Sentiment Apprehension in Human-Robot Interaction with NAO

Jie Shen, Ognjen Rudovic, Shiyang Cheng
Department of Computing
Imperial College London
London, U.K.
{js1907, o.rudovic, shiyang.cheng11}@imperial.ac.uk

Maja Pantic
Department of Computing
Imperial College London, U.K.
EEMCS, Univ. Twente, N.L.
maja@imperial.ac.uk

***Abstract*—The ability of robots to interact in a socially intelligent manner with humans is the core of human-robot interaction (HRI). The quality of this interaction is typically measured in terms of how it is engaging to the users either reflected in duration of time users spend interacting with a robot, or their self-reports on engagement during the interaction. In contrast to existing studies that analyze the influence of robots' ability to mimic affective states (happy or sad) of users on their engagement, in this paper we study the influence of sentiment apprehension by robots (i.e., robot's ability to reason about the user's attitudes such as judgment / liking) on the user engagement. Specifically, we present the findings from our pilot study on the effect of sentiment apprehension in HRI using NAO robot. In this study, we analyzed two versions of mimicry game: in the first, NAO was solely mimicking facial expressions of the users, while in the second he was also providing a feedback based on the sentiment apprehension. A total of 32 participants (7 female, 25 male) were recruited for this experiment, and the results show that the participants in the second group spent more time interacting with the robot and played more rounds of the mimicry game. After experiencing both versions of the game, ratings given by the participants indicate (with 99% confidence) that the game with sentiment apprehension is more engaging than the baseline version.**

Keywords—sentiment analysis; human-robot interaction; facial expression recognition

I. INTRODUCTION

Despite recent technological advances in design of humanoid robots [1], most of these robots still lack the ability to apprehend the sentiment, affective states and intentions of the user. However this is one of the key factors driving their ability to display a socially acceptable behavior, the lack of which prevents them from engaging in a truly natural interaction with users. On the other hand, 'social' robots designed to recognize facial expression and audio cues might provide a much more interesting and engaging social interaction [2]. This can benefit applications from automated tutors [3], entertainment robots [4], but also medical applications where social robots are used as assistive tools in, for instance, treatment of children with autism [5].

The main prerequisite for these 'social' robots is the ability to sustain long-term interactions. To this end, robots need be able to understand the user's sentiment and affective states in order to display socially intelligent behavior, a key requirement for engaging with humans [2]. To facilitate the engagement, the main idea is to endow the robots with ability to apprehend affective states (such as the six basic emotions) and sentiment of the user during interaction. This is typically attempted by integrating within the robot architecture the computer vision and machine learning algorithms for automated analysis of user's facial expressions, vocal verbal and/or non-verbal signals), or both. These allow the robot to, for instance, mirror the facial expression of a user in the course of a communicative task, leading to the shared feeling of empathy of the user towards the robot [6]. This, in turn, facilitates a more engaging interaction between users and a robot.

To the best of our knowledge, existing studies on HRI using social robots focus mainly on affect expression and apprehension through the robot mirroring of (facial) expressions of emotional states (such as happy or sad, for instance) of the users. While these works focus on the influence of human affect in HRI to measure engagement [2], empathy [6], and so on, there are no studies showing the effects of human sentiment and its influence on engagement in HRI. Sentiment refers to affective attitudes, which may be reflected in user's judgment or evaluations of certain situations [7]. In humans, the ability to apprehend sentiment of other people is a result of more complex cognitive processes than those involved in recognition of, for instance, six basic emotions [7], and it's a sign of a more socially intelligent behavior. Therefore, robots with ability to apprehend user's sentiment, such as his / her evaluation of the robot's performance in the target task, and, furthermore, convey that to the users, are expected to produce more enduring and engaging interactions with humans.

To analyze the influence of sentiment apprehension in HRI, in this paper we present a pilot study using the NAO robot [10]. To this end, a game called Mimic-Me was developed. The baseline version of the game involves NAO mimicking the human player's facial expression using a combination of body gestures and audio cues. We implemented a multi-modal dialogue model enabling NAO to interact with a player in a naturalistic way using natural language, head movement and facial expressions. The facial expression recognition engine is built upon the discriminative response map fitting (DRMF) facial point tracker [12]. The output of this tracker serves as

This work has been supported by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA).

input features to the Support Vector Machine (SVM) classifier, trained to recognize the player's facial expression. In the experimental version of the game, a sentiment apprehension stage is added after each game-play session, where NAO estimates whether the player likes the game-play experience or not in the previous session by recognizing positive emotions. The robot then gives a simple verbal feedback in case of a positive result.

