Nonverbal Behavior Cue for Recognizing Human Personality Traits in Human-Robot Social Interaction*

Zhihao Shen, Armagan Elibol, and Nak Young Chong
School of Information Science
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa 923-1292, Japan
{shenzhihao, aelibol, nakyoung}@jaist.ac.jp

Abstract— **In parallel to breathtaking advancements in Robotics, more and more researchers have been focusing on enhancing the quality of human-robot interaction (HRI) by endowing the robot with the abilities to understand its user's intention, emotion, and many others. The personality traits can be defined as human characters that can affect the behaviors of the speaker and listener, and the impressions about each other. In this paper, we proposed a new framework that enables the robot to easily extract the participants' visual features such as gaze, head motion, and body motion as well as the vocal features such as pitch, energy, and Mel-Frequency Cepstral Coefficient (MFCC). The experiments were designed based on an idea that the robot is an individual during the interaction, therefore, the interaction data were extracted without external devices except for the robot itself. The Pepper robot posed a series of questions and recorded the habitual behaviors of each participant, meanwhile, whose personality traits were assessed by a questionnaire. At last, a linear regression model can be trained with the participants' habitual behaviors and the personality traits label. For simplicity, we used the binary labels to indicate that the participant is high or low on each trait. And the experimental results showed the promising performance on inferring personality traits with the user's simple social cues during social communication with the robot toward a long-term human-robot partnership.**

Index Terms— **personality traits, human-robot interaction, social cue, regression model**

I. INTRODUCTION

The robots are designed to perform many tasks in numerous fields, thereinto, the home service robots have become increasingly important in the era of population aging. They will play important roles in looking after kids, accompanying older people, taking care of patients, entertainment, etc. Some humanoid robot platforms, such as Pepper, Nao, ASIMO, have applied the synchronized verbal and non-verbal behaviors to achieve a better interaction with their users. However, the synchronized behaviors have been focused on attracting the users' attention and making the users engage with them effectively. These behaviors were designed without considering the user's behaviors, feelings, and engagement between human and robot. And also, there is a serious lack of strategies to enable a robot to comprehend its user's feelings, behaviors, and thoughts and, in order to interact with the user in a natural manner. Piles of studies are needed to enable the robot to fulfill these extremely arduous tasks.

It is known that the personality traits reflect the individual differences in characteristic patterns of thought, feelings, and behaviors [1]. During the human-human interaction, the personality traits affect the humans' actual behavior and their comprehension of the behavioral responses of the person they are interacting with. Furthermore, human can understand the characteristics of their interacting person and then adapt their behaviors to enhance the engagement consciously. The social robots are designed as real and affinities partners to work with, entertain and help its user in their daily life. Similarly, as a partner, the robot should capture and analyze the details of the coherent social cues of its users to boost the engagement during HRI.

It has been shown that there is a strong connection between personalities and behaviors have shown in HRI in previous studies [2], [3]. Specifically, in [4], the author investigated how the Nao robots were perceived as smarter on the basis of their personality (introversion or extroversion) and profession (CEO, pharmacist, or teacher). The fact is that the robot that acted as an extroverted teacher are more likely to be perceived intelligent. On the contrary, if a robot acted as the introverted CEO, which is perceived more intelligent. Some researches also focused on the complementary attraction [8], [9], they mentioned that some people enjoy more talking with the people having complementary personalities to themselves. In [5], they made a robot that appears in two different personalities, extroversion and introversion, and investigated if people prefer a robot that is aligned with their personalities by self-assessment. Also, there is a related study [7] that reveals the relationship between the engagement and the user's personality traits, where they used an introverted robot and an extroverted robot separately to interact with each participant. The extroverted robot was designed to speak louder and faster and with lots of hand gestures. The introverted robot sounded less energetic and displayed fewer gestures. Some evidence also showed the relation between emotion and personality traits. What one observed at a specific moment is emotion. A helpful analogy

*This work was supported by the EU-Japan coordinated R&D project on "Culture Aware Robots and Environmental Sensor Systems for Elderly Support," commissioned by the Ministry of Internal Affairs and Communications of Japan and EC Horizon 2020 Research and Innovation Programme under grant agreement No. 737858.

for comparing the two things is "personality is to emotion as the climate is to weather" [10].

