THE EFFECTS OF ROBOT VOICES AND APPEARANCES ON USERS' EMOTION RECOGNITION AND SUBJECTIVE PERCEPTION

SANGJIN KO[1], JACLYN BARNES[2], JIAYUAN DONG[1], CHUNGHYUK PARK [3], AYANNA HOWARD [4] AND MYOUNGHOON JEON[12*]

[1]*Grado Department of Industrial and Systems Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA*
[2]*Department of Computer Science, Michigan Technological University, Houghton, MI, USA*
[3]*Department of Biomedical Engineering, George Washington University, Washington DC, USA*
[4]*Department of Electrical and Computer Engineering, Ohio State University, Columbus, OH, USA*

[*]**Corresponding author**: Myounghoon Jeon **Tel.**: +1-540-231-3510, **E-mail:** myounghoonjeon@vt.edu

As the influence of social robots in people's daily lives grows, research on understanding people's perception of robots including sociability, trust, acceptance, and preference becomes more pervasive. Research has considered visual, vocal, or tactile cues to express robots' emotions, whereas little research has provided a holistic view in examining the interactions among different factors influencing emotion perception. We investigated multiple facets of user perception on robots during a conversational task by varying the robots' voice types, appearances, and emotions. In our experiment, twenty participants interacted with two robots having four different voice types. While participants were reading fairy tales to the robot, the robot gave vocal feedback with seven emotions and the participants evaluated the robot's profiles through post surveys. The results indicate that 1) the accuracy of emotion perception differed depending on presented emotions, 2) a regular human voice showed higher user preferences and naturalness, 3) but a characterized voice was more appropriate for expressing emotions with significantly higher accuracy in emotion perception, and 4) participants showed significantly higher emotion recognition accuracy with the animal robot than the humanoid robot. A follow-up study (*N*=10) with voice-only conditions confirmed that the importance of embodiment. The results from this study could provide the guidelines needed to design social robots that consider emotional aspects in conversations between robots and users.

Keywords: Social Robots; Conversational Agent; Emotive Voices; User Perception; User Preference

1. Introduction

As robots have become prevalent in people's daily lives, expectations for social robots have increased, which has brought numerous studies regarding Human-Robot Interaction (HRI). Robots are expected to play social roles such as a caregiver or companion that might serve as a friend or family member. In this regard, many studies have been conducted to facilitate richer and more natural interaction following human social norms. One of the ways of making the interaction more natural is attributing human characteristics to robots, called anthropomorphism (Schilhab, 2002). It can be humanlike appearance (i.e., superficial human characteristics) or humanlike mind (i.e., essential human characteristics) (Waytz, Heafner, & Epley, 2014). Some researchers have focused more on external design aspects (e.g., DiSalvo, Gemperle, Forlizzi & Kiesler, 2002), whereas others have investigated more on human mind (e.g., Epley, Waytz, & Cacioppo, 2007; Waytz et al., 2014).

Focusing on the appearance and behavior, research has been conducted on interactions between robots and users via multiple modalities incorporating variations in appearances, facial expressions, gestures, verbal communications, non-verbal sounds, and movements (Fong, Nourbakhsh, & Dautenhahn, 2003; Nabe et al., 2006; Nonaka, Inoue, Arai, & Mae, 2004). These modalities convey a wealth of information, influence user perception, and engage in establishing unique relationships between robots and users.

Focusing on the mental state, specifically on emotions, research has been conducted to see which factors influence user perception of robots' emotions. Although these studies have considered robots' facial expressions, voice (speech), body language, and posture as critical factors, the majority of emotion recognition research in HRI has focused on facial expressions (Calvo & D'Mello, 2010; Schirmer & Adolphs, 2017). Consequently, there has been little research on integrating both superficial and essential characteristics in one study to see interactions among the factors. A few exploratory studies have shown mixed results (Eyssel, De Ruiter, Kuchenbrandt, Bobinger, & Hegel, 2012; Eyssel, Kuchenbrandt, Hegel, & de Ruiter, 2012; McGinn & Torre, 2019; Nass, Foehr, Brave, & Somoza, 2001). As such, to fill this research gap, we investigated the effects of various factors–robots' appearances (robot types), voice types, and emotions on users' perception–clarity, characteristics, naturalness, and preference, as well as emotion recognition accuracy.

2. Related Work

2.1 *Emotion Taxonomy, Expression, and Perception*

There have been different theories proposed and studies conducted about (1) emotion classification, (2) emotion expression and (3) emotion perception in multiple domains, including psychology, psychiatry, neuroscience, and HRI research.

Largely, there are two types of emotion classification, including a dimensional approach and a categorical approach. In the dimensional approach, the circumplex model has been widely used with arousal and valence dimensions (Russell, 1980; Russell, 2017). An individual emotional state can be positioned on the Cartesian coordinate depending on the levels of arousal and valence. In the categorical approach, researchers often assume that people have basic emotions. Ekman's six basic emotions (Ekman and Cordaro, 2011) (happiness, sadness, fear, anger, surprise, and disgust) have been one of the most widely mentioned emotion sets in emotion-related research in Psychology, Human Factors, Affective Computing, and HRI (Cakmak, Hoffman, & Thomaz, 2016; Calvo & D'Mello, 2010; Reisenzein et al., 2013). Basic emotions (Ekman and Cordaro, 2011) are known to have unique features such as signal, physiology, and antecedent events, and common characteristics with other emotions such as rapid onset, short duration, unbidden occurrence, automatic appraisal, and coherence among responses. Ekman (1992) argued that these basic emotions are expressed and recognized cross-culturally. However, there has been still much criticism about the basic emotion theories (Ortony, 2021). See (Jeon, 2017) for more discussions on generic taxonomy and theories about emotions in the context of Human Factors and Human-Computer Interaction. In our everyday lives, we typically describe our emotional states using categorical terms, rather than dimensional terms; for example, during a conversation, people usually express happy feelings as "happiness" (categorical) but not "an emotion that is high arousal with positive valence" (dimensional). Therefore, we provided the emotional states using the categorical approach in the present study. Research also shows that these basic emotions are pervasive over the world (Ekman & Cordaro, 2011). In addition to Ekman's six emotions, we added 'anticipation', one of the Pluchik's basic emotions (1980) because the passage of our stories included anticipation. With the addition of anticipation, we were able to have the second positive emotional state in our study in addition to happiness.

In terms of emotion expression, Darwin and Prodger (1998) proposed three causal origins of expressions; immediate benefits (e.g., increasing one's body size to intimidate an opponent), effective communications (e.g., lowering one's body to signal submission), and vestigial byproducts that may not serve a useful role (e.g., trembling in fear). Previous studies also showed that emotion expressions exhibited useful functions (e.g., widening eyes to maximize the visual field during fear) and emotional vocal expressions effectively manipulated the behavior of perceivers (Bachorowski & Owren, 2003; Susskind et al., 2008). Among these, the current study focuses more on the effective communications and vocal expressions of emotion.

Emotion perception is the identification of emotionally salient information in the environment, including verbal (lexico-semantic) and nonverbal (intonational, facial, visual, and body movement) cues to the emotions of other people (Phillips, 2003). Emotion is one of the perceptual representations of social cues along with intentionality and eye direction (Decety, 2010; Mitchell & Phillips, 2015). In line with this, human social and emotional behaviors are highly intertwined (Beer & Ochsner, 2006). 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 (Frith & Frith, 2006).

As provided from previous theories, emotion expression and emotion perception play a critical role in human-robot interactions and are widely studied in a range of disciplines. Researchers commonly argue that these emotion-related expressions and perceptions can be achieved through both visual and auditory stimuli. However, previous studies have been dominated by facial emotions and other modalities such as vocal and tactile processing have been less frequently considered (Calvo & D'Mello, 2010; Schirmer & Adolphs, 2017). In this regard, in our work, we focused more on auditory stimuli by including various emotive voices, representing seven different emotions and investigated the differences in users' emotion perception.

2.2 *User Perception on Robots from Embodiment, Appearance, and Sounds*

There have been studies focused on examining the impact of robots' embodiment, appearance, and auditory displays on HRI.

The physical embodiment of robots could impact user perception positively and promote HRI in many social situations. With the embodiment, social robots brought many benefits to user experience. For example, participants reported higher satisfaction in the shopping mall (Sakai et al., 2021) and higher enjoyment while playing a chess game (Pereira et al., 2008) with the physical embodied robots than the disembodied ones. Many research studies also suggested that the embodiment of social robot engaged longer interaction duration (Rodriguez-Lizundia et al., 2015), increased human empathy towards the robots (Kwak et al., 2013; Seo et al., 2015), and enhanced compliance with robots' instruction and made the interaction more natural than the virtual or simulated ones (Li, 2015). Because the presence of the social robot played an important role in HRI, we used physical robots to emit sounds instead of using just a speaker in the present study.

The appearance of robots was considered as an important factor of user perception to support interaction since anthropomorphism allows people to give robots lifelike qualities (e.g., intentions, emotions, etc.) (Seo, Geiskkovitch, Nakane, King, & Young, 2015; Sharma, Hildebrandt, Newman, Young, & Eskicioglu, 2013). Barnes, FakhrHosseini, Jeon, Park, and Howard (2017) and FakhrHosseini, Hilliger, Barnes, Jeon, Park, and Howard (2017) showed that participants preferred robots which resemble animals or humans over imaginary creatures or robots highly deviating from existing creatures. Barnes et al. (2017) compared five different robots (Robosapien, Pleo, Zoomer, Romo, and Mindstorm) which are humanoid, zoomorphic, fantastical, and mechanistic. Participants showed different user perception across robots but similar patterns before and after interacting with robots. Another study (Saint-Aimé, Le-Pevedic, Duhaut, & Shibata, 2007) suggested that a companion robot requires a certain level of emotional expression for a good interaction to occur with children. Also, people accept and trust robots more when the robots show some emotional activities (Lowe, Barakova, Billing, & Broekens, 2016).