The outcome of the experiment shows that NAO's ability to apprehend player's sentiment makes the HRI experience more engaging, and as a result, the participant's willingness to spend more time playing with NAO.

In what follows, we review existing works on affect analysis in HRI. We then describe the employed robot architecture, and the design of the "Mimic-Me" game. This is followed by the description of the experimental design, and conclusions of this study.

II. Releated Works

Social HRI has been extensively studied over the last decade. A detailed review can be found in [1]. The robot's ability to 'understand' human affect, and, in particular, ways to improve user's engagement, have been a subject of many studies in HRI.

For instance, [3] presented a motivational system that implements 'emotions', 'drives' and facial expressions analysis into a robot in order to facilitate target HRI. The goal of this (motivational) robot is to generate an analogous interaction for a robot-human dyad as for an infant-caretaker dyad. The study showed that such system positively influenced each other (human and robot) to establish and maintain social interactions.

Ref. [2] studied children engagement in a naturalistic scenario in which children play chess with the iCat, a robot companion. In this study, several causes and effects of engagement are modeled: features related to the user's non-verbal behaviour, the task and the companion's affective reactions are identified to predict the children's engagement. Their results show that the multimodal integration of task and social interaction-based features outperforms those based solely on non-verbal behaviour.

The study in [6] focused on the impact of robot's mirroring of facial expressions of the user, to determine their influence on empathy of a human towards a robot and perceived subjective performance during interaction with a robot head. The result of the study supported the hypothesis that the robot behavior during interaction heavily influences the extent of empathy by a human towards a robot and perceived subjective task-performance.

Ref. [8] reports on a study of human subjects with a robot designed to mimic human conversational gaze behavior in collaborative conversation. The authors show that users engage in mutual gaze with these robots, direct their gaze to them during turns in the conversation, and follow their commands when asked to perform tasks. While talking heads were capable of capturing users' attention very often, the use of head movements together with gaze changes captured the user's attention more often, showing preference for embodied dynamic robots rather than static.

Ref. [9] proposes an experimental design to facilitate engagement during HRI between a humanoid robot and children with Cerebral Palsy (a motor impairment where repetitive exercise plays a key role in rehabilitation). The children are asked to mimic four different actions performed by NAO, as a part of their therapy. However, no results on engagement were reported.

All these works aim at studying the influence of users' engagement during HRI based on 'social' robots with different levels of affect sensing. They show that, in their studies, it helps to improve the engagement of the users with the robots. Yet, none of these studies addressed the influence of sentiment apprehension (different from mimicking emotion expressions) on and during HRI, which is the scope of the work presented in this paper.

III. The "Mimic-Me" Game

"Mimic-Me" (shown in Fig. 1) consists of an interactive game played with the NAO humanoid robot [10]. The game involves the robot 'mimicking' the player's facial expression using a combination of body gestures and audio cues [11]. We implement a multimodal dialogue model enabling the robot to interact with player in a naturalistic way using only natural language, head movement and facial expressions. The facial expression recognition engine is built upon the discriminative response map fitting (DRMF) facial point tracker described in [12] and [13]. Using the 3D point distribution model (PDM) [12] shape parameters as features, a support vector machine (SVM) classifier is trained to recognize the player's facial expression. The "Mimic-Me" game is implemented as a modular, loosely coupled, software system using the HCI^2 Framework [14]. As a result, the system and its modules can be easily reused and / or extended to facilitate further studies in the area of human-robot / human-computer interactions.