In light of these, it is absolutely necessary to endow the social robots with the capability of inferring human's personality traits. This work is aimed at developing an efficient algorithm that measures the user's "Big Five"[6] personality dimensions exploring the user's social cues and the robot's perceptual capabilities.

A. Problem Statement

In [11], they explained the problems in human personality traits analysis, and the methods used to infer personality traits with frequently-used non-verbal features. In order to address the aforementioned points, we focus on solving the following question:

- How can we extract non-verbal features as easy as possible?

The non-verbal feature extraction is the most essential and hardest part of inferring personality traits. However, in the aforementioned research, they often used many external devices to record videos and audios. Then, complicated models were proposed to detect the face and upper body movements of the participant for extracting non-verbal features. The body activity was described by using Motion energy images (MEI) [12]. And a standard state-space formulation was applied in Visual Focus of Attention (VFOA) [13] for detecting both pose and the location of user's head. To analyze the videos, high computational cost and time were required. For example in [5], they had to localize the robot and participants on the videos. In this research, we used the Pepper robot's camera to extract each participant's visual features such as head motion, gaze, body motion in real-time. The vocal non-verbal features were extracted from the audio that was also acquired by Pepper.

II. Feature Representation

A. Methodology

We can classify human's daily activities in the domestic environment into three categories: (1) human-human interaction; (2) human-robot interaction; (3) non-interactive activities. It should be noted that there are several factors which may affect the user's actions, such as age, gender, occupation, living alone or living with someone (and their relationship), etc. Some previous researches have been done for inferring personality traits in meetings [14] and identifying emergent leaders [15]. Inspired by the previously mentioned researches, we leverage their ideas in inferring the user's personality traits during HRI.

We use a semi-humanoid robot Pepper manufactured by SoftBank Robotics in this research. Pepper is equipped with more than 300 applications and has four microphones in its head, two high definition cameras in the mouth and forehead, and a 3D depth sensor behind the eyes.

Fig. 1 illustrates the three steps of how we design and train our model. In Step a) first of all, we will prepare a questionnaire for assessing the personality traits of each participant; Step b) the interaction video and audio will be recorded by pepper for extracting the non-verbal features including the visual and vocal features which will be introduced in the following section; and Step c) we will evaluate the machine learning model that was trained with feature data from Step b by using personality traits labels from Step a.

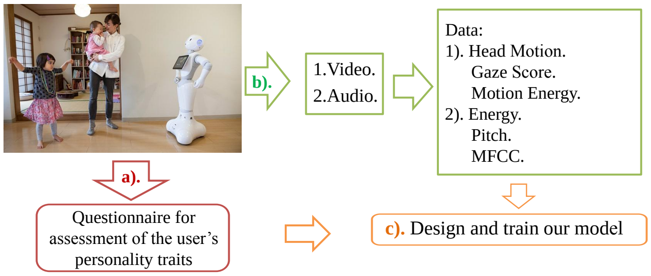

b).
1.Video.
2.Audio.
Data:
1). Head Motion.
Gaze Score.
Motion Energy.
2). Energy.
Pitch.
MFCC.
a).
Questionnaire for assessment of the user's personality traits
c). Design and train our model

Fig. 1. The Pipeline of The Model for Inferring Personality Traits in Human-Robot Interaction

B. Personality Traits

In the most of existing research on personalities, authors mainly discussed the Big-Five Personality Traits model [17] including Agreeableness, Extroversion, Conscientiousness, Openness to experience, and Emotional Stability which have been commonly used as descriptors of personality in psychology. In order to gain some intuitive insights, Table I shows that how people react to each personality trait when s/he is high or low in each trait.

TABLE I
Big-Five Personality Traits

Big-Five	**High on this trait**	**Low on this trait**
Extroversion	Enjoy meeting new people Enjoy being attention center Easy to make new friends	Prefer solitude Do not talk much Think things through
Agreeableness	Care about others Feel concern for others Enjoy helping others	Do not interest in others Do not care about others Insult and belittle others
Conscientiousness	Spend time preparing Pay attention to details Enjoy having a schedule	Make messes Do not take care of things Delay to finish tasks
Emotional Stability	Do not worry much Deal well with stress Rarely feel sad or depressed	Worry about many things Experience a lot of stress Get upset easily
Openness	Enjoy tackling challenges Like abstract concepts Open to trying new things	Do not enjoy new things Resist new ideas Not very imaginative

Some questionnaires have been developed for measuring the Big-Five personality traits, like Ten-Item Personality Inventory (TIPI) [16], Big-Five Factor Markers from the International Personality Item Pool (IPIP) [18], and many others. IPIP Big-Five Factor Markers provides questions relatively easier to answer than TIPI questionnaire e.g., answering "I am the life of the party" rather than grading "Extroverted, enthusiastic".

We are going to briefly introduce IPIP questionnaires. There are a total of fifty questions and ten questions for each trait. Moreover, there are positive-scored items and reverse-scored items of each trait. Users are asked to answer fifty items and rely on how they think the items are about themselves based on the scale of 1 to 5 where for the positive-scored items 1 = Disagree, 3 = Neutral and 5 = Agree, for the reverse-scored items 5 = Disagree, 3 = Neutral and 1 = Agree. Then, the mean values of the ten items makes up each trait. Currently, we simply measure whether the user is high or low on each trait *e.g.,* if a user's score on extroversion is higher than the mean score of all participants on extroversion, the user is regarded as extrovert. Otherwise, the user is an introvert. Therefore, we use binary trait labels for training our model. If a score of one trait is not less than the mean score of all participants on this trait, it is assigned the value 1. Otherwise, it is considered 0. Finally, a machine learning model are trained with the non-verbal features and the associated binary label.

C. Non-verbal Features

We use the experience of the previous research [19]. All features are briefly summarized in Table II. We define three visual features that include the user's head Motion, gaze score, upper body motion energy, as well as three vocal features such as pitch, energy, and MFCC. In this study, We limit our attention to the case of human-robot verbal interaction that relates to the communication between the participant and the robot. Therefore, each feature is observed and extracted while the participant or the robot is talking.

TABLE II
NON-VERBAL FEATURE REPRESENTATION

F1	Head Motion	HM1	Users move head while talking
		HM1$_b$	Binarized HM1
		HM2	Users move head while pepper is talking
F2	Gaze Score	GS1	Users' gaze score while talking
		GS1$_b$	Binarized GS1
		GS2	Users' gaze score while pepper is talking
F3	Motion Energy	ME1	Users move body while talking
		ME1$_b$	Binarized ME1
		ME2	Users move body while pepper is talking
F4	Pitch	Pn	Normalized pitch
		P$_b$	Binarized pitch
F5	Energy	En	Normalized energy
		E$_b$	Binarized energy
F6	MFCC	MFCCO	One of the thirteen MFCC vectors
		MFCCO$_b$	Binarized MFCCO

1) Head Motion

We use Pepper's camera to capture the images for face detection. The rotation invariant multi-view face detection method [20] is used to detect the user's face. In their research, all face data were taken from different viewpoints and categorized based on the variant appearance. Each view category was used to train the weak classifier which was configured as Look-Up-Table of Haar feature. The nested cascade face detector was constructed with these weak classifiers by real AdaBoost algorithm. A fast multi-view face detection system was also developed for making the face detector rotation invariant. Once a face is detected, the sub-window of the face is inputted to the weak classifier to calculate the user's head pose which is a 3D orientation including the yaw angle, pitch angle, and roll angle of the user's face. Then, we calculate the Manhattan distance of two adjacent head pose to represent the Head Motion.

2) Gaze Score

Once the position of the user's face was found, it is easy to detect the points of eyes. However, in our experiment, we did not use a high resolution camera for reducing the computational cost. And this makes it more challenging to detect the gaze when people were relatively far from the camera thus robot. At least 20 pixels of the face width in the image are needed for the face detection. In practice, each participant is required to sit in front of the camera within 2 meters by considering the resolution of the Pepper's camera. In such cases, the human eyes were represented by only a few pixels in the image. The gaze direction is briefly represented by the face plane in the form of yaw and pitch. In practice, it is hard to detect the face if the Tilt/Pan degree is more than +/- 20 degrees (0 deg means a face facing the camera). Therefore, the gaze score is measured between 0 and 1 by mapping the square sum of the yaw and pitch to the maximum Tilt/Pan degree.

3) Motion Energy

The motion energy is considered as another important visual feature which can be obtained from the whole HRI. The concept of the motion energy is the proportion of the different pixels of two successive frames that relative to the total number of pixels, which means the current frame need to be calculated how many pixels are different from the previous frame. Note that all three visual features will be normalized among all participants. Three visual features are shown as F1, F2, and F3, respectively, in Table II. The binarized visual features *HM1*$_b$, *GS1*$_b$, and *ME1*$_b$ are calculated by judging whether the values of *HM1*, *GS1*, and *ME1* are larger than 0 or not.

4) Pitch and Energy

The vocal features are important speech signals of each participant. Three well-known prosodic features used in this research are pitch, energy, and MFCC. Pitch is interpreted as the frequency of the sound by ear and brain. The frequency is produced by the vibration of vocal cords. If the frequency is high, human ears will interpret the sound as a high pitch, while the low frequency is perceived low pitch. The pitch and energy have been widely used in emotion estimation. For example, people will use a higher pitch and lower energy voice to talk when they feel fear. However, the actual audio signal is constantly changing. Therefore, usually the signal of a short period of time is simply considered as time invariant. The short term pitch sequences in the time domain [21] can be extracted by the auto-correlation function (given in Eq. 1) easily:

$$acf(\tau) = \sum_{i=1}^{N-1-\tau} s(i)s(i+\tau), (0 \leq \tau < N) \qquad (1)$$

where the $acf(\tau)$ is the auto-correlation function, the audio signal of one frame is represented by $s(i)$, N is the length of the current frame, τ is used as a delay.

The pitch of each frame is calculated by dividing the sampling frequency by the location of the second peak of auto-correlation functions (the first peak is when τ is 0). Then, we slide the window to the next frame until the end of the audio signal.

The normalized energy of each frame was calculated by using the following equation:

$$Energy = \frac{1}{N}\sum_{i=1}^{N} s(i)^2 \tag{2}$$

the $s(i)$ is one frame of the audio signal and N is the length of the current frame.

5) Mel-Frequency Cepstral Coefficient

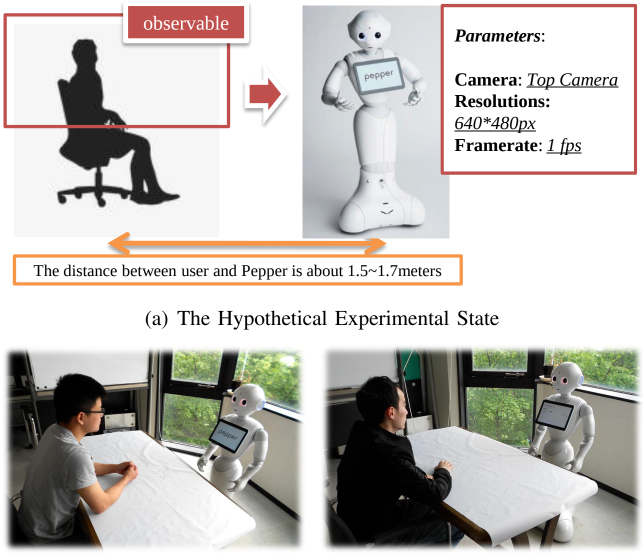
The incoming sound vibrated different spots of the cochlea to stimulate different nerves which informed our brain that some frequencies are present. Mel-Frequency Cepstral Coefficient (MFCC) [24] that is well known in the speech recognition area [25] was proposed based on this concept. We are going to briefly explain how to calculate the MFCC in the following.

Firstly, the audio signal is split into short-time frames and a window function such as the Hamming window is applied to each frame. Then, we calculate the periodogram-based power spectral by squaring the result (usually 512 points) of the Fast Fourier Transform on each frame and keep the first 256 coefficients. A set of triangular filters (the standard value is 26) are used to calculate the periodogram power spectral and we take the logarithm of the 26 energies. Finally, the Discrete Cosine Transform (DCT) is used to calculate the 26 log filterbank energies and keep the lower 13 coefficients.

MFCCO is one vector of the thirteen MFCC vectors. *m_MFCC* is the mean of the thirteen MFCC vectors. We normalize three vocal features pitch, energy, and each MFCC vectors among the whole dataset. The binary features of Pitch P_b and Energy E_b are calculated by estimating their trend *e.g.*, if the pitch of the current frame tends to increase or stable when it is compared to the previous frame, we assign 1. Otherwise, we assign 0. We generate the binary feature of *MFCCO* by judging the value whether it is greater than 0 or not. We assign 1 if the value is not smaller than 0. Otherwise, we assign 0.

III. Experimental Design

In the following, we will show you the outcomes of HRI experiments that includes 12 participants with the total speaking time of 2009 seconds. It would be good if the robots are able to understand its user's personality traits on short notice soon after a conversation started. On the other hand, we assume that the user's behaviors do not vary within the short notice. For these purposes, the visual and audio signal is divided into 30-sec long clips. Each clip is overlapped by 50%, which means that the first half of each clip is same as the last half of the previous clip. By doing this, we can generate more data to use for training the system. We designed short and simple conversations including four questions such as "Hello, I am Pepper, nice to meet you. Can you introduce yourself?", "It is nice weather today. What is the weather like in your hometown these days? Can you describe your hometown to me? I want to know more about you and your hometown.", etc. Four questions were saved into a text file. Pepper can say any string of characters by its NAOqi ALTextToSpeech API, which is based on speech synthesizers. During the experiment, Pepper only speaks in English. However, participants can speak any language they want, such as English, Italian, Chinese, etc. All participants were asked to sit about 1.5 to 1.7 meters away from Pepper. We used the Pepper's forehead camera whose resolution is 640 × 480 pixels, and the frame rate is set to 1 fps. Our reasoning is that if the frame rate is too high, the movements will be refined too much and the computational cost will increase drastically. If the frame rate is too low, some subtle movements will not be able to detect. Therefore, we set the frame rate to 1 fps. The upper body of each participant can be captured by the Pepper's camera. The hypothetical scenario and parameters can be seen in Fig. 2(a). All participants know the conversations before starting the experiment. Fig. 2(b) shows the real experimental environment.



(a) The Hypothetical Experimental State

(b) The Real Experimental Environment

Fig. 2. The Experimental Design

The framework of the feature extraction can be seen in Fig. 3. Note that if the participant pauses for 5 seconds, the robot will deem that the participant stopped talking, then it will start the next question. However, when Pepper tries to extract visual features, there must be no movement of Pepper's head. Since Motion Energy is calculated as a proportion of the moving pixels in the current frame, *e.g.*, if Pepper moves its camera, which causes the changes of the background, it may cause improper data collection. For this reason, in the current experimental setup, Pepper stayed still resulting that it seemed not very natural to participants.

The audio was recorded in 16,000 Hz by the microphone mounted on the front of Pepper's head. We process the audio separately after each participant completes his/her conversation with Pepper. It should be noted that the fan noise was also recorded in the audio. Therefore, we need to remove the noise before extracting vocal features. The last five seconds of the vocal information was removed as the participants stopped talking already. Similarly, we also remove the data of each visual feature that was extracted in the last five seconds, which aims to make sure that both visual and vocal features are synchronized.

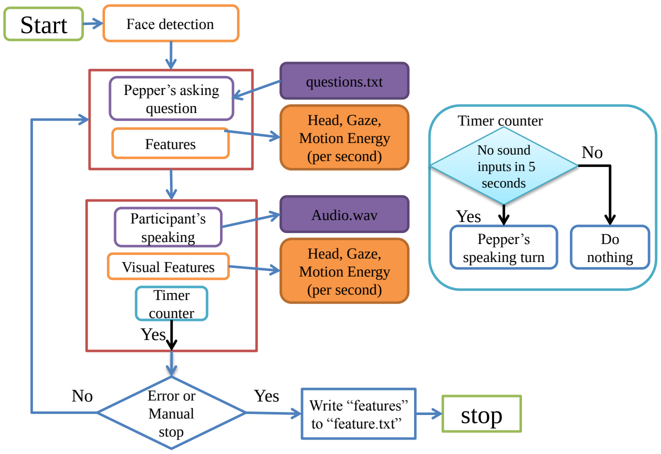


Fig. 3. The pipeline of Feature Extraction

In [26], they present the existing methods of previous studies for inferring an emergent leader such as SVM, Naive Bayes, Hidden Markov model, etc. In this paper, A regression model mentioned in [22], [23] is used to infer personality traits from non-verbal features. Leave-one-out method is used to evaluate the classification performance. Every time one feature is used for testing, and the rest of the features are used for training. The results are the ratio of the correct classified data of each feature. On the other hand, a cross-validation scheme is used to optimize the ridge parameters in the ridge regression model. The regression parameters are estimated by using:

$$\omega = (X^T X + \gamma I)^{-1} X^T y \tag{3}$$

where X is the feature matrix, γ is the ridge parameter, I is an identity matrix, and y is the binary score of the estimation result from questionnaