Liu_X_T_2023

By: Xiaozhen Liu

As of: Jun 25, 2023 1:48:52 AM
14,851 words - 127 matches - 55 sources

Similarity Index

13%

Mode: Similarity Report

paper text:

The Effects of a Humanoid Robot's Non-lexical Vocalization on Emotion Recognition and Robot Perception Xiaozhen Liu

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of Master of Science in Industrial and Systems Engineering 14

Thesis Myounghoon Jeon, Chair Sol Lim EunJeong Cheon May 5, 2023 Blacksburg, VA Keywords: Human-centered computing, Human robot interaction (HRI), Interaction techniques, Auditory feedback, Robotics, Robotic components The Effects of a Humanoid Robot's Non-lexical Vocalization and Musical Sound on Emotion Recognition and Robot Perception Xiaozhen Liu ABSTRACT As robots have become more pervasive in our everyday life, social aspects of robots have attracted researchers' attention. Because emotions play a key role in social interactions, research has been conducted on conveying emotions via speech, whereas little research has focused on the effects of non-speech sounds on users' robot perception. We conducted a within-subjects exploratory study with 40 young adults to investigate the effects of non-speech sounds (regular voice, characterized voice, musical sound, and no sound) and basic emotions (anger, fear, happiness, sadness, and surprise) on user perception. While listening to the fairytale with the participant, a humanoid robot (Pepper) responded to the story with a recorded emotional sound with a gesture.

Participants showed significantly higher emotion recognition accuracy from **the** regular voice **than** from other sounds. **The** 2

confusion matrix showed that happiness and sadness had the highest emotion recognition accuracy, which aligns with the previous research. Regular voice also induced higher trust, naturalness, and preference compared to other sounds. Interestingly, musical sound mostly showed lower perceptions than no sound. A further exploratory study was conducted with an additional 49 young people to investigate the effect of regular non-verbal voices (female voices and male voices) and basic emotions (happiness, sadness, anger, and relief) on user perception. We also further explored the impact of participants' gender on emotion and social perception toward robot Pepper. While listening to a fairy tale with the participants, a humanoid robot (Pepper) responded to the story with gestures and emotional voices. Participants showed

significantly higher emotion recognition accuracy and social perception from the voice + Gesture condition than Gesture only conditions. The confusion matrix showed that happiness and sadness had the highest emotion recognition accuracy, which aligns with the previous research. Interestingly, participants felt more discomfort and anthropomorphism in male voices compared to female voices. Male participants were more likely to feel uncomfortable when interacting with Pepper. In contrast, female participants were more likely to feel warm. However,

the gender of the robot **voice or the gender of the** 36

participant did not affect the accuracy of emotion recognition. Results are discussed with social robot design guidelines for emotional cues and future research directions. Exploring the influence of emotional perception on non-speech sounds and non-linguistic sounds Xiaozhen Liu GENERAL AUDIENCE ABSTRACT As robots increasingly appear in people's lives as functional assistants or for entertainment, there are more and more scenarios in which people interact with robots. More research on human- robot interaction is being proposed to help develop more natural ways of interaction. Our study focuses on the effects of emotions conveyed by a humanoid robot’s non-speech sounds

on people’s perception about **the robot and** its emotions. **The** results **of** 45

our experiments show that the accuracy of emotion recognition of regular voices is significantly higher than that of music and robot-like voices and elicits higher trust, naturalness, and preference. The

gender of the robot 's voice **or** the gender **of the participant** did not affect **the** 22

accuracy of emotion recognition. People are now not inclined to traditional stereotypes of robotic voices (e.g., like old movies), and expressing emotions with music and gestures mostly shows a lower perception. Happiness and sadness were identified with the highest accuracy among the emotions we studied. Participants felt more discomfort and human-likeness in the male voices than in female voices. Male participants were more likely to feel uncomfortable when

interacting with the humanoid **robot** , while female **participants were more likely to** 46

feel warm. Our study discusses design guidelines and future research directions for emotional cues in social robots. ACKNOWLEDGMENTS First,

I would like to acknowledge and **thank my advisor** , Dr. Myounghoon Jeon **for the continuous support of my** Master **study and research. His** guidance **and** 15

suggestions guided me through every stage

of my master's thesis. His patience, encouragement, and tremendous **knowledge** 35

guided me through all the time of research, conference (HCII) and Journal paper (under revision*). Your mentorship has helped me a lot in my professional research years.

I would like to thank Dr . Sol Lim **and Dr** . EunJeong Cheon **for** accepting **my** 40

invitation to join my thesis committee and offering recommendations, comments, and time

to help me finish **this** thesis. **I would** also **like to thank** the Virginia Tech **and** 42

Grado Department of Industrial and Systems Engineering community for providing great resources and environment to pursue a degree.

Last but not the **least, I would like to thank my family** 31

that supported me all the time and all the way through my education. * Study 1 has been submitted to a journal special issue, “Sound in HRI” as below and is currently under revision. Liu, X., Dong, J., & Jeon, M. (under revision). Robots’ “Woohoo” and “Argh” can enhance users’ emotional and social perceptions: An exploratory study on non-lexical vocalizations and non-linguistic sounds. ACM Transactions on Human-Robot Interaction. iv TABLE OF CONTENTS

1. INTRODUCTION 1.1 MOTIVATION 1.1.1 RESEARCH OBJECTIVES With **the** advancement **of** 34

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, which we call, semantic-free utterances (SFUs). The SFUs used in the present study include non-lexical vocalizations (e.g., an exclamation like “Argh”) or non-linguistic sounds (e.g., musical sounds like humming) [10-13]. Taken together, we should also consider the complexity of behavioral synchronization in the relationship among robots' appearance, body gestures, and SFUs. 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 [14]. We tried to explore how the robot’s SFUs 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** 44

’s emotional states with various non-speech auditory cue types (regular human voice, robotically characterized (or metallic) voice, musical sounds, and no sounds) and

basic emotions (anger, fear, happiness, sadness, and surprise). Participants **were** 4

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 55

. 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

[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** 27

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 this 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 SFUs and gestures. 2.2 Robot Emotions The purpose of social robots is to actively connect with humans in order to accomplish their own social objectives [21]. Emotional engagement between humans and robots makes communications between the two more natural. The capacity to perceive feelings and emotions is an essential data source 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** 47

[23]. To make more emotional robots, researchers have tried to implement robots that can (1) detect human emotions, (2) perceive their own feelings, 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 feelings 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 Semantic-free utterances can be described as an auditory means of interaction for machines expressing emotions and intentions consisting of vocalizations and sounds without semantic content [1]. Semantic-free utterances have four typical types:

Gibberish Speech (GS), Non-Linguistic Utterances (NLUs), Musical Utterances (Mus), and Paralinguistic Utterances (Pus 25

). Mus and NLUs are the types we focused on in the current study [1]. 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 utterances use simple intonation such as "Argh" or "uh-huh" as a carrier to convey a semantic meaning [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** 38