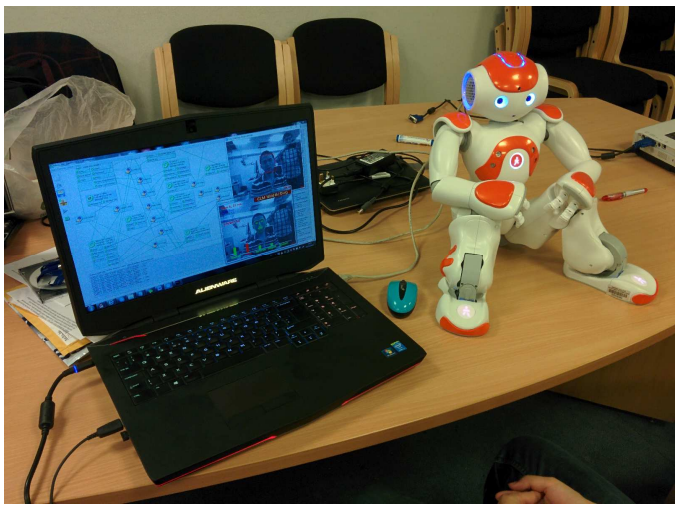

Fig. 1. The hardware setup of the "Mimic-Me" systems: NAO robot on the right and the host computer on the left.

A. Dialogue Model

In the baseline version of the "Mimic-Me" game, each game-play session is divided into two stages. In the first stage (game-play initiation stage), the robot initiates dialogue with the potential player and tries to engage him / her. If the person agrees to play, the actual game-play session (i.e., facial expression recognition and "mimicking") is conducted in the second stage (expression imitation stage). The game is played repeatedly until the player decides to stop. In particular, the robot always asks the player whether to play another round after each game-play session. An additional sentimental apprehension stage is added as the third stage in the experimental version of the game. At this stage, the robot continues to recognize the player's facial expression after the game-play session ends and, based on the recognition result, provides a feedback to the player commenting on its own performance (i.e., whether the play likes the interaction experience).

1) Game-play initiation stage: As shown in Fig. 2, the dialogue of this stage consists of 4 states. Once the initialization is finished, the dialogue enters the 'Awaiting Player' state, in which the robot starts the face tracking module. A Viola-Jones face detector [15] is used to detect face(s) in the scene. Once a face is detected, the face tracking module enables the robot to track the player's face with its eye-gaze.

After a stable detection of the player's face, the robot first introduces itself ('S1' in Fig. 2), then asks the player whether he / she wants to play the game ('Q1' in the Fig. 2), and the dialogue enters the 'Expecting Answer' state. The robot's onboard voice recognition engine is used to recognize 'yes' or 'no' uttered by the player. A positive answer from the player marks the completion of the game-play initiation stage and triggers the start of the expression imitation stage. Upon receiving a negative answer ('no'), the robot prompts the player to call it when he / she is interested in playing ('S2' in the figure) and the dialogue enters the 'Expecting U1' state. In this state, the voice recognition engine is configured to recognize 'Hi', 'Hello' or 'NAO' with an indefinite timeout. Once any of the words is recognized, the robot asks again whether the player wants to play the 'Mimic-Me' game and the dialogue re-enters the 'Expecting Answer' state.

The face tracker runs continuously in all states. Once the tracked face is lost, the dialogue manager assumes the player has left the scene and triggers the dialogue to move back to the 'Awaiting Player' state.

2) Expression imitation stage: The dialogue of this stage (as illustrated in Fig. 3) contains 2 states. After initialization, the dialogue enters the 'Expecting <Start2>' state to wait for the completion of the game-play initiation stage. Once the signal is received, the dialogue manager instructs the player to display a facial expression ('S4' in the figure) and activates the facial expression recognition (FER) component. The FER component output the recognition result as one of seven possible values, corresponding to 'neutral' plus the 6 universal facial expressions (anger, disgust, sadness, fear, happiness, and surprise), respectively. If no expression is detected within a timeout period, 'neutral' is given as the default output.

