Robots' "Woohoo" and "Argh" can enhance users' emotional and social perceptions: An exploratory study on non-lexical vocalizations and non-linguistic sounds

XIAOZHEN LIU (0009-0005-8675-9935)

Virginia Polytechnic Institute and State University

JIAYUAN DONG (0000-0002-7253-8868)

Virginia Polytechnic Institute and State University

MYOUNGHOON JEON (0000-0003-2908-671X)

Virginia Polytechnic Institute and State University

As robots have become more pervasive in our everyday life, social aspects of robots have attracted researchers' attention. Because emotions play a crucial role in social interactions, research has been conducted on conveying emotions via speech. Our study sought to investigate the synchronization of multimodal interaction in human-robot interaction (HRI). We conducted a within-subjects exploratory study with 40 participants to investigate the effects of non-speech sounds (natural voice, synthesized voice, musical sound, and no sound) and basic emotions (anger, fear, happiness, sadness, and surprise) on user perception with emotional body gestures of an anthropomorphic robot (Pepper). While listening to a fairytale with the participant, a humanoid robot responded to the story with a recorded emotional non-speech sounds and gestures. Participants showed significantly higher emotion recognition accuracy from the natural voice than from other sounds. The confusion matrix showed that happiness and sadness had the highest emotion recognition accuracy, which is in line with previous research. The natural voice also induced higher trust, naturalness, and preference, compared to other sounds. Interestingly, the musical sound mostly showed lower perception ratings, even compared to the no sound. Results are discussed with design guidelines for emotional cues from social robots and future research directions.

CCS CONCEPTS •**Human-centered computing~Human computer interaction (HCI)~Interaction techniques~Auditory feedback** •**Computer systems organization~Embedded and cyber-physical systems~Robotics~Robotic components**

Keywords and Phrases: Human-robot interaction, Emotion perception, Non-speech sounds, Robot voice

1 INTRODUCTION

With the advancement of automation and artificial intelligence, smart speakers and social robots have become popular. They are often designed to interact with users in a more social way. Just as emotions are important in human-human interactions, emotions are crucial in social interactions between a human and a robot. Robots typically convey their emotional states in the form of speech, even though natural language processing (NLP) still remains challenging in human-robot interaction (HRI) [1]. However, there is a myriad of methods to deliver emotions in addition to speech, such as facial expressions [2,3], body motions [4,5], or gestures [6-9]. In emotional situations, people also generate non-speech sounds. In the present study non-speech sounds include non-linguistic utterances (e.g., musical sounds like humming) or non-lexical vocalizations (e.g., an exclamation like "Argh") [10-13]. Taken together, we should also consider the complexity of behavioral synchronization in the relationship among robots' appearance, body gestures, speech and non-speech. With the higher complexity of the robot, users' expectations can increase [1], while the risk of their discovery of the robot's limitations and disengagement from the interaction also increases

2573-9522/2023/1-ART1 $15.00
https://doi.org/10.1145/3626185

[14]. We tried to explore how the robot's non-speech sounds can more naturally and accurately convey emotional information when combined with an anthropomorphic robot's body motions and gestures.

In this exploratory study, we examined people's social perceptions of a robot's emotional states with various non-speech auditory cue types (natural human voice, robotically synthesized (or metallic) voice, musical sounds, and no sounds) and basic emotions (anger, fear, happiness, sadness, and surprise). Participants were asked to listen to four fairytales from a computer voice with a humanoid robot, Pepper, and determine its emotional state when Pepper made a reaction to the specific part of the fairytale using one of the non-speech sounds. The present study is expected to provide a more comprehensive understanding of how non-speech sounds can enhance human-robot interaction (HRI) in conveying robot emotions and influencing social perceptions towards a humanoid robot.

2 RELATED WORK

2.1 Robot Anthropomorphism

In the setting of HRI and cooperation, social and interpersonal abilities are important. The capacity of social communication should be appropriate for the context in which a robot functions and would require unique skills depending on the application areas. Thus, the development of social skills for robots is costly. However, when robots have these social skills, people may begin to accept the concept of a robot companion [15]. Anthropomorphism can facilitate this social relationship. Researchers have identified five categories of characteristics that can be used to measure the degree of anthropomorphism in social robots, including appearance, behavior, cognition, emotion, and morality [16]. Human-like appearance can be considered as superficial characteristics, whereas human-like mind (i.e., cognition and emotion) can be considered as essential human characteristics [17]. The more human-like a robot seems, the more it is perceived as intelligent [2]. Due to their physical likeness to people, anthropomorphic robots may be more efficient in conveying emotional responses while interacting with humans [18]. This is why we used an anthropomorphic robot in the present study. Research shows that human emotional reactions are influenced by the movement features of robots [19]. People consider a robot to have better communicative competencies when it makes hand and arm movements while speaking [20]. As such, robots' appearance and behavior components interact with the human-like mind, including emotions. Therefore, the present study investigates one of the essential characteristics – *emotional* aspects of a social robot's anthropomorphism with the robot's non-speech sounds and gestures.

2.2 Robot Emotions

The purpose of social robots is to actively connect with humans to accomplish their own social objectives [21]. Emotional engagement between humans and robots makes communications between the two more natural. The capacity to design the robot to express its emotional states is an essential step for the theory of mind [15]. Emotions are so contagious even in human-robot interaction that people's valence rating, arousal rating, and task performance are impacted by robots' emotions because they develop empathy towards robots [22]. People preferred an emotional robot over a neutral robot, although they rated the emotional robot to have a lower speech intelligence, and they also recognized emotions faster in the emotional robot condition than in the neutral condition [23]. To make more emotional robots, researchers have tried to implement

robots that can (1) detect human emotions, (2) pretend to perceive their own emotional states, and (3) express their emotions. Numerous research studies have been conducted to decipher human emotions by analyzing noises, facial expressions, or physical contact [24]. These affect detection technologies have also been applied to robot development. Even though a robot's capability of perceiving its own emotional states is contradictory, there have been a few attempts [e.g., 25]. There are many ways for robots to express their own emotional states: facial expressions, affective prosody, music, and gestures [26, 27, 10]. The emotional interpretation of a behavior is higher than the emotional interpretation of an utterance; the entire interpretation is enhanced when the behavior and utterance have a consistent meaning [28]. When using new interactive robots with little prior experience, people will quickly decode and perceive meaningful, familiar, emotionally charged content using social cues [29]. People's emotional interpretations are category-specific, and utterances are subject to a "magnet effect" where people are attracted to typical emotional interpretations (basic emotions such as happiness, anger, and sadness) [29]. In the current study, users' emotional perceptions are examined using non-speech stimuli, such as non-lexical vocalizations and non-linguistic sounds that reflect five different emotions.

2.3 Robot's Emotional Expressions Using Sounds

There are multiple ways of auditory interaction for a machine expressing its emotions and intentions consisting of vocalizations and sounds without semantic content [1]. The present study focuses on Non-Linguistic Utterances (NLUs) and Non-Lexical Vocalizations (NLVs). Non-Linguistic Utterances (NLUs) are technologically generated non-speech phrases that use non-speech sounds to convey information, similar to sonification and auditory icons [1, 30]. Non-lexical vocalizations such as "Argh" or "uh-huh" can also convey information [31]. Emotional speech expression may successfully influence the behaviors of the perceiver [11], and particularly, prosodic expression is an effective communication route for humans and robots to communicate emotions [12]. Vocal emotions may be conveyed by hyper syncopation management of speech prosody [13] or brief non-speech vocalizations [32], often known as emotional bursts [33]. Emotional prosody may be characterized by modifying the acoustic properties of speech, such as intensity, frequency, and fundamental frequency [13]. In the absence of context, emotional outbursts are capable of conveying a distinct emotional meaning [32]. Emotional bursts are characterized as short, emotionally charged non-speech utterances that feature distinct non-speech noises and interjections with a phonological structure [33]. Compared to phonetic prosody, non-speech vocalizations are believed to convey more basic emotional expressions and aural simulations of facial emotions [34]. Multiple studies have investigated the accuracy with which listeners discern vocal emotions using phonetic prosody, with the accuracy varied by emotion types [35]. In terms of speech prosodic cues, the most accurate emotions recognized were sadness and disgust [36]. Fear, anger, and happiness are the most often misinterpreted vocal emotions [37]. The blended sonification model can also be used to improve emotional communication in auditory conditions (music and speech emotion perception) [38], and the addition of robot movements can further enhance the communication of emotional expressions [39].

