The Effects of Robot Appearances, Voice Types, and Emotions on Emotion Perception Accuracy and Subjective Perception on Robots

Sangjin Ko, Xiaozhen Liu, Jake Mamros, Emily Lawson, Haley Swaim,
Chengkai Yao, and Myounghoon Jeon(✉)

Department of Industrial and Systems Engineering, Mind Music Machine Lab, Virginia Tech, Blacksburg, VA, USA
{sangjinko,xliu26,jakem17,emily03,haley1,yaock,
myounghoonjeon}@vt.edu

Abstract. In human-robot interaction, natural and intuitive communication between robot and human is one of the most important research topics. Emotion plays a crucial role to make natural and social interactions. Research has focused more on robots' appearances and facial emotional expressions, but little research has investigated robots' voices and their mixed effects with robot types and different emotions for users to perceive robots' emotional states. In this study, anthropomorphic and zoomorphic robots, four different voice types, and seven different emotional voices were used as mixed factors to discuss how these influence users' perception on robots' emotional expression and other characteristics. Sixteen participants were asked to read fairy tales to robots and determine robots' emotional states when the robots verbally responded. Overall, the anthropomorphic robot (Nao) was preferred over the zoomorphic robot (Pleo), but this appearance did not influence emotion recognition accuracy or other robot characteristics. Participants showed lower accuracy in recognizing negative emotions with high arousal: anger, fear, and disgust. TTS was rated lower than other human voices in all robot characteristics, such as warmth, honesty, trustworthiness, and naturalness. Implications and design directions are discussed with the results.

Keywords: Human-robot interaction · Emotion recognition · Robot appearance · Robot voice

1 Introduction

Research on social robots has sharply increased. To design social robots, it is important to consider key variables to influence robots' sociability, trust, and acceptance. Previous studies in human-robot social interaction discovered that users expect natural and intuitive communication with robots [1–3]. Emotion is a critical component to enable this natural and intuitive communication. To express emotions effectively and accurately, robots can utilize many different sensory cues. However, research has focused more on facial expressions [4] and a comprehensive study on the correlation between different

C. Stephanidis et al. (Eds.): HCII 2020, LNCS 12424, pp. 174–193, 2020.
https://doi.org/10.1007/978-3-030-60117-1_13

factors and emotion perception is still rudimentary. In this paper, an exploratory experiment was conducted to study users' perception on robots' emotional states with mixed factors, including robot appearances, voice types, and emotions. Participants were asked to read fairy tales to robots and determine robots' emotional states when the robots made a comment.

1.1 Related Work

Companion Robots and Natural Communication

The prospect of introducing companion robots into an individual's daily life has received a significant focus from those working and conducting research in the field of Human-Robot Interaction. The development and integration of effective robot companions hold a tremendous degree of promise, given the potential for robots to assist people or even perform tasks that exceed their capabilities.

Previous studies have indicated broad support for the concept of companions, with participants viewing a robot's roles as an assistant, machine, or servant to conform to their expectations of the robot's function [1]. A much greater support and confidence was expressed for robots to be charged with performing household tasks as opposed to tasks dealing with children or animals [1]. A robot companion's ability to communicate is as significant as establishing the social context under which such a companion should operate. Humanlike communication is a desired trait [1]. Studies have similarly shown that the natural language interface of a robot receives more attention in comparison to its functionality, suggesting that the communicative behavior may be a more critical component of the system [3]. Proposed criteria by which to evaluate communicative behavior includes the ability of a robot to detect communication partners and pay attention to them, as well as its comprehension of speech, gestures, and its surrounding environment so as to understand an assigned task [3]. Such criteria revolve around maximizing the social aptness of robot companions so that they may interact and carry out tasks in a natural way.

Expressing one's own emotions and reading others' emotions is also critical for facilitating this natural interaction. To express emotions effectively and accurately, a number of verbal (e.g., voice style, accent, gender, and affective prosody) and nonverbal (e.g., appearance, facial expression, gesture, and movement) cues can be used.

In the current study, we explored the scenario where our participants served as a storyteller and our robots were emotionally empathized with them and responded to the story. We specifically considered the robots' ability to convey emotions, which is a critical part of human-like communication.

Form Factor of Robots

In the design of robots, there are two typical forms of design; one is anthropo-morphic, and the other is zoomorphic [5]. Each one has its unique characteristics and deals with different tasks. In studies on robots' form factor (or appearance) and users' perception, anthropomorphic and zoomorphic robots were preferred over machine-like ones or imaginary creatures [7–9]. Anthropomorphic and zoomorphic robots may have different working scenarios. The more a robot's appearance is human-like shaped, the more intelligent people think it is [10]. Also, during the interaction with humans, anthropomorphic robots may be more able to convey emotional expressions more effectively because their appearance is similar to humans [5]. However, the influence of facial expression of anthropomorphic robots on users' perceptions are sometimes controversial, maybe due to the Uncanny Valley [6]. In other words, a robot's suitability to being like a human or an animal is highly dependent on what kind of task it has and what intelligence level it wants users to perceive. Preference to an anthropomorphic or zoomorphic robot is influenced by many complex factors. In our work, we investigated the effects of voice types on each robot, by applying both qualitative and quantitative measures, examining user perception from broader perspectives.

Robots' Emotion Expression
In terms of emotional expression, emotional vocal expressions can effectively influence the behavior of perceivers [11]. Research explains a robot's emotion expression process in relation to communication theory: 1) a robot's internal state drives expressions, 2) specific robot behaviors are related to specific user reactions, and 3) the situation is an important driver of emotion expressions [12]. Emotion perception is an important source of information about the theory of mind and emotions can be perceived from facial expressions, voices, and whole-body movements [13]. As mentioned, emotion expression and emotion perception play a critical role in human-robot interaction and are widely studied in a range of disciplines. However, previous studies have been dominated by robots' facial emotions and other modalities such as vocal and tactile processes have been less frequently considered [14, 15]. The present study focused more on auditory stimuli by including various emotive voices, representing seven different emotions and investigated the differences in users' emotion perception.