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 also 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 Study 1 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** 54

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 can have more effects on users' emotion recognition and social perceptions towards the humanoid robot between non-lexical vocalizations (e.g., "argh") and non-linguistic sounds (e.g., harsh sound effects)? • RQ 3. When the vocal sound is exaggerated with robotic and metallic sound effects ("characterized sounds" in our experiment), does it improve users' social perceptions? or degenerate their 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** 49

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** college students. **Our** 4

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 (regular voice for non-lexical vocalization (human utterances), characterized voice (robotic utterances) for non-lexical vocalization, musical sound, and no sound) combined with

five basic emotions (anger, fear, happiness, sadness, and surprise 13

). The 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** 51

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. 2.5 The Study 2 and Research Questions The study 1 above shows that the regular voice appeared to have higher impacts on emotion and social perceptions towards the humanoid robots than the other sounds. Even though the experimental protocol is promising and provides interesting implications, Study1 also has limitations. Only male voices were used in emotion expressions in study 2. The results might be different depending on the gender of the voice. The recorded videos of Pepper presented to participants might also have influenced the results. Emotions from different dimensions of the circumplex model were investigated in study 2 instead of basic emotions. A

systematic study of the relationship between arousal and valence emotion dimensions was conducted. In study 1, the result of non-linguistic sounds (musical sounds) might not increase positive perceptions towards a humanoid robot or even harm social perceptions. Participants felt unnatural that a humanoid robot generated a musical piece. Therefore, in Study 2, musical sounds were removed. Gender is a cue that might influence the emotion perception. Furthermore, limit the confusion among participants. We aim to investigate the effect of gender-specific non-lexical vocalizations and gestures only of the robot on the emotion recognition and perception of the humanoid robot. New research questions have been developed for Study 2. • RQ 1. How will different emotions influence users' recognition accuracy of the humanoid robot? • RQ 2. Can adding non-lexical sounds improve users' emotion recognition and social perceptions towards the humanoid robot? • RQ 3. How will the gender of the robot voice have an impact on users' emotion recognition and social perceptions towards the humanoid robot? • RQ 4. How will participants' gender have an impact on their emotion recognition and social perceptions towards the humanoid robot? 3. Study 1 Study 1 attempted to convey the robot's emotions using the system's non-verbal voice. Study 1 also investigated whether musical sounds affected the user's emotional perception of the robot and compared it with non-lexical speech sounds. To this end, Study 1 was designed to have a humanoid robot generate four different sound conditions (regular voice, characterized voice, musical sound, and no sound) to examine the impact of robots' non-lexical vocalizations and non-verbal sounds (music sounds in this study) on the user's emotional recognition accuracy and perception of humanoid robots.

3.1 METHOD 3.1.1 Participants Forty **university** undergraduate and graduate **students (age** 37

range 19-27) participated in the experiment. Participants were compensated with $10-$15 ($10 per hour). Twenty-one

participants identified themselves as male and the other nineteen **participants identified as female** 1

. The experiment took at most 1.5 hours.

All participants agreed to participate after reviewing the consent form approved by the Institutional Review Board (IRB VT 4

). 3.1.2 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. 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 2

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). Each sound has been used in this study and four example videos

are provided on the web for other researchers and educators to 4

have a better understanding: https://osf.io/8rhs4/. The non-lexical vocalization samples, which made the regular voice condition, were from the pre-recorded and pre-validated male sounds from the study performed at the University of California, Berkeley [49]. 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 characterized utterances. The robotically characterized utterances (characterized voice) were made by applying the Flanger and Flangus synthetic/robotic effect to the same regular voice using the Fruity Loops Studio software to make more traditional robot sounds. The source sounds were edited to be 3-5 seconds long.

Flanging is a form of phase cancellation created by combining multiple, variously delayed copies of the input sound. Flangus increases **the** width of **the** 20

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 edited to 1-3 seconds long after the story paused, and unnecessary silences were removed to ensure the robot conveyed immediate feedback. 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 13

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** 4

). A pilot test was conducted for regular voice, characterized voice, and musical sounds based on Ekman's six basic emotions. Eight college students had been recruited and only the sounds with above 80 percent accuracy have been selected and used. 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 Sounds Emotion Regular Voice Characterized Voice Musical Sound No sound (Gestures-only) Happiness "Woohoo" "58% Elation + 25% Triumph + 8% Ecstasy + 8% Embarrassment" [49] "Woohoo" Delightful and upbeat sound made with multiple instruments [52] Fast, positive gesture with elation and open arms. Sadness Crying "75% Sadness + 17% Distress + 8% disgust" [49] Crying Low-pitched, descending melodies "Piano minor chords" [54] Slow, negative gesture. Body dropped and shrunk. Shoulders bowed.

Hands kept lower than their normal positions, hands closed or moving slowly. The 13

face covered with two hands. Fear Scream "67 Fear + 17% Pain + 8% Ecstasy + 8% Surprise(positive)" [49] Scream Spooky, high-pitch sound "Tremolo string sound" [54] Fast gesture with shock and unexpectedness. Surprise Short Gasp "75% Surprise(negative) + 17% Fear + 8% Realization" [49] Short Gasp Atonal, harmonic sound. "Ascending fuzzy keyboard (Orchestration bang)" [37] Hands on hips, largely recognized as an angry gesture. Anger "Argh" "83% Anger + 8% Disgust + 8% Amusement" [49] "Argh" Short, noisy, electrical sound "Distorted percussive guitar chords" [53] Arms rigged/tensed/clenched and out to side while robot looks around apprehensively and moves body backwards. 3.1.3

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** 1

a set of corresponding gestures. A single participant participated in each session.

After the consent form procedure, each participant interacted with 4

20 trials. Participants were given video clips that have a TTS- generated story as a background sound. The Pepper robot in the video listened to the story with the participant. Note that the TTS voice without any embodiment read a story, not Pepper. Throughout one story, Pepper had 5 gestural responses with sounds. After each emotion expression has been 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** 30

and each sound type condition was also counterbalanced. 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** Figure **3: An example of** the **story the participant** listened **The** 3

participants

were asked to fill out **the questionnaire after** watching **each** gesture **and** 48

sound reaction generated from Pepper [MH5] , the questionnaire provided five times for each story. After each response from the robot, the questionnaire [45] was administered

for measuring the accuracy of emotion perception and trust (Warmth, Honesty, Trustworthiness), naturalness (Natural, Robot-like), and preference (Likability, Attractiveness) of the robot. The questionnaire consisted of free response questions , 7- point Likert scale questions, and single-choice questions . Seven relevant questions were asked, and each question was rated on a 7-point Likert scale (1: lowest, 7: highest 2

). This questionnaire was used in the previous robot voice study [45]. Table 3:

The list of questions and types in questionnaires Category Question (Type) Post-comment questionnaire 1. What emotion do you feel the robot expressed 3

? (Choose from Anger, Fear, Happiness, Sadness, Surprise, and other) 2.