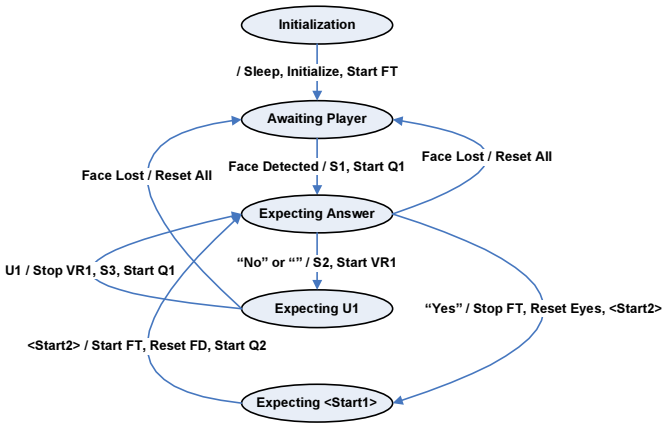

Fig. 2. Dialogue model of the game-play initiation stage. Within the figure: 'FT' stands for 'Face Tracking', 'FD' stands for 'Face Detection', 'S1' stands for the statement 'Hello! My name is NAO. Nice to meet you!', 'S2' stands for the statement 'OK, please just call me when you are interested.', 'S3' stands for the statement 'Hello again! I am still here.', 'Q1' stands for the question 'Do you want to play a game with me?', 'VR1' stands for the voice recognition session used to recognize 'U1', and 'U1' stands for player utterance of 'Hello', 'Hi', or 'NAO'.

Upon receiving a non-neutral FER result, the dialogue manager instructs the robot to invoke the manually composed animation sequence ('A*' in Fig. 3) conveying the same emotional state. Since the NAO robot does not support physical facial articulation, the emotional state is expressed through a combination of body movement [11], non-verbal audio cues (such as laughter in case of 'happiness'), and colour change / flashings of the robot's eye LED. An example of the body gestures we used to express different emotional states is given in Fig. 4.

After the animation is displayed, depending on the version, the system either (in the baseline version) moves back to the game-play initiation stage or (in the experimental version) enters the sentiment apprehension stage.

3) Sentiment apprehension stage: At this stage, the robot continues to recognize the player's facial expression, and based on the result, tries to deduce whether the player likes or dislikes the game. This information is potentially useful for the iterative optimization of the FER algorithm. Nonetheless, in the current version, the result is only utilised by the robot to provide a verbal feedback commenting on its own performance.

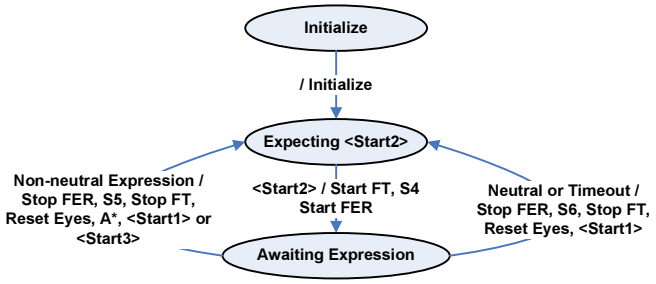

Fig. 3. Dialogue model of the expression imitation stage. Within the figure: 'FT' stands for 'Face Tracking', 'FER' stands for 'Facial Expression Recognition', 'S4' stands for the statement 'OK, Please make a face, I will try to mimic your expression.', 'S5' stands for the statement 'Now it's my turn, watch me!', 'S6' stands for the statement 'Sorry, I can't do a poker face like you.', and 'A*' stands for the body animation corresponding to anger, disgust, fear, happiness, sadness, or surprise.

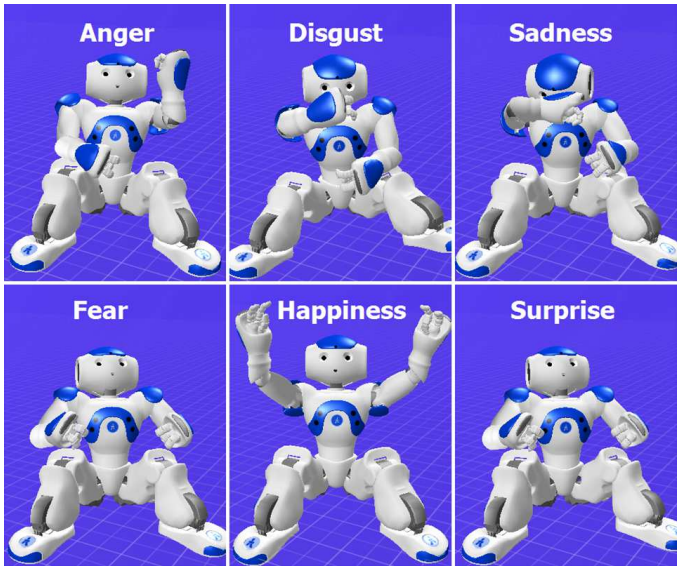

Fig. 4. The robot's body animation used to express different types of emotions.

