First Impressions of a NAO Robot: Evaluating Emotions Through E4 Data and Questionnaires

Karla Conn Welch, Natalie Warning, Rohit Narayanan, Prasad Nethala, Ha Do, Saipruthvi Vanaparthy, Shaye Allen, and Sabrina Daisey
Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY, USA

Introduction

Affect-sensitive human-robot interaction (HRI) will allow the robot to vary its behavior and respond to changes in the subject's mood and attentiveness. This form of HRI can be applied to the treatment of autism by having robots serve as social skills mediators who can stimulate and encourage conversation while monitoring the subject through non-invasive sensors.

Successful HRI treatment of autism will rely upon the user's acceptance of the robot. This study focused on initial reactions to a NAO robot (Fig. 1) to determine the most appropriate type of motions and voice the robot should display to be an affect-sensitive social skills aid.

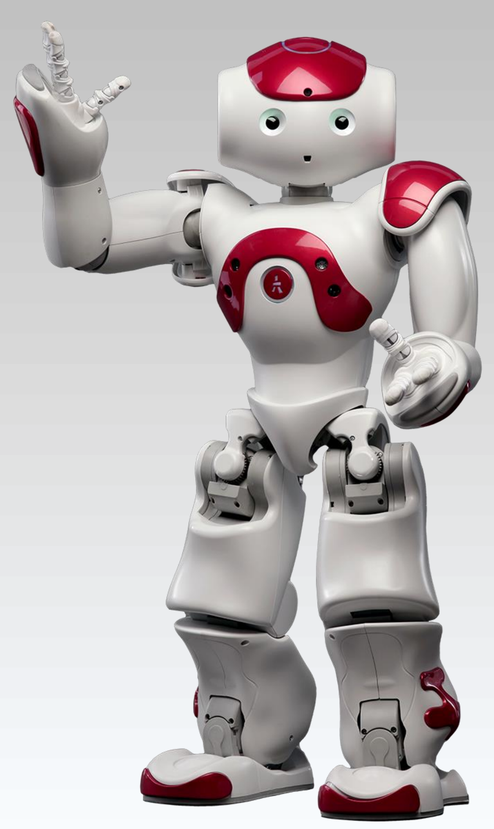

Figure 1. NAO Robot

Experiment Design

Two text-to-speech voice options and two styles of motion were tested in the following condition combinations:

- **Condition A – Default NAO voice & smooth motions**
- **Condition B – Amazon Polly Justin voice & smooth motions**
- **Condition C – Default NAO voice & jerky motions**
- **Condition D – Amazon Polly Justin voice & jerky motions**

The NAO default text-to-speech is similar to a robotic child's voice, and the Justin voice from Amazon Web Services, Polly, is similar to a young boy's voice. The jerky motions were defined from how the robot tends to "jerk" into positions when transitioning. The smooth motions were programmed with a short delay before each motion to allow for smoother transitions with less jerking.

The subjects wore an Empatica E4 sensor (Fig. 2) on their wrist and were video recorded. Subjects interacted with the robot through a scripted conversation for all four conditions, presented in a Latin Squares sequence to account for order effects. Between conditions, subjects completed the Godspeed [1] and RoSAS [2] questionnaires to assess their reaction to qualities of the robot including factors that influenced perceived safety, anthropomorphism, and discomfort. The Godspeed survey rates factors on a scale from 1 to 5 [1]. The RoSAS survey rates factors on a scale from 1 to 7, with 1 = definitely not associated and 7 = definitely associated [2].

Figure 2. Empatica E4

The user acceptance study was conducted with 38 college-age adult subjects. All subjects were asked to rate their experience with NAO robots. Of the 38 subjects, 17 responded as having little to no experience with the NAO or robotics. This smaller group was 70.6% female (12/17 subjects), and the average age was 22.1 years old with a standard deviation of 3.93 years.