The effects of robots' voices have also been investigated in relation to user perception. These studies have employed different types of sounds, such as human voices, TTS voices, and beeping sounds in conjunction with various robots having different form factors. Research showed that participants assumed that a human voice was more capable than a TTS voice, and they anthropomorphized robots with human voices (Sims et al., 2009; Walters, Syrdal, Koay, Dautenhahn, & Te Boekhorst, 2008). Similar to the pattern in user perception on robots' appearances, people showed a tendency to prefer interacting with robots similar to themselves in voice characteristics, including human-like speech style and accent, and gender (Eyssel, De Ruiter, Kuchenbrandt, Bobinger, & Hegel, 2012; Eyssel, Kuchenbrandt, Hegel, & de Ruiter, 2012). A recent exploratory study (McGinn & Torre, 2019) showed that gender and naturalness of vocal manipulations strongly affected user perception.

Although various aspects of user perception from visual and auditory cues have been examined through exploratory studies, many of them focused more on users' preferences based on subjective self-report measures (e.g., Barnes et al., 2017; FakhrHosseini et al., 2017). To tackle these issues, in our work, we applied both qualitative and quantitative measures by examining user perception from broader perspectives.

2.3 *Emotions in HRI and Emotive Voices*

An effective HRI could be achieved or improved by involving an appropriate emotional communication from social robots (Liu et al., 2016). Regarding previous empirical studies on emotive communications in HRI, diverse aspects of communication such as gesture, appearance, style of speech, prosody, and context have been investigated. Implementing emotional features to social robots might enhance children's learning skills and engaged the learning process. Conti et al. (2020) in their storytelling environment showed that children can memorize more details of a tale if the robot narrates with an expressive social behavior, even compared to the static inexpressive human storyteller. Also, the emotional appearance of robots was proposed for creating a more suitably moral agent (Coeckelbergh & Technology, 2010) or providing interactive interventions for children with autism spectrum disorder (ASD) (e.g., Barnes, Park, Howard, & Jeon, 2020; Bevill, Park, Kim, Lee, Rennie, Jeon, & Howard, 2016). With the results from previous studies, we considered emotion as an indispensable factor in HRI.

To investigate the impact of emotion expressions in HRI, there have been various research projects regarding emotional conversations that are driven by either internal states, behaviors, or situations (Feldmaier, Stimpfl, & Diepold, 2017; Jung, 2017; Song & Yamada, 2017). These studies were based on communication theories about emotion expressions: 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 (Fischer, Jung, & Jensen, 2019).

Regarding emotive voices on social robots, crucial features such as the style of speech, gender, and prosody have been widely investigated through exploratory studies in HRI. FakhrHosseini et al. (2017) emphasized the importance of the congruency between anthropomorphism in the appearances and the style of speech. Their study showed that only when the human-like robot speaks with emotional expressions, participants perceive the robot as their social companion. Kishi et al. (2013) showed that the integration of dynamic emotional expressions and movements made the humanoid robot more attractive, more favorable, more useful, and less mechanical-like. Gender stereotypes were also examined with the explicit gender (from name and voice) and implicit gender (from personality) in a previous study (Bryant, Bornstein, & Howard, 2020; Kraus, Kraus, Baumann, & Minker, 2018). For example, in Kruas et al.'s study, no gender stereotypes were found for the explicit gender, but implicit gender showed a strong effect on trust and likability in the stereotypical male task. Participants perceived that the male personality robot (dominant, confident and assertive utterances) is more trustable, reliable, and competent than the female personality robot (agreeable and warm utterances), while the female personality robot is more likable. A social robot's voice type could also play a critical role in emotive conversation. Eyssel et al. (2012) examined the effects of vocal cues that reflected both the gender of robot voices (male, female) and voice types (robot-like, human-like). It showed a human voice was rated more likable than the synthetic voice. Jeon and Rayan (2011) examined the effects of expressing affective prosody from a zoomorphic robot (Pleo) and showed a higher accuracy of emotion perception in a physical one than a virtual one. Half of the participants mentioned that the human voice generated from the zoomorphic robot was awkward and a characterized or a cartoon-like voice might be more appropriate. Recently, Ko, Liu, Mamros, Lawson, Swaim, Yao, and Jeon (2020) have investigated the effects of different voice types with two types of robots (same as in the present study) on robot emotion perception. Text-to-speech (TTS) condition showed significantly lower emotion recognition accuracy than other human voices, but the robot type (humanoid vs. animal) did not influence emotion recognition accuracy or other robot perceptions. However, in their study the voice was recorded by *female* students, not voice experts, which might have led to different results from the present study.

Overall, emotive voice associated with social robots is still veiled in various aspects such as acoustic characteristics, voice types, gender, and prosody. Since previous studies found contrasting results toward voice types in social robots, we narrowed down the scope and focused on the differences in emotion recognition accuracy and user perception on four different voice types in the present study.

2.4 *Research Questions and Hypotheses*

From this background, we tried to attain a deeper understanding of the effects of robot types, voice types, and emotion types on users' perception towards robots and their emotions. Especially, we aimed to answer the research questions as follows:

- RQ1: How do robot types, voices, emotions, and their interactions have impacts on participants' recognition of different robots' emotional states?
    - H1a: There will be no effects of robot types on emotion recognition accuracy (Ko et al., 2020).
    - H1b: Participants will show higher emotion recognition accuracy in the human voice over TTS voice (Ko et al., 2020).
    - H1c: There will be no emotion recognition accuracy difference between regular human and characterized human voices (Ko et al., 2020).
    - H1d: Different emotions will show different emotion recognition accuracy (Jeon & Rayan, 2011; Ko et al., 2020).
- RQ2: How do robot types, voices, and their interactions have impacts on participants' perception of robots' warmth, honesty, and trustworthiness?

- H2a: Participants will show higher ratings on the humanoid robot than the animal robot in warmth, honesty, and trustworthiness ratings (Barnes et al., 2017; Hosseini et al., 2017).
- H2b: There will be differences in warmth, honesty, and trustworthiness ratings among the different voice conditions (Ko et al., 2020).

- RQ3: How do robot types, voices, and their interactions have impacts on participants' preference of robots?
  - H3a: Participants will prefer the humanoid robot over the animal robot (Ko et al., 2020).
  - H3b: Participants will prefer the human voice over TTS (Eyssel, De Ruiter, Kuchenbrandt, Bobinger, & Hegel, 2012; Eyssel, Kuchenbrandt, Hegel, & de Ruiter, 2012).
  - H3c: There will be no preference difference between regular human and characterized human voices (Ko et al., 2020).

To address these research questions, we conducted an experimental study with young adults. Our participants read the two fairy tales to two types of robots each (human-like and animal-like). The robots made emotional comments using four different voices (regular human, characterized human-like, characterized animal-like, and TTS) with seven emotions (six basic emotions + anticipation).

3. Method

3.1 *Experimental Design*

Twenty university students participated in the study (Age: M = 22.1, SD = 2.97). Twelve participants identified as male and the other eight participants identified as female. Participants were ethnically diverse (6 Asians, 1 Hispanic, 11 Caucasian, and 2 Multiracial). Participants participated in the experiment for at most 2 hours and participants were compensated with \$20 (\$10 per hour). All participants agreed to participate after reviewing the consent form approved by the university Institutional Review Board (IRB).

A 2 (robots) × 4 (voice types) × 7 (emotions) within-subjects design was applied. Therefore, 8 different combinations of robots and voice types were provided to each participant with all 7 emotions. Two social robots, NAO and Pleo, were used in the experiment. Four voice types were referred to two Characterized voices (NAO and Pleo), a Regular voice, and a TTS voice. There were two human voices and two TTS engines (Group A and Group B in Table 2) used. They were alternatively mapped to both robots and both stories across participants. More details were explained in the Procedure section.

3.2 *Robotic Systems and Stimuli*

Two robots, NAO and Pleo, having different appearances and features were employed in the experiment (Figure 1). We used these two robots, which represent a humanoid robot and zoomorphic robot each, to contrast the effects that robotic appearance has on people's emotion perception. NAO is a small-size humanoid robot (Height: 57.4 cm, Length: 27.4 cm, Width 31 cm) having similarity to human and Pleo is a zoomorphic robot (Height: 20.3 cm, Length: 38.1 cm, Width 10.2 cm) which looks like a little dinosaur. Both robots played recorded auditory feedback, which were emotive utterances, to participants following the storylines. The task selected to provide structure to the interaction and a more realistic context for conversational emotions was to read fairy tales to the robots. Two different stories ("The three little pigs" and "The boy who cried wolf") were used in this experiment. These two stories are simple narratives with easy vocabulary and globally well-known so that participants can easily read to the robots even if they are not native speakers. Crucially, we could include all of the emotions we wished to study within the framework of each story. Fairy tales seemed fitting given the childlike appearances of both robots and are suitable for use with a broad range of other populations for replication of the present study.

Four voice types were created for seven emotional expressions. We first categorized different voice types as a TTS voice and a recorded human voice. The human voices were provided by two male voice actors 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 TTS voices were generated using text-to-speech (Williams, Watts, MacLeod, & Mathews, 1988) engines. Microsoft's David voice and the iOS Alex voice were used, which were provided by default with the respective operating systems. These TTS voices included no emotional information beyond the words themselves. Characterized voices for each NAO and Pleo were designed to exaggerate emotional expressions with the robots' characters. These characterized voices were provided by voice actors who majored in performing arts while envisioning the characteristics of robots from their appearances. Direction for the characterization process, vocal performances, and recording was provided by a professional voice actor and professor of theatre who teaches voice and acting in the Department of Visual and Performing Arts. To control for gender effects, only characteristically male voices were used. While the same control effect could have been achieved using female voices, male voices were chosen based on the availability of the actors while designing the study. The example recordings of each voice type are provided on the web for other researchers and educators to get an idea of what participants heard during evaluation: https://osf.io/m8h64/.