1.2 Research Questions

From this background, we tried to attain a deeper understanding of the effects of robots' appearances, voices, and emotion types on users' perception about robots and their emotions. More specifically, we were interested in the following research questions:

- How can robot appearances, voices, emotion types, and their interactions influence people's perception of robots' emotional states?
- How can robot appearances and voices, and their interactions influence people's perception of robots' characteristics?
- How can robot appearances and voices, and their interactions influence people's preference on robots?

To answer these research questions, we conducted a preliminary empirical experiment in which young adults (college students) interacted with two robots (human-like and animal-like) using four different voices (regular human, characterized human-like, characterized animal-like, text-to-speech) and seven emotions (six basic emotions + anticipation). We collected our participants' emotion recognition accuracy and other subjective perception on robots.

2 Method

2.1 Participants

Sixteen university students participated in the study (Age: M = 23.5, SD = 3.97). Six participants identified themselves as male and the other ten participants identified as female. Participants were ethnically diverse (3 Asians, 2 Hispanic, 9 Caucasians, 1 Middle easterners, and 1 Africans). Participants participated in the experiment for at most 2 h and participants were compensated with $20 ($10 per hour). All participants agreed to participate after reviewing the consent form approved by the VT IRB.

2.2 Robotic Systems and Stimuli

Two robots, NAO and Pleo, having different appearances and features were employed in the experiment (Fig. 1). We used these two robots, which represent an anthropomorphic robot and zoomorphic robot each, to contrast the effects that robotic appearance has on people's emotion perception. NAO is a small-size anthropomorphic robot (Height: 22.6 inch, Length: 10.8 inch, Width 12.2 inch) having similarity to human and Pleo is a zoomorphic robot (Height: 8 inch Length: 15 inch, Width 4 inch) which looks like a little dinosaur. Both robots played recorded auditory feedback, which were emotive utterances, to participants following the storylines. Two different stories ("The three little pigs" and "The boy who cried wolf") were used in this experiment.

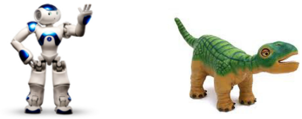


Fig. 1. Pictures of robots (NAO, Pleo)

Four voice types were created for seven emotional expressions. We first categorized different voice types as a synthesized voice (text-to-speech or TTS voice) and a recorded human voice. The human voices were provided by two female native speakers in our research group and all the voices were speaking American English with American accents. Next, the recorded human voice was subdivided into three categories that included a regular voice and a characterized voice for each robot (i.e., characterized NAO voice and characterized Pleo voice). The characterized voices for NAO and Pleo were designed to exaggerate emotional expressions with the robots' characters while the female speakers envisioned the characteristics of robots from their appearances.

The TTS voices were generated using text-to-speech [16] engines. Microsoft's female voice and the iOS female voice were used, which were provided by default with the respective operating systems. These TTS voices included no emotional information beyond the words themselves.

Seven different emotions were presented throughout each story including Ekman's six basic emotions. The six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) were chosen for their prevalence in psychology. In addition to them, the seventh emotion, anticipation, was chosen for its similarity to fear and surprise [17]. Its inclusion allows us the opportunity to see if participants can discern an emotion that is not traditionally regarded as a basic emotion and to gauge confusion between emotions with subtle differences. The seven emotions were fit into both stories ("The three little pigs" and "The boy who cried wolf") as depicted in Table 1.

Table 1. Dialogues in stories for presenting different emotions

Presented emotions	Robots' utterance in a story	
	The boy who cried wolf	The three little pigs
Anger	That's not nice!	They shouldn't tease him like that
Anticipation	This should be good.	I wonder what's going to happen!
Disgust	Gross!	He can't want to EAT them!
Fear	He's going to eat the sheep!	Oh no!
Happiness	That sounds nice!	Good!
Sadness	All his sheep are gone	He destroyed their homes
Surprise	Why didn't they help?	Woah, that's fast!

2.3 Design and Procedure

A 2 (robots) × 4 (voice types) × 7 (emotions) within-subject design was applied. Therefore, 8 different combinations of robots and voice types were provided to each participant with all 7 emotions. The presented order and the number of each combination were counterbalanced such that 1) each combination was almost equally presented about 20 times across participants in total and 2) levels of each treatment were presented at least once to each participant. Therefore, each participant interacted with all 8 conditions of robots and voice types and all 7 presented emotions. The 8 conditions were separated into two sessions to help participants recall and compare four different conditions. In each condition, the participant was instructed to read the script in front of a robot and listen to the emotional comment from the robot at various points in the story. The whole procedure including each step and the experiment environment are depicted in Figs. 2, 3 and 4 below.

The participants were asked to fill out several questionnaires after listening to each comment generated from the robot, after finishing reading each full story, and after experiencing four conditions. Specifically, after each response to seven emotions, each condition, and each session, the surveys were conducted for measuring the accuracy of emotion recognition and robot characteristics (Warmth, Honesty, Trustworthiness), naturalness (Natural, Human-like, Robot-like) and preferences (Likability, Attractiveness) of presented emotions. The questionnaire consisted of open questions, seven-point Likert scales (1: Lowest, 7: Highest), and single-choice questions. (Table 2).