As shown in Fig. 5, the dialogue model of the sentiment apprehension stage is very similar to that of the previous stage. In this pilot study, the sentiment of the player is deduced using very simple heuristics. Namely the robot assumes the player likes the game if a smile or laughter is spotted.

B. *Facial Expression Recognition*

The "Mimic-Me" game's facial expression recognition component is based on the discriminative response map fitting (DRMF) facial point tracker presented in [12] and [13].

Using the location of the tracked facial landmarks as input features, we have trained a multi-class support vector machine (SVM) classifier [16] to perform facial expression recognition. A one-against-one approach has been used because it is one of the most commonly used methods for expression classification [17]. In particular, the k-class SVM classifier consists of k(k-1)/2 binary SVM classifiers, each trained using examples from only two classes to find a hyper-plain maximizing the margin between them. To classify unseen data, all binary classifiers are used and the overall decision is derived using majority vote. We use Radial Basis Function (RBF) as the kernel of SVM classifiers:

$$K(x,\bar{x}) = \exp(\frac{-\| x-\bar{x} \|^2}{2\sigma^2})$$

An empirical grid search was performed over the parameter space (cost parameters and σ) to find the best parameter configuration for the SVM.

In our current implementation, the classifier has been trained on the Multi-PIE database [18]. Specifically, around 3500 images from subjects 1-170 have been used as training examples. The feature vector consists of the 3D location of the 66 facial landmarks tracked by the FROG facial point tracker. Nonetheless, to eliminate unwanted influence of rigid motion and scaling, the faces are first registered to frontal pose before the feature vectors are calculated. Due to the content limitation of the Multi-PIE database, we have only trained a 5-class SVM capable of distinguishing between neutral, smile (happiness), scream (fear), surprise, and disgust. This is because the Multi-PIE database does not contain sufficient amount of examples of to sadness and anger. Nonetheless, both the method and the FER component we developed can be easily extended to recognize these expressions by retraining the SVM with more examples.

C. *Implementation*

The Mimic-Me game has been implemented as a loosely-coupled modular software system using the HCI^2 Framework [14] with some fine-grained action sequences programmed using NAO's own graphical development environment (Choregraphe). The overall structure of the Mimic-Me game is illustrated in Fig. 6. This system is constructed from the high-level dialogue management modules (3 modules are developed, each corresponding to one stage described in subsection II.A), the modules used for the video capturing ('NAO Vision'), face detection ('Face Detector') and facial expression recognition ('FER Component'), and the modules to invoke the action sequences programmed using Choregraphe ('NAO TTS' and 'NAO Communicator').

IV. Experiment Protocol

A total of 32 participants (7 female, 25 male) have been recruited for this study. In the experiment, the participants were randomly assigned to either the control group, which played the baseline version of the game, or the experimental group, which played the experimental version with of game with sentiment apprehension.

Participants in both groups were asked to play the game repeatedly for as many rounds as they like. We then measured the length of time each participant spent interacting with the robot and the number of rounds he / she played. The participants were also asked to rate the game-play experience in terms of engagement level within the range of 0 to 10.

Afterward, participants in both groups were invited to play the other version of the game. Then, the participants were asked to rate that whether and to what extent they agree (or disagree) to the statement that the experimental version of the game is more engaging than the baseline version. The range of the answer is from -5 (strongly disagree) to 0 (neutral) to +5 (strongly agree).

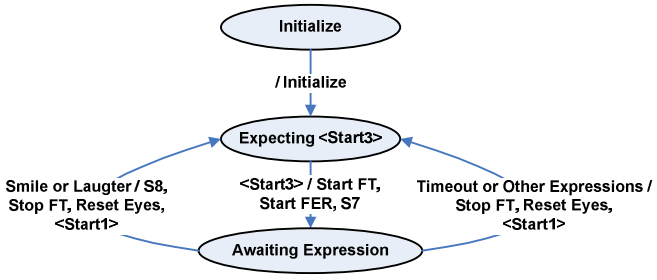

Fig. 5. Dialogue model of the sentiment apprehension stage. Within the figure: 'FT' stands for 'Face Tracking', 'FER' stands for 'Facial Expression Recognition', 'S7' stands for the statement 'I hope I did well in mimicking your expression.', and 'S8' stands for the statement 'And I think I did, thank you!'.