Musical utterances of emotion have also been explored for many years, and a large number of studies have been published. A study showed that with music and rhythm, emotions can be categorized with precision (e.g., happiness, sadness, fear, etc.) [13]. In their study, the emotion recognition accuracy was higher when facial expression and music were presented together than when facial expression or music was

individually used [13]. In a similar line, the created music can express more emotion than the robot's facial expression, and when mixed with the robot's facial expression, the robot's emotion could be amplified [40]. This finding is consistent with that of the prior studies [e.g., 41]. Another study showed that non-speech audio was more trustworthy than text-to-speech technology [42].

2.4 The Current Study and Research Questions

There have been attempts to use non-speech sounds to convey robots' emotions, but those have been sparse and more research is still required. Many studies investigated affective prosody, which is a part (or form) of the speech [12, 13, 43]. Some studies used other types of robots (e.g., animal [34]) than humanoid robots. In speech interactions, characterized speech showed users' different emotion recognition and social perceptions compared to natural speech [44]. Musical sounds were also used, but it was not investigated if and to what extent the use of musical sounds can influence users' emotion perception of a robot compared to vocal sounds.

From this context, we aimed to investigate the effects of a robot's non-lexical vocalizations and non-linguistic sounds (musical sounds in the present study) on users' emotion recognition and perception of the humanoid robot. More specifically, we tried to answer the following research questions.

- RQ 1. Can adding non-speech sounds improve users' emotion recognition and social perceptions towards the humanoid robot?
- RQ 2. Which sounds (non-lexical vocalizations (e.g., "argh") vs. non-linguistic sounds (e.g., harsh sound effects)) can have more effects on users' emotion recognition and social perceptions towards the humanoid robot?
- RQ 3. Does exaggerating the vocal sound with robotic and metallic sound effects ("synthesized sounds" in our experiment) improve or degenerate users' social perceptions towards the humanoid robot?
- RQ 4. Will different emotions influence users' recognition accuracy from the non-speech sounds of the humanoid robot?

A previous study shows that people understand human utterances better than animal and technological utterances in different robotic appearances [45]. The morphology of the robot has an impact on the users' judgment of the appropriateness of the discourse, with human utterances being more appropriate for humanoid robots and animal utterances being considered more appropriate for animal robots [45]. Users prefer to express perceptions in terms of basic emotions and rarely explain them in terms of more subtle emotions [46]. It is not easy to express social intentions with non-verbal sounds out of context [47]. To address these research questions, we conducted an exploratory experiment with university students. Our participant and a humanoid robot, "Pepper", listened to the four fairytale stories together from the computer system using Text-to-Speech (TTS) voice. While listening, Pepper emotionally responded five times in the middle of each story using four different sound conditions (natural voice for non-lexical vocalization (human utterances), synthesized voice (robotic utterances) for non-lexical vocalization, musical sound, and no sound) combined with five basic emotions (anger, fear, happiness, sadness, and surprise).

The present study has unique contributions in terms of both theoretical and practical aspects. In the current study, we tested users' emotion recognition and social perceptions towards an *anthropomorphic robot* when it provides different types of *non-speech sounds with body movements and gestures* in the

context of storytelling. The no sound condition served as a baseline condition to understand the unique effects of different sounds. A systematic investigation into the effects of different types of non-speech sounds will advance not only robot emotion research, but also sound and auditory display research. The outcomes of the present study will contribute to the design of emotional cues for a humanoid robot to make it more sociable and trustable.

3 METHOD

3.1 Participants

Forty university undergraduate and graduate students (age range: 19-27 years old) participated in the experiment. Participants were compensated with $10 per hour. Twenty-one participants identified themselves as male, and the other nineteen participants identified as female. The experiment took at most 1.5 hours. All participants agreed to participate after reviewing the consent form approved by the Virginia Tech Institutional Review Board (IRB).

3.2 Equipment and Stimuli

A humanoid robot, Pepper, was employed in the experiment (Figure 1). Pepper is a big-size humanoid robot (Height: 4 ft, Length: 17 in, Width 19 in) having similarity to human appearance. According to the storyline, the robot played a recorded sound with gestural feedback which provided emotional cues to the participants. Modified video recordings of a Pepper robot were used for the entire experiment; The experiment was conducted in a lab where participants watched the pre-recorded videos and answered questionnaires. Four different stories ("The Three Little Pigs", "The Boy Who Cried Wolf", "Beauty and the Beast" and "Little Red Riding Hood") were used in this experiment. Listeners can make specific responses only when they have sufficient narrative information [48]. The stories we chose are simple narratives with easy vocabulary and globally well-known so that participants can easily understand. More importantly, all these four stories included all of the emotions we examined in the present study.

Figure 1: "Pepper" robot

Three sound types were created for five emotional expressions (Table 2). All sound files and video files used in this study are provided on the web for other researchers and educators to have a better understanding: https://osf.io/8rhs4/ (please use the Chrome browser to view the files or download the files to play if they are not working on other browsers). The non-lexical vocalization samples, which made the natural voice

condition, were from the pre-recorded and pre-validated male sounds from the study performed at the University of California, Berkeley [49]. We used a library of sound bursts created from 2032 emotional sound bursts generated from the Berkeley Lab [49]. We selected the sounds from the interactive online map where the complex, high-dimensional emotional space conveyed by the sounds was visualized. The scales of the vocalization we chose for our experiment are shown in Table 2. In human-robot interaction, robots can benefit from attractive speech features [50], and it is crucial to pay attention to the robot's speech features when creating synthesized utterances. The robotically synthesized utterances (synthesized voice) were made by applying the Flanger and Flangus synthetic/robotic effect to the same natural voice using the Fruity Loops Studio software to make more traditional robot sounds. Flanging is a form of phase cancellation created by combining multiple, variously delayed copies of the input sound. Flangus increases the width of the stereo audio using complex flange effects and unison mode. In simpler terms, using Flanger overlays multiple instances of the same sound on itself, with all instances having different delay values. Flangus also uses Flanging techniques but instead uses multiple instances of the source sound to increase the width of the stereo effect, while keeping all instances playing in the same key (unison). Unnecessary silence or noise was cut out and the pitch of the sound was edited to fit the robot. The sound editing also included frequency manipulation (cutting out the extremes). The musical sounds (sounds like earcons [51] - a short musical motif) were selected from the sound pool, which was designed and validated for the auditory emoticon studies [52, 53, 54]. The gestures of the robot in five emotional expressions were created by using Choregraphe, a software program coming with the Pepper robot. Gestures were made based on the previous research shown in Table 1. In the videos, the robot's responses were inserted with a 1-3 second long gap after the story paused.

Table 1: "Pepper" robot Gestures

Pepper	Happy	Sad	Surprise	Anger	Fear
Description/ Procedure	Fast, positive gesture with elation and open arms. Arms move [6,7].	Slow, negative gesture. Body dropped and shrunk. Shoulders bowed. Hands kept lower than their normal positions, hands closed or moving slowly. The face covered with two hands [6,7].	Fast gesture with shock and unexpectedness [7].	Arms Akimbo - Hands on hips, largely recognized as an angry gesture [7,8].	Arms rigged/tensed/clenched and out to side while robot looks around apprehensively and moves body backwards [7,9].
Picture					

Voice quality and speaker personality can potentially influence language attitudes [55]; we used a synthetic speech to tell the story to eliminate the possibility that participants were influenced by the speaker's attitude and the diversity of the language they spoke. The storytelling voice was generated using the text-to-speech (TTS) engine provided by Amazon. The female voice Ivy (US English) was used because it sounds like a little girl's voice, which fits fairytales. Ivy's voice was used for the stories by default volume, rate, and pitch.

Each story presented five different emotions from Ekman's six basic emotions [56] (anger, fear, happiness, sadness, surprise, and disgust). A pilot test was conducted for natural voice, synthesized voice, and musical sounds based on Ekman's six basic emotions. Eight university students in the age range 19-22 years old were recruited, and only the sounds with above 80 percent accuracy were selected and used for the actual experiment. The average recognition accuracy of surprise, sadness, anger, and fear was 87.5%. Happiness had 100% and disgust had 25% accuracy. The participants recruited for the voice pilot test did not participate in the actual experiment. The disgust emotion has been omitted from our investigation because no musical sounds passed a pilot test with 80 percent accuracy. The five emotions (anger, fear, happiness, sadness, and surprise) were fit into four stories each ("The Three Little Pigs", "The Boy Who Cried Wolf", "Beauty and the Beast", and "Little Red Riding Hood").