How clearly did the robot express this emotion? (1-7 Likert scale 4

) 3. Any other thoughts on the voice? (Optional, Open question)

Post-condition questionnaire 1. How likable is the voice? (1-7 Likert scale) 2. How attractive is the voice? (1-7 Likert scale) 3. How warm is the voice? (1-7 Likert scale) 4. How honest is the voice? (1-7 Likert scale) 5. How trustworthy is the voice? (1-7 Likert scale) 6. How natural does the voice sound? (1-7 Likert scale) 7. How robotic does the voice sound? (1-7 Likert scale 3

) 3.2. Results 3.2.1 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; $X2(4, 39) = 62.35$, $p <.01$, $W = .40$) and among the four sound types (Figure 4b; $X2(3, 39) = 9.23$, $p =.03$, $W = .08$). 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 regular voice condition and least accurately in the no sound condition. The pairwise comparisons of emotion recognition accuracy were analyzed with Wilcoxon Signed Rank Test. The emotion recognition accuracy in the characterized voice

condition (M = 0 .56, **SD = 0** .50) **was significantly** higher **than in the** no sound **condition (M = 0** .38, **SD = 0** 11

.48; Z(39) = 419.50, p =.01). The regular voice

condition (M = 0 .56, **SD = 0** .50) **was** also **significantly** higher **than the** no sound **condition** 11

(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** 4

< .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** 10

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** 2

. Figure 5 shows

clarity over emotions and sound types. There were statistically significant differences in clarity 1

among the five emotions (Figure 5a; F(4,36) = 4.95, p < .01, ηp2 = .35) and among the four sound types (Figure 5b; F(3,27) = 22.95, p <.01, ηp2 = .72) (see Table 5 for detailed statistics). No

significant difference was found in the interaction between emotions and 6

sound types. The average rating score of clarity was perceived the highest by participants in the regular voice condition. The clarity score of the regular 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 5

.70) conditions. In addition, the average rating score of clarity

was significantly higher in the characterized voices (M = 5.29, SD = 1 .70) than in the 33

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 < 0 .0083, error bars represent standard errors 6

). 3.2.2 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** 18

[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 10

warmth

results were analyzed with a 5 (Emotions) **x 4** (Sound Types) **Repeated Measures ANOVA** 2

. Figure 6 shows warmth rating scores

over emotions and sound **types. There were** statistically **significant differences in** 1

warmth among the five emotions (Figure 6a; F(4,136) =16.13, p <.01, ηp2 = .32) and among the four sound types (Figure 6b; F(3,102) = 13.20, p <.01, ηp2 = .28). No significance was found in the interaction between emotions and sound types. The average rating score of warmth was perceived significantly

higher in the regular voice (**M = 4** .14, **SD = 1** .61) and **the no** sound (**M = 4** .18, **SD = 1** 26

.59) conditions than the characterized voice (

M = 3 .48, **SD = 1** .63) **and** musical **sound (M = 3** .32, **SD = 1** 5

.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 < 0** .0083, **error bars represent standard errors** 6

). Figure 7 shows honesty rating scores over sound types. There were statistically significant differences in warmth among the four sound types (Figure 7a; F(3,102) = 16.57, p <.01, ηp2 = .24) and among the four sound types (Figure 7b; F(4,136) = 3.52, p <.01, ηp2 = .09). No

significant difference was found for **the interaction** effects **between emotions and** 6

sound types. Participants rated sounds in the characterized voice (

M = 4 .45, **SD = 1** .54), no **sound (M = 4** .52, **SD = 1** .53), **and** regular voice (**M = 4** .94, **SD = 1** 5

.50) conditions significantly higher in honesty

than in the musical sound **condition (M = 3** .89, **SD = 1** .44). In addition, **the** 7

average rating score of honesty

was perceived significantly higher in the regular voice **condition than in the** 12

characterized 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** 1

trustworthiness among the five emotions (Figure 8a; F(4,136) =4.13, p <.01, ηp2 = .11) and among the four sound types (Figure 8b; F(3,102) = 9.38, p <.01, ηp2 = .22). No

significant difference was found in the interaction between emotions and 6

sound types. Participants rated sounds in the no sound (M = 4.46, SD = 1.54) and regular 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** 5

average rating score of trustworthiness

was perceived significantly higher in the regular voice **condition than the** 12

characterized voice and 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 < 0** .0083, **error bars represent standard errors** 6

). 3.2.3 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 10

naturalness

results were analyzed with a 5 (Emotions) **x 4** (Sound Types) **Repeated Measures ANOVA** 2

. Figure 9 shows naturalness rating scores over sound types (Figure 9a; F(3,99) = 30.00, p <.01, ηp2 = .48) and among the four sound types (Figure 9b; F(4,132) = 3.52, p <.01, ηp2 = .10). No

significant difference was found for **the interaction** effects **between emotions and** 6

sound types. Participants rated sounds in the no sound (

M = 4 .39, **SD = 1** .66) **and** regular voice (**M = 5** .03, **SD = 1** .59) **conditions** 22

significantly higher

than in the musical sound **condition (M = 3** .26, **SD = 1** .63). **The** average rating score **of** 7

naturalness was also perceived

significantly higher by participants **in the** regular **condition than in the** 12

characterized 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** characterized voice **condition** 12

. (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, ηp2 = .45). No

significant difference was found either for **the main effect of** emotions or for **the** 15

interaction effects between emotions and sound types. Participants rated sounds in no sound (M = 4.30, SD = 1.78) and regular voice (

M = 3 .13, **SD = 1** .63) conditions **significantly** lower **than** in **the** musical **sound condition (M** = 5.10, **SD = 1** .58). **The** 5

characterized voice (

M = 4 .71, **SD = 1** .70) **was** also **significantly** lower **than the** musical **sound** 5

. The average rating score of robot- likeness was perceived significantly lower

in the regular **voice condition than in the** characterized **voice** and no sound **conditions** 53

. Moreover, the average rating score of robot- likeness was perceived marginally lower in the no sound condition

than the characterized voice condition. **Figure** 10. **The rating scores of** 2

robot-likeness over emotions and sound conditions (*: p < .0083, error bars represent standard errors). 3.2.4 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 10

preferences rating

results were analyzed with a 5 (Emotions) **x 4** (Sound Types) **Repeated Measures ANOVA** 2

. Figure 11 shows likability rating scores

over emotions and sound **types. There were** statistically **significant differences in** 1

likability among the five emotions (Figure 11a; F(4,136) = 11.46, p < .01, ηp2 = .23) and among the four sound types (Figure 11b; F(3,102) = 19.84, p < .01, ηp2 = .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** regular voice **condition (M = 4** .73, 5

SD = 1

.71) had a significantly higher likeability rating score than both the characterized 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** 1

attractiveness among the five emotions (Figure 12a; F(4,136) = 10.20, p <.01, ηp2 = .23) and among the four sound types (Figure 12b; F(3,102) = 13.37, p <.01, ηp2 = .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** regular voice **condition (M = 4** .62, **SD = 1** 5