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. Ekman's basic emotions have four negative emotions (anger, disgust, fear, and sadness), but have only one positive emotion (happiness); surprise can be either. A previous study showed that valence might influence people's emotion recognition accuracy (Ko et al., 2020). In Bänziger, Grandjean, and Scherer's study (2009), participants were examined to recognized emotions, and the emotion recognition results showed a higher emotion recognition accuracy score on positive emotions, such as happiness, than the negative emotions, such as anxiety, sadness, and disgust. To make a balance between positive and negative emotions, the seventh emotion, anticipation, was chosen from Plutchik's eight basic emotions (1980). Its inclusion allowed us to add one more positive emotion in addition to happiness. The seven emotions fit into both stories ("The three little pigs" and "The boy who cried wolf") as depicted in Table 1. The content of these emotional phrases was not considered as an experimental factor in the present study because all participants received the same treatments (eight combinations of robots and voice types) during the study.

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!

3.3 *Procedure*

A single participant participated in each session. Note that this study was completed before the COVID pandemic. Thus, there was no COVID-relevant procedure. After the consent form procedure, 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 each. The presented order of each condition was counterbalanced. In each condition, the participant was instructed to read the script aloud in front of a robot and wait for and listen to the robots' emotional comments at various points in the story, which were marked down in the given script. Before reading the script and listen to the robots, participants were explained about all possible voice types they would interact with during the experiment. All voice clips were embedded in each robot and the voice was triggered by a remote controller which was controlled by an experimenter. Participants were aware that the robots were not acting autonomously. Other than vocal communication, the participants did not do any physical interaction with the robot.

The experimental environment (upper) and the whole procedure including each step (lower) are depicted in Figure 1.

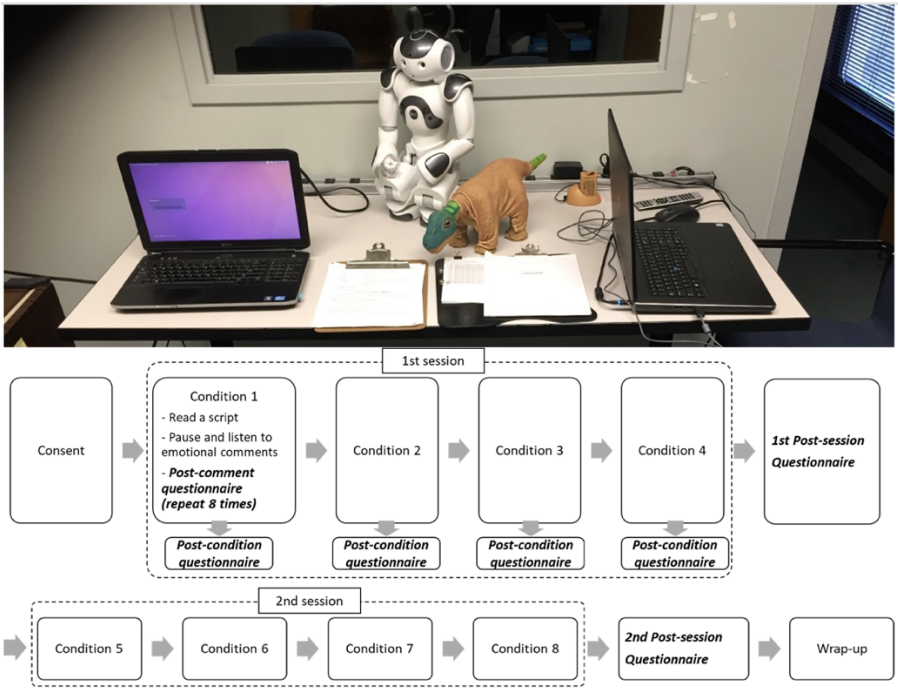


Figure 1. Experiment settings with NAO (left) and Pleo (right) (upper) and experimental procedure including each step (lower).

The participants were asked to fill out several questionnaires after listening to each comment generated from the robots, 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 perception and characteristics (Warmth, Honesty, Trustworthiness), naturalness and preferences (Likability, Attractiveness) of presented emotions. The questionnaire consisted of open questions, seven-point Likert scales, and single-choice questions. Related questions were asked and each category was rated using a 1 to 7 Likert-scale (1: Lowest, 7: Highest) (Appendix A).

Presented orders for emotions in the two stories were different but the order in each story was fixed to maintain the storylines. Two different stories having the same 7 emotions presented and two different voice groups having the same characteristics but recorded by different voice actors and two different TTS engines were employed to generalize the results. Each participant experienced both human voice actors and both TTS sounds. The examples of the presented order are depicted in Table 2. To validate the equivalence in emotion recognition accuracy, clarity, suitability, and preference, after the experiment, the results were analyzed (Table 3) showing similar results in all categories. The experiment took 2 hours at most as approved by IRB. Most participants completed it within 1.5 to 2 hours.

Table 2. 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	Characterized Pleo	TTS	Characterized NAO	Regular
	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	Characterized Pleo	TTS	Characterized NAO	Regular
	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*
***Group A and Group B had the same characteristics but were recorded by different voice actors and TTS engines*

Table 3. Accuracy, clarity, suitability, and preference over stories and voice groups.

		Accuracy	Clarity	Suitability	Preference
Story	The Boy Who Cried Wolf	57.0%	5.13	4.64	4.10
	The Three Little Pigs	56.1%	5.25	4.78	4.38
Voice Group	Group A	58.6%	5.05	4.53	4.16
	Group B	53.0%	5.11	4.68	4.33

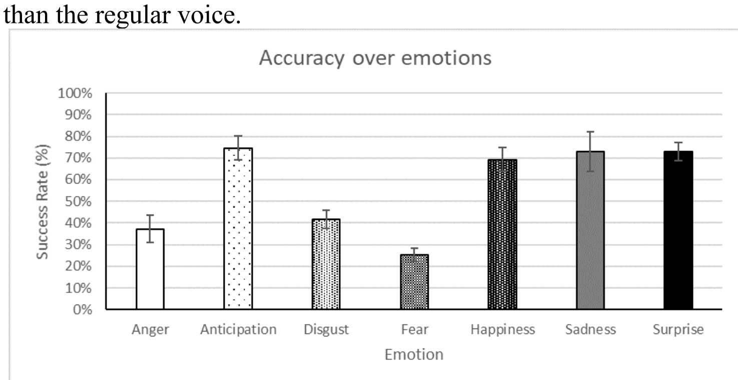
4. Results

4.1 *Data Collection*

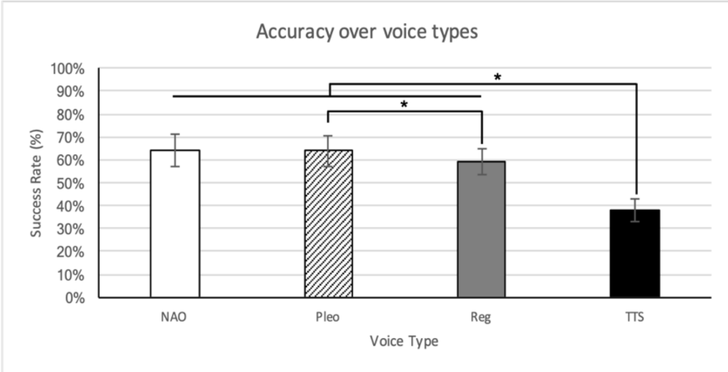
The answer to open questions regarding emotions was interpreted by two examiners. Each examiner categorized all the answers into seven pre-defined emotions or marked as 'indistinguishable' if the answers do not fall into any categories. Two examiners worked independently, and the inter-rater reliability test showed the high coefficient value of Cronbach Alpha using variance (=0.86). If interpretations from examiners were different, a third examiner reviewed the answers and decided which emotion the answer fell into.

4.2 *Emotion Perception: Accuracy, Clarity, Suitability, and Features*

First, the emotion recognition accuracy, defined as the proportion of correct emotion answers, was analyzed. Figure 2 and Table 4 show the descriptive statistics of emotion recognition accuracy across presented emotions, voice types, and robots. Regarding presented emotions, anger, disgust, and fear showed lower accuracies than positive emotions, such as anticipation and happiness. The accuracies for anger, disgust, and fear were 37.5%, 41.9%, and 25.6%, which were all lower than 50%. These three extreme conditions were excluded in statistical analysis to minimize the effects of biased data sets. Results were analyzed with the aligned rank transform (ART) (Wobbrock, Findlater, Gergle, & Higgins, 2011) for factorial analyses since there are 3 factors (Robots, Voice Types, and Emotions) and dependent variable (1: correct, 0: wrong) is not normally distributed. To apply ART, we first computed residuals and estimated effects for all main and interaction effects. After computing aligned response, we assigned averaged ranks. With this data, we could perform a full-factorial repeated measures analysis of variance (ANOVA) following the guidelines of Wobbrock et al. (2011). The ART allowed analyzing the aligned-ranked data with a 2 (Robots) x 4 (Voice Types) x 4 (Emotions) repeated measures ANOVA and testing all main effects and interaction effects. The result revealed a statistically significant difference across robots and voice types. However, there was no significant interaction effect between robots and voice types. For the multiple comparisons among voice types, paired-samples t-tests were conducted. All pairwise comparisons applied a Bonferroni adjustment to control for Type-I error in this study, which meant that we used more conservative alpha levels (critical alpha level = .0083 (0.05/6)). Participants recognized emotions more accurately with Pleo than NAO. Participants showed significantly lower emotion recognition accuracy in the TTS voice than all other three voice types. Moreover, the characterized Pleo voice showed significantly higher emotion recognition accuracy than the regular voice.



(a)



(b)

Figure 2. Accuracy of perceiving emotions over emotions (a) and voice types (b) (*: p <0.0083)

Table 4. Statistics for emotion recognition (accuracy).