Table 2: "Pepper" Robot Sound Types

Emotion	Natural Voice	Synthesized Voice	Musical Sound	No sound (Gestures-only)
Happiness	"Woohoo" "58% Elation + 25% Triumph + 8% Ecstasy + 8% Embarrassment" [49] Reaction duration: 3.2s	"Woohoo" Reaction duration: 3.3s	Delightful and upbeat sound made with multiple instruments [52] Reaction duration: 8.1s	Fast, positive gesture with elation and open arms. Arms move [6,7].
Sadness	Crying "75% Sadness + 17% Distress + 8% disgust" [49] Reaction duration: 10.4s	Crying Reaction duration: 10.5s	Low-pitched, descending melodies "Piano minor chords" [54] Reaction duration: 11.4s	Slow, negative gesture. Body dropped and shrunk. Shoulders bowed. Hands kept lower than their normal positions, hands closed or moving slowly. The face covered with two hands [6,7].
Fear	Scream "67 Fear + 17% Pain + 8% Ecstasy + 8% Surprise(positive)" [49] Reaction duration: 4.9s	Scream Reaction duration: 4.9s	Spooky, high-pitch sound "Tremolo string sound" [54] Reaction duration: 5.3s	Arms rigged/tensed/clenched and out to side while robot looks around apprehensively and moves body backwards [7,9].
Surprise	Short Gasp "75% Surprise(negative) + 17% Fear + 8% Realization" [49] Reaction duration: 7.2s	Short Gasp Reaction duration: 6s	Atonal, harmonic sound. "Ascending fuzzy keyboard (Orchestration bang)" [37] Reaction duration: 7s	Fast gesture with shock and unexpectedness [7].
Anger	"Argh" "83% Anger + 8% Disgust + 8% Amusement" [49] Reaction duration: 3.3s	"Argh" Reaction duration: 3.3s	Short, noisy, electrical sound "Distorted percussive guitar chords" [53] Reaction duration: 4.6s	Arms Akimbo - Hands on hips, largely recognized as an angry gesture [7,8].

3.3 Experimental Design and Procedure:

A 5 (emotions) x 4 (sounds) within-subjects design was applied to this experiment. Twenty different combinations of emotions and sound types were provided to each participant with a set of corresponding gestures.

A single participant participated in each session. After the consent form procedure, each participant interacted with 20 trials. Participants were given video clips that had a text-to-speech (TTS)-generated story as a background sound. In the video clips, Pepper listened to the story with the participant. Note that the TTS voice without any embodiment read a story, not Pepper. Participants were notified that the narrative voice is a background storyteller and Pepper was listening to the stories with the participant. Throughout one story, Pepper had five gestural responses with sounds. After each emotion expression was presented, participants were asked about perceived emotions and the characteristics of the robot through an online questionnaire. The order of the four stories was counterbalanced across participants. The mapping between the story and each sound type condition was also counterbalanced across participants. The presentation order of emotions in each story was not counterbalanced because of the storyline. The whole sample procedure is depicted in Figures 2 and 3 below.

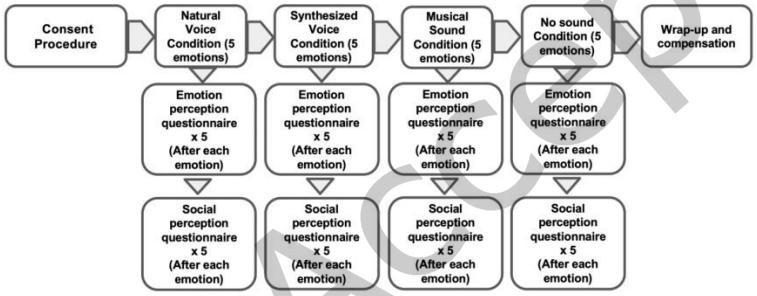


Figure 2: The flow diagram of the example procedure

THE THREE LITTLE PIGS
RESEARCHER VERSION
11/29/2021

Today, I'm going to read you the story of "The Three Little Pigs" Once upon a time there were three little pigs. When it came time for the pigs to build themselves homes, one little pig built his home out of straw. The second pig built his home out of twigs and sticks. These two pigs were lazy and had wanted to build their homes quickly so they could spend the rest of their day playing rather than working. The third little pig toiled hard all day in the sun and built himself a fine home of bricks. The two little pigs laughed at the third little pig for working so hard. They chided him for wasting his time and danced past his work to show off how much fun they were having.

ROBOT (anger)

One day, a big bad wolf saw the two little pigs out in the sun dancing and playing. He thought to himself, "What a tasty meal those pigs will be!" and he began to chase them. All the wolf could imagine was how tasty the pigs would be. The two pigs ran and hid inside their homes. So, the big bad wolf went over to the first home made of straw. He huffed, and he puffed, and he blew the house down in mere minutes.

ROBOT (surprise)

Figure 3: An example of the story the participant listened

The participants were asked to fill out the two questionnaires (emotion perception and social perception: Table 3) after watching each gesture and sound reaction generated by Pepper. Both questionnaires were provided five times for each story. After each response from the robot, the questionnaires [44] were administered to measure the accuracy of emotion perception; preference (Likability, Attractiveness), trust (Warmth, Honesty, Trustworthiness), and naturalness (Natural, Robot-like) of the robot. The questionnaires consisted of single-choice questions, 7-point Likert scale questions, and free-response questions. These questionnaires were used in the previous robot voice study [44].

Table 3: The list of questions and types in questionnaires

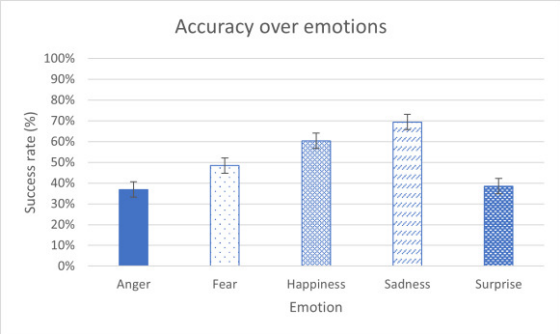
Category	Question (Type)
Emotion perception questionnaire	1. What emotion do you feel the robot expressed? (Choose from Anger, Fear, Happiness, Sadness, Surprise, and other) 2. How clearly did the robot express this emotion? (1-7 Likert scale) 3. Any other thoughts on the voice? (Optional, Open question)
Social perception questionnaire	1. How likable is the voice/gesture? (1-7 Likert scale) 2. How attractive is the voice/gesture? (1-7 Likert scale) 3. How warm is the voice/gesture? (1-7 Likert scale) 4. How honest is the voice/gesture? (1-7 Likert scale) 5. How trustworthy is the voice/gesture? (1-7 Likert scale) 6. How natural is the voice/gesture? (1-7 Likert scale) 7. How robotic is the voice/gesture? (1-7 Likert scale)

4 RESULTS

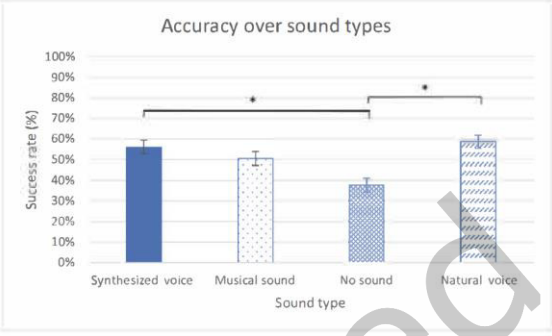
4.1 Emotion Perception: Accuracy and Clarity

The emotion perception accuracy was determined by the number of correct answers of emotion recognitions over the total number of answers as a percentage. Participants' answers from their perceived emotion were compared with the actual emotions present in the experiment (1: correct, 0: wrong). Because of this binary input for emotion recognition accuracy data, the results were analyzed with a 5 (Emotions) x 4 (Sound Types) Friedman Test testing main effects and Kendall's W Test computing the effect size. Figure 4 shows accuracy over emotions and sound types and interaction between emotions and sound types. There were statistically significant differences among the five emotions (Figure 4a; $X^2(4, 39) = 62.35$, $p < .01$, $W = .40$) and among the four sound types (Figure 4b; $X^2(3, 39) = 9.23$, $p = .03$, $W = .08$). Sadness was perceived most accurately by participants with a percentage of 69.4% followed by happiness, fear, anger, and surprise with percentages of 60.3%, 48.4%, 38.5%, and 36.9% each. Other than happiness and sadness, the emotion recognition accuracy of other emotions was lower than 50%. Participants perceived emotions most accurately in the natural voice condition and least accurately in the no sound condition. The pairwise comparisons of emotion recognition accuracy were analyzed with the Wilcoxon Signed Rank Test. The emotion recognition accuracy in the synthesized voice condition ($M = 0.56$, $SD = 0.50$) was significantly higher than in the no sound condition ($M = 0.38$, $SD = 0.48$; $Z(39) = 419.50$, $p = .01$). The natural voice condition ($M = 0.56$, $SD = 0.50$) was also significantly higher than the no sound condition ($Z(39) = 161.50$, $p = .01$). With all sound conditions, sadness was highly accurately recognized, whereas with the no sound

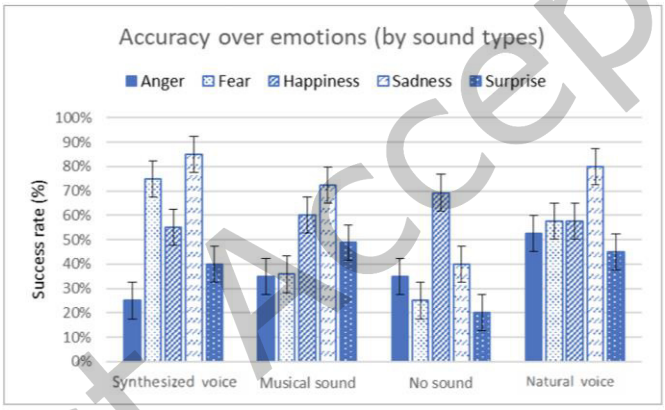
condition, sadness was not highly recognized (Figure 4c). With the no sound condition, only happiness was highly recognized, and the other emotions were recognized by less than 50%.