.73) had a significantly higher likeability rating score than both the characterized 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). 3.3 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 regular voice appeared to have higher impacts on emotion and social perceptions towards the humanoid robots than the other sounds. 3.3.1 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 adding musical sounds to a humanoid robot could diminish clarity even compared to the no sound (i.e., gesture- only). The results of the emotion recognition accuracy also showed

significant differences in emotions , sound types, **and the interaction between emotions and** 21

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 characterized voice, musical voice, and regular voice which is also consistent with a prior study that sadness conveyed the best performance through sound [59]. A prior study proposed the idea that 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 1

fairytales [44]. Negative emotions with a high arousal level include anger, and fear. It's possible that the individuals didn't anticipate these strong, unpleasant feelings from the fairy stories. However, it cannot be validated from the present study and thus, it requires further research. Table 13 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** 52

[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** 2

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 characterized voice, musical sound, and regular voice conditions were all significantly higher than the no sound condition, which makes sense because multimodal cues can enhance the emotion recognition performances [21]. Participants showed higher clarity

scores in the regular **voice** condition **than** the **other three** sound **types** significantly. **The** characterized **voice** also received **significantly higher** clarity **scores than the** 2

musical and no sound conditions from participants.

These results suggest that a regular or **characterized voice might be more appropriate** 4

than the musical sound in expressing emotions accurately during HRI. 3.3.2 Trust, Naturalness, and Preferences We can infer a clear trend in all three categories of social perceptions - trust, naturalness, and preference ratings. Regular voice was the highest in all measures, whereas musical sound was the lowest in all measures. In terms of the trust scale, regular voice and no sound showed relatively higher ratings in all three ratings, including warmth, honesty, and trustworthiness. Musical sound was the lowest. People can interpret the human voice from the robot as conveying emotions and also are highly attuned to non-speech clues, so both voice conditions might be considered “credible” by our participants [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 may imply that musical sound does not fit a humanoid robot to express its emotional states. On the other hand, the no sound condition showed higher naturalness than the characterized voice. Therefore, we can infer that a robotically characterized voice is not anymore what people expect to hear from a humanoid robot in this era. They seem to be already accustomed to more human-like voices from smart speakers (e.g., Alexa) and smartphones (e.g., Siri). In the same line, the musical sound showed the highest robot-likeness with the lowest naturalness. In the last perceptions of likability and attractiveness, participants showed the highest rating scores for regular voice, followed by no sound. The characterized voice and musical sound were not what people expected to hear from a humanoid robot. In conclusion, people seem to prefer the human-like regular voice with humanoid robots [66] and rather prefer no sound over characterized 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. 3.3.3 Revisiting RQs and Extracting

Design Guidelines Based on the results of **the present study, we** can provide design **guidelines for** emotional cues **of** 6

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 regular voice, robotically characterized voice, or musical sounds (RQ1). · Using a regular voice will significantly increase emotion recognition accuracy than other sounds (RQ2). · Using a regular 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 characterized voice for a humanoid robot does not seem to be effective in increasing social perceptions, compared to a regular 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 characterized 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). 3.4 Limitations and Future Work Even though the experimental protocol is promising and provides interesting implications, this study also 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** 1

characterized voice can be designed in different ways (e.g., FM synthesis or vocoding). Musical sounds for the same emotion can also vary in unlimited ways. Therefore, the generalizability of this study's results can be limited. In the future, alternative methods can be adopted using a free software program and compared even in one sound 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. 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** 1 **the results**

. 4. Study 2 Study 2 systematically investigated the influence of emotional non-lexical sounds of different genders, which helped to understand robotic emotion sonification research and design. Study 2 further investigated the effects of how people perceive robot emotions. Also, instead of using the recorded videos of Pepper, the participants experienced the physical robot in person in Study 2.

4.1. METHOD 4.1.1 Participants Forty-nine university students 32

(age: Mean = 21.49, SD = 2.9) participated in the experiment. Twenty-

six participants identified themselves as male, and the other twenty-three **participants identified as female** . Twenty-four **participants** 1

experienced male voice group; twenty-five participants experienced a female voice group. The experiment took at most 1

hour. All participants agreed to participate after reviewing the consent form approved by the VT **Institutional Review Board (IRB** 4

). 4.1.2 Equipment and Stimuli A humanoid robot, Pepper, same as Experiment 1 has been employed in Experiment 2 (Figure 1). According to the storyline, the robot expressed a recorded sound with gestural feedback which provides emotional cues

to the participants. **Two different stories** ("The three little pigs" **and** "The boy who cried wolf") **used in Experiment** 2. **Two** 2

sound types were created for four emotional expressions. The female and male non- lexical vocalization samples were from the pre-recorded and pre-validated sounds from the study performed at the University of California, Berkeley [34]. Only one gender's voice was provided to one vocalization sample form University of California, Berkeley. Therefore, a male and female graduate students reproduced the same sounds based on chosen vocalization samples. The gestures of the robot in five emotional expressions were created by using Choregraph, a software program coming with the Pepper robot. Gestures were made based on the previous research and Experiment 1 as shown in Table 1. The same storytelling voice and setting has been used in Experiment 2 as Experiment 1. Each story presented four different emotions from the circumplex model [38] (

high arousal positive valence emotion: happiness, **high arousal negative valence** emotion: **anger,** 9
low arousal positive valence emotion: **relief, and low arousal negative valence** emotion: **sadness**

). A pilot test conducted for male and female voices and musical sounds, and the only sounds with above 85 percent accuracy have been selected for use in the experiment. The four emotions (

high arousal positive valence emotion: happiness, **high arousal negative valence** emotion: **anger,** 9
low arousal positive valence emotion: **relief, and low arousal negative valence** emotion: **sadness**

) were fit into two stories each (“The three little pigs”, and “The boy who cried wolf”). Table 14: “Pepper” Robot Sounds Circumplex Model Emotion Male Voice Female Voice

High arousal positive valence Happiness “Hohohoo” “Hohohoo” **High arousal negative valence Anger** 9
“Argh” “Argh” **Low arousal positive valence Relief** “Ehaa” “Ehaa” **Low arousal negative valence**
Sadness

Crying Crying 4.1.3 Experimental Design and Procedure A 4 (emotions) x 2 (sound conditions) Mixed design applied to this experiment. emotions and sound conditions were a within-subjects variable. In addition, only in the Voice + Gesture condition, the robot voice gender was used as a between-subjects variable. Order of the fairy tales has been counterbalanced across participants and the order of Voice + Gesture or Gesture-only conditions were also counterbalanced. A single participant participated in each session. After the consent form procedure, participants were welcomed and provided information about Pepper robot. A demographic questionnaire provided to participants to collect age, gender and region. A TTS-generated story played from a laptop and the Pepper robot listened to the story with the participant. Throughout one story, Pepper had 4 gestural responses with voice + gesture condition or gesture only condition. After each emotion expression has been presented, participants were asked about perceived emotions of the robot through an online questionnaire. After each story has been presented, participants were asked about perceived social perception questionnaires (Social Characteristics, RoSAS and Godspeed) of the robot through an online questionnaire. The order of the three stories has been counterbalanced across participants. The mapping between the story and each sound type condition has also been counterbalanced. The presentation order of emotions in each story has not been counterbalanced because of the storyline. The whole sample procedure is depicted in Figure 13. Figure 13: The flow diagram of the example procedure