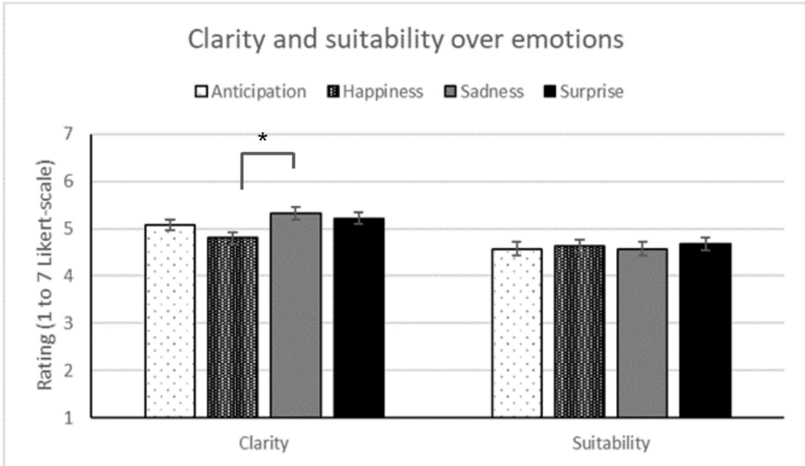
Measures	Conditions		Statistics
Accuracy (%)	Main Effect for Robots		$F(1, 607) = 4.27, p = .0393$
	NAO Robot $M = 0.68, SD = 0.47$	Pleo Robot $M = 0.76, SD = 0.43$	
	Main Effect for Voice Types		$F(3, 607) = 16.07, p < .0001$
	Characterized NAO $M = 0.64, SD = 0.48$	TTS $M = 0.38, SD = 0.49$	$t(19) = 5.78, p < .0001$
	Characterized Pleo $M = 0.64, SD = 0.48$	TTS $M = 0.38, SD = 0.49$	$t(19) = 6.15, p < .0001$
	Regular $M = 0.59, SD = 0.49$	TTS $M = 0.38, SD = 0.49$	$t(19) = 3.34, p = .0009$
	Characterized Pleo $M = 0.64, SD = 0.48$	Regular $M = 0.59, SD = 0.49$	$t(19) = 2.80, p = .0053$

Table 5 shows the confusion matrix between presented and perceived emotions. Anger was mostly misclassified as sadness (32.50%), disgust was mostly misclassified as surprise (18.75%) or undistinguished (14.38%), and fear was mostly misclassified as anticipation (28.75%). Interestingly, 21.25% of happiness was also undistinguished even though it showed higher emotion recognition accuracy than anger, disgust, and fear did.

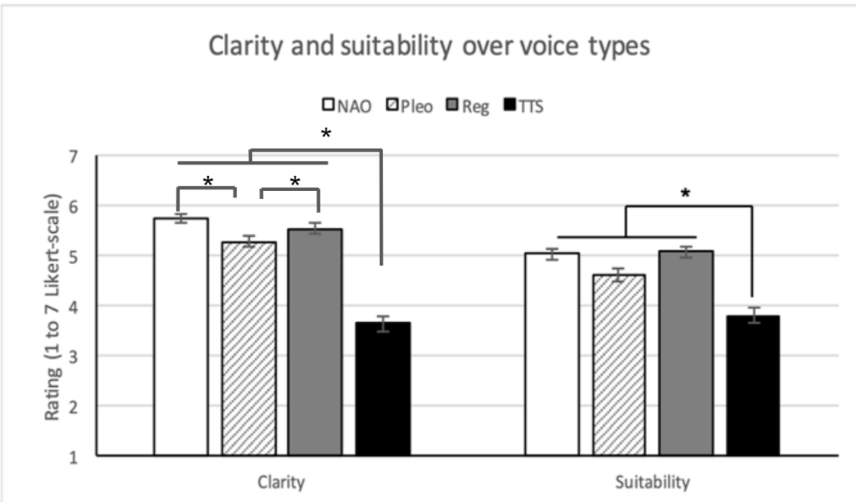
Table 5. The confusion matrix between presented and perceived emotions (grey: most misclassified)

Perceived \ Presented		anger	anticipation	disgust	fear	happiness	sadness	surprise
anger	Count	60	1	7	6	0	7	5
	Col %	37.50	0.63	4.38	3.75	0.00	4.38	3.13
anticipation	Count	15	120	14	46	13	2	11
	Col %	9.38	75.00	8.75	28.75	8.13	1.25	6.88
disgust	Count	8	1	67	0	2	0	0
	Col %	5.00	0.63	41.88	0.00	1.25	0.00	0.00
fear	Count	0	0	14	41	0	0	1
	Col %	0.00	0.00	8.75	25.63	0.00	0.00	0.63
happiness	Count	1	9	1	0	111	1	3
	Col %	0.63	5.63	0.63	0.00	69.38	0.63	1.88
sadness	Count	52	1	4	27	0	118	9
	Col %	32.50	0.63	2.50	16.88	0.00	73.75	5.63
surprise	Count	5	2	30	10	0	7	117
	Col %	3.13	1.25	18.75	6.25	0.00	4.38	73.13
indistinguishable	Count	19	26	23	30	34	25	14
	Col %	11.88	16.25	14.38	18.75	21.25	15.63	8.75

Second, clarity and suitability of perceived emotions over robots, voice types, and presented emotions were computed with the results as shown in Figure 3 and Table 6. Clarity and suitability were rated using a 1 to 7 Likert-scale (1: Lowest, 7: Highest). We considered only responses with correctly recognized emotions. The clarity and suitability scores were measured for the present emotions; therefore, participants had to first recognize the emotions correctly to have their rating scores to be considered for the clarity and suitability measurements without bias. Overall, there were differences found in clarity over emotions and voice types and suitability over voice types. For robots, there were no significant differences found in both categories. Results were analyzed with a 2 (Robot) x 4 (Voice Type) x 7 (Emotions) repeated measures analysis of variance (ANOVA). The result revealed a statistically significant difference in clarity ratings among voice types and presented emotions. For the multiple comparisons among voice types, paired-samples t-tests were conducted. The TTS voice had a significantly lower clarity rating than the characterized and regular voices. In addition, the characterized Pleo voice had a significantly lower clarity rating than the characterized NAO and regular voices. Participants reported Sadness as having a significantly higher clarity rating than Happiness. There was also a significant interaction effect between voice types and presented emotions. It is assumed that the relatively too low rating score of TTS voice compared to the other three voices caused the interaction effects. In suitability ratings, the result revealed a statistically significant difference among voice types. There were no significant interaction effects between emotions and voice types. For the multiple comparisons among voice types, paired-samples t-tests were conducted. Participants showed significantly lower rating scores in the TTS voice than all other three voice types.



(a)



(b)

Figure 3. The rating scores of clarity and suitability over emotions (a) and voice types (b) (*: p <0.05).

Table 6. Statistics for clarity and suitability.

Measures	Conditions		Statistics
Clarity	Main Effect for Voice Types		$F(3, 52.86) = 18.32$, $p < .0001$
	Characterized NAO $M = 5.61$, $SD = 1.05$	TTS $M = 3.63$, $SD = 1.67$	$t(19) = 9.89$, $p < .0001$
	Characterized Pleo $M = 5.10$, $SD = 1.38$		$t(19) = 6.52$, $p < .0001$
	Regular $M = 5.76$, $SD = 1.22$		$t(19) = 11.36$, $p < .0001$
	Characterized NAO $M = 5.61$, $SD = 1.05$	Characterized Pleo $M = 5.10$, $SD = 1.38$	$t(19) = 3.39$, $p = .0010$
	Regular $M = 5.76$, $SD = 1.22$		$t(19) = 3.82$, $p = .0002$
	Main Effect for Emotions		$F(6, 115.1) = 3.25$, $p = .0055$
	Sadness $M = 5.41$, $SD = 1.47$	Happiness $M = 5.00$, $SD = 1.45$	$t(19) = 2.02$, p = .0456
	Interaction between Voice Types and Emotions		$F(18, 312.3) = 2.77$, $p = .0002$
Suitability	Main Effect for Voice Types		$F(3, 57.58) = 6.59$, $p = .0007$
	Characterized NAO $M = 5.02$, $SD = 1.59$	TTS $M = 3.79$, $SD = 1.63$	$t(19) = 3.96$, $p = .0002$
	Characterized Pleo $M = 4.61$, $SD = 1.77$		$t(19) = 3.07$, $p = .0032$
	Regular $M = 5.07$, $SD = 1.47$		$t(19) = 3.86$, $p = .0003$

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. After reading through each participant's answer, we categorized their comments into different feature groups. Terms used in the participant's answers that fell into specific features were counted. Most of the emotions were perceived from tone by 29.53%, words by 19.29%, and pitch by 17.72%. For each emotion, speech tone highly influenced perceiving anger (29.58%), anticipation (32.12%), happiness (32.56%), sadness (32.89%), and surprise (27.97%). Different from these emotions, disgust was mostly perceived by words (26.19%). Fear was perceived by different features such as pitch (24.49%), words (22.45%), and tone (20.41%).

Table 7. The result of surveys on features that used to perceive emotions. (grey: most used)

Feature		Anger	Anticipation	Disgust	Fear	Happiness	Sadness	Surprise	Total
Context	Count*	2	9	1	3	6	8	7	36
	Col %**	2.82%	6.57%	1.19%	6.12%	4.65%	5.37%	4.90%	4.72%
Familiarity	Count	3	7	5	7	5	9	6	42
	Col %	4.23%	5.11%	5.95%	14.29%	3.88%	6.04%	4.20%	5.51%
Length	Count			7		2	4	4	17
	Col %	0.00%	0.00%	8.33%	0.00%	1.55%	2.68%	2.80%	2.23%
Loudness	Count	8	5	4	2	3	3	5	30
	Col %	11.27%	3.65%	4.76%	4.08%	2.33%	2.01%	3.50%	3.94%
Mood	Count	3	5	5	1	8	6	6	34
	Col %	4.23%	3.65%	5.95%	2.04%	6.20%	4.03%	4.20%	4.46%
Pitch	Count	12	26	10	12	26	31	18	135
	Col %	16.90%	18.98%	11.90%	24.49%	20.16%	20.81%	12.59%	17.72%
Pronunciation	Count	4	1	3	2	1	4	8	23
	Col %	5.63%	0.73%	3.57%	4.08%	0.78%	2.68%	5.59%	3.02%
Speed	Count	2	5	4	1	2	15	9	38
	Col %	2.82%	3.65%	4.76%	2.04%	1.55%	10.07%	6.29%	4.99%

Tone	Count	21	44	19	10	42	49	40	225
	Col %	29.58%	32.12%	22.62%	20.41%	32.56%	32.89%	27.97%	29.53%
Words	Count	9	28	22	11	27	14	36	147
	Col %	12.68%	20.44%	26.19%	22.45%	20.93%	9.40%	25.17%	19.29%
Vague	Count	7	7	4		7	6	4	35
	Col %	9.86%	5.11%	4.76%	0.00%	5.43%	4.03%	2.80%	4.59%
Total	Count	71	137	84	49	129	149	143	762
	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*

4.3 *Characteristics: Warmth, Honesty, and Trustworthiness*

Figure 4 and Table 8 showed the rating scores in warmth, honesty, and trustworthiness over voice types and robots. For robots, there were no significant differences found in three categories. Because by definition, emotions are short-lasting "states", not long-lasting "traits", the factor emotion was not analyzed in the following perception sections. Results were analyzed with a 2 (Robot) x 4 (Voice Type) repeated measures analysis of variance (ANOVA). First, the result revealed a statistically significant difference in warmth 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. In all three categories, the results commonly showed the lowest score in a TTS voice. Also, there were no significant differences among the characterized NAO, Pleo, and regular voices.