Machine Learning Analysis

The E4 wearable sensor collects data on six signals, and four were used for the analysis of the user acceptance study: electrodermal activity (EDA), blood volume pulse (BVP), skin temperature (SKT), and heart rate (HR). The E4 physiological data is compiled and fed into a machine learning algorithm by the following process:

1. Raw E4 data converted to .csv file using a Jupyter function [3]
2. Condition start and end timestamps found using manual video analysis
3. Appropriate filters were used to divide tonic and phasic portions of EDA as well as on the raw BVP signal
4. Segmented data into 20-second blocks
5. From each block, calculated 12 features [4]
   a. EDA Tonic Mean, Standard Deviation
   b. EDA Phasic Peak Rate, Max Amplitude, Mean Amplitude
   c. BVP Peak Mean Amplitude, Max Amplitude
   d. HR Mean, Standard Deviation
   e. SKT Mean, Slope, Standard Deviation
6. Divide features by baseline (lowest-deviation 20-second block)
7. Normalize all feature segments between zero and one
8. Each E4 segment is labeled with the condition in which it was originally recorded
9. Segments are fed into a machine learning module to calculate results from five separate algorithms:
   a. Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and Gradient Boosting

Machine Learning Results

After E4 data processing and analysis, an F1-score was calculated for each algorithm and condition. The F1-score is a weighted average of precision and recall, ranging from zero to one and calculates how well the algorithm can classify the condition. A desirable model will be able to classify all conditions at a rate better than chance; for a binary classification better than chance would be F1 > 0.5 in both categories. Accuracy represents the total number of correct predictions over total number of predictions made by the machine learning analysis.

By the gradient boosting model, an accuracy of 0.81 was returned for a binary classification of B vs D and 0.77 for B vs C. Accuracies were over 0.7 for all pairs of conditions. The results for a four-way classification of all conditions, for a desirable F1 > 0.25 for each condition, are shown in Table 1. Again, the gradient boosting model produced the highest F1-scores and accuracy.

Model	F1-score A	F1-score B	F1-score C	F1-score D	Accuracy
Logistic Regression	0.34	0.25	0.26	0.34	0.29
Support Vector Machine	0.26	0.34	0.11	0.29	0.28
Decision Tree	0.59	0.53	0.62	0.33	0.53
Random Forest	0.68	0.66	0.65	0.63	0.65
Gradient Boosting	**0.81**	**0.71**	**0.72**	**0.75**	**0.75**

Table 1. Machine Learning Results

Survey Results

Using the Godspeed and RoSAS questionnaire data of the 17 subjects who had little to no experience with the NAO or robotics, a T-test was used to determine statistically significant differences, shown by *p*-values (<0.05) in Table 2. Table 3 shows the results of a Friedman test and a Nemenyi post-hoc test to determine significant relationships between all four conditions, where significance was determined if the value was greater than 7.81 at α = 0.05.

Condition	A & B	A & C	A & D	B & C	B & D	C & D
Perceived Safety		0.00412		0.00007	0.00475	
Anthropomorphism		0.02832		0.00840		
Awkward		0.04082	0.02181			
Strange		0.01310	0.01705	0.01397	0.01806	

Table 2. Godspeed and RoSAS Questionnaire T-test Results

Factor	Ranking	Test Value	Significant Relationship
Perceived Safety	B>A>D>C	19.5	B and C
Anthropomorphism	B>A>D>C	9.54	B and C
Discomfort	D>C>A>B	10.5	B and D

Table 3. Friedman and Nemenyi post-hoc Significant Relationship Results

Conclusions

This study of initial user acceptance to the NAO robots provided significant results to indicate that differences in voice type and smoothness of motions during HRI provokes different physiological responses among adults around college age. In future studies, these results will help determine the most appropriate behaviors the NAO robots should present (e.g., perceived as least provoking of discomfort, most human-like, and safest) during affect-sensitive social skills interventions with children diagnosed with autism spectrum disorder.

References

[1] C. Bartneck, D. Kulic, and E. Croft, "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots," *IJSR*, 2009.
[2] C. Carpinella, M. Perez, A. Wyman, and S. Stroessner, "The Robotic Social Attributes Scale (RoSAS): Development and Validation," *Proc. of ACM/IEEE HRI*, 2017.
[3] A. Bizzego, A. Battisti, G. Gabrieli, G. Esposito, and C. Furlanello, "pyphysio: A physiological signal processing library for data science approaches in physiology," *SoftwareX*, 2019.
[4] K. Welch, U. Lahiri, N. Sarkar, Z. Warren, W. Stone, and C. Liu, "Affect-Sensitive Computing and Autism," *Affective Computing and Interaction*, 2011.

Acknowledgements

The authors want to thank the adult subjects who participated in this study. Additionally, thanks goes to the Norton Children's Autism Center and all research staff who contributed insights for future research based on this study. This project was supported in part by NIH and NSF through grant SCH #1838808 and NSF through grant EPSCoR-OIA #1849213.