(a) emotion recognition accuracy over emotions



(b) emotion recognition accuracy over sounds

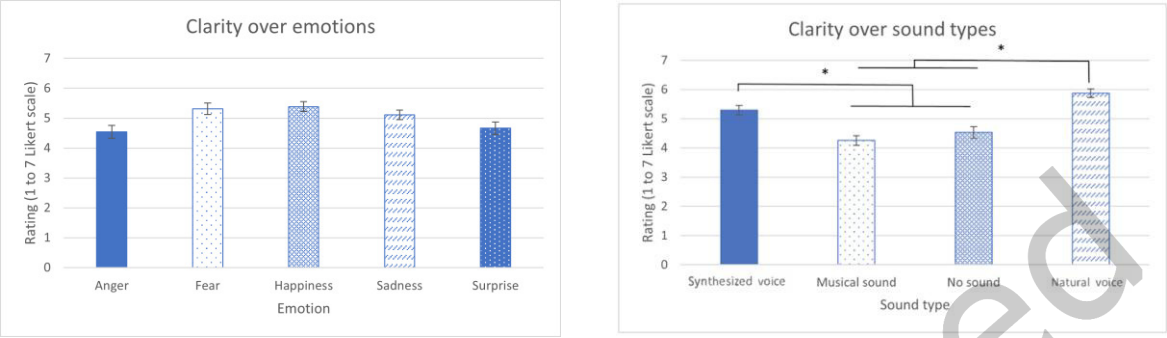


(c) emotion recognition accuracy over emotions by sounds

Figure 4: Accuracy of perceiving emotions over emotions, sound types, and emotions by sound types (*: $p < .05$, error bars represent standard errors).

The clarity of perceived emotion was rated by the participants based on a 7-point Likert scale (1: lowest, 7: highest). Only the clarity ratings of emotions that were perceived correctly by the participants were considered in the present study. The clarity results were analyzed with a 5 (Emotions) x 4 (Sound Types) Repeated Measures ANOVA. Figure 5 shows clarity over emotions and sound types. There were statistically significant differences in clarity among the five emotions (Figure 5a; $F(4,36) = 4.95$, $p < .01$, $\eta_p^2 = .35$) and among the four sound types (Figure 5b; $F(3,27) = 22.95$, $p < .01$, $\eta_p^2 = .72$) . No significant difference was found in the interaction between emotions and sound types. The average rating score of clarity was perceived the highest by participants in the natural voice condition. The clarity score of the natural voice ($M = 5.87$, SD

= 1.23) was significantly higher than the musical (M = 4.26, SD = 1.78) and no sound (M = 4.53, SD = 1.70) conditions. In addition, the average rating score of clarity was significantly higher in the synthesized voices (M = 5.29, SD = 1.70) than in the musical sound and no sound conditions.



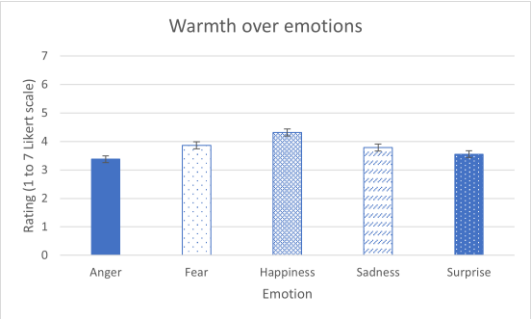
(a) clarity ratings over emotions (b) clarity ratings over sounds

Figure 5. The rating scores of clarity over emotions and sound conditions (*: $p < .0083$, error bars represent standard errors).

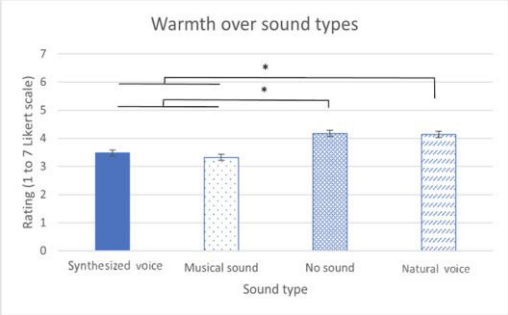
4.2 Trust Characteristics: Warmth, Honesty, and Trustworthiness

For social perceptions, we focused on analyzing the results based on sound types because we were interested in investigating the effects of different non-speech sounds on users' social perception towards the humanoid robot. The social perceptions data were all normally distributed, and Greenhouse-Geisser correction was applied for sphericity violation if needed. All pairwise comparisons for the sound types in the subjective ratings were analyzed with the Bonferroni correction (α = 0.05/6 = 0.0083). Previous studies have demonstrated that F-test (used in ANOVA or ANCOVA) was robust to violations of the interval data assumption and could be used to conduct statistical tests at the scale level of data using at least 5-point Likert response format with no resulting bias [57, 58]. Therefore, we used ANOVA for the analysis of social perception measures.

The trust characteristics of sound types were rated by the participants based on a 7-point Likert scale (1: lowest, 7: highest). The warmth results were analyzed with a 5 (Emotions) x 4 (Sound Types) Repeated Measures ANOVA. Figure 6 shows warmth rating scores over emotions and sound types. There were statistically significant differences in warmth among the five emotions (Figure 6a; $F(4,136) = 16.13$, $p < .01$, $\eta_p^2 = .32$) and among the four sound types (Figure 6b; $F(3,102) = 13.20$, $p < .01$, $\eta_p^2 = .28$). No significant difference was found in the interaction between emotions and sound types. The average rating score of warmth was perceived significantly higher in the natural voice (M = 4.14, SD = 1.61) and the no sound (M = 4.18, SD = 1.59) conditions than the synthesized voice (M = 3.48, SD = 1.63) and musical sound (M = 3.32, SD = 1.49) conditions respectively.



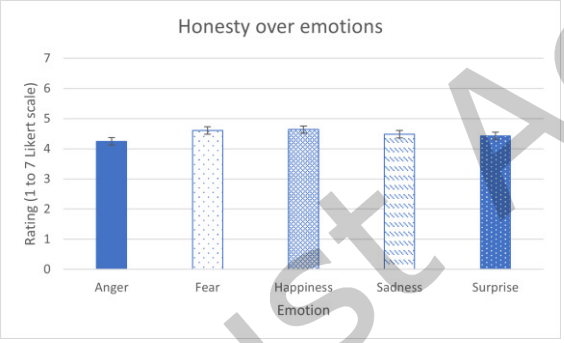
(a) warmth ratings over emotions



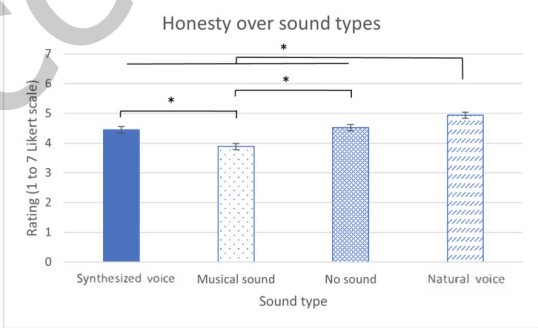
(b) warmth ratings over sounds

Figure 6. The rating scores of warmth over emotions and sound conditions (*: $p < .0083$, error bars represent standard errors).

Figure 7 shows honesty rating scores over sound types. There were statistically significant differences in honesty among the five emotions (Figure 7a; $F(3,102) = 16.57$, $p < .01$, $\eta_p^2 = .24$) and among the four sound types (Figure 7b; $F(4,136) = 3.52$, $p < .01$, $\eta_p^2 = .09$). No significant difference was found for the interaction effects between emotions and sound types. Participants rated sounds in the synthesized voice ($M = 4.45$, $SD = 1.54$), no sound ($M = 4.52$, $SD = 1.53$), and natural voice ($M = 4.94$, $SD = 1.50$) conditions significantly higher in honesty than in the musical sound condition ($M = 3.89$, $SD = 1.44$). In addition, the average rating score of honesty was perceived significantly higher in the natural voice condition than in the synthesized voice and no sound condition.