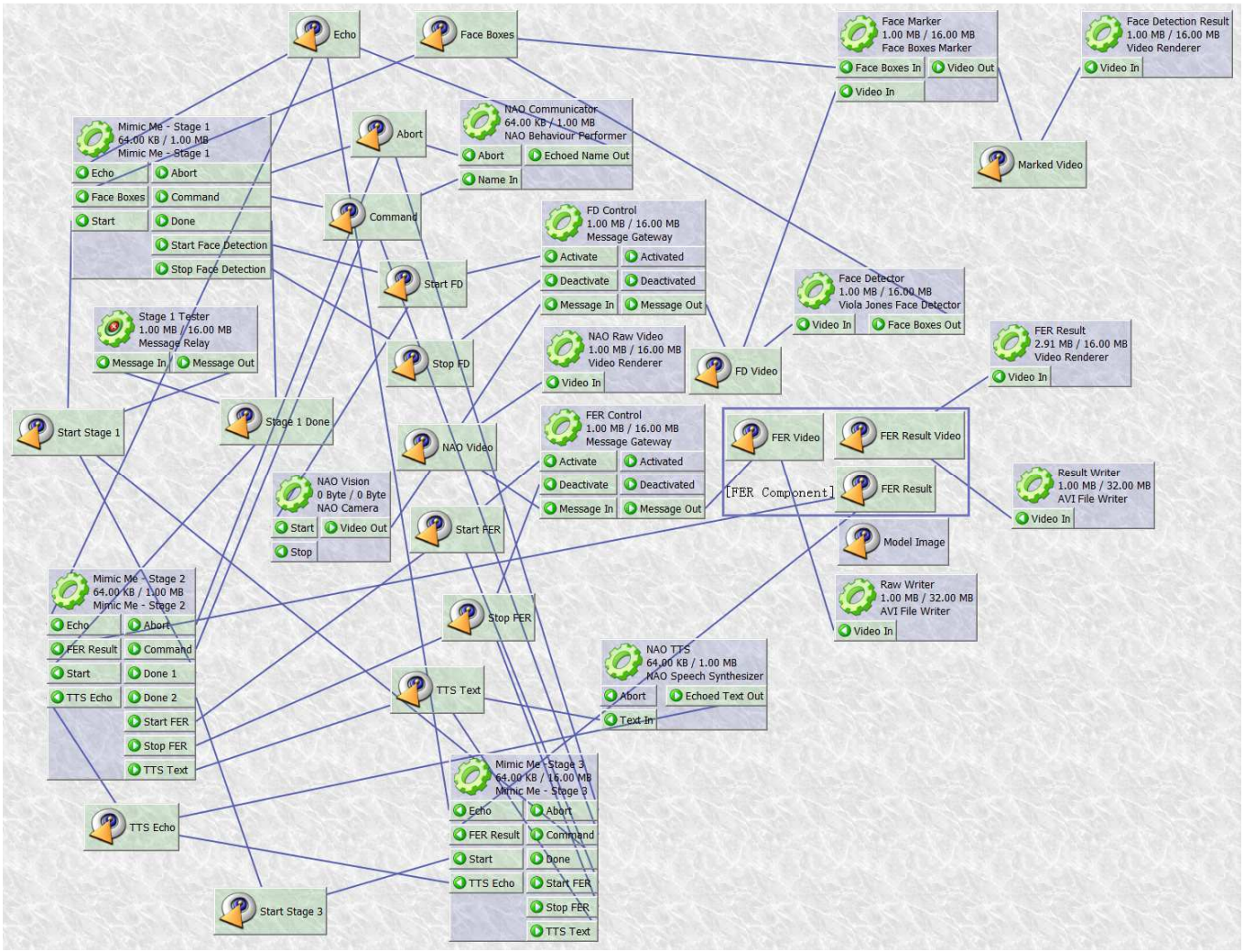

Fig. 6. Overall system structure of the Mimic-Me game. Shown in the figure is the experimental version of the system. Also note that the FER component has been implemented as an external module, hence is labelled differently than the other modules.