The participants asked to fill out the **questionnaires after** watching **each** 1

gesture and sound reaction generated from the robot Pepper, and the questionnaire has been provided four times for each story. After each response from the robot, the questionnaires have been administered for measuring the accuracy of emotion perception. Accuracy questionnaire has single-choice questions about emotions and free responses that participants can type in emotion they perceived, which was not in options. Social Characteristics, RoSAS and Godspeed questionnaires used to measure Likability, Attractiveness, Warmth, Honesty, Trustworthiness, Naturalness, Robot-likeness, Competence, Warmth, Discomfort,

Anthropomorphism, Animacy, Likeability, Perceived intelligence , Relaxed, Calm **and** Surprised **of the robot** Pepper. **The** 41

Social Characteristics questionnaire has 1-7 point Likert scale questions. Relevant questions were asked in Godspeed and RoSAS questionnaires, and each question

was rated on a 5-point Likert scale (1: lowest, 5 : highest). **The** 29

selected Godspeed and RoSAS questionnaires are widely used

in Human Factors **and Human-Robot Interaction research** to measure impressions **of** the **robots** 43

and agents. Table 15:

The list of questions and types in questionnaires Category **Question (Type) Post-comment questionnaire 1. What emotion do you feel the robot expressed** 3

? (Choose from Anger, Happiness, Sadness, Relief, and other) 2. Any other thoughts on the voice? (Optional, Open question) 4.2 RESULTS 4.2.1 Emotion Perception: Accuracy 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). The results were analyzed with a 4 (Emotions) x 2 (Sound Conditions) repeated measures ANOVA, testing main and interaction effects. Figure 14 shows accuracy over emotions and sound types and interaction between emotions and sound types. A conservative alpha level (α: 0.05/6 = p < 0.0083) was applied for emotion accuracy recognition. There is a main effect of emotion (F(2.401, 115.269) = 38.060, p <.0001,$\eta 2_p$ = 0.442)There were interaction effect between emotions and voice

conditions (F(2.715, 127.587) = 39.777, p <.0001,$\eta 2p$ = 0.457). There were statistically significant differences among the four emotions. With all voice conditions happiness was highly accurately recognized in 96.93% accuracy, followed by sadness, anger, and relief with percentages of 90.81%, 68.37%, and 56.12% each. Happiness and sadness accuracy were significantly higher than anger and relief. The pairwise comparisons of emotion recognition accuracy were conducted with voice condition with paired samples t-tests with the Bonferroni correction. A conservative alpha level (α: 0.05/4 = p < 0.0125) was applied for voice condition focused on the relationship between voice condition and emotion recognition. The emotion recognition accuracy in the voice + Gesture condition was significantly higher than gesture only condition in anger (t(48) = 6.516, p < .001), relief (t(48) = 13.682, p < .001) and sadness (t(48) = 2.449, p = .009).

There were no statistically significant differences in **main** effect (**F(1** , 47) = **0** .003, **p** 7

= .957, $\eta 2p$ = 0.000) and interaction effect (F(2.487, 116.905) = 0.36, p = 0.744, $\eta 2p$ = 0.008) among emotion recognition accuracy and voice gender.

There were no statistically significant differences in **main** effect (**F(1** , 47) = **0** .003, **p** 7

= .957, $\eta 2p$ = 0.000) and interaction effect (F(2.552, 244.947) = 1.178, p = 0.316, $\eta 2p$ = 0.01) among emotion recognition accuracy and participant gender. (a) emotion recognition accuracy over emotions (α: 0.05/6 = p < 0.0083) (b) emotion recognition accuracy over voice condition (α: 0.05/4 = p < 0.0125) (c) emotion recognition accuracy over voice gender (α: 0.05/4 = p < 0.0125) (d) emotion recognition accuracy over participant gender (α: 0.05/4 = p < 0.0125) Figure 14:

Accuracy of perceiving emotions over emotions , sound types, **and** emotions by sound **types (*:** 4
p

< .05, error bars represent standard errors). 4.2.2 Social Perception: Social Characteristics For social perceptions, we focused on analyzing the results based on sound condition, voice gender and participant gender because we were interested in investigating these effects on social perception towards the humanoid robot. The social characteristics of robot reaction were rated by the participants based

on a 7-point Likert scale (1: lowest, 7: highest). The 10

social characteristics results of sound conditions were analyzed with paired samples t-tests. Voice + Gesture condition showed significantly higher likability (t(48) = 2.809, p = .004), warmth (t(48) = 2.005, p = .025), honesty (t(48) = 2.396, p =

.01), trustworthiness (t(48) = 2.361, p = .011) and naturalness (t(48) =3.203, p = .001) in than Gesture-Only condition. Voice + Gesture condition showed significantly lower Robot Likeness (t(48) = -2.896, p = .003) than Gesture-Only condition. There were no significant differences between robot voice gender in social characteristics from

independent samples t-tests. There were also **no significant differences between** 39

participants' gender in social characteristics from independent samples t-tests. (a) Rating score over voice condition in Social Characteristics (p < 0.05) (b) Rating score over voice gender in Social Characteristics (p < 0.05) (c) Rating score over participant gender in Social Characteristics (p < 0.05) Figure 15: The rating scores of social characteristics over voice condition, voice gender and participant gender (error bars represent standard errors). 4.2.3 Social Perception: RoSAS We continued our investigation about social perceptions towards the humanoid robot focused on analyzing the results based on sound condition, voice gender and participant gender through the RoSAS questionnaire. The RoSAS results of robot reactions were rated by the participants

based on a 5-point Likert scale (1: lowest, 5: highest). The 17

RoSAS rating results of sound conditions were analyzed with paired samples t-tests. Voice + Gesture condition showed significantly higher competence (t(48) = 3.931, p < 0.001) and warmth (t(48) = 6.283, p < 0.001) than Gesture-Only condition. The RoSAS rating results of robot voice gender and participants' gender were analyzed with independent samples t-tests respectively. The average rating score of discomfort (t(47) = 1.56, p = .007) was perceived significantly higher in male voice than female voice. The average rating score of warmth (t(47) =-0.574, p = .008) was perceived significantly higher in female participants than male participants. The average rating score of discomfort (t(47) = 1.198, p = .002) was perceived significantly higher in male participants than female participants. (a) Rating score over sound condition in RoSAS (p < 0.05) (b) Rating score over voice gender in RoSAS (p < 0.05) (c) Rating score over participant gender in RoSAS (p < 0.05) Figure 16: The rating scores of RoSAS over sound condition, voice gender and participant gender (error bars represent standard errors). 4.2.4 Social Perception: Godspeed We continued our investigation about social perceptions towards the humanoid robot focused on analyzing the results based on sound condition, voice gender and participant gender through the Godspeed questionnaire. The Godspeed results of robot reactions were rated by the participants