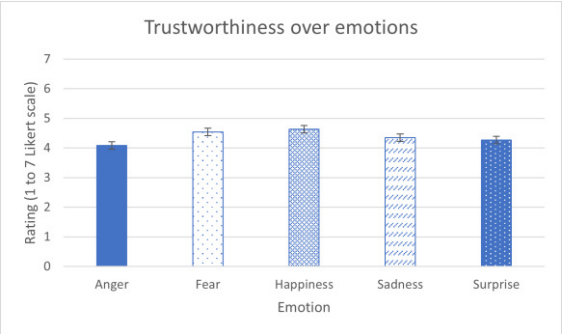
(a) honesty ratings over emotions



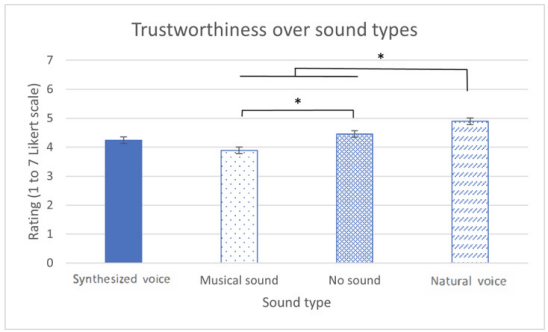
(b) honesty ratings over sounds

Figure 7. The rating scores of honesty over emotions and sound conditions (*: $p < .0083$, error bars represent standard errors).

Figure 8 shows trustworthiness rating scores over emotions and sound types. There were statistically significant differences in trustworthiness among the five emotions (Figure 8a; $F(4,136) = 4.13$, $p < .01$, $\eta_p^2 = .11$) and among the four sound types (Figure 8b; $F(3,102) = 9.38$, $p < .01$, $\eta_p^2 = .22$). No significant difference was found in the interaction between emotions and sound types. Participants rated sounds in the no sound ($M = 4.46$, $SD = 1.54$) and natural voice ($M = 4.90$, $SD = 1.53$) conditions significantly higher than in the musical sound condition ($M = 3.90$, $SD = 1.60$). In addition, the average rating score of trustworthiness was perceived significantly higher in the natural voice condition than the no sound condition.



(a) trustworthiness ratings over emotions

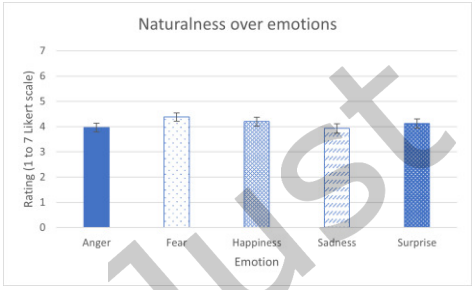


(b) trustworthiness ratings over sounds

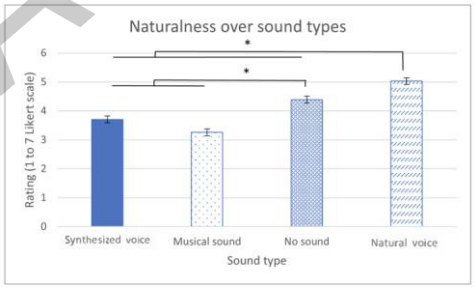
Figure 8. The rating scores of trustworthiness over emotions and sound conditions (*: $p < .0083$, error bars represent standard errors).

4.3 Naturalness: Naturalness and Robot-likeness

The naturalness ratings of sound types were rated by the participants based on a 7-point Likert scale (1: lowest, 7: highest). The naturalness results were analyzed with a 5 (Emotions) x 4 (Sound Types) Repeated Measures ANOVA. Figure 9 shows naturalness rating scores among the five emotions (Figure 9a; $F(3,99) = 30.00$, $p < .01$, $\eta_p^2 = .48$) and among the four sound types (Figure 9b; $F(4,132) = 3.52$, $p < .01$, $\eta_p^2 = .10$). No significant difference was found for the interaction effects between emotions and sound types. Participants rated sounds in the no sound ($M = 4.39$, $SD = 1.66$) and natural voice ($M = 5.03$, $SD = 1.59$) conditions significantly higher than in the musical sound condition ($M = 3.26$, $SD = 1.63$). The average rating score of naturalness was also perceived significantly higher by participants in the natural condition than in the synthesized voice ($M = 3.71$, $SD = 1.73$) and no sound conditions. Moreover, the average rating score of naturalness was perceived significantly higher by participants in the no sound condition than in the synthesized voice condition.



(a) naturalness ratings over emotions



(b) naturalness ratings over sounds

Figure 9. The rating scores of naturalness over emotions and sound conditions (*: $p < .0083$, error bars represent standard errors).

Figure 10 shows robot-likeness rating scores over sound types ($F(3,102) = 27.24$, $p < .01$, $\eta_p^2 = .45$). No significant difference was found either for the main effect of emotions or for the interaction effects between emotions and sound types. Participants rated sounds in no sound ($M = 4.30$, $SD = 1.78$) and natural voice ($M = 3.13$, $SD = 1.63$) conditions significantly lower than in the musical sound condition ($M = 5.10$, $SD = 1.58$).

The synthesized voice (M = 4.71, SD = 1.70) was also significantly lower than the musical sound. The average rating score of robot-likeness was perceived significantly lower in the natural voice condition than in the synthesized voice and no sound conditions. Moreover, the average rating score of robot-likeness was perceived marginally lower in the no sound condition than the synthesized voice condition.

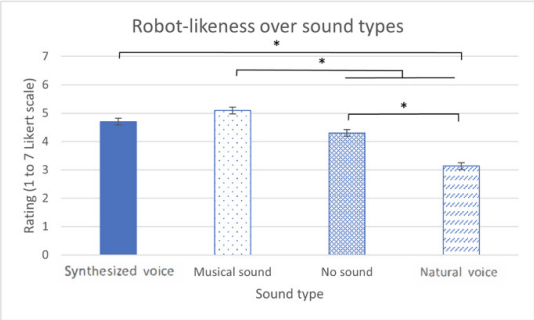
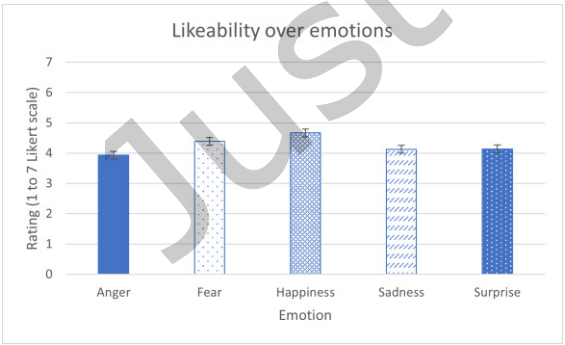


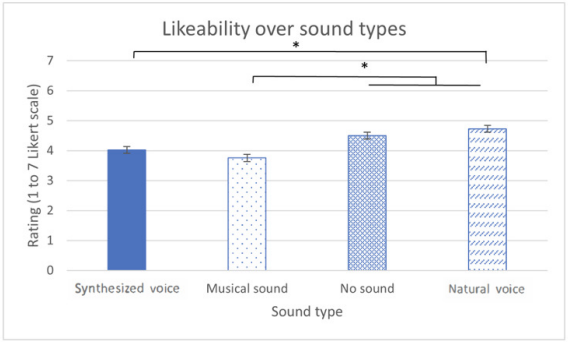
Figure 10. The rating scores of robot-likeness over sound conditions (*: $p < .0083$, error bars represent standard errors).

4.4 Preferences: Likability and Attractiveness

The preferences ratings of sound types were rated by the participants based on a 7-point Likert scale (1: lowest, 7: highest). The preferences rating results were analyzed with a 5 (Emotions) x 4 (Sound Types) Repeated Measures ANOVA. Figure 11 shows likability rating scores over emotions and sound types. There were statistically significant differences in likability among the five emotions (Figure 11a; $F(4,136) = 11.46$, $p < .01$, $\eta_p^2 = .23$) and among the four sound types (Figure 11b; $F(3,102) = 19.84$, $p < .01$, $\eta_p^2 = .25$). No significance was found in the interaction between emotions and sound types. Participants rated the no sound condition ($M = 4.50$, $SD = 1.59$) significantly higher than the musical sound condition ($M = 3.76$, $SD = 1.56$). In addition, the natural voice condition ($M = 4.73$, $SD = 1.71$) had a significantly a higher likeability rating score than both the synthesized voice ($M = 4.02$, $SD = 1.76$) and the musical sound conditions.