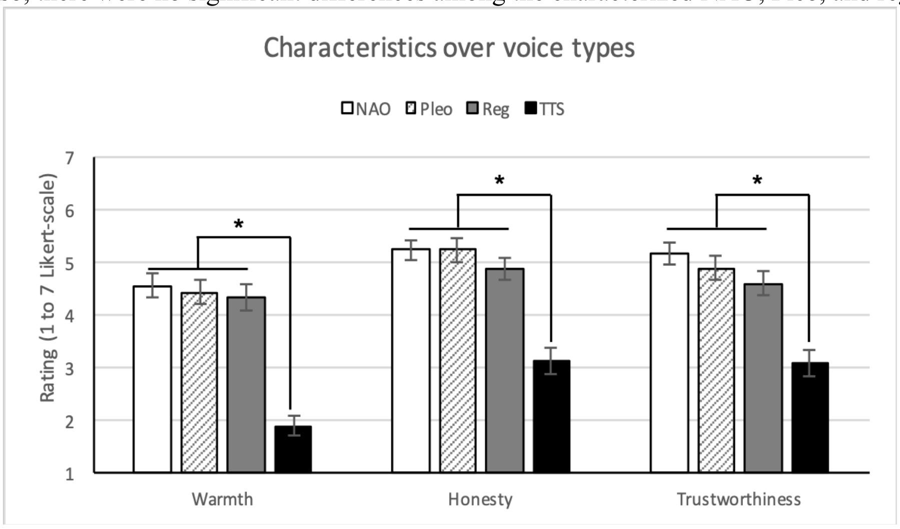


Figure 4. The rating scores of characteristics (*: p <0.05).

Table 8. Statistics for characteristics (warmth, honesty, trustworthiness).

Measures	Conditions		Statistics
Warmth	Main Effect for Voice Types		$F(3, 57) = 33.84, p < .0001, \eta_p^2 = .640$
	Characterized NAO $M = 4.55, SD = 1.52$	TTS $M = 1.88, SD = 1.18$	$t(19) = 7.48, p < .0001$
	Characterized Pleo $M = 4.32, SD = 1.55$		$t(19) = 7.14, p < .0001$
	Regular $M = 4.33, SD = 1.49$		$t(19) = 7.14, p < .0001$
Honesty	Main Effect for Voice Types		$F(3, 57) = 32.24, p = < .0001, \eta_p^2 = .630$
	Characterized NAO $M = 5.23, SD = 1.19$	TTS $M = 3.10, SD = 1.60$	$t(19) = 6.67, p < .0001$
	Characterized Pleo $M = 5.23, SD = 1.40$		$t(19) = 6.87, p < .0001$
	Regular $M = 4.88, SD = 1.34$		$t(19) = 5.70, p < .0001$

Trustworthiness	Main Effect for Voice Types		$F(3, 57) = 20.19$, $p < .0001$, $\eta_p^2 = .515$
	Characterized NAO $M = 5.15$, $SD = 1.33$	TTS $M = 3.08$, $SD = 1.54$	$t(19) = 5.61$, $p < .0001$
	Characterized Pleo $M = 4.88$, $SD = 1.44$		$t(19) = 5.11$, $p < .0001$
	Regular $M = 4.58$, $SD = 1.45$		$t(19) = 4.17$, $p < .0001$

4.4 ***Naturalness***

Figure 5 and Table 9 showed the rating scores in naturalness over voice types and robots. For voice types, the regular voice showed the highest scores in naturalness. For robots, there were no significant differences found in both categories.

Results were analyzed with a 2 (Robot) x 4 (Voice Type) repeated measures analysis of variance (ANOVA). Since there was no interaction effect between robots and voice types, paired-samples t-tests were conducted for the multiple comparisons among voice types. First, the result revealed a statistically significant difference in the rating scores in naturalness among voice types. Participants showed significantly lower rating scores in the TTS voice than all other three voice types. The regular voice showed significantly higher rating scores than the characterized Pleo voice.

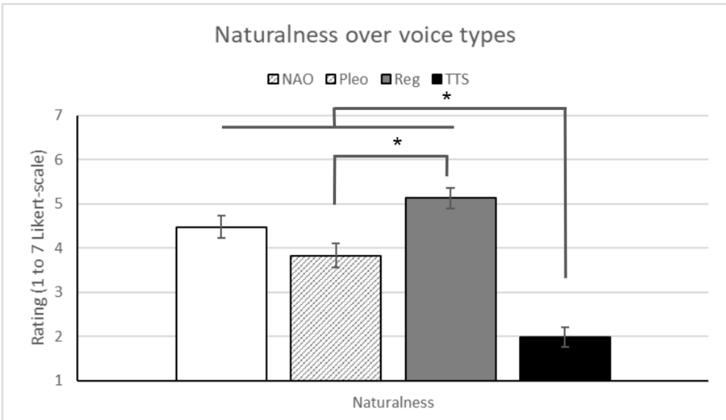


Figure 5. The rating scores of naturalness (*: $p < 0.05$).

Table 9. Statistics for naturalness.

Measures	Conditions		Statistics
Naturalness	Main Effect for Voice Types		$F(3, 57) = 37.67$, $p < .0001$, $\eta_p^2 = .665$
	Characterized NAO $M = 4.48$, $SD = 1.58$	TTS $M = 1.98$, $SD = 1.40$	$t(19) = 6.75$, $p < .0001$
	Characterized Pleo $M = 3.83$, $SD = 1.71$		$t(19) = 5.09$ $p < .0001$
	Regular $M = 5.13$, $SD = 1.42$		$t(19) = 8.49$ $p < .0001$
	Characterized Pleo $M = 3.83$, $SD = 1.71$	Regular $M = 5.13$, $SD = 1.42$	$t(19) = 3.45$, $p = .0011$

4.5 ***Preferences: Likability and Attractiveness***

Figure 6 and Table 10 showed the rating scores in likability and attractiveness over voice types and robots. Among voice types, the TTS voice commonly showed the lowest rating scores in both categories. For robots, there were no significant differences found in both categories.

Results were analyzed with a 2 (Robot) x 4 (Voice Type) repeated measures analysis of variance (ANOVA). First, the result revealed a statistically significant difference in likability 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 lower rating scores in the TTS voice than all other three voice types. Next, the result revealed a statistically significant difference in attractiveness 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. Same as shown in a likability category, participants showed significantly lower rating scores in the TTS voice than all other three voice types. The regular voice showed significantly higher rating scores than the characterized Pleo voice.

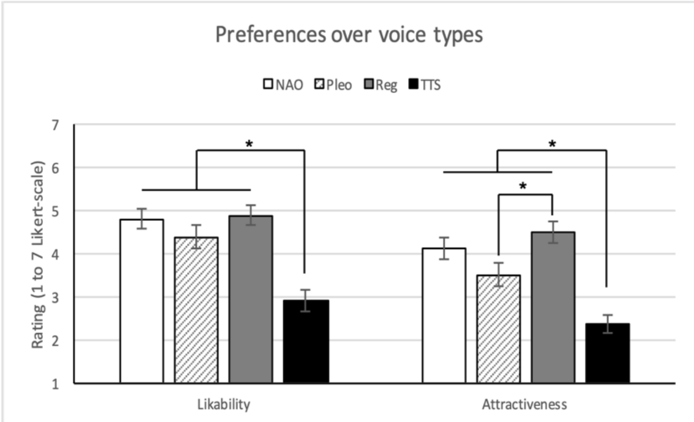


Figure 6. The rating scores of preferences (*: p <0.05).

Table 10. Statistics for preferences (likability, attractiveness).

Measures	Conditions		Statistics
Likability	Main Effect for Voice Types		$F(3, 57) = 18.91$, $p < .0001$, $\eta_p^2 = .499$
	Characterized NAO $M = 4.80$, $SD = 1.44$	TTS $M = 2.90$, $SD = 1.57$	$t(19) = 4.84$, $p < .0001$
	Characterized Pleo $M = 4.38$, $SD = 1.64$		$t(19) = 3.90$, $p = .0003$
	Regular $M = 4.88$, $SD = 1.42$		$t(19) = 5.19$, $p < .0001$
Attractiveness	Main Effect for Voice Types		$F(3, 57) = 18.65$, $p = < .0001$, $\eta_p^2 = .495$
	Characterized NAO $M = 4.10$, $SD = 1.53$	TTS $M = 2.38$, $SD = 1.33$	$t(19) = 4.85$, $p < .0001$
	Characterized Pleo $M = 3.50$, $SD = 1.63$		$t(19) = 3.18$, $p = .0025$
	Regular $M = 4.50$, $SD = 1.53$		$t(19) = 6.14$, $p < .0001$
	Characterized Pleo $M = 3.50$, $SD = 1.63$	Regular $M = 4.50$, $SD = 1.53$	$t(19) = 2.97$, $p = .0045$

5. Discussions

In the experiment, 20 participants experienced verbal interactions with robots while reading scripts of fairy tales to robots. Humanoid and zoomorphic robots used four different voice types and seven emotions were presented to participants through robots' verbal comments. Each participant interacted with all 8 conditions of robots and voice types and all 7 presented emotions. 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 participant filled out the questionnaire after listening to each emotional comment, completing each condition and completing 4 conditions. The emotion recognition accuracy and subjective ratings such as characteristics, naturalness, and user preferences were measured.