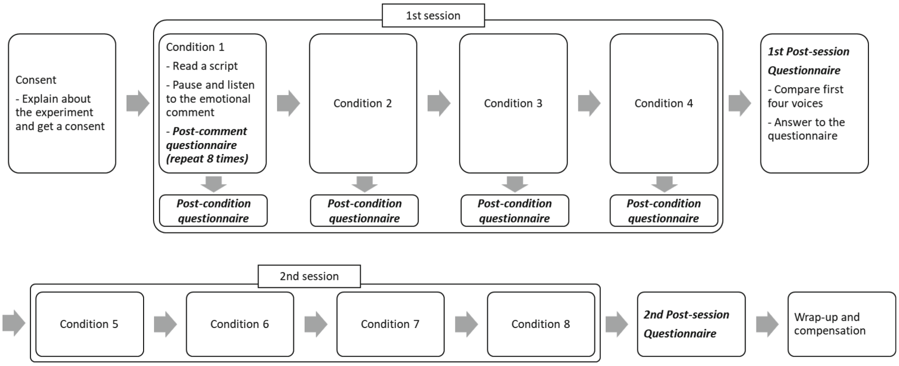
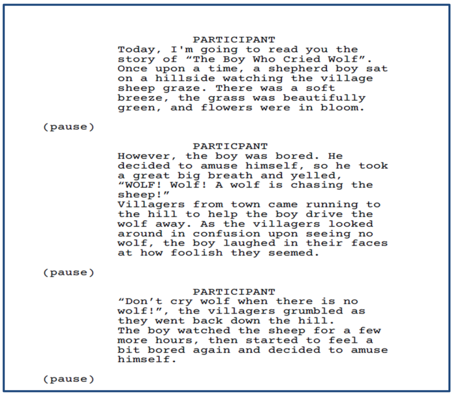

1st session
Consent
- Explain about the experiment and get a consent
Condition 1
- Read a script
- Pause and listen to the emotional comment
- *Post-comment questionnaire (repeat 8 times)*
Condition 2
Condition 3
Condition 4
Post-condition questionnaire
Post-condition questionnaire
Post-condition questionnaire
Post-condition questionnaire
1st Post-session Questionnaire
- Compare first four voices
- Answer to the questionnaire
2nd session
Condition 5
Condition 6
Condition 7
Condition 8
2nd Post-session Questionnaire
Wrap-up and compensation

Fig. 2. The flow diagram of the procedure


```
                    PARTICIPANT
          Today, I'm going to read you the
          story of “The Boy Who Cried Wolf”.
          Once upon a time, a shepherd boy sat
          on a hillside watching the village
          sheep graze. There was a soft
          breeze, the grass was beautifully
          green, and flowers were in bloom.
(pause)

                    PARTICPANT
          However, the boy was bored. He
          decided to amuse himself, so he took
          a great big breath and yelled,
          “WOLF! Wolf! A wolf is chasing the
          sheep!”
          Villagers from town came running to
          the hill to help the boy drive the
          wolf away. As the villagers looked
          around in confusion upon seeing no
          wolf, the boy laughed in their faces
          at how foolish they seemed.
(pause)

                    PARTICIPANT
          “Don’t cry wolf when there is no
          wolf!”, the villagers grumbled as
          they went back down the hill.
          The boy watched the sheep for a few
          more hours, then started to feel a
          bit bored again and decided to amuse
          himself.
(pause)
```

Fig. 3. An example of part story the participant read (The Boy Who Cried Wolf)

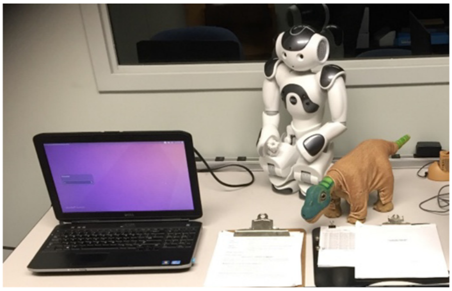


Fig. 4. Experimental setting

Table 2. The list of questions and types in questionnaires

Category	Question (Type)
Post-comment questionnaire	1. What emotion do you feel the robot expressed? (Open question) 2. What characteristics of the voice brought to mind that emotion? (Open question) 3. How clearly did the robot express this emotion? (1–7 Likert scale) How suitable was this emotion coming from the robot? (1–7 Likert scale)
Post-condition questionnaire	1. How likable is the voice? (1–7 Likert scale) 2. How attractive is the voice? (1–7 Likert scale) 3. How warm is the voice? (1–7 Likert scale) 4. How honest is the voice? (1–7 Likert scale) 5. How trustworthy is the voice? (1–7 Likert scale) 6. How natural does the voice sound? (1–7 Likert scale) 7. How human does the voice sound? (1–7 Likert scale) 8. How robotic does the voice sound? (1–7 Likert scale)
Post-session questionnaire	1. Thoughts about 1st, 2nd, 3rd, and 4th voices (Open question) 2. Which story was your favorite? (Open question) 3. What is your sex? (Open question) 4. What is your age? (Open question) 5. What is your race and/or ethnicity? (Multiple-choice, Open question)

Presented orders of emotions in the two stories were different, but the order in each story was fixed to maintain the storylines. To generalize the results, we employed two different stories having the same 7 emotions presented and two different voice groups having the same characteristics but recorded by different female speakers and two different female text-to-speech (TTS) engines. The examples of the presented order are depicted in Table 3. To validate the equivalence in accuracy, clarity, suitability, and preference of the two stories and two voice groups, the results were analyzed as below (Table 4), showing similar results in all categories.

3 Results

3.1 Data Collection

The answers to open questions regarding emotions were interpreted by two examiners. Each examiner categorized all the answers into seven pre-defined emotions or marked as ‘indistinguishable’ if the answers do not fall into any categories. Two examiners worked independently, and the inter-rater reliability test showed that 87.8% (787/896) of the results were consistent with the high coefficient value of Cronbach Alpha using variance (=0.96). If interpretations from examiners were different, a third examiner reviewed the answers and decided which emotion the answer fell into.