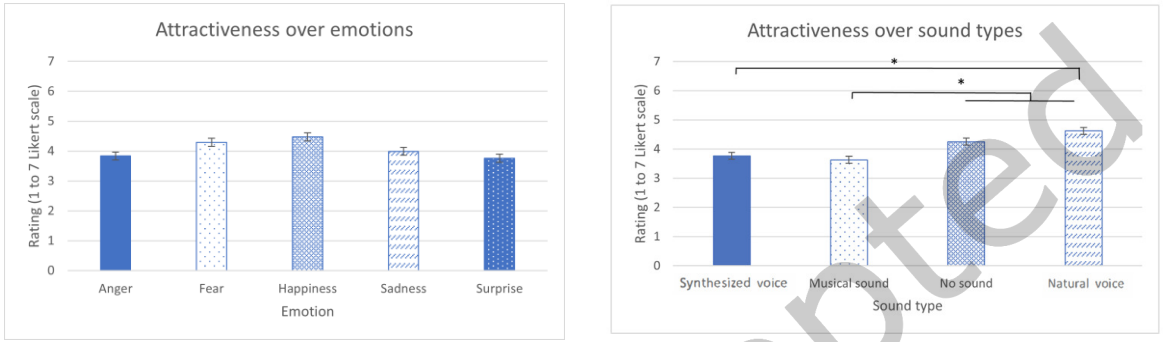
(a) likability ratings over emotions



(b) likability ratings over sounds

Figure 11. The rating scores of likeability over emotions and sound conditions (*: $p < .0083$, error bars represent standard errors).

Figure 12 shows attractiveness rating scores over emotions and sound types. There were statistically significant differences in attractiveness among the five emotions (Figure 12a; $F(4,136) = 10.20$, $p < .01$, $\eta_p^2 = .23$) and among the four sound types (Figure 12b; $F(3,102) = 13.37$, $p < .01$, $\eta_p^2 = .28$). No significance was found in the interaction between emotions and sound types. Participants rated the no sound condition ($M = 4.25$, $SD = 1.62$) significantly higher than the musical sound condition ($M = 3.63$, $SD = 1.60$). In addition, the natural voice condition ($M = 4.62$, $SD = 1.73$) had a significantly a higher likeability rating score than both the synthesized voice ($M = 3.77$, $SD = 1.86$) and the musical sound conditions.



(a) attractiveness ratings over emotions (b) attractiveness ratings over sounds

Figure 12. The rating scores of attractiveness over emotions and sound conditions (*: $p < .0083$, error bars represent standard errors).

5 DISCUSSION

A preliminary study was conducted to obtain a comprehensive view of the influence of emotion types, non-speech sounds, and gestures on users' perceptions of robot emotions and other characteristics. Overall, results showed that the natural voice appeared to have higher impacts on emotion and social perceptions towards the humanoid robots than the other sounds.

5.1 Accuracy and Clarity

As predicted, the accuracy of emotion recognition was significantly higher in all sound conditions than in the no sound condition. However, unexpectedly, the accuracy of emotion recognition of musical sounds is inferior to that of other human vocal sounds. Musical sounds also showed the lowest clarity among all the conditions. This might suggest that incorporating musical sounds into a humanoid robot Pepper could potentially reduce clarity even compared to the no sound (i.e., gesture-only) in our experiment setup. The results of the emotion recognition accuracy also showed significant differences in emotions, sound types, and the interaction between emotions and sound types. Happiness demonstrated significantly higher accuracy than other emotions in the no sound (gesture only); this result is consistent with the prior study that joy is best expressed through color and movement [59]. Sadness demonstrated significantly higher accuracy than anger, fear, and surprise in the synthesized voice, musical voice, and natural voice which is also consistent with a prior study that sadness conveyed the best performance through sound [59]. A prior study proposed the idea of the inclusion of sound helps compensate for the perception of emotions [60]. It is in line with the

previous study [36], which showed that sadness was one of the most accurately recognized emotions in the affective prosody experiment. Taken together, we may cautiously infer that sadness can be easily induced by both speech and non-speech sounds. In addition, happiness showed significantly higher recognition accuracy than anger and fear. In our previous study about robot speech [44], we also showed that anger and fear showed lower emotion recognition accuracy, whereas happiness and sadness showed higher emotion recognition accuracy. It might be because happiness and sadness are more common emotional states the participants can expect from the fairytales [44]. Anger and fear are all negative-high arousal emotions. People might not expect these types of high-strength, negative emotions from fairytales. However, it cannot be validated from the present study and thus, it requires further research.

Table 4 shows that anger was mostly misclassified as surprise (16.9%) and surprise was mostly misclassified as fear (30.4%). Interestingly, fear was mostly misclassified as happiness (21.4%) and happiness was mostly misclassified as fear (17.0%) when those two emotions were supposed to have opposing emotional valence. A previous study has shown that happiness and anger are easily misclassified because of the acoustic characteristics of similar higher pitch and faster speech rate [34]. Although sadness had the highest accuracy among the emotions, 30.6% of sadness was also not recognized correctly by the participants.

Table 13. The confusion matrix between presented and perceived emotions

Perceived	**Presented**	**Anger**	**Fear**	**Happiness**	**Sadness**	**Surprise**
Anger	Count	59	4	1	3	18
	Col %	36.9%	2.5%	0.6%	1.9%	11.2%
Fear	Count	22	77	27	15	49
	Col %	13.8%	48.4%	17.0%	9.4%	30.4%
Happiness	Count	17	34	96	7	8
	Col %	10.6%	21.4%	60.4%	4.4%	5.0%
Sadness	Count	14	4	4	111	5

	Col %	8.8%	2.5%	2.5%	69.4%	3.1%
Surprise	Count	27	25	15	8	62
	Col %	16.9%	15.7%	9.4%	5.0%	38.5%
Other	Count	21	15	16	16	19
	Col %	13.1%	9.4%	10.1%	10.0%	11.8%

The emotion recognition accuracy in the synthesized voice, musical sound, and natural voice conditions were all significantly higher than in the no sound condition, which was expected because multimodal cues can enhance the emotion recognition performances [24]. Participants showed higher clarity scores in the natural voice condition than the other three sound types significantly. The synthesized voice also received significantly higher clarity scores than the musical and no sound conditions from participants. The results of the study indicate that, in the context of our experiment design, a natural or synthesized voice may be more suitable than musical sounds for accurately expressing emotions.

5.2 Trust, Naturalness, and Preferences

We can infer a clear trend in all three categories of social perceptions - trust, naturalness, and preference ratings. Natural voice was the highest in all measures, whereas musical sound was the lowest in all measures. In terms of the trust scale, natural voice and no sound showed relatively higher ratings in all three ratings, including warmth, honesty, and trustworthiness. Musical sound was the lowest. When the humanoid robot had natural voice with gestures, people could interpret it as conveying emotions and might have considered it "credible" [61-63]. Also, proper gestures alone (i.e., no sound condition) could play a crucial role in conveying meaning, guiding, leading, and building rapport among discussants [64]. On the other hand, literature shows that when the interface is noisy and gimmicky, users are frustrated by it [65]. The inappropriate use of musical sounds for a humanoid robot might even harm users' social perceptions and trust towards the robot. Interestingly, the musical sound condition showed the lowest rating in natural and the highest rating in the robot-likeness scale. There seems to be a general belief that the use of movement and music together can express more emotions than facial expression or music alone [13]. But this was not supported by our study. It implies that musical sound may not be suitable for a humanoid robot to express its emotional states effectively in our experiment design. On the other hand, the no sound condition showed higher naturalness than the synthesized voice. Therefore, we can infer that a robotically synthesized voice is not what participants expected to hear from our experiment scenario with the humanoid robot Pepper. With the pervasive use of voice user interfaces as voice assistants in smart devices, they are getting closer to people's natural voices, and participants may prefer to hear natural human voices from the humanoid robot. In the same line, the musical sound showed the highest robot-likeness with the lowest naturalness. In the

likability and attractiveness scales, participants showed the highest rating scores for natural voice, followed by no sound. The synthesized voice and musical sound differed from what participants expected to hear from a humanoid robot Pepper in our experiment scenario. In conclusion, people seem to prefer the human-like natural voice with humanoid robots [66] and rather prefer no sound over synthesized or musical sounds when they provide sufficient body gestures. Lastly, it was not our focus in the present study, but with the happy emotion, participants felt the highest warmth, trust, preference, and attractiveness, which we could readily anticipate.

5.3 Revisiting RQs and Extracting Design Guidelines

Based on the results of the present study, we can provide design guidelines for emotional cues of a humanoid robot as follows. Of course, depending on users, robot types, tasks, and situations, the guidelines may vary.