Referring to the research questions and hypotheses in Section 2.4, the results are listed as follows:

- RQ1:
    - H1a (rejected): A significantly higher emotion recognition accuracy was reported from Pleo robot than NAO robot.

- H1b (supported): The TTS voice showed significantly lower emotion recognition accuracy than the characterized NAO, characterized Pleo, and regular voices.
- H1c (rejected): The characterized Pleo voice showed significantly higher emotion recognition accuracy than the regular voice.
- H1d (supported): Anger, disgust, and fear had significantly lower emotion recognition accuracy with lower rating scores in clarity and suitability than other emotions.

- RQ2:
  - H2a (rejected): No significant difference was found among robot types for different characteristics ratings.
  - H2b (supported): The TTS voice showed significantly lower rating scores in warmth, honesty and trustworthiness than the characterized NAO, characterized Pleo, and regular voices; and the regular voice showed significantly higher rating scores in naturalness than the characterized Pleo and TTS voices.
- RQ3:
  - H3a (rejected): There were no significant differences found in both likeability and attractiveness ratings for robot types.
  - H3b (supported): The regular voice showed significantly higher rating scores in attractiveness than the TTS voice.
  - H3c (rejected): The regular voice also showed significantly higher rating scores in attractiveness than the characterized Pleo voice.

The critical points and explanations in each category are described below by dependent variables.

5.1 *Accuracy, Clarity, and Suitability*

The result showed that the emotion recognition accuracy significantly differed depending on presented emotions (H1d). As shown in Table 5, overall, unpleasant emotions with high arousal levels such as anger, disgust and fear showed significantly lower emotion recognition accuracy than other emotions such as anticipation, happiness, surprise and sadness did. There might be possible explanations about why some emotions were not accurately perceived. First, the emotion recognition accuracy results aligned with our previous study (Ko et al., 2020) that negative emotions received lower emotion recognition accuracy than positive emotions. Those two fairy tales used in the experiments were well-known for children and thus, participants might expect pleasant emotions more than unpleasant emotions. The most misclassified three emotions were all unpleasant emotions with high arousal levels (Russell, 1980). Next, the intensity of emotions might be different, which causes inequivalence among emotions. For example, among auditory stimuli used in the experiment, the intensity of unpleasant emotions might be lower than the one of positive emotions. Lastly, the mixed result was possible because there were many emotions presented through auditory cues. As shown in (Birkholz, Martin, Willmes, Kröger, & Neuschaefer-Rube, 2015), although emotion recognition can be fairly accurate when listeners choose from a limited set of emotion categories, agreement drops significantly as more categories of emotion become available. Note that in our experiment, the participants freely guessed each emotion without preset options. Also, fewer emotions can be perceived from voice (Cordaro, Keltner, Tshering, Wangchuk, & Flynn, 2016) compared to facial expressions.

For voice types (H1b & H1c), as expected, the TTS voice showed significantly lower emotion recognition accuracy than all other human voice types—characterized NAO, characterized Pleo, and regular voices—did. Furthermore, the TTS voice also showed significantly lower rating scores in clarity and suitability. It suggests that these TTS voices are inappropriate for emotive expressions since the intended emotions might not be delivered correctly to listeners even though they have the same semantic content. Instead, recorded human voices such as characterized NAO, characterized Pleo, and regular voices are more suitable for robots to express emotive voices and deliver emotions correctly. Most interestingly, the characterized Pleo voice showed significantly higher emotion recognition accuracy than the regular voices did. There was a possibility that these results suggest that a characterized voice might be more appropriate for emotive expressions delivering intended emotions more accurately and facilitating the interactions than just a regular voice. However, because only characterized Pleo voice showed a higher emotion recognition accuracy in the present study, more research should be conducted to determine if characterized voice types are more effective than the regular voice in expressing the emotions more accurately. It also suggests that there may be value in creating TTS engines that exaggerate emotional characterization for use in contexts where highly recognizable emotional signals are desired. Mimicking a natural speaking style may not be the optimal approach for delivering emotional information via synthetic speech from a robot. The results provide additional guidance on designing robot speech to deliver different emotions more effectively. As shown, other emotions can be sufficiently conveyed by affective tones, but disgust and fear require more semantic contents.

For robot types (H1a), NAO showed significantly lower emotion recognition accuracy than Pleo for happiness (NAO: $M = 0.61$, $SD = 0.49$; Pleo $M = 0.76$, $SD = 0.43$, $p < .05$). However, there was no difference between voice types of the two robots. We can cautiously infer that the participants might expect happy expressions from Pleo more than Nao and it caused higher emotion recognition accuracy in happiness. According to the previous findings (Díaz, Nuño, Saez-Pons, Pardo, & Angulo, 2011; Fraune, Sherrin, Sabanović, & Smith, 2015; Haring, Watanabe, & Mougenot, 2013), people perceive that Pleo manifested positive emotions (e.g., Love, Grateful) more than NAO (e.g., Uneasy, Fear). However, to the best of our knowledge, the relationships between perceived emotions (e.g., Happiness) and robots' appearances have not been comprehensively studied. The overall underlying cognitive process of recognizing emotions from form factors should be investigated in the future.

5.2 ***Characteristics, Naturalness, and Preferences***

Surprisingly, no significant difference was found on participants' perception of robot's characteristics and preferences (H2a & H3a). This result might suggest that participants perceived both robots as similar, or they evaluated the auditory portion of the social robots more than the embodiment and appearance regarding the ratings for each category. Because participants reported a significantly higher emotion recognition accuracy in Pleo than NAO robot, this might imply that performance and perception might not always be congruent. In the results, the TTS voice showed the significantly lowest rating scores across all characteristics and preferences including likability, attractiveness, warmth, honesty, and trustworthiness (H2b & H3b). The TTS voice showed a significantly lower rating score in the naturalness feature and the result might be because it had basically a flat voice without variations in pitch and speed. Other recorded voices such as characterized and regular voices having intonations and variations in speech showed significantly higher scores in the naturalness rating than the TTS voice.

A regular voice showed significantly higher rating scores in naturalness and attractiveness than a characterized Pleo voice (H3c). The results indicate that a regular voice might be more suitable for general use with higher user preferences and naturalness than characterized or TTS voices.

Overall, these results indicate that the characterized voice might lead to the highest emotion recognition accuracy, but the regular voice is the most preferred. It is assumed that characterized voices might be appropriate for emotional expressions. On the other hand, regular voices which show the highest attractiveness and naturalness might be suitable for general use. For example, for the first stage of human-robot interaction, regular voices might be appropriate to facilitate the interaction. However, for the next step for in-depth and emotion-related interactions, a characterized voice might be helpful to express emotional states and establish a unique relationship between users and robots since this stage involves personal familiarity with the other person and strong emotional commitment to the relationship (Lewis & Weigert, 1985). To further generalize our results, more experiments are required to consider possible other variables.

5.3 ***Anecdotal Findings***

Interestingly, there were no significant effects of the appearance of robots on all dependent variables except for emotion recognition accuracy. This might be because the given tasks were mostly focused on conversation which requires reading aloud and listening to verbal feedback but were not relevant to visual cues as much as auditory cues. According to (Frith & Frith, 2006), emotions are perceived by facial expressions and whole body movements instead of fixed features such as appearances, but these dynamic visual cues were not applied in this experiment.

There were interesting comments on auditory feedback from participants. A participant said, *"(P2) The final robot seemed to be happy at the start of the wolf story. My brain was saying it shouldn't be that but that's all my emotions were getting"*, which indicates the individual differences in expectation. Other comments such as *"(P15) The robots sounded more surprised/happier than showing signs of any other emotion"* and *"(P18) When Pleo would say "What!" in a shocked tone, it was easy to recognize his surprise in both the natural sounding voice and robotic sounding voice,"* which showed that the intensity of emotions could vary for different participants.

5.4 ***Limitations***

There are limitations and improvements that need to be considered in the next experiment to broaden this study and draw more reliable results. First, twenty participants may not be enough to generalize the results of the present study. We plan to replicate the study with more participants and expand it to other populations (e.g., children and older adults). Because the present study includes multiple factors (robot types, emotions, and voice types), a different approach of statistical tests could be used (e.g., a linear mixed effect model), to investigate the effects of multiple factors on one measurement. In the future study, we will explore more appropriate statistical tests for further analysis.

The equivalence among the intensity of emotions should be secured. We used one of the most widely used emotion sets, Ekman's basic emotions, but the result showed that some of them were not clearly distinguished by participants. The present study excluded the selected negative emotions with poor emotion recognition accuracy due to potential biases, but again using a different statistical model or analysis will help us understand the deviation. Using the only two phrases for each emotion might have provided biases to the participants' emotion recognition. Also, it may not be sufficient to ensure the generalizability of the finding. Depending on the content of the phrase, emotional semantics or strength might have been changed. However, as our results indicated, even with those same phrases, the participants showed significantly different emotion recognition accuracy depending on the robot type and voice type. In future research, we will diversify the phrases more with the similar length. The order of presentation might also have influenced the participants' responses. However, it is an intrinsic limitation because we were not able to change the storyline every time. If we randomly change the order of emotions without the context of the story, the experiment might lack external validity. We believe that people perceive emotions in the context.