Table 3. Examples of the presented order

PID	Start	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Trial 7	Trial 8
1	Robot	NAO	Pleo	NAO	Pleo	NAO	Pleo	NAO	Pleo
	Voice type	Regular	Characterized NAO	TTS	Characterized Pleo	TTS	Regular	Characterized Pleo	Characterized NAO
	Story*	Pigs	Wolf	Pigs	Wolf	Pigs	Wolf	Pigs	Wolf
	Voice group	Group A	Group A	Group A	Group A	Group B	Group B	Group B	Group B
2	Robot	Pleo	NAO	Pleo	NAO	Pleo	NAO	Pleo	NAO
	Voice type	Characterized NAO	Characterized Pleo	Regular	TTS	Regular	Characterized NAO	TTS	Characterized Pleo
	Story*	Pigs	Wolf	Pigs	Wolf	Pigs	Wolf	Pigs	Wolf
	Voice group	Group B	Group B	Group B	Group B	Group A	Group A	Group A	Group A

**Pigs: The three little pigs, Wolf: The boy who cried wolf*

Table 4. Accuracy, clarity, suitability, and preference over stories and voice groups

		Accuracy	Clarity	Suitability	Preference
Story	The boy who cried wolf	51.2%	4.96	4.93	4.35
	The three little pigs	59.3%	5.27	5.24	4.66
Voice group	Group A	57.6%	5.11	5.17	4.66
	Group B	53.1%	5.12	5.01	4.37

3.2 Emotion Perception: Accuracy, Clarity, Suitability, and Features

First, the accuracy of emotion perception, defined as the proportion of correct emotion answers, was analyzed. Figure 5 and Table 5 show the inferential statistics of accuracy across presented emotions, voice types, and robots. Regarding presented emotions, anger, disgust, and fear showed significantly lower accuracies (below chance level) than other emotions. Therefore, we removed these three emotions from further accuracy analyses. Results were analyzed with the aligned rank transform (ART) [43] for nonparametric factorial analyses since there are 3 factors (Robots, Voice Types, and Emotions) and dependent variable (1: correct, 0: wrong) is not normally distributed. The ART allowed analyzing the aligned-ranked data with a 2 (Robots) × 4 (Voice Types) × 4 (Emotions) repeated measures analysis of variance (ANOVA) and testing all main effects and interaction effects.

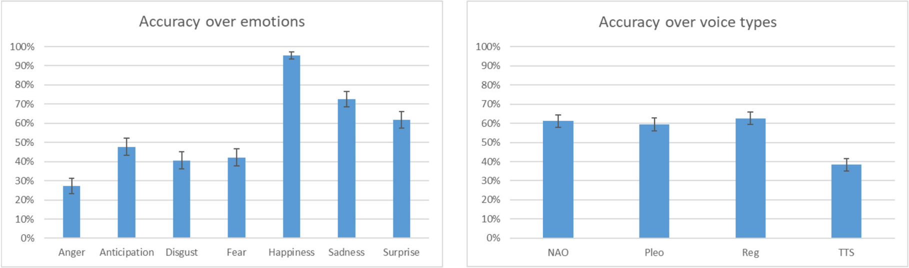


Fig. 5. Accuracy of perceiving emotions over emotions and voice types

For accuracy, there was no significant difference between Nao ($M = 57.1\%$, SD correct $= 0.5\%$) and Pleo ($M = 53.6\%$, SD correct $= 0.5\%$). The result revealed a statistically significant difference across voice types. However, there was significant interaction effect between emotions and voice types. For the multiple comparisons among voice types, paired-samples t-tests were conducted. All pairwise comparisons in this comparison applied a Bonferroni adjustment to control for Type-I error, which meant that we used more conservative alpha levels (critical alpha level = .0083 (0.05/6)). Participants showed significantly lower accuracy in a TTS voice than all other three voice types.

Table 5. Statistics for emotion perception (accuracy, clarity, suitability)

Measures	Conditions		Statistics
Accuracy (%)	Main effect for voice types		$F(3, 839) = 16.08, p < .0001$
	Interaction between voice types and emotions		$F(18, 839) = 3.39, p < .0001$
	Characterized NAO $M = 0.61$, SD $= 0.48$	TTS $M = 0.38$, $SD = 0.48$	$t(839) = 90.65$, $p < .0001$
	Characterized Pleo $M = 0.59$, SD $= 0.49$		$t(839) = 80.69$, $p < .0001$
	Regular $M = 0.62$, SD $= 0.48$		$t(839) = 97.33$, $p < .0001$
Clarity	Main effect for robots		$F(1, 838) = 4.9321, p = .0266$
	NAO $M = 5.22$, SD $= 1.77$	Pleo $M = 5.00$, $SD = 1.78$	$t(838) = 2.22$, $p = .0266$
	Main effect for voice types		$F(3, 838) = 99.40$, $p < .0001$
	Characterized NAO $M = 5.75$, SD $= 1.21$	TTS $M = 3.58$, $SD = 2.13$	$t(838) = 14.91$, $p < .0001$
	Characterized Pleo $M = 5.55$, SD $= 1.36$		$t(838) = 13.56$, $p < .0001$
	Regular $M = 5.56$, SD $= 1.31$		$t(838) = 13.65$, $p < .0001$
	Main effect for emotions		$F(6, 838) = 3.90$, $p = .0007$
	Disgust $M = 4.78$, SD $= 1.89$	Surprise $M = 5.54$, $SD = 1.70$	$t(838) = 3.91$, $p = .0001$
	Happiness $M = 4.89$, SD $= 1.70$		$t(838) = 3.35$, $p < .0008$

Table 6 shows how participants misclassified emotions.