- Adding non-speech sounds to a humanoid robot's emotional gesture can increase emotion recognition accuracy regardless of the type of sounds, including natural voice, robotically synthesized voice, or musical sounds (RQ1).
- Using a natural voice will significantly increase emotion recognition accuracy than other sounds (RQ2).
- Using a natural voice will increase trust, naturalness, and preference towards, at least a humanoid robot (RQ2).
- Using musical sounds might not increase positive perceptions towards a humanoid robot or even harm social perceptions. Thus, careful design is required (RQ2).
- Using a robotically synthesized voice for a humanoid robot does not seem to be effective in increasing social perceptions, compared to a natural voice (RQ3).
- Depending on the emotion type, different sounds may enhance the perception accuracy differently (e.g., fear recognition accuracy is uniquely high only in the synthesized voice, whereas sadness recognition accuracy is generally high in all sound conditions) (RQ4).
- Robot gestures without any non-speech sounds can also convey a certain emotional state (e.g., happiness with 70% accuracy).

6 LIMITATIONS AND FUTURE WORK

Even though the experimental protocol is promising and provides interesting implications, this study still has limitations. First, only limited sound design approaches were used. For example, only male voices were used as emotion expressions in this study. Depending on the gender of the voice, the results might be different. Also, the synthesized voice can be designed in different ways (e.g., frequency modulation synthesis or vocoding). Musical sounds for the same emotion can also vary in unlimited ways. The duration of the sounds for different emotions varied. The vocalizations of the same emotion also had different durations for different conditions. Different stimuli duration could impact participants' social and emotional perception toward the robot. Pepper's reactions in the videos were also not instant, and there was a gap. This delay in reaction time may also influence participants' responses and may decrease their satisfaction [71]. According to research, 750 ms is the optimal time point for users to get the best subjective and psychological experience of feedback, and the threshold limit time for users to wait for feedback is 1850 ms, beyond which the level of arousal and valence of users' emotions may decrease [72]. Therefore, the generalizability of this study's results can be limited. In the future, alternative sound design methods can be adopted using a free software

program and compared across emotions and even in the same emotion category. Second, only one humanoid robot was used in this experiment. The results might vary depending on the robot appearance and form factors (e.g., animal or mechanical robots). Third, to see the independent effects of the non-speech sounds, the sounds from Pepper were not accompanied by any speech. When these non-speech sounds are combined with speech, people might perceive stronger emotions and higher social perceptions towards the robot. Next, presenting the questionnaire in the middle of the story might influence the participants' perception and break the flow. It could be addressed in the future study. Lastly, the experiment was designed during the COVID-19 pandemic, and thus, the experiment used the video format for Pepper (however, note that the actual experiment was conducted after the COVID-19 protocol was lifted by the university IRB so that participants did not go through an additional screening phase). The recorded videos of Pepper presented to participants might also have influenced the results (e.g., emotion recognition accuracy and clarity) [43]. The quality of the synthesized speech may negatively affect their perception of naturalness [67] and should be addressed in the next study by having a physical robot with participants in person. The post-condition questionnaire used in this study was designed by researchers in the previous robot voice studies [44, 68]. This questionnaire provided interesting results; however, in the future, validated questionnaires will also be used, such as the Godspeed [69] and Robotic Social Attributes Scale [70]. In future work, more participants with diverse cultural backgrounds should be recruited to generalize the results.

ACKNOWLEDGEMENT

The authors would like to thank Shiven Saxena for the design and post-processing of the sounds.

REFERENCES

[1] Yilmazyildiz, S., Read, R., Belpeame, T., & Verhelst, W. (2016). Review of semantic-free utterances in social human-robot interaction. International Journal of Human-Computer Interaction, 32(1), 63-85.

[2] Seyama, J. I., & Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. Presence: Teleoperators and virtual environments, 16(4), 337-351.

[3] Dong, J., Santiago-Anaya, A., & Jeon, M. (2022). Facial expressions increase emotion recognition accuracy and clarity on a humanoid robot without adding the uncanny valley. Proceedings of the Human Factors and Ergonomics Society's 2022 International Annual Meeting (HFES2022), Atlanta, GA, October 10 - 14.

[4] 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. In 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)(pp. 293-300).

[5] Kishi, T., Kojima, T., Endo, N., Destephe, M., Otani, T., Jamone, L., ... & Takanishi, A. (2013). Impression survey of the emotion expression humanoid robot with mental model based dynamic emotions. In 2013 IEEE International Conference on Robotics and Automation (pp. 1663-1668). IEEE.

[6] Pelikan, H.R., Broth, M., and Keevallik, L. 2020. "Are you sad, Cozmo?". *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*.

[7] Noroozi, F., Corneanu, C. A., Kaminska, D., Sapinski, T., Escalera, S., & Anbarjafari, G. (2021). Survey on emotional body gesture recognition. IEEE Transactions on Affective Computing, 12(2), 505-523.

[8] Cabibihan, J.-J., So, W.-C., and Pramanik, S. 2012. Human-recognizable robotic gestures. *IEEE Transactions on Autonomous Mental Development 4*, 4, 305–314.

[9] Embgen, S., Luber, M., Becker-Asano, C., Ragni, M., Evers, V., and Arras, K.O. 2012. Robot-specific social cues in emotional body language. *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*.

[10] Schirmer, A., Adolphs, R. 2017. Emotion perception from face, voice, and touch: Comparisons and convergence. Trends Cogn Sci 21(3), 216-228

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

[12] Savery, R., Rose, R., and Weinberg, G. 2019. Establishing human-robot trust through music-driven robotic emotion prosody and gesture. *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*.

[13] Laukka, P., Elfenbein, H. A., Söder, N., Nordström, H., Althoff, J., Chui, W., et al. 2013. Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. *Frontiers in Psychology, 4*, 353.

[14] Ros, R., Nalin, M., Wood, R., Baxter, P., Looije, R., Demiris, Y., Belpaeme, T., Giusti, A., & Pozzi, C. (2011). Child-robot interaction in the wild. *Proceedings of the 13th International Conference on Multimodal Interfaces*, 335–342

[15] Dautenhahn, K. 2007. Socially intelligent robots: Dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences 362*, 1480, 679–704.

[16] David, D., Hayotte, M., Thérouanne, P., d'Arripe-Longueville, F., & Milhabet, I. 2022. Development and validation of a social robot anthropomorphism scale (SRA) in a french sample. International Journal of Human-Computer Studies, 162, 102802.

[17] 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.

[18] Lohse, M., Hegel, F., Swadzba, A., Rohlfing, K., Wachsmuth, S., Wrede, B. 2007. What can I do for you? Appearance and application of robots. In: Proceedings of AISB, Vol. 7, pp. 121-126.

[19] Castellano, G., Leite, I., Pereira, A., Martinho, C., Paiva, A., and McOwan, P.W. 2009. Affect recognition for interactive companions: Challenges and design in real world scenarios. *Journal on Multimodal User Interfaces 3*, 1-2, 89–98.

[20] Salem M, Rohlfing K, Kopp S, Joublin F. 2011. A friendly gesture: investigating the effect of multimodal robot behavior in human–robot interaction. In: 2011 RO-MAN 2011, pp 247–252.

[21] Breazeal, C. 2003. Toward sociable robots. *Robotics and Autonomous Systems 42*, 3-4, 167–175.

[22] McColl, D., Hong, A., Hatakeyama, N., Nejat, G., & Benhabib, B. 2016. A survey of autonomous human affect detection methods for social robots engaged in natural HRI. Journal of Intelligent & Robotic Systems, 82(1), 101-133.

[23]Borutta, I., Sosnowski, S., Zehetleitner, M., Bischof, N., & Kuhnlenz, K. 2009. Generating artificial smile variations based on a psychological system-theoretic approach. In RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication (pp. 245-250). IEEE.

[24] Calvo, 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

[25] Kuehn, J., & Haddadin, S. 2016. An artificial robot nervous system to teach robots how to feel pain and reflexively react to potentially damaging contacts. IEEE Robotics and Automation Letters, 2(1), 72-79.

[26] Jee, E.-S., Kim, C.H., Park, S.-Y., and Lee, K.-W. 2007. Composition of musical sound expressing an emotion of robot based on musical factors. *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*.

[27] Blow, M. P., Dautenhahn, K., Appleby, A., Nehaniv, C. L. & Lee, D. 2006. Perception of robot smiles and dimensions for human–robot interaction design. In Proc. 15th IEEE Int. Symp. on Robot and Human Interactive Communication (RO-MAN06), Hatfield, UK, 6–8 September 2006, pp. 469–474.

