EMOTIVE VOICE ACCEPTANCE IN HUMAN-ROBOT INTERACTION

Jaclyn Barnes, Erin Richie, Qiyue Lin, Myounghoon Jeon

Mind Music Machine Lab
Michigan Technological University
1400 Townsend Ave., Houghton, MI 49931 USA
{jaclynb, enrichie, qiyuel, mjeon}@mtu.edu

Chung Hyuk Park

Assistive Robotics and Tele-Medicine (ART-Med) Lab
The George Washington University
2121 Eye Street, NW Washington, DC 20052 USA
chpark@gwu.edu

ABSTRACT

Social robots are expected to interact with people in a natural and socially appropriate manner. Often, this will involve communicating by speech. An experiment with two social robots investigated how participants accepted different voices used by the robots and how they interpreted emotions expressed in those voices. A humanoid and a dinosaur robot used recorded human voices and synthesized voices to express seven different emotions while commenting on fairy tales read by the participants. The preliminary results of this study suggest that monotone synthesized voices are not well-suited to emotion-rich interactions, while regular speaking voices and expressive character voices both have relative strengths. We hope that this preliminary study can spark lively discussions on emotive robot voices in the ICAD community.

1. INTRODUCTION

As robots develop more prominent roles as assistants and companions, the nuances of communication between robots and the people they interact with deserve careful consideration. While body language, gestures, expression, and other social cues can convey a wealth of information in social interactions, here we focus particularly on voice and emotion. The ability of a person to understand not only what a robot is saying, but what is meant by a statement based on context, word usage, and tone of voice is critical. Emotion is ever present in human interactions and, whether intended or not, will necessarily also be present in human-robot interactions. This research investigates the capability of participants to discern emotional responses in context from a robot, their understandings of these utterances, and their opinions of vocalization the robot uses to respond to the participant as they read stories. By investigating these factors, we hope to find valuable correlations that can help future companion and assistive robots required to interact emotionally to do so in a more meaningful and easily interpretable way.

2. BACKGROUND

There have been some studies related to acceptance of the sounds used by robots, though they have not received as much attention as physical appearance has. In a study comparing participants' textual responses to a mechanical robot using a low affect human voice, a synthesized voice, and beeping sounds, participants gave significantly more commands to complete the same tasks under the synthetic voice condition than either of the other conditions. The researchers suggested that the robot with the human voice was assumed to be more capable than the synthetically voiced robot and therefore believed to need less instruction, while participants may have assumed that the robot that did not itself speak could not understand the speech of others [1]. Further supporting the idea that people anthropomorphize robots with human voices and relate to them with human expectations, a mechanical robot with a human voice was approached more closely than when the same robot was using synthetic voices [2]. Human voiced robots of the same gender as participants were anthropomorphized more than robots of the opposite gender or with a synthesized voice. Robots with a human voice were also considered more likeable [3], [4]. All of this suggests a general preference to interacting with robots similar to oneself. This is further supported by a study showing New Zealanders preferred interacting with robots that had the local accent rather than an American or British accent [5].

Coeckelbergh argues that including the appearance of emotion in robots is necessary to create suitably moral agents [6]. While emotional expression is complex, detection of emotion based on voice alone is possible, though not entirely reliable. Scherer's review of vocal emotion research found accuracy recognizing emotion in vocal portrayals tends to be around 60%, well above chance [7]. Our study is unique in focusing on emotional expression of robots and for including voice actor designed character voices (similar to what a voice actor would create for an animated character, for instance), which have more exaggerated emotional expression than typical speaking voices.

3. METHODS

3.1. Participants

Nine students of Michigan Technological University were recruited from the university psychology study participant pool. The participants included two females and seven males with an average age of 19. All participants self-identified as White/Caucasian.