Second, clarity and suitability of perceived emotions over robots, voice types, and presented emotions were analyzed as shown in Fig. 6. Clarity and suitability were rated using a 1 to 7 Likert-scale (1: Lowest, 7: Highest). Again, only answers that correctly recognized emotions were considered. Overall, there were differences found in clarity over emotions and voice types. For robots, there were no significant differences found in both clarity and suitability categories. Results were analyzed with a 2 (Robot) × 4 (Voice Type) × 7 (Emotions) repeated measures analysis of variance (ANOVA). The result revealed a statistically significant difference in clarity ratings over robots, voice types, and presented emotions. Nao showed significantly higher clarity rating than

Table 6. The confusion matrix between presented and perceived emotions

Perceived		Presented						
		Anger	Anticipation	Disgust	Fear	Happiness	Sadness	Surprise
Anger	Count	35	0	12	0	0	0	2
	Col %	27.3%	0.0%	9.4%	0.0%	0.0%	0.0%	1.6%
Anticipation	Count	0	61	3	0	5	0	8
	Col %	0.0%	47.7%	2.3%	0.0%	3.9%	0.0%	6.3%
Disgust	Count	2	0	52	0	0	0	0
	Col %	1.6%	0.0%	40.6%	0.0%	0.0%	0.0%	0.0%
Fear	Count	2	1	15	54	0	0	5
	Col %	1.6%	0.8%	11.7%	42.2%	0.0%	0.0%	3.9%
Happiness	Count	0	46	1	1	96	0	0
	Col %	0.0%	35.9%	0.8%	0.8%	75.0%	0.0%	0.0%
Sadness	Count	51	0	5	18	0	93	0
	Col %	39.8%	0.0%	3.9%	14.1%	0.0%	72.7%	0.0%
Surprise	Count	0	4	4	18	1	2	79
	Col %	0.0%	3.1%	3.1%	14.1%	0.8%	1.6%	61.7%
Indistinguishable	Count	38	16	36	37	26	33	34
	Col %	29.7%	12.5%	28.1%	28.9%	20.3%	25.8%	26.6%

Pleo. For the multiple comparisons among voice types, paired-samples t-tests were conducted and the result is shown in Table 5. TTS showed significantly lower clarity rating than the other three voice types. For the multiple comparisons among seven emotions, paired-samples t-tests were conducted. All pairwise comparisons in this item applied a Bonferroni adjustment to control for Type-I error, with an alpha levels = .0023 (0.05/21). Surprise showed significantly higher clarity rating than disgust and happiness. For suitability ratings, the result revealed that TTS showed significantly lower score than the other three voice types. No other differences were found.

Finally, the features by which to perceive emotions were analyzed with the results as shown in Table 7. The answers were collected from an open question ("What characteristics of the voice brought to mind that emotion?") and the number of occurrences of words was counted. Each participant was allowed to provide multiple answers for each comment. Most of the emotions were perceived from tone by 40.9%, pitch by 15.6%, and context by 12.4%.

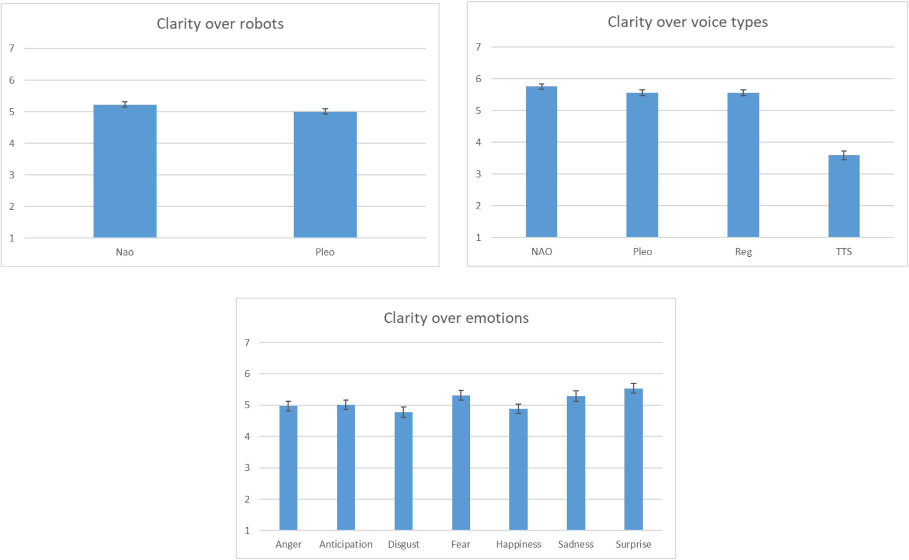


Fig. 6. The rating scores of clarity over robots, voice types, and emotions

3.3 Characteristics: Warmth, Honesty, and Trustworthiness

Figure 7 and Table 8 show the rating scores in trustworthiness over voice types and robots. Results were analyzed with a 2 (Robot) × 4 (Voice Type) repeated measures analysis of variance (ANOVA). For robots, there were no significant differences found in three categories. The result revealed a statistically significant difference in trustworthiness among voice types. There was no interaction effect between robots and voice types. For the multiple comparisons among voice types, paired-samples t-tests were conducted. Warmth and Honesty showed the exactly same pattern as trustworthiness (i.e., no other differences except for voice types with TTS being significantly lower).

3.4 Naturalness: Natural, Human-like, and Robot-like

Figure 8 and Table 9 show the rating scores in "robot-like" over voice types. Results were analyzed with a 2 (Robot) × 4 (Voice Type) repeated measures analysis of variance (ANOVA). For robots, there were no significant differences found in all three categories. The result revealed a statistically significant difference in the rating scores in "robot-like" among voice types. There was no interaction effect between robots and voice types. For the multiple comparisons among voice types, paired-samples t-tests were conducted. Participants showed significantly higher rating scores for TTS than all other three voice types. In addition, characterized Pleo showed significantly higher robot-likeness rating than regular voice. Natural and Human-like showed the exactly opposite pattern as Robot-like (i.e., TTS was significantly lower than others).