Next, the characteristics of voice types should be more specifically studied to figure out which factors cause differences. In this study, characterized NAO and Pleo voices were generated by voice actors to exploit their expertise. It was a first attempt to produce the voice that well expresses the characteristics of NAO and Pleo. Regarding the emotion recognition accuracy results, participants reported a significantly higher emotion recognition accuracy in the characterized Pleo voice (but not in the characterized Nao voice) than the regular voice. The reason for this result might be that different appearances of the robots (animal versus humanoid) impacted participants' emotion recognition, because participants recognized emotions significantly more accurately in the Pleo robot than the NAO robot. In the follow-up study (Appendix B), participants reported a higher emotion recognition accuracy in both characterized voices (NAO and Pleo) than the regular voice. In the next experiment, the acoustic characteristics with specific physical properties (e.g., frequency range, speed, intensity) will also be considered when the representative voice types were designed so that the influential factors for different voice types will be investigated in depth. This approach will enable us to quantify the relationship between voice parameters and perception effects and model the robot voices. The gender effects will also be investigated. In this experiment, only male voices were used to control the gender effect and female voices were not included. We will design female voices for all four voice types and compare the gender differences in the following experiment.

There might have been some novelty effects. The participants did not have any previous opportunity to interact with or see the robots used in the present study. To minimize any novelty effects, the orders of the robots and voice types were counterbalanced across participants. Therefore, while interacting with the robots, the plausible novelty effects might have been reduced. We also had a standardized introductory section and minimized features used in the experiment (i.e., we used only the "speech" function and did not use other features, such as moving robot arms or its head). We are conducting separate experiments to see the effects of robot gestures and facial expressions. Taking all together of these experiments, we will be able to see the separate and overall effects.

6. Future Work

Throughout this study, various aspects of social robots such as appearances, emotive expressions, and voice types were investigated. Based on the results and experimental settings, follow-up studies will be conducted with two complementary approaches. First, the research scope will be narrowed down to focus more on the acoustic characteristics of voice types having distinct features. This approach will help in-depth understanding in emotive and interactive robotic systems and developing computational models for emotional and conversational human-robot interactions. Gender-specific factors such as the user's gender and the gender of robot voice will also be considered based on the previous result (Eyssel, Kuchenbrandt, et al., 2012). Meanwhile, other factors such as ages and modalities will be included to widen the research scope to investigate the multiple influential factors. As provided from previous studies (Fong et al., 2003; Nabe et al., 2006; Nonaka et al., 2004), considering that the interactions take place via various modalities, facial expressions, gaze and gestures (Ham, Cuijpers, & Cabibihan, 2015), and even non-verbal sounds can be included as independent variables. The results will provide a design guideline for emotional and trustworthy robots, especially employing emotive expressions and facilitate the relationship between people and social robots such as assistive robots, voice assistants, and any other conversational agents.

Acknowledgements

This work was partly supported by National Institutes of Health (US) (No.1 R01 HD082914-01).

References

Bachorowski, J. A., & Owren, M. J. (2003). Sounds of emotion: Production and perception of affect-related vocal acoustics. *Annals of the New York Academy of Sciences, 1000*(1), 244-265.

Bänziger, T., Grandjean, D., & Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT). *Emotion*, 9(5), 691

Barnes, J., Hosseini, S. M. F., Jeon, M., Park, C. H., & Howard, A. M. (2017). *The influence of robot design on acceptance of social robots.* Paper presented at the 2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI).

Barnes, J. A., Park, C. H., Howard, A., & Jeon, M. (2020). Child-Robot Interaction in a Musical Dance Game: An Exploratory Comparison Study between Typically Developing Children and Children with Autism. *International Journal of Human–Computer Interaction*, 1-18.

Barnes, J., Richie, E., Lin, Q., Jeon, M., & Park, C. H. (2018). *Emotive Voice Acceptance in Human-Robot Interaction.* Paper presented at the Proceedings of the 24th International Conference on Auditory Display.

Beer, J. S., & Ochsner, K. N. (2006). Social cognition: a multi level analysis. *Brain research, 1079*(1), 98-105.

Bevill, R., Park, C. H., Kim, H. J., Lee, J. W., Rennie, A., Jeon, M., and Howard, A. M. (2016) Interactive robotic framework for multi-sensory therapy for children with autism spectrum disorder. In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 421-422). IEEE.

Birkholz, P., Martin, L., Willmes, K., Kröger, B. J., & Neuschaefer-Rube, C. (2015). The contribution of phonation type to the perception of vocal emotions in German: an articulatory synthesis study. *The Journal of the Acoustical Society of America, 137*(3), 1503-1512.

Bryant, D., Bornstein, J., & Howard, A. (2020). *Why Should We Gender? The Effect of Robot Gendering and Occupational Stereotypes on Human Trust and Perceived Competency*. Paper presented at the ACM/IEEE International Conference on Human-Robot Interaction (HRI), Cambridge, UK.

Cakmak, M., Hoffman, G., & Thomaz, A. (2016). Computational Human-Robot Interaction. *Foundations and Trends in Robotics, 4*(2-3), 104-223. doi:10.1561/2300000049

Calvo, R. A., & D'Mello, S. (2010). Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing, 1*(1), 18-37. doi:10.1109/t-affc.2010.1

Coeckelbergh, M. J. E., & Technology, I. (2010). Moral appearances: emotions, robots, and human morality. *12*(3), 235-241.

Conti, D., Cirasa, C., Di Nuovo, S., & Di Nuovo, A. (2020). "Robot, tell me a tale!": A social robot as tool for teachers in kindergarten. *Interaction Studies, 21*(2), 220-242.

Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D., & Flynn, L. M. (2016). The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion, 16*(1), 117.

Darwin, C., & Prodger, P. (1998). *The expression of the emotions in man and animals*: Oxford University Press, USA.

Decety, J. (2010). The neurodevelopment of empathy in humans. *Developmental neuroscience, 32*(4), 257-267.

Díaz, M., Nuño, N., Saez-Pons, J., Pardo, D. E., & Angulo, C. (2011). Building up child-robot relationship for therapeutic purposes: From initial attraction towards long-term social engagement. In Face and Gesture 2011 (pp. 927-932). IEEE.

DiSalvo, C.F., Gemperle, F., Forlizzi, J., & Kiesler, S.B. (2002). All robots are not created equal: the design and perception of humanoid robot heads. In *Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, DIS*, 321– 326.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6*(3-4), 169-200.

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion review*, *3*(4), 364-370

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review, 114*(4), 864-886.

Eyssel, F., De Ruiter, L., Kuchenbrandt, D., Bobinger, S., & Hegel, F. (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.* Paper presented at the 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI).

Eyssel, F., Kuchenbrandt, D., Hegel, F., & de Ruiter, L. (2012). *Activating elicited agent knowledge: How robot and user features shape the perception of social robots.* Paper presented at the 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication.

FakhrHosseini, S. M., Hilliger, S., Barnes, J., Jeon, M., Park, C. H., & Howard, A. M. (2017). *Love at first sight: Mere exposure to robot appearance leaves impressions similar to interactions with physical robots.* Paper presented at the 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).

FakhrHosseini, S. M., Lettinga, D., Vasey, E., Zheng, Z., Jeon, M., Park, C. H., & Howard, A. M. (2017). *Both "look and feel" matter: Essential factors for robotic companionship.* Paper presented at the 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).

Feldmaier, J., Stimpfl, M., & Diepold, K. (2017). *Development of an Emotion-Competent SLAM Agent.* Paper presented at the Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction.

Fischer, K., Jung, M., & Jensen, L. C. (2019). *Emotion Expression in HRI – When and Why*. Paper presented at the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI).

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems, 42*(3-4), 143-166.

Fraune, M. R., Sherrin, S., Sabanović, S., & Smith, E. R. (2015). Rabble of robots effects: Number and type of robots modulates attitudes, emotions, and stereotypes. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (pp. 109-116).

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. Neuron, 50(4), 531-534.

Ham, J., Cuijpers, R. H., & Cabibihan, J. J. (2015). Combining robotic persuasive strategies: the persuasive power of a storytelling robot that uses gazing and gestures. *International Journal of Social Robotics, 7*(4), 479-487.

Haring, K. S., Watanabe, K., & Mougenot, C. (2013). The influence of robot appearance on assessment. In 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 131-132). IEEE.

Jeon, M., & Rayan, I. A. (2011). *The effect of physical embodiment of an animal robot on affective prosody recognition.* Paper presented at the International Conference on Human-Computer Interaction.

Jung, M. F. (2017). *Affective grounding in human-robot interaction.* Paper presented at the 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI.

Kishi, T., Kojima, T., Endo, N., Destephe, M., Otani, T., Jamone, L., . . . Cosentino, S. (2013). *Impression survey of the emotion expression humanoid robot with mental model based dynamic emotions.* Paper presented at the 2013 IEEE International Conference on Robotics and Automation.

Ko, S., Liu, X., Mamros, J., Lawson, E., Swaim, H., Yao, C., & Jeon, M. (2020, July). The Effects of Robot Appearances, Voice Types, and Emotions on Emotion Perception Accuracy and Subjective Perception on Robots. In *International Conference on Human-Computer Interaction* (pp. 174-193). Springer, Cham.

Kraus, M., Kraus, J., Baumann, M., & Minker, W. (2018). *Effects of Gender Stereotypes on Trust and Likability in Spoken Human-Robot Interaction.* Paper presented at the Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).

Kwak, S. S., Kim, Y., Kim, E., Shin, C., & Cho, K. (2013, August). What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot. In *2013 IEEE RO-MAN* (pp. 180-185). IEEE.

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. *Social forces, 63*(4), 967-985.

Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77, 23-37.

Liu, Z. T., Pan, F. F., Wu, M., Cao, W. H., Chen, L. F., Xu, J. P., ... & Zhou, M. T. (2016, July). A multimodal emotional communication based humans-robots interaction system. In *2016 35th Chinese Control Conference (CCC)* (pp. 6363-6368). IEEE.

Lowe, R., Barakova, E., Billing, E., & Broekens, J. (2016). Grounding emotions in robots–An introduction to the special issue: Sage Publications Sage UK: London, England.