V. Results and Discussion

As objective measures, the rounds of "Mimic-Me" game played by the participants and the time they spent on interaction with NAO are shown in Fig. 7A and Fig. 7B, respectively. There is a clear trend that the participants in the experimental group played more rounds of the game and spent more time interacting with the robot than the participants in the control group. On average, participants in the experimental group played 7.94 rounds of the game and spent 389 second with the robot, while their counterparts in the control group only played 5.75 rounds and spent 228 second. According to t-test, these differences are statistically significance with a confidence level of 93% and 99%, respectively.

The statistics of the participants' subjective rating of the engagement level is illustrated in Fig. 7C. Contrary to the observable differences in the objective measures, the perceived engagement levels rated by participants in both groups tend not to differ. Given that interacting with humanoid robot is still not a common experience, we suspect that the lack of difference in the subjective ratings may be an artifact caused by the robot's own novelty factor.

Once the participants experienced both versions of the game, both groups tend to agree that the experimental version with sentiment apprehension is at least somewhat more engaging than the baseline version. This is illustrated by the mostly positive ratings shown in Fig. 7D. When the ratings from both groups are pooled, the t-test indicates that the experimental version of the game was preferred by the participants (in terms of engagement) with a confidence level of 99%. Fig. 7D also shows that the participants in the experimental group preferred the experimental version slightly more than those in the control group. Specifically, participants in the experimental group gave a median rating of +3, while the participants in the control group only gave a median rating of +2 (though the difference is not statistically significant). This conforms to our earlier hypothesis that the robot's novelty factor has played a role in affecting the participants' subject rating on the game's engagement level.

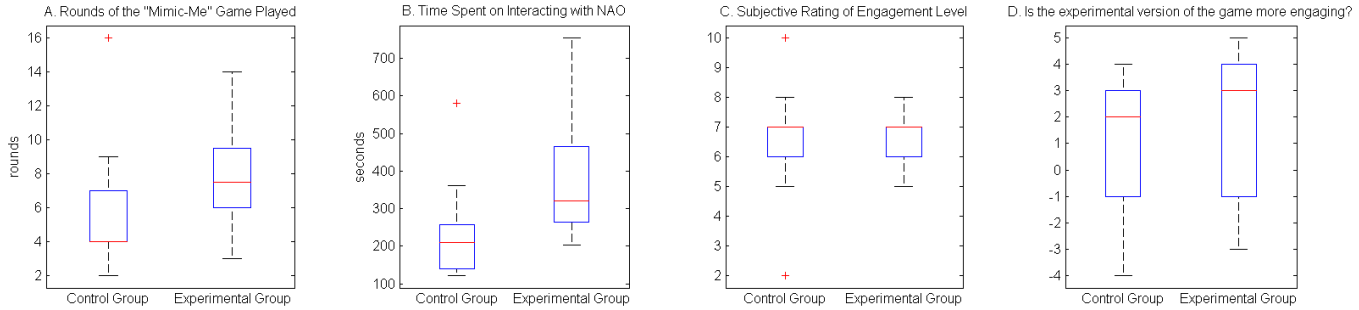

Fig. 7. Experiment results: A shows the number of rounds played by the participants in both groups. B shows the participants' time spent on interacting with NAO. C shows the the participants' subjective rating of the game's engagement level. D shows the participant's subjective rating on whether and to what extent they agree (or disagree) that the experimental version of the game is more engaging than the baseline version (-5 = strongly disagree, 0 = neutral, +5 = strongly agree).

With all the evidences, it is safe to conclude that the additional sentiment apprehension stage has had a positive effect in making the game more engaging to the audience.