Table 7. The result of surveys on features that used to perceive emotions

Feature		Anger	Anticipation	Disgust	Fear	Happiness	Sadness	Surprise	Total
Contents	Count*	1	1	1	2	4	1	8	25
	Col %**	1.3%	1.0%	1.3%	1.9%	2.1%	0.4%	5.2%	2.0%
Context	Count	14	20	19	11	25	22	16	153
	Col %	18.7%	20.0%	23.8%	10.4%	13.0%	9.4%	10.3%	12.4%
Familiarity	Count	0	4	2	10	9	9	5	56
	Col %	0.0%	4.0%	2.5%	9.4%	4.7%	3.8%	3.2%	4.5%
Indistinguishable	Count	0	0	0	1	0	0	1	8
	Col %	0.0%	0.0%	0.0%	0.9%	0.0%	0.0%	0.6%	0.6%
Length	Count	0	0	2	1	6	2	3	19
	Col %	0.0%	0.0%	2.5%	0.9%	3.1%	0.9%	1.9%	1.5%
Loudness	Count	1	0	0	2	5	5	1	19
	Col %	1.3%	0.0%	0.0%	1.9%	2.6%	2.1%	0.6%	1.5%
Mood	Count	6	6	3	2	9	15	4	58
	Col %	8.0%	6.0%	3.8%	1.9%	4.7%	6.4%	2.6%	4.7%
Pitch	Count	14	10	8	22	29	45	25	192
	Col %	18.7%	10.0%	10.0%	20.8%	15.1%	19.2%	16.1%	15.6%
Pronunciation	Count	13	17	12	16	12	23	18	141
	Col %	17.3%	17.0%	15.0%	15.1%	6.3%	9.8%	11.6%	11.5%

(continued)

Table 7. (*continued*)

Feature		Anger	Anticipation	Disgust	Fear	Happiness	Sadness	Surprise	Total
Speed	Count	2	2	2	5	9	13	9	57
	Col %	2.7%	2.0%	2.5%	4.7%	4.7%	5.6%	5.8%	4.6%
Tone	Count	25	40	31	34	84	99	65	503
	Col %	33.3%	40.0%	38.8%	32.1%	43.8%	42.3%	41.9%	40.9%
Total	Count	75	100	80	106	192	234	155	1231
	Col %	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

* The total number of answers

** The proportion of the count in each column

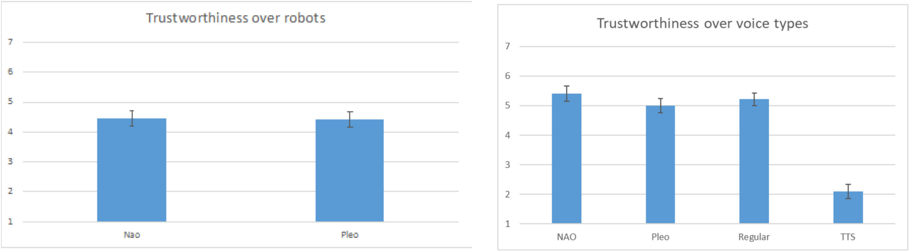


Fig. 7. The rating scores of trustworthiness over voice types

Table 8. Statistics for characteristics (trustworthiness)

Measures	Conditions		Statistics
Trustworthiness	Main effect for voice types		$F(3, 112.1) = 45.54$, $p < .0001$
	Characterized NAO $M = 5.40$, $SD = 1.49$	TTS $M = 2.09$, $SD = 1.37$	$t(112.1) = 10.07$, $p < .0001$
	Characterized Pleo $M = 5.00$, $SD = 1.39$		$t(112.1) = 8.83$, $p < .0001$
	Regular $M = 5.21$, $SD = 1.21$		$t(112.1) = 9.56$, $p < .0001$

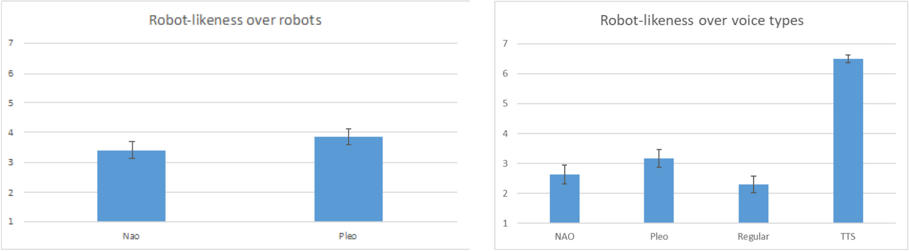


Fig. 8. The rating scores of robot-like over voice types

3.5 Preferences: Likability and Attractiveness

Figures 9, 10 and Table 10 showed the rating scores in "likability" and "attractiveness" over robots and voice types. Results were analyzed with a 2 (Robot) × 4 (Voice Type) repeated measures analysis of variance (ANOVA). For "likability", participants showed significantly higher rating scores for characterized Nao than characterized Pleo. The result also revealed a statistically significant difference in the rating scores over voice types. There was no interaction effect between robots and voice types. For the multiple comparisons among voice types, paired-samples t-tests were conducted. Participants showed significantly lower rating scores for TTS than all other three voice types. For "attractiveness", the result revealed a statistically significant difference in the rating

Table 9. Statistics for naturalness (natural, human-like, robot-like)

Measures	Conditions		Statistics
Robot-like	Main effect for voice types		$F(3, 98.7) = 55.07$, $p < .0001$
	Characterized Nao $M = 2.63$, $SD = 1.69$	TTS $M = 6.50$, $SD = 0.68$	$t(104.1) = 10.73$, $p < .0001$
	Characterized Pleo $M = 3.16$, $SD = 1.62$		$t(104.0) = 9.24$, $p < .0001$
	Regular $M = 2.30$, $SD = 1.48$		$t(90.9) = 10.87$, $p < .0001$
	Characterized Pleo $M = 3.16$, $SD = 1.62$	Regular $M = 2.30$, $SD = 1.48$	$t(90.9) = 2.38$, $p = .0195$

scores over voice types. For the multiple comparisons among voice types, paired-samples t-tests were conducted. Participants showed significantly lower rating scores for TTS than all other three voice types. Also, participants showed significantly lower rating scores for characterized Pleo than either characterized NAO or regular voice.