3.2. Robots

We utilized two robots: Nao (Figure 1) and Pleo (Figure 2). Nao is a humanoid robot with rounded toy-like styling. Pleo is a dinosaur shaped robot with a cute, pet-like or toy appearance. Nao and Pleo were chosen as the robots to conduct the experiment in order to contrast the effect that robotic appearance has on voice acceptance. It is possible that participants may consider one voice more suited to a humanoid

robot than to an animoid robot or vice versa. A TV remote was used to control Pleo with IR signals and Nao was controlled from a nearby laptop over Wi-Fi. Participants were advised of the researcher's control of the robot during the consent process as the activation of the robots was conspicuous and thus the Wizard of Oz method was not effective in our particular set up.

Figure 1: Nao

Figure 2: Pleo

3.3. Voices

Different vocalizations were utilized to determine what impact, if any, a different voice or speaking style brings to interpretation and acceptance of emotional responses. A pair of young adult male voice actors with Midwestern American accents provided six of the voices used. Each actor recorded their lines in their regular speaking voice, a character voice envisioned for Nao, and a character voice for Pleo. Two synthesized voices from popular text-to-speech engines were also used: the Microsoft David voice and the iOS Alex voice.
These conditions were chosen because they were perceived to cover a spectrum of expressiveness and appropriateness. A regular human speaking voice is a baseline that may be both expected of a robot and also highly emotive. In contrast, a character voice is likely to be very expressive, but may not be expected from a robot. Meanwhile, a synthesized voice may be expected of a robot, but may not be expected to be as highly emotive. These considerations of expectation versus emotiveness suggest a variety of possible interpretations of what is a suitable voice for a robot. As the research is still in progress, we will presently only consider results for one of the voice actors, Thomas, and the Alex synthesized voice, hereafter referred to as the human voices and synthesized voice for simplicity

3.4. Emotions

Seven emotions were tested in the study: happiness sadness, anger, fear, disgust, surprise, and anticipation. Six of the seven emotions are Ekman's six basic emotions and were chosen for their prevalence in psychology [8]. The seventh condition, anticipation, was chosen for its similarity to fear and surprise. Its inclusion allows us the opportunity to see if participants can discern an emotion that is not traditionally regarded as basic and to gauge confusion between emotions with subtle differences. When participants were asked to identity emotions, the questions were open-ended to avoid participants using process of elimination to "guess" the emotion from a list over their own intuition or expectation.

3.5. Scripts

The stories chosen were adaptations of the traditional fairy tales "The Boy Who Cried Wolf" and "The Three Little Pigs" crafted specifically for this experiment. Fairy tales were selected because they naturally evoke a wide range of emotions that would be difficult to contrive in a game or other activity. Adjustments were made to the stories to accommodate certain emotions (e.g., an evocative description of the wolf was used to prompt disgust). These particular fairy tales were chosen for their similar cadence and structure, so while the emotions may appear in different orders, the stories themselves are fairly similar to minimize confounding variables. The exact words spoken by the robots differed in each story to be contextually relevant. The utterances are listed in Table 1.

Table 1: Robot utterances and corresponding emotions

The Boy Who Cried Wolf	**The Three Little Pigs**
That sounds nice. *(Happiness)*	I wonder what's going to happen! *(Anticipation)*
That's not nice! *(Anger)*	They shouldn't tease him like that. *(Anger)*
This should be good. *(Anticipation)*	He can't want to EAT them! *(Disgust)*
Gross! *(Disgust)*	Woah, that's fast! *(Surprise)*
What? Why didn't they help? *(Surprise)*	He destroyed their homes. *(Sadness)*
He's going to eat the sheep! *(Fear)*	Good. *(Happiness)*
All his sheep are gone. *(Sadness)*	Oh no! *(Fear)*

3.6. Conditions

The experiment was designed so that each participant would interact with all of the eight voices, two stories, and two robots, but not all combinations thereof. Alternating robots and stories, each participant went through four conditions with questionnaires, then completed a group questionnaire for those four voices, before repeating the procedure for the final four conditions and another group questionnaire. Each group of four conditions included all three voices (regular speaking voice, character voices for both Nao and Pleo) from a single voice actor and one synthesized voice. Conditions were counterbalanced between participants to counteract order effects. While the character voices were created with a particular robot in mind, we used them with both robots during the experiment for thoroughness. A number of participants ran out of time during the experiment and did not encounter all of the intended conditions.