McGinn, C., & Torre, I. (2019). *Can you Tell the Robot by the Voice? An Exploratory Study on the Role of Voice in the Perception of Robots.* Paper presented at the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI).

Mitchell, R. L., & Phillips, L. H. (2015). The overlapping relationship between emotion perception and theory of mind. *Neuropsychologia, 70*, 1-10.

Nabe, S., Cowley, S. J., Kanda, T., Hiraki, K., Ishiguro, H., & Hagita, N. (2006). *Robots as social mediators: coding for engineers.* Paper presented at the ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication.

Nass, C., Foehr, U., Brave, S., & Somoza, M. (2001). *The effects of emotion of voice in synthesized and recorded speech.* Paper presented at the Proceedings of the AAAI symposium emotional and intelligent II: The tangled knot of social cognition.

Nonaka, S., Inoue, K., Arai, T., & Mae, Y. (2004). *Evaluation of human sense of security for coexisting robots using virtual reality. 1st report: evaluation of pick and place motion of humanoid robots.* Paper presented at the IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004.

Ortony, A. (2021). Are All "Basic Emotions" Emotions? A Problem for the (Basic) Emotions Construct. *Perspectives on Psychological Science*, 1745691620985415. https://doi.org/10.1177/1745691620985415

Pereira, A., Martinho, C., Leite, I., & Paiva, A. (2008, May). iCat, the chess player: the influence of embodiment in the enjoyment of a game. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 3* (pp. 1253-1256).

Phillips, M. L. (2003). Understanding the neurobiology of emotion perception: implications for psychiatry. *The British Journal of Psychiatry, 182*(3), 190-192.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion* (pp. 3-33). Academic press.

Reisenzein, R., Hudlicka, E., Dastani, M., Gratch, J., Hindriks, K., Lorini, E., & Meyer, J.-J. C. (2013). Computational Modeling of Emotion: Toward Improving the Inter- and Intradisciplinary Exchange. *IEEE Transactions on Affective Computing, 4*(3), 246-266. doi:10.1109/t-affc.2013.14

Rodriguez-Lizundia, E., Marcos, S., Zalama, E., Gómez-García-Bermejo, J., & Gordaliza, A. (2015). A bellboy robot: Study of the effects of robot behaviour on user engagement and comfort. *International Journal of Human-Computer Studies*, 82, 83-95.

Russel, J. (1980). A circumplex model of emotions. *Journal of Personality and Social Psychology*, *39*, 1161-1178.

Russell, J. A. (2017). Cross-cultural similarities and differences in affective processing and expression. In M. Jeon (Ed.), *Emotions and affect in human factors and human-computer interaction* (pp. 123-141). Academic Press.

Saint-Aimé, S., Le-Pevedic, B., Duhaut, D., & Shibata, T. (2007). *EmotiRob: companion robot project.* Paper presented at the RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication.

Sakai, K., Nakamura, Y., Yoshikawa, Y., & Ishiguro, H. (2021). Effect of robot embodiment on satisfaction with recommendations in shopping malls. *IEEE Robotics and Automation Letters*, 7(1), 366-372.

Schilhab, T. S. S. (2002). Anthropomorphism and mental state attribution. *Animal Behavior*. Academic Press.

Schirmer, A., & Adolphs, R. (2017). Emotion Perception from Face, Voice, and Touch: Comparisons and Convergence. *Trends Cogn Sci, 21*(3), 216-228. doi:10.1016/j.tics.2017.01.001

Seo, S. H., Geiskkovitch, D., Nakane, M., King, C., & Young, J. E. (2015). *Poor thing! Would you feel sorry for a simulated robot? A comparison of empathy toward a physical and a simulated robot.* Paper presented at the 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI).

Sharma, M., Hildebrandt, D., Newman, G., Young, J. E., & Eskicioglu, R. (2013). *Communicating affect via flight path: exploring use of the laban effort system for designing affective locomotion paths.* Paper presented at the Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction.

Sims, V. K., Chin, M. G., Lum, H. C., Upham-Ellis, L., Ballion, T., & Lagattuta, N. C. (2009). *Robots' auditory cues are subject to anthropomorphism.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Song, S., & Yamada, S. (2017). *Expressing emotions through color, sound, and vibration with an appearance-constrained social robot.* Paper presented at the Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction.

Susskind, J. M., Lee, D. H., Cusi, A., Feiman, R., Grabski, W., & Anderson, A. K. (2008). Expressing fear enhances sensory acquisition. *Nature neuroscience, 11*(7), 843.

Walters, M. L., Syrdal, D. S., Koay, K. L., Dautenhahn, K., & Te Boekhorst, R. (2008). *Human approach distances to a mechanical-looking robot with different robot voice styles.* Paper presented at the RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113–117.

Williams, J. M. G., Watts, F. N., MacLeod, C., & Mathews, A. (1988). *Cognitive psychology and emotional disorders*: John Wiley & Sons.

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). *The aligned rank transform for nonparametric factorial analyses using only anova procedures.* Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.

Appendix A. Questionnaires

- Post-comment questionnaire
  - What emotion do you feel the robot expressed? (Open question)
  - What characteristics of the voice brought to mind that emotion? (Open question)
  - 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
  - How likable is the voice? (1-7 Likert scale)
  - How attractive is the voice? (1-7 Likert scale)
  - How warm is the voice? (1-7 Likert scale)
  - How honest is the voice? (1-7 Likert scale)
  - How trustworthy is the voice? (1-7 Likert scale)
  - How natural does the voice sound? (1-7 Likert scale)

- Post-session questionnaire
  - Thoughts about 1st, 2nd, 3rd, and 4th voices (Open question)
  - Which story was your favorite? (Open question)
  - What is your sex? (Open question)
  - What is your age? (Open question)
  - What is your race and/or ethnicity? (Multiple-choice, Open question)

Appendix B. Voice Types Validation Study

To further investigate the impact of robot embodiment on participants' perception towards different voice types, we conducted a follow-up validation study for voice types only. Based on the results of the main study, TTS voice showed significantly lower score on the most subjective ratings. Therefore, this validation study used only human voices, which made a 3 (Voice Types) by 7 (Emotions) within-subjects design. Ten new participants (Age: M = 22.5, SD = 4.12) were recruited for the follow-up study. Six participants identified as male and four participants identified as female with 5 Asians, 4 Caucasian, and 1 Hispanic. They listened to all recordings and evaluated three voice types: Characterized NAO voice, Characterized Pleo voice, and Regular human voice. Because the suitability rating subjectively determined how suitable the voice types were on a certain robot, we excluded the scale in the validation study because there was no robot or physical embodiment involved with this follow-up study.

B1. Accuracy

Following the main study, the emotion recognition accuracy data were transformed with the aligned rank transform (ART) (Wobbrock, Findlater, Gergle, & Higgins, 2011). Then, the aligned-ranked data were analyzed with a 3 (Voice Types) x 7 (Emotions) repeated measures ANOVA, followed by paired samples t-tests with a Bonferroni correction for pairwise comparisons. A significant difference was found in the main effects of voice types, $F(2, 18) = 11.68$, $p < .001$, $\eta_p^2 = .567$ emotions, $F(6, 54) = 4.61$, $p < .001$, $\eta_p^2 = .339$ and the interaction effect between voice types and emotions, $F(12, 108) = 4.48$, $p < .001$, $\eta_p^2 = .342$. The average accuracy of emotion recognition in both characterized voices (NAO and Pleo) were significantly higher than the regular voice. The average accuracy was significantly higher in happiness (65.7%), sadness (77.6%), and surprise (67.6%) than anger (41.6%), disgust (37.6%), and fear (37.1%), which is similar to the main study. However, the average accuracy of anticipation (58.9%) was much lower compared to the percentage of the main study (75%). It might not be appropriate to compare the absolute percentage between the main study and the follow-up study because of different population and different number of participants. However, the average emotion recognition accuracy of the main study (56.61%) is numerically higher than that of the follow-up study (55.16%). The emotion recognition accuracy of the four emotions (happiness, anticipation, surprise, and disgust) was numerically higher in the main study than in the follow-up study. This might imply that when the voice is presented with embodied robots, emotion recognition accuracy might increase depending on different emotions. Further analysis of the interaction effects showed that the accuracy of emotion recognition was higher when characterized voices were paired with emotions that are positive and high arousal, such as happiness and surprise, or negative and low arousal, such as sadness than the regular voices paired with the emotions with opposite valence and arousal, such as anger, disgust, and fear. These results might suggest that the characterized voices improve participants' emotion recognition capabilities for certain emotions compared to regular human voices when there was no physical embodiment.

B2. Other Subjective Ratings

The results from other subjective ratings of this validation study were similar to the results in the main study. The main effect of voice types was found significant in the scale of warmth, $F(2, 832) = 3.65$, $p = .0466$; , $\eta_p^2 = .297$; trustworthiness, $F(2, 832) = 5.38$, $p = .0147$, $\eta_p^2 = .375$; naturalness, $F(2, 832) = 17.57$, $p < .0001$, $\eta_p^2 = .664$; likeability, $F(2, 832) = 10.20$, $p = .0011$, $\eta_p^2 = .532$; and attractiveness, $F(2, 832) = 12.42$, $p = .0004$, $\eta_p^2 = .586$.

Participants rated higher scores of warmth, and trustworthiness in regular voices than just the characterized Pleo voice. However, participants reported higher scores of naturalness, likeability, and attractiveness in regular voices than both characterized NAO and characterized Pleo voices. Note that in the main study, regular voice did not show higher scores of warmth and trustworthiness than the characterized voices. This might suggest that the appearance and embodiment of the robots can improve participants' perception toward the characterized voice positively such as increasing the warmth and trustworthiness of the robot. It is interesting to see that the validation study results of naturalness aligned with the results in the main study because it might imply that naturalness did not necessarily influence warmth and trustworthiness of the robot.

In sum, when there is embodiment of the robots, overall, people may recognize the same voice's emotions better. Also, they may perceive the characterized voice more positively (e.g., warm and trustworthy). The results of the validation study once again revealed the importance of the robot appearance and embodiment in HRI.