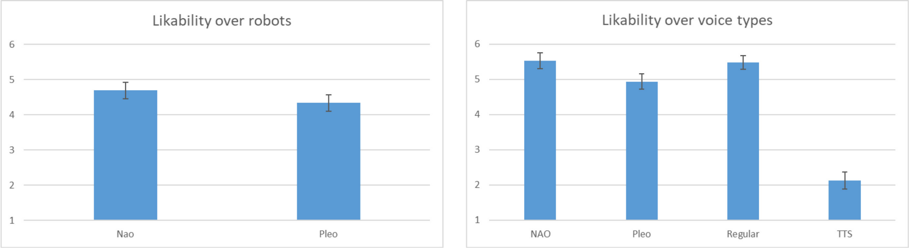


Fig. 9. The rating scores of likability over robots and voice types

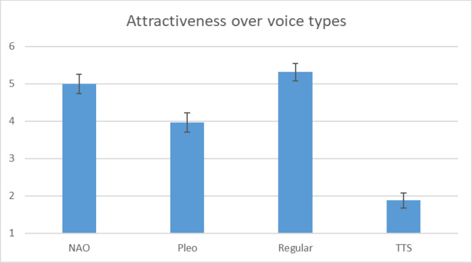


Fig. 10. The rating scores of attractiveness over voice types

Table 10. Statistics for naturalness (likability, attractiveness)

Measures	Conditions		Statistics
Likability	Main effect for robots		$F(3,\, 112.1) = 6.33$, $p = .0146$
	NAO $M = 4.69$, SD $= 1.85$	Pleo $M = 4.33$, $SD = 1.87$	$t(112.1) = 2.52$, $p = .0146$
	Main effect for voice types		$F(3,\, 112.1) = 40.32$, $p < .0001$
	Characterized NAO $M = 5.53$, $SD = 1.24$	TTS $M = 2.12$, $SD = 1.36$	$t(112.1) = 9.32$, $p < .0001$
	Characterized Pleo $M = 4.93$, $SD = 1.21$		$t(112.1) = 7.78$, $p < .0001$
	Regular $M = 5.48$, $SD = 1.09$		$t(112.1) = 9.56$, $p < .0001$
Attractiveness	Main effect for voice types		$F(3,\, 112.1) = 35.43$, $p < .0001$
	Characterized NAO $M = 5.00$, $SD = 1.48$	TTS $M = 1.87$, $SD = 1.09$	$t(112.1) = 8.36$, $p < .0001$
	Characterized Pleo $M = 3.96$, $SD = 1.46$		$t(112.1) = 5.66$, $p < .0001$
	Regular $M = 5.31$, $SD = 1.28$		$t(112.1) = 9.39$, $p < .0001$
	Characterized NAO $M = 5.00$, $SD = 1.48$	Characterized Pleo $M = 3.96$, $SD = 1.46$	$t(112.1) = 2.72$, $p = .0085$
	Regular $M = 5.31$, $SD = 1.28$		$t(112.1) = 3.74$, $p < .0004$

4 Discussion

To get a holistic picture of the effects of robot appearances, voices, and emotions types on users' perception on robots' emotions and characteristics, we conducted a preliminary study. Overall, results showed that the effects of voice types (human vs. TTS) seem to be larger than those of robot appearances on multiple dependent variables.

For emotion recognition accuracy, robot appearances did not show a significant difference between anthropomorphic (Nao) and zoomorphic (Pleo) robots. As expected, TTS showed significantly lower emotion recognition accuracy than other three human voice types. However, there were no differences in accuracy among the three voice types (characterized Nao, characterized Pleo, and regular). Also, there were no differences among the three human voice types for clarity and suitability. Taken together, this might imply the potential for using characterized voice for different purposes where appropriate (e.g., for children) without degrading emotion recognition accuracy, as long as it is a human voice. However, the result shows that the emotion recognition accuracy significantly varies depending on the expressed emotions. As shown in Fig. 5, happiness, sadness and surprise showed relatively higher accuracy than anger, disgust, and fear. Anticipation was placed in between. This might happen because happiness, sadness, and surprise are more common emotional states the participants can expect from the fairy tales. Anger, disgust, and fear are all negative-high arousal emotions. The participants might not expect these types of high strength, negative emotions from the fairy tales. However, the relationship between accuracy and each emotion shown in the present study is not in line with the results of the previous study [e.g., 18]. The difference might stem from different experimental settings (e.g., emotional words, prosody, context given by fairy tales, etc.). Thus, more iterative research is required to unpack the underlying mechanisms. For the misclassified emotions, valence showed a big impact. Based on the confusion matrix, anger (negative) was mostly misclassified as sadness (negative) (39.8%) and anticipation (positive) was mostly misclassified as happiness (positive) (35.9%). Based on the participants' self-report, most of the emotions were perceived from tone by 40.9%, pitch by 15.6%, and context by 12.4%, which shows that affective prosody is more critical than the content itself.

For robot characteristics, there was no statistically significant difference between the two robots, but there were differences between all human voices and TTS. We can cautiously infer that people did not perceive any differences among the regular, characterized Nao and characterized Pleo in terms of warmth, honesty, and trustworthiness.