3.7. Experiment

To begin session, each participant was welcomed and given a consent form. The experiment was explained verbally in addition to in writing within the consent form. Participants were also introduced briefly to the robots. The participants were given scripts for the stories indicating their lines and pauses for them to wait before continuing the story. At these pauses, the researcher activated the robot's response to the point in the story.
After each of the robot's responses, participants were asked what emotion they believed the robot was trying to convey. They were then asked to rate on a linear scale from 1 to 7 the clarity with which they felt the emotion was expressed and the suitability of the response coming from the robot. Each of the seven emotions appeared once in each story. After a story was completed, there was a set of linear scale questions evaluating

Table 2: Participant ratings of voice characteristics; ratings on a 7-point linear scale

	Likability		Attractiveness		Warmth		Honesty		Trustworthiness		Naturalness		Humanness		Roboticism	
Voice	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Nao Voice	4.10	1.60	3.50	1.43	3.40	1.07	4.40	0.52	4.30	0.95	3.20	1.48	4.00	1.25	3.10	1.10
Pleo Voice	4.10	1.66	3.50	1.27	4.00	1.63	4.70	1.16	4.70	1.34	3.00	1.33	3.50	1.08	3.50	1.08
Regular Voice	4.10	1.37	4.30	1.25	3.90	0.99	4.30	0.95	4.40	0.84	4.40	1.26	4.80	0.92	3.10	1.52
Synthesized Voice	3.18	1.60	2.73	1.10	1.91	1.04	3.82	1.17	3.45	1.29	1.64	1.03	1.36	0.92	6.55	1.21

properties of the voice overall. The process repeated for the next three conditions, with the robot and story alternated and the voice changed. At the end of four conditions, the participant completed a questionnaire that included basic impressions of one of the robots, evaluations of the voices, and demographics information. Voices were labeled A, B, C, and D respectively to avoid influencing participants. Finally, the procedure was repeated with four more conditions and the group questionnaire again.

The researcher conducting the experiment verbally asked all of the questions and noted the participant's responses in a form on a laptop rather than having the participant filling out the survey in writing, with one exception. The survey was coded with a participant ID and information about the setup of each iteration of the test (robot, voice, and story). With the researcher already at the computer to take notes and control robots, this design allowed researchers to record responses with key information or notes as needed. This design also required the questions be asked aloud, which created a more conversational atmosphere. We anticipated that participants would be more likely to give longer or more thorough answers when speaking as opposed to writing and could feel more comfortable asking clarifying questions.

There was a group of questions that were not done orally between the participant and the experimenter. Open-ended prompts for general thoughts on each of the voices were completed by the participants in writing when time permitted to allow us to directly quote participants.

4. PRELIMINARY STUDY RESULTS

On a seven point scale from strongly dislike to strongly like, participants rated the human regular speaking voice highest (M = 5.11; SD = 1.45), followed by the Pleo character voice (M = 4; SD = 2.06), the Nao character voice (M = 3.78; SD = 1.64), and finally the synthesized voice (M = 3; SD = 1.41). A single factor ANOVA comparing ratings approached statistical significance, but did not reach the conventional level ($F(3,32)$ = 2.48, p = 0.08) due to the small sample size. Four participants listed the regular voice as their favorite, 3 listed the Pleo voice, 1 participant selected the Nao voice, and 1 the synthesized voice. Chosen least favorite voices were the synthesized voice (4), the Pleo voice (2), the Nao voice (2), and the regular speaking voice (1). In contrast, however, slightly more participants (4) believed the Pleo voice to be more emotionally expressive than the regular voice (3). Only 1 participant considered the Nao voice most expressive and no one chose the synthesized voice as most expressive.

