing, Pausing, and the Voice*. Syddansk Universitet, 88–90. https://my.eventbuizz.com/assets/editorImages/1575286883-SEFOS_2019__proceedings_programme.pdf

[5] E. J. Lee, C. Nass, and S. Brave. 2000. Can computer-generated speech have gender?: An experimental test of gender stereotype. In *CHI EA '00: CHI '00 Extended Abstracts on Human Factors in Computing Systems*. 289–290. https://doi.org/10.1145/633292.633461

[6] S. X. Liu, E. Arredondo, H. Mieczkowski, J. Hancock, and B. Reeves. 2021. A picture is (still) worth a thousand words: The impact of appearance and characteristic narratives on people's perceptions of social robots. In *Handbook of Computational Social Science*. Routledge, 324–342. https://doi.org/10.4324/9781003024583-23

[7] Conor McGinn and Ilaria Torre. 2019. Can you Tell the Robot by the Voice? An Exploratory Study on the Role of Voice in the Perception of Robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 211–221. https://doi.org/10.1007/s10503-011-9215-x

[8] Andreea Niculescu, Betsy van Dijk, Anton Nijholt, and Swee Lan See. 2011. The influence of voice pitch on the evaluation of a social robot receptionist. In *2011 International Conference on User Science and Engineering (i-USEr)*. 18–23. https://doi.org/10.1109/iUSEr.2011.6150529

[9] E. Phillips, D. Ullman, M. M. A. de Graaf, and B. F. Malle. 2017. What does a robot look like?: A multi-site examination of user expectations about robot appearance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications Sage CA, Los Angeles, CA, 1215–1219. https://journals.sagepub.com/doi/10.1177/1541931213601786

[10] B. Reeves, J. Hancock, and X. Liu. 2020. Social robots are like real people: First impressions, attributes, and stereotyping of social robots. *Technology, Mind, and Behavior* 1, 1 (2020). https://doi.org/10.1037/tmb0000018

[11] Michihiro Shimada and Takayuki Kanda. 2012. What is the appropriate speech rate for a communication robot? *Interaction Studies* 13 (01 2012). https://doi.org/10.1075/is.13.3.05shi

[12] S. Song, J. Baba, J. Nakanishi, Y. Yoshikawa, and H. Ishiguro. 2020. Mind The Voice!: Effect of Robot Voice Pitch, Robot Voice Gender, and User Gender on User Perception of Teleoperated Robots. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'20 Extended Abstracts)* (New York, NY, USA). ACM, Honolulu, HI, USA, 10. https://doi.org/10.1145/3334480.3382988

[13] S. J. Sutton, P. Foulkes, D. Kirk, and S. Lawson. 2019. Voice as a design material: Sociophonetic inspired design strategies in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3290605.3300833

[14] Sean Thomas, Ylva Ferstl, Rachel McDonnell, and Cathy Ennis. 2022. Investigating how speech and animation realism influence the perceived personality of virtual characters and agents. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 11–20. https://doi.org/10.1109/VR51125.2022.00018

[15] I. Torre and L. White. 2021. Trust in vocal human-robot interaction: Implications for robot voice design. In *Voice Attractiveness. Prosody, Phonology and Phonetics*, B. Weiss, J. Trouvain, M. Barkat-Defradas, and J. J. Ohala (Eds.). Springer, Singapore. https://doi.org/10.1007/978-981-15-6627-1_16

[16] Mertes Silvan Janowski Kathrin Weitz Katharina Jacoby Nori & André Elisabeth van Rijn, Pol. 2024. Giving Robots a Voice: Human-in-the-Loop Voice Creation and Open-Ended Labeling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 584. https://doi.org/10.1145/3613904.3642038

[17] J. Wuth, P. Correa, and T. et al. Núñez. 2021. The role of speech technology in user perception and context acquisition in HRI. *International Journal of Social Robotics* 13 (2021), 949–968. https://doi.org/10.1007/s12369-020-00682-5