Similarly, for naturalness, there was no statistically significant difference between the two robots even though participants consistently showed a tendency to perceive higher natural ($M = 4.39$, $SD = 2.18$ vs. $M = 3.88$, $SD = \mathit{1.93}$), higher human-like ($M = 4.73$, $SD = 2.07$ vs. $M = 4.2$ $SD = 1.9$), and lower robot-like ($M = 3.42$, $SD = 2.22$ vs. $M = 3.87$, $SD = 2.16$) from Nao, compared to Pleo. As expected, there were significant differences in these ratings between all human voices and TTS.

Finally, participants liked Nao significantly more than Pleo from the two robot types. They gave the highest rating to characterized Nao voice, followed by Regular, characterized Pleo, and TTS, even though only TTS was significantly different from other voice types. Participants also gave higher attractiveness rating to Nao ($M = 4.25$, $SD = 1.96$) than Pleo ($M = 3.83$, $SD = 1.82$), which did not reach the statistical significance level due to large variance. All three human voices were significantly more attractive than TTS. Also, both characterized Nao and regular voice were significantly more attractive than TTS. Again, this shows the potential for use of the characterized voice, at least, for anthropomorphic robots.

This exploratory study can provide practical guidelines for the voice design of various robots and further research studies. People seemed to generally perceive higher preference for an anthropomorphic robot compared a zoomorphic robot, which is in line with literature [8]. However, using either characterized or regular human voice did influence neither people's emotion recognition nor their perception about robot characteristics, such as warmth, honesty, and trustworthiness, as well as naturalness. Therefore, this study supports using human voice as a medium to express robots' emotions with a different voice design choice, depending on users, goal, and context.

5 Limitations and Future Work

The results of this experiment have been limited by several factors. First, only female voice was used in this study. Depending on the gender of the voice, the results might be different. Second, the sample size was small and not sufficient to draw a firm conclusion. Due to the COVID-19, the experiment was not run as much as planned. In future work, more participants with diversity should be recruited to generalize the results. Another limitation includes that the questionnaire for emotion recognition was an open-ended, which caused considerable confusion and lower accuracy rate. In future work, a questionnaire with more specific emotion options can be provided with additional open-ended input. Finally, the different speaker systems of different robots might also have influenced on the result (e.g., clarity) and should be addressed in the next study.

References

1. Dautenhahn, K., Woods, S., Kaouri, C., Walters, M.L., Koay, K.L., Werry, I.: What is a robot companion-friend, assistant or butler? In: Proceedings of the International Conference on Intelligent Robots and Systems 2005, IEEE/RSJ, pp. 1192–1197 (2005)
2. Vu, C., Cross, M., Bickmore, T., Gruber, A., Campbell, L.: U.S. Patent No. 8,935,006. In: U.S. Patent and Trademark Office, Washington, DC (2015)
3. Wrede, B., et al.: Research issues for designing robot companions: BIRON as a case study (2004)
4. Schirmer, A., Adolphs, R.: Emotion perception from face, voice, and touch: comparisons and convergence. Trends Cogn. Sci. **21**(3), 216–228 (2017)
5. Lohse, M., Hegel, F., Swadzba, A., Rohlfing, K., Wachsmuth, S., Wrede, B.: What can I do for you? Appearance and application of robots. Proc. AISB **7**, 121–126 (2007)
6. Seyama, I., Nagayama, S.: The uncanny valley: effect of realism on the impression of artificial human faces. Teleoperators Virtual Environ. **16**(4), 337–351 (2007)
7. Li, D., Rau, P., Li, Y.: A cross-cultural study: effect of robot appearance and task. Int. J. Soc. Robot. **2**(2), 175–186 (2010)
8. Hosseini, F., Hilliger, S., Barnes, J., Jeon, M., Park, H., Howard, M.: Love at first sight: mere exposure to robot appearance leaves impressions similar to interactions with physical robots. In: Proceedings of the 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 615–620. IEEE (2017)
9. Barnes, J., FakhrHosseini, M., Jeon, M., Park, H., Howard, A.: The influence of robot design on acceptance of social robots. In: Proceedings of the 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), pp. 51–55. IEEE (2017)
10. Hegel, F., Krach, S., Kircher, T., Wrede, B., Sagerer, G.: Understanding social robots: a user study on anthropomorphism. In: Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, Roman, pp. 574–579. IEEE (2008)
11. Bachorowski, J., Owren, M.: Sounds of emotion: production and perception of affect-related vocal acoustics. Ann. New York Acad. Sci. **1000**(1), 244–265 (2003)
12. Fischer, K., Jung, M., Jensen, L.: Emotion expression in HRI – when and why. In: Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (2019)
13. Frith, D., Frith, U.: The neural basis of mentalizing. Neuron **50**(4), 531–534 (2006)
14. Calvo, A., D'Mello, S.: Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans. Affect. Comput. **1**(1), 18–37 (2010)

15. Schirmer, A., Adolphs, R.: Emotion perception from face, voice, and touch: Comparisons and convergence. Trends Cogn. Sci **21**(3), 216–228 (2017)
16. Williams, G., Watts, N., MacLeod, C., Mathews, A.: Cognitive Psychology and Emotional Disorders. John Wiley & Sons, Oxford (1988)
17. Barnes, J., Richie, E., Lin, Q., Jeon, M., Park, H.: Emotive voice acceptance in human-robot interaction. In: Proceedings of the 24th International Conference on Auditory Display (2018)
18. Jeon, Myounghoon, Rayan, Infantdani A.: The effect of physical embodiment of an animal robot on affective prosody recognition. In: Jacko, Julie A. (ed.) HCI 2011. LNCS, vol. 6762, pp. 523–532. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21605-3_57