based on a 5-point Likert scale (1: lowest, 5: highest). The 17

Godspeed rating results of sound conditions were analyzed with pirated samples t-tests. Voice + Gesture condition showed significantly higher average rating score in anthropomorphism (t(48) = 2.146, p = .081), animacy (t(48) = 2.146, p < .001), likeability (t(48) = 2.254, p = .014), and surprised (t(48) = 2.574, p = .007) than Gesture-Only condition. The Godspeed rating

results of voice gender and participants' gender were analyzed with independent samples t-tests. The average rating score of anthropomorphism (t(47) = 1.378, p = .022) was perceived significantly higher in male voice than female voice. There were no significant differences between participants' gender and social perceptions in Godspeed. (1) Rating score over sound condition in Godspeed (p < 0.05) (2) Rating score over participant gender in Godspeed (p < 0.05) (3) Rating score over participant gender in Godspeed (p < 0.05) Figure 17: The rating scores of Godspeed over sound condition, voice gender and participant gender (error bars represent standard errors). 4.3. DISCUSSION A systematic study was conducted to obtain a comprehensive view of the effects of emotion types, non-lexical sounds, robot voice gender, and participants' gender on users' perceptions of robot emotions and other social perceptions. Overall, results showed that the Voice + Gesture condition appeared to have higher impacts on emotion recognition and social perception towards the humanoid robots than the Gesture-Only condition. 4.3.1 Emotion Perception: Accuracy The results of the emotion recognition accuracy showed

significant differences in emotions , voice conditions, **and the interaction between emotions and** 21

voice conditions. Emotion recognition in anger, relief and sadness were significantly higher in Voice + Gesture condition than Gesture-Only condition. This result is consistent with the prior study that when emotions were presented unimodally, participants were less confident and accurate in categorizing expressions [71]. Multimodal interactions increase the information provided to compensate for the ambiguity and deficiencies of unimodal [71] and thus, increase the accuracy of emotion perception. There was no significant difference between voice conditions in happiness because Gesture-Only condition also had high emotion recognition. At the same time, the gesture- language multimodal system is the best pairing compared to the face-gesture and face-speech systems [72]. This result is consistent with previous research that happiness can best express through color and movement [59]. Happiness and sadness recognition accuracy were significantly higher than anger and relief. In summary, we can cautiously infer that happiness and sadness can be easily recognized by not only non-lexical Voice + Gesture but also Gesture only. This corroborates a previous study on robotic speech [44], which found higher recognition accuracy for happiness and sadness, but lower recognition accuracy for anger. Table 15 shows that relief was mostly misclassified as happiness (19.39%). A previous study showed the head position facing upward increased the correct recognition of some emotional expressions (pride, happiness and excitement), while the head position facing downward increased the correct recognition of other emotional expressions (anger, sadness) [73]; that head-up was always evaluated as more arousal than head-straight or head-down. [73] The head position of relief (high valence, low arousal) gesture design was a little upward and thus, could lead to misclassification as happiness (high valence, high arousal). There were no statistically significant differences in main effect and interaction effect among emotion recognition accuracy and robot voice or participant gender.

These results suggest that voice + gesture **might be more appropriate** 4

than gesture-only in expressing emotions accurately in HRI. Table 15.

The confusion matrix between presented and perceived emotions Perceived **Presented Anger** **Happiness** Relief **Sadness Count** 67 1 3 **0** 3

Anger Col % 68.37% 1.02% 3.06% 0% Count 3 95 19 2 Happiness Col % 3.06% 96.93% 19.39% 2.04% Count 4 0 55 0 Relief Col % 4.08% 0% 56.12% 0% Count 2 0 5 89 Sadness Col % 2.04% 0% 5.10% 90.81% Count 22 2 16 7 Other Col % 22.45% 2.04% 16.33% 7.14% 4.3.2 Social Perception: sound condition, voice gender and participant gender We can infer a clear trend of social perception in three categories - sound condition, robot voice gender, and participant gender. Voice + Gesture condition conveyed higher scores of positive social perceptions than Gesture-Only condition in likability, warmth, honesty, trustworthiness, naturalness, competence, anthropomorphism, animacy, likeability, and surprise. This makes sense because multimodal cues can enhance and complement information [24] while potentially reinforcing positive social perceptions. The lack of information may have a negative impact on social perception which leads to a significantly higher robot-likeness rating score in Gesture-Only condition. Any hint of gender, no matter how small, can trigger stereotypes about that gender [75].

Men are perceived to possess more instrumental attributes (i.e., self -directed, **goal-oriented** **characteristics** 24

) [76] and to be more dominant and influential than women [77]. Stereotypes about the gender of voices extend to robot interactions and impact social perceptions. This might explain that our participants felt more discomfort and anthropomorphism in male voices because male voices may have a more dominant and aggressive orientation. On the other hand, the previous study suggests that females form significantly more rapport and are more persistent in their interactions with robotic partners than males [74]. In our study, female participants felt more warmth interacting with Pepper, while male participants were more likely to feel uncomfortable. 3.3.3 Revisiting RQs and Extracting

Design Guidelines Based on the results of **the present study, we** can provide design **guidelines** **for** emotional cues **of** 6

a humanoid robot as follows. Of course, depending on users, robot types, tasks, and situations, the guidelines may vary. • Happiness and sadness show significantly higher recognition accuracy than relief and anger of the humanoid robot (RQ1). • Adding non-linguistic sounds to a humanoid robot's emotional gesture can increase emotion recognition accuracy regardless of the type of voice gender and participant gender (RQ2). • Adding non-linguistic sounds will significantly increase positive social perception than gesture-only (RQ2). • Adding non-linguistic sounds will significantly decrease negative social perception than gesture-only (RQ2). • Male robot voice significantly increased discomfort and anthropomorphism than female voice in social perceptions (RQ3). • Female participants were feeling significantly more warmth towards the

humanoid robot than male participants (RQ4). • Male participants were feeling significantly more discomfort towards the humanoid robot than female participants (RQ4). • Robot gestures without any non-speech sounds can also convey a certain emotional state (e.g., happiness with 96.93% accuracy, sadness with 83.67% accuracy). 4.4 Limitations and Future Work Although the experiment provides interesting implications, this study also has limitations. First, only a humanoid robot was used in this experiment. Results may vary depending on the appearance and form factors of the robot (e.g., animal or mechanical robot). Second, limited gesture design approaches were used. For example, only one set of gestures were used for male and female robot voices as emotion expressions. If we map specific gender gestures related to robot voice gender, the results might be different. Next, based on Pepper's appearance, participants may expect the child's voice, and we can explore how the child's voice affects our results in future work. Lastly, pepper's eye lights can display different colors. Since some emotions can be perceived better through color [59], we can explore the impact of color cues on emotion perception and social perception. In future work, we will recruit more participants with different cultural backgrounds to generalize the findings. 5. CONCLUSION

An exploratory study was conducted to comprehensively **understand the** effects **of** 50

emotion types, nonverbal sounds, and gestures on users' perception of robot emotions and other characteristics. The accuracy of emotion recognition for non-speech of regular voice was significantly higher than that of robotically characterized voice or musical sounds in all sound conditions. Inappropriate robotic characterization of voice or musical sound phases does not improve emotion recognition's accuracy and social perception. It sometimes decreases the accuracy and social perception of emotion recognition. Regardless of the conditions, some emotions, such as happiness and sadness, were always significantly more recognized than others.