[28] Read, R., & Belpaeme, T. (2014). Situational context directs how people affectively interpret robotic non-linguistic utterances. Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction.

[29] Read, R., & Belpaeme, T. (2016). People interpret robotic non-linguistic utterances categorically. International Journal of Social Robotics, 8(1), 31–50.

[30] Wolfe, H., Peljhan, M., & Visell, Y. (2020). Singing robots: how embodiment affects emotional responses to non-linguistic utterances. IEEE Transactions on Affective Computing, 11(2), 284–295.

[31] Ward, N., 2004. Pragmatic functions of prosodic features in non-lexical utterances, *In SP-2004*, 325-328.

[32] Juslin, P. N., & Laukka, P. 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129*(5), 770.

[33] Koeda, M., Belin, P., Hama, T., Masuda, T., Matsuura, M., & Okubo, Y. 2013. Cross-cultural differences in the processing of non-verbal affective vocalizations by Japanese and Canadian listeners. *Frontiers in Psychology, 4,*105.

[34] Yilmazyildiz, S., Henderickx, D., Vanderborght, B., Verhelst, W., Soetens, E., & Lefeber, D. 2013. Multi-modal emotion expression for affective human–robot interaction. Proceedings of the workshop on affective social speech signals.

[35] Liu, T., Pinheiro, A. P., Deng, G., Nestor, P. G., McCarley, R. W., & Niznikiewicz, M. A. 2012. Electrophysiological insights into processing nonverbal emotional vocalizations. *NeuroReport, 23*(2), 108–112.

[36] Schröder, M. (2003). Experimental study of affect bursts. *Speech Communication, 40*(1), 99–116.

[37] Vasconcelos, M., Dias, M., Soares, A. P., & Pinheiro, A. P. 2017. What is the melody of that voice? Probing unbiased recognition accuracy of nonverbal vocalizations with the Montreal Affective Voices. *Journal of Nonverbal Behavior,41*(3), 239–267.

[38] Emma, F., & Roberto, B. (2021). Perceptual evaluation of blended sonification of mechanical robot sounds produced by emotionally expressive gestures: augmenting consequential sounds to improve non-verbal robot communication. *International Journal of Social Robotics*, 1-16, 1–16.

[39] Zahray L, Savery R, Syrkett L, Weinberg G (2020) Robot gesture sonification to enhance awareness of robot status and enjoyment of interaction. In: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp 978–985.

[40] Devillers, L., Vidrascu, L., and Lamel, L. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks 18*, 4, 407–422.

[41] Jee, E.-S., Kim, C. H., Park, S.-Y., & Lee, K.-W. (2007). Composition of musical sound expressing an emotion of robot based on musical factors. In Proceedings of the 16th international symposium on robot and human interactive communication, 637–641.

[42] Komatsu, T., Yamada, S., Kobayashi, K., Funakoshi, K., and Nakano, M. 2010. Artificial subtle expressions. *Proceedings of the 28th international*

conference on Human factors in computing systems - CHI '10.

[43] Jeon, M., & Rayan, I. A. 2011. The effect of physical embodiment of an animal robot on affective prosody recognition. In International Conference on Human-Computer Interaction (pp. 523-532). Springer, Berlin, Heidelberg.

[44] Ko, S., Liu, X., Mamros, J., Lawson, E., Swaim, H., Yao, C., & Jeon, M. 2020. 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.

[45] Read R, Belpaeme T (2010) Interpreting non-linguistic utterances by robots: studying the influence of physical appearance. In: Proceedings of the 3rd international workshop on affective interaction in natural environments (AFFINE 2010) at ACM multimedia 2010. ACM, Firenze, pp 65–70

[46] Read R, Belpaeme T (2012) How to use non-linguistic utterances to convey emotion in child–robot interaction. In: Proceedings of the 7th international conference on human–robot interaction (HRI'12). ACM/IEEE, Boston, pp 219–220.

[47] Fernandez De Gorostiza luengo, J., Alonso Martin, F., Castro-Gonzalez, A., & Salichs, M. A. (2017). Sound synthesis for communicating nonverbal expressive cues. *Ieee Access*, *5*.

[48] Bavelas, J. B., L. Coates, and T. Johnson (2000). Listeners as co-narrators. Journal of Personality and Social Psychology 79(6), 941–952.).

[49] Cowen, A. S., Elfenbein, H. A., Laukka, P., & Keltner, D. 2019. Mapping 24 emotions conveyed by brief human vocalization. American Psychologist, 74(6), 698–712.

[50] Fischer, K, Niebuhr, O, Jensen, L. C. and Bodenhagen, L. (2020): Speech Melody Matters – How robots can profit from using charismatic speech. ACM Transactions on Human-Robot Interaction 9, 1, Article 4: 1-21

[51] Blattner, M. M., Sumikawa, D. A., & Greenberg, R. M. 1989. Earcons and icons: Their structure and common design principles. Human–Computer Interaction, 4(1), 11-44.

[52] Jeon, M., Heo, U., Ahn, J. H., & Kim, J. (2008). Emotional palette: Affective user experience elements for product design according to user segmentation. Proceedings of the 6th International Conference of Cognitive Science (ICCS2008), 600-603.

[53] Sterkenburg, J., Jeon, M., & Plummer, C. 2014. Auditory emoticons: Iterative design and acoustic characteristics of emotional auditory icons and earcons. In International Conference on Human-Computer Interaction (pp. 633-640). Springer, Cham.

[54] Jeon, M., Lee, J. H., Sterkenburg, J., & Plummer, C. 2015. Cultural differences in preference of auditory emoticons: USA and South Korea. Georgia Institute of Technology.

[55] Fischer, K., & Niebuhr, O. (2020). Studying language attitudes using robots. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*.

[56] Ekman, P. 1992. Are there basic emotions? Psychological Review, 99(3), 550–553.

[57] Carifio, J. & Perla, R. (2007). Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. Journal of Social Sciences, 2, 106-116.

[58] Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. Advances in health sciences education, 15(5), 625-632.

[59] Loffler, D.; Schmidt, N.; Tscharn, R.; 2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI) Chicago, IL, USA 2018 March 5 - 2018 March 8. 2018 13th Acm/ieee International Conference on Human-Robot Interaction (hri). In *Multimodal Expression of Artificial Emotion in Social Robots Using Color, Motion and Sound*; ACM, 2018; pp 334–343.

[60] Adrian B. Latupeirissa, Claudio Panariello, & Roberto Bresin. (2020, June 17). Exploring emotion perception in sonic HRI

[61] Breazeal C, Scassellati B. 1999. How to build robots that make friends and influence people. In: Proceedings of the 1999 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 858–863

[62] Rani P, Sarkar N. 2004. Emotion-sensitive robots - a new paradigm for human-robot interaction. In: Proceedings of the 4th IEEE/RAS international conference on humanoid robots, vol 1, pp 149–167

[63] Breazeal C, Aryananda L. 2002. Recognition of affective communicative intent in robot-directed speech. Auton Robots 12(1):83–104

[64] Duncan, S. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology 23*, 2, 283–292.

[65] Jennifer, P., Yvonne, R., & Helen, S. 2002. Interaction design: beyond human-computer interaction. NY: Wiley.

[66] Ray C, Mondada F, Siegwart R. 2008. What do people expect from robots? In: Proceedings of the 2008 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 3816–3821

[67] Yilmazyildiz, S., Latacz, L., Mattheyses, W., & Verhelst, W. (2010). Expressive gibberish speech synthesis for affective human-computer interaction. In International Conference on Text, Speech and Dialogue (pp. 584-590). Springer, Berlin, Heidelberg.

[68] Ko, S., Barnes, J., Dong, J., Park, C.H., Howard A., & Jeon, M. (in press). The effects of robot voices and appearances on users emotion recognition and subjective perception, *International Journal of Humanoid Robotics*

[69] Bartneck, C., Kulić Dana, Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International Journal of Social Robotics, 1(1), 71–81.

[70] Carpinella, C. M.; Wyman, A. B.; Perez, M. A.; Stroessner, S. J.; 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI) Vienna, Austria 2017 March 6 - 2017 March 9. 2017 12th Acm/ieee International Conference on Human-Robot Interaction (hri. In The Robotic Social Attributes Scale (rosas): Development and Validation; ACM, 2017; pp 254–262.

[71] Yang, E., & Dorneich, M. C. (2015). The effect of time delay on emotion, arousal, and satisfaction in human-robot interaction. Proceedings

of the Human Factors and Ergonomics Society Annual Meeting, 59(1), 443–447.

[72] Wang, J., Li, Y., Yang, S., Dong, S., & Li, J. (2023). Waiting experience: Optimization of feedback mechanism of voice user interfaces based on time perception. IEEE Access, 11, 21241–21251.