The average likability scores of all human voices were equal and were greater than the average score of the synthesized voice. The synthesized voice ranked highest in average roboticness and nothing else. The Pleo voice was highest in warmth, honesty, and trustworthiness. The regular speaking voice had the greatest average score in attractiveness, naturalness, and humanness. See Table 2 for details.

Since participants used their own words to describe the emotions they heard, similar emotions were grouped for the purposes of data analysis. For example, curiosity and excitement where combined under "anticipation". While the actors had been asked to portray seven emotions (anger, anticipation, disgust, fear, happiness, sadness, and surprise), participants' responses included 24 emotions after consolidation. The most common unintended response was "Neutral" with 30 occurrences, followed by "Confusion" with 13. The vast majority (22/30) of the "Neutral" responses were for the relatively uninflected synthesized voice.

The emotions with the best recognition across conditions was fear (68.29% accurate recognition), followed by anticipation (65.85%), happiness (65.85%), sadness (56.10%), anger (53.66%), disgust (43.90%), and surprise (34.15%). However, the emotion with the highest average rated clarity was sadness (5.04). The average clarity ratings of the other emotions were: anger (4.86), surprise (4.86), anticipation (4.85), disgust (4.78), fear (4.75), and happiness (4.69). The emotions of the Nao voice were accurately recognized most reliably (67.14% of the time), then Pleo (61.43%), regular (60.00%), and synthesized (35.06%).

There were some robot effects. Utterances from the Pleo robot (irrespective of what voice was being used) were accurately recognized 62.86% of the time, while those from the Nao robot were recognized 48.30% of the time. Conversely, the average ratings for clarity and suitability of correctly identified emotions were higher for Nao (4.86 and 4.90, respectively) than for Pleo (4.81 and 4.33).

Participants remarked that the Nao voice was "a little weird for the robot…", "goofy like a bad cartoon", and "very creepy, like a puppet voice". The Pleo voice was called "childlike", "almost like a kid", "sounded like a child", "very high pitched", and "annoying and high pitch". The regular speaking voice was repeatedly noted as sounding older than the other two human voices: "Older tone (early 30s)", "sounds like a young adult", and "very mature". The apparent age of the voice conflicted with the appearance of the robot and the task of the experiment: "I thought this voice did not really go with both the context and the robot itself." and "I didn't feel like that voice should be listening to this type of story". Some also remarked that it was "very clear" and "showed emotion". Finally, the synthesized voice was also considered clear, but monotone and robotic. "Liked that it was very clear but it was monotone and unemotional[.]" "The voice was very monotonous and robotic, and made it feel cold and uninviting[.]"

5. DISCUSSION

Despite a number of negative comments, the human regular speaking voice was the highest rated. Many of the comments were related to the age of the voice being discordant with the appearance of the robots and the usual age for fairy tales. This suggests that when rating the voices, participants may have been considering them largely independent of the context of the experiment. However, when context was a factor, it was significant.

Unsurprisingly, the synthesized voice was poorly regarded. While many participants considered it robotic both in their comments and in the ratings of roboticness, they were unhappy with the monotonous tone and lack of expression. What is stereotypically suitable, such a robotic voice, apparently is not a good choice for social robots in emotional interactions.

The Pleo voice was highest in warmth, honesty, and trustworthiness, perhaps because the voice was the most child-like. Meanwhile, the regular speaking voice was highest in attractiveness, naturalness, and humanness. They were also the highest rated of the voices, the top two favorite voices, and equally likable. This suggests that the voices could be used to evoke different feelings or impressions without significantly undermining acceptance. For example, using the Pleo voice might be more effective than the regular speaking voice in a situation requiring warmth and trust, such as comforting a nervous pediatric patient.