VI. Conclusion

In this paper, we presented the result from our pilot study in the effect of sentiment apprehension in human-robot interaction (HRI) using the NAO robot. Our experiment shows that the robot's ability to understand sentiment renders the HRI experience more engaging, and as a result, the participants' willingness to spend more time playing with the robot. For this study, a "Mimic-Me" game was developed. The core game-play involves the NAO robot "mimicking" the human player's facial expression using a combination of body gestures and audio cues. In the experimental version, a sentiment apprehension stage is added after each game-play session. At this stage, the robot estimates whether the player likes or dislikes the game-play experience by recognizing positive emotions. A verbal feedback would be given in case of a positive result. A total of 32 participants were recruited for our experiment. The results show that on average the participants playing the experimental version of the game spent more time interacting with the robot and played more rounds of the "Mimic-Me" game. Both results are statistical significant with 95% and 93% confidence, respectively. After experiencing both versions of game, subject ratings given by the participants also indicate (with 99% confidence) that the game with sentiment apprehension is more engaging.

References

[1] D. Feil-Seifer and M. J. Matari´c, "Human robothuman–robot interaction (hri) interactioninteraction human robot," in Encyclopedia of complexity and systems science. Springer, 2009, pp. 4643–4659.

[2] G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. W. McOwan, "Detecting user engagement with a robot companion using task and social interaction-based features," in ACM, 2009, pp. 119–126.

[3] C. Breazeal, "Regulating human-robot interaction using emotions,drives, and facial expressions," in Proceedings of Autonomous Agents, vol. 98, 1998, pp. 14–21.

[4] R. C. Arkin and L. Moshkina, "Affect in human-robot interaction,"GEORGIA INST. OF TECH. ATLANTA. DTIC Document, Tech. Rep., 2014.

[5] B. Scassellati, H. Admoni, and M. Mataric, "Robots for use in autism research," Annual review of biomedical engineering, vol. 14, pp. 275–294, 2012.

[6] B. Gonsior, S. Sosnowski, C. Mayer, J. Blume, B. Radig, D. Wollherr, and K. K¨uhnlenz, "Improving aspects of empathy and subjective performance for hri through mirroring facial expressions," in RO-MAN, 2011. IEEE, 2011, pp. 350–356.

[7] K. Ahmad, Affective Computing and Sentiment Analysis: Emotion, Metaphor and Terminology. Springer, 2011.

[8] C. L. Sidner, C. D. Kidd, C. Lee, and N. Lesh, "Where to look: a study of human-robot engagement," in Proceedings of the 9th international conference on Intelligent user interfaces. ACM, 2004, pp. 78–84.

[9] N. Abdul Malik, H. Yussof, F. A. Hanapiah, and S. J. Anne, "Human robot interaction (hri) between a humanoid robot and children with cerebral palsy: Experimental framework and measure of engagement," in Biomedical Engineering and Sciences (IECBES), 2014 IEEE Conference on. IEEE, 2014, pp. 430–435.

[10] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, "Mechatronic Design of NAO Humanoid", 2009 IEEE International Conference on Robotics and Automation (ICRA'09), pp. 769-774. 2009.

[11] H. Gunes, C. Shan, S. Chen, and YingLi Tian, "Bodily Expression for Automatic Affect Recognition", In Advances in Emotion Recognition, A. Konar, A. Chakraborty (Eds.), Wiley-Blackwell, 2012.

[12] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, "Robust Discriminative Response Map Fitting with Constrained Local Models", 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013). Portland, Oregon, USA, June 2013.

[13] S. Cheng, A. Asthana, S. Zafeiriou1, J. Shen and M. Pantic, "Real-Time Generic Facial Tracking in the Wild with CUDA", ACM Multimedia Systems, 2014.

[14] J. Shen and M. Pantic, "HCI^2 Framework: A Software Framework for Multimodal Human-Computer Interaction Systems", in IEEE Transactions on Cybernetics, vol.43, no.6, pp.1593-1606, Dec. 2013.

[15] P. Viola, and M. J. Jones, "Robust real-time face detection", International Journal of Computer Vision, vol. 57, no. 2, pp. 137-154, 2004.

[16] CC Chang, and CJ Lin, "LIBSVM: a library for support vector machines", in ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, article no. 27, 2007.

[17] S. Moore, and R. Bowden, "Local binary patterns for multi-view facial expression recognition", Computer Vision and Image Understanding, vol. 115, no. 4, pp. 541-558, 2011.

[18] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE", Image and Vision Computing, vol. 28, no. 5, pp. 807-813, 2010.