There was no statistically **significant difference between the** accuracy rate of **emotion recognition** and the gender **of the** 8

robot's voice or the gender of the participants. The regular sound was the highest in the trust, nature, and preference rating measures, while the musical sound was the lowest in all measures. The overall performance of characteristic sounds and musical sounds is not as good as regular sounds and gestures only; people may not expect to hear characteristic sounds and musical sounds from humanoid robots. Participants felt more discomfort and anthropomorphism in male voices compared to female voices. Male participants were more likely to feel uncomfortable from interactions with Pepper, while female participants were more likely to feel warmth. This finding has instructive suggestions for future robot design about user gender and robot voices. In-person interaction in study 2 helped participants focus more into the experiment and reduce the confusion participants may feel from background story and robot voices compared with study 1. 6. DESIGN GUIDELINES Based on the results of the present two studies, we can provide design guidelines for emotional cues of a humanoid robot as follows: • Adding non-speech sounds to a humanoid robot’s emotional gesture can increase emotion

recognition accuracy and social perceptions regardless of the type of sounds, including regular voice, robotically characterized voice, or musical sounds; or gender of sounds in regular voice. • Using a regular voice will significantly increase emotion recognition accuracy and social perceptions than robotically characterized voice, or musical sounds. • Using musical sounds and robotically characterized voices for a humanoid robot does not seem to be effective in increasing social perceptions, compared to a regular voice. • Depending on the emotion type, different sounds may enhance the perception accuracy differently. • Male robot voices significantly increase discomfort and anthropomorphism than female voices in social perceptions. • Female participants were feeling significantly more warmth towards the humanoid robot than male participants. • Male participants were feeling significantly more discomfort towards the humanoid robot than female participants. • Robot gestures without any non-speech sounds can also convey a certain emotional state. • In person robot interactions will improve the accuracy of emotion recognition, compared to recorded videos. (Same gestures in both studies, e.g., happiness with 70% accuracy in study 1, happiness with 96.93% accuracy in study 2.) 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]: Paul B. and De G., "The psychology of mul-timodal perception", Crossmodal space and crossmodal attention, pp. 141-177, 2004. [72]: Kessous, L., Castellano, G., & Caridakis, G. (2010). Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. Journal on Multimodal User Interfaces, 3(1-2), 33–48. https://doi.org/10.1007/s12193-009-0025-5 [73]: A. Beck, L. Cañamero and K. A. Bard, "Towards an Affect Space for robots to display emotional body language," 19th International Symposium in Robot and Human Interactive Communication, Viareggio, Italy, 2010, pp. 464-469, doi: 10.1109/ROMAN.2010.5598649. [74]: Nichola L., Erin W., and Heather P., 2016. Effects of Voice-Adaptation and Social Dialogue on Perceptions of a Robotic Learning Companion. In The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI '16). IEEE Press, 255–262. [75]: Nass C, Moon Y, Green N. Are machines gender neutral? gender-stereotypic responses to computers with voices. Journal of applied social psychology. 1997;27(10):864-876. doi:10.1111/j.1559-1816.1997.tb00275.x [76]: Rosenkrantz, P. S., Vogel, S. R., Bee, H., Broverman, I. K., & Broverman, D. M. (1968). Sex role stereotypes and self concepts in college students. Journal of Consulting and Clinical Psychology, 32, 287-295. [77]: Eagly, A. H., & Wood, W. (1982). Inferred sex differences in status as a determinant of gender stereotypes about social influence. Journal of Personality and Social Psychology, 43, 915-928.

APPENDIX A Study 1 Story:

THE THREE LITTLE PIGS Study 1 Story: **THE BOY WHO CRIED WOLF** 4

Study 1 Story: Little Red Riding Hood Study 1 Story:Beauty and the Beast Study 2 Story: THE THREE LITTLE PIGS THE THREE LITTLE PIGS RESEARCHER VERSION 12/25/2022 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. The frightened little pig quickly ran to the second pig's home, made

of sticks. The big bad wolf ran over **and** again huffed **and** puffed and **blew the house down** 28

in hardly any time at all. ROBOT (sad) The two little pigs were terrified and ran as quickly as they could. They pounded on the door of the third pig's home, which was made of bricks, and were let in before the wolf could catch them. The big bad wolf took a deep breath, ready to blow the house down. But no matter how he huffed and puffed, he couldn't blow the house down! He tried and tried for hours, but the home was strong, and the pigs were safe inside. ROBOT (relief) The wolf saw a chimney above the home and hatched a plan to crawl down the chimney and snatch the pigs up. The wolf crawled down the chimney and had the pigs in his sights, ready to make them his meal. However, the pigs were very clever, and they had laid a trap at the bottom of the chimney. As the wolf lunged out, he was caught in the net and the pigs were safe! The two little pigs who were lazy felt terrible for the mistakes they had made. They went back and built their homes of brick so they could live happily ever after. ROBOT (happy) Study 2 Story: THE BOY WHO CRIED WOLF Study 1: Emotion Accuracy Questionnaire Study 1: Social Perception Questionnaire Study 2: Demographic Questionnaire Study 2: Emotion Perception Questionnaire Study 2: Social Characteristic Questionnaire Study 2: RoSAS Competence Questionnaire Study 2: RoSAS Warmth Questionnaire Study 2: RoSAS Discomfort Questionnaire Study 2: Godspeed Questionnaire 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70

sources:

1 133 words / 1% - Internet from 08-Jul-2022 12:00AM

link.springer.com

2 156 words / 1% - Crossref
Sangjin ko, Jaclyn Barnes, Jiayuan Dong, Chunghyuk Park, Ayanna Howard, Myounghoon Jeon. "The effects of robot voices and appearances on users emotion recognition and subjective perception", International Journal of Humanoid Robotics, 2023

3 147 words / 1% - Crossref
"HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence", Springer Science and Business Media LLC, 2020

4 143 words / 1% - from 19-May-2023 12:00AM
digitalcommons.mtu.edu

5 119 words / 1% - Internet
Worrall, David. "Sonic Information Design: Proceedings of the 22nd Annual International Conference on Auditory Display", Georgia Institute of Technology, 2016

6 97 words / 1% - Crossref
Scott Zieger, Jiayuan Dong, Skye Taylor, Caitlyn Sanford, Myounghoon Jeon. "Happiness and high reliability develop affective trust in in-vehicle agents", Frontiers in Psychology, 2023

7 39 words / < 1% match - Internet from 27-Aug-2019 12:00AM
link.springer.com

8 11 words / < 1% match - from 30-Mar-2023 12:00AM
link.springer.com

9 59 words / < 1% match - Crossref
Hatice Gunes, Maja Pantic. "Automatic, Dimensional and Continuous Emotion Recognition", International Journal of Synthetic Emotions, 2010

10 55 words / < 1% match - Internet from 11-Dec-2022 12:00AM
digitalcommons.lsu.edu

11 50 words / < 1% match - Internet from 13-Dec-2022 12:00AM
atrium.lib.uoguelph.ca

12 40 words / < 1% match - Internet from 06-May-2019 12:00AM
tel.archives-ouvertes.fr

13 35 words / < 1% match - Internet from 19-Aug-2020 12:00AM
mafiadoc.com

14 31 words / < 1% match - Internet from 13-Jun-2014 12:00AM
www.coursehero.com

15 27 words / < 1% match - Internet from 14-Nov-2019 12:00AM
ir.soken.ac.jp

16 24 words / < 1% match - Crossref
Dayle David, Meggy Hayotte, Pierre Thérouanne, Fabienne d'Arripe-Longueville, Isabelle Milhabet. "DEVELOPMENT AND VALIDATION OF A SOCIAL ROBOT ANTHROPOMORPHISM SCALE (SRA) IN A FRENCH SAMPLE", International Journal of Human-Computer Studies, 2022

17 24 words / < 1% match - Internet from 19-Oct-2022 12:00AM
pureadmin.qub.ac.uk

18 23 words / < 1% match - Crossref
Jiaqi Ma, Brian L. Smith, Michael D. Fontaine. "Comparison of In-Vehicle Auditory Public Traffic Information With Roadside Dynamic Message Signs", Journal of Intelligent Transportation Systems, 2015

19 22 words / < 1% match - Internet from 15-Jan-2023 12:00AM
diposit.ub.edu

20 21 words / < 1% match - Internet from 21-Mar-2022 12:00AM
kupdf.net

21 20 words / < 1% match - Crossref
"Workshops at 18th International Conference on Intelligent Environments (IE2022)", IOS Press, 2022

22 19 words / < 1% match - Internet from 23-Nov-2022 12:00AM
tsukuba.repo.nii.ac.jp

23 14 words / < 1% match - ProQuest
Crumpton, Joseph John. "Use of vocal prosody to express emotions in robotic speech.", Proquest, 2015.

24 14 words / < 1% match - Internet from 09-Oct-2015 12:00AM
www.researchgate.net

25 14 words / < 1% match - Internet from 25-Mar-2020 12:00AM
www.tandfonline.com

26 13 words / < 1% match - Internet from 22-Dec-2022 12:00AM
www.acrwebsite.org

27 12 words / < 1% match - Internet from 20-Jan-2023 12:00AM
core.ac.uk

28 12 words / < 1% match - Internet
Christopher PROWANT. "My Language Passport : An Evaluation Method for Elementary and Junior High School English Classes with Instructor Feedback", 鳴門教育大学小学校英語教育センター, 2019

29 12 words / < 1% match - Internet from 30-Mar-2020 12:00AM
prism.ucalgary.ca

30 11 words / < 1% match - ProQuest
Huang, Ying. "Gestural Learning in Children with Autism", The Chinese University of Hong Kong (Hong Kong), 2020

31 11 words / < 1% match - Internet from 08-Aug-2022 12:00AM
academicworks.cuny.edu

32 11 words / < 1% match - Internet from 23-Nov-2017 12:00AM
era.library.ualberta.ca

33 11 words / < 1% match - Internet from 25-Nov-2022 12:00AM
ir.canterbury.ac.nz

34 11 words / < 1% match - Internet from 19-Nov-2022 12:00AM
openresearch-repository.anu.edu.au

35 10 words / < 1% match - ProQuest
Alnajjar, Sara Sadiq. "Identity and Attitudes in Pennsylvania, Germany, and Norway: Toward Better Policy Framing for Wind Energy", Lehigh University, 2020

36 10 words / < 1% match - from 09-May-2023 12:00AM
ediss.uni-goettingen.de

37 10 words / < 1% match - Internet from 18-Mar-2022 12:00AM
pure.mpg.de

38 10 words / < 1% match - Internet from 25-Nov-2022 12:00AM
repositorium.sdum.uminho.pt

39 10 words / < 1% match - Internet

Vela, Lori E.. "Investigating the Effect of Humor Communication Skills Training on Pro-social and Anti-social Humor Styles, Cognitive Learning, Self-efficacy, Motivation, and Humor Use", 'West Virginia University Libraries', 2013

40 10 words / < 1% match - Internet from 10-Jan-2023 12:00AM
scholarsjunction.msstate.edu

41 10 words / < 1% match - Internet from 17-Dec-2021 12:00AM
uwspace.uwaterloo.ca

42 10 words / < 1% match - Internet from 26-Sep-2022 12:00AM
vtechworks.lib.vt.edu

43 9 words / < 1% match - Crossref
"Social Robotics", Springer Science and Business Media LLC, 2017

44 9 words / < 1% match - Crossref
"Social Robotics", Springer Science and Business Media LLC, 2020

45 9 words / < 1% match - Crossref
"Social Robotics", Springer Science and Business Media LLC, 2021

46 9 words / < 1% match - Crossref
Marlena R. Fraune, Selma Šabanović, Takayuki Kanda. "Human Group Presence, Group Characteristics, and Group Norms Affect Human-Robot Interaction in Naturalistic Settings", Frontiers in Robotics and AI, 2019

47 9 words / < 1% match - Crossref
Terry Libkuman, Charles Stabler, Hajime Otani. "Arousal, valence, and memory for detail", Memory, 2004

48 9 words / < 1% match - Internet from 23-Sep-2022 12:00AM
cim.lim.di.unimi.it

49 9 words / < 1% match - Internet from 15-Sep-2022 12:00AM
dokumen.pub

50 9 words / < 1% match - Internet from 12-Sep-2018 12:00AM
propertibazar.com

51 9 words / < 1% match - Internet
Simones, Lilian. "The Roles of Gesture in Piano Teaching and Learning"

52 9 words / < 1% match - Internet from 14-Jun-2020 12:00AM
rd.springer.com

53 9 words / < 1% match - Internet from 18-Jan-2014 12:00AM
tampub.uta.fi

54 9 words / < 1% match - from 26-May-2023 12:00AM
uu.diva-portal.org

55 9 words / < 1% match - Internet from 28-Apr-2016 12:00AM
www.ncbi.nlm.nih.gov