The emotion recognition scores present a rather unclear picture. The differences in recognition between emotions suggests that some emotions can be adequately conveyed by voice, but not all. This is consistent with the findings of Scherer [7]. To improve emotional communication, it might make sense to focus on improving non-vocal emotional cues (e.g., body language) for emotions that are poorly recognized by voice. There is no clear pattern in the data between the recognition of particular emotions and their rated clarity. Analysis of the effects of the individual robots is also similarly unclear. Given the relatively small sample size at this stage, it is hard to come to any firm conclusions regarding these ambiguous factors.

6. FUTURE WORK

Some participants' responses were based on the wording of the utterance rather than the tone of voice. Some instances of respondent suitability scores may also be tied to this effect of phrasing over tone. While it is certainly a phenomenon that will be inescapable in naturalistic applications, it should be kept in mind while analyzing the results of experiments involving a relatively small number of interactions and be considered while designing future studies as well. Designing the study with different responses for certain emotional cues may work to mitigate this effect's impact on the data.

Given the homogeny of our current participant group, similar work with older adults, children, and people of color should be conducted. Additionally, based on the findings of [3], [4], it would also be advisable to examine gender differences in more diverse pool of participants. It would also be worthwhile to follow up with a comparison of male and female voices in the same context as all of the voices in this study were male. Finally, future versions of this study should track the English language proficiency of participants.

7. CONCLUSION

These findings are preliminary, so we should be cautious of drawing conclusions too strongly. However, in designing this study and working through the preliminary data, it is apparent that the area of properly designing vocal emotional expression in robotics is filled with intriguing avenues of investigation and possible implications. The field of social robotics continues to grow and with it the necessity for robots to be not only functional but emotive as well.

8. ACKNOWLEDGMENT

Our thanks to our volunteer voice actors, Thomas Klonowski and Robert Mampe, and director Patricia Helsel for so generously creating the voices used in this study. This project was supported by the National Institutes of Health under grant No. 1 R01 HD082914-01.

9. REFERENCES

[1] V. K. Sims, M. G. Chin, H. C. Lum, L. Upham-Ellis, T. Ballion, and N. C. Lagattuta, "Robots' Auditory Cues are Subject to Anthropomorphism," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 53, no. 18, pp. 1418–1421, 2009.

[2] M. L. Walters, D. S. Syrdal, K. L. Koay, K. Dautenhahn, and R. Te Boekhorst, "Human approach distances to a mechanical-looking robot with different robot voice styles," *Proc. 17th IEEE Int. Symp. Robot Hum. Interact. Commun. RO-MAN*, pp. 707–712, 2008.

[3] F. Eyssel, D. Kuchenbrandt, S. Bobinger, L. de Ruiter, and F. Hegel, "'If you sound like me, you must be more human,'" *Proc. 7th ACM/IEEE Int. Conf. Human-Robot Interact.*, p. 125, 2012.

[4] F. Eyssel, D. Kuchenbrandt, F. Hegel, and L. De Ruiter, "Activating elicited agent knowledge: How robot and user features shape the perception of social robots," *Proc. - IEEE Int. Work. Robot Hum. Interact. Commun.*, pp. 851–857, 2012.

[5] R. Tamagawa, C. I. Watson, I. H. Kuo, B. A. Macdonald, and E. Broadbent, "The effects of synthesized voice accents on user perceptions of robots," *Int. J. Soc. Robot.*, vol. 3, no. 3, pp. 253–262, 2011.

[6] M. Coeckelbergh, "Moral appearances: Emotions, robots, and human morality," *Ethics Inf. Technol.*, vol. 12, no. 3, pp. 235–241, 2010.

[7] K. R. Scherer, "Expression of emotion in voice and music," *J. Voice*, vol. 9, no. 3, pp. 235–248, 1995.

[8] P. Ekman and H. Oster, "Facial expressions of emotion," *Annu. Rev. Psychol.*, vol. 30, no. 1, pp. 527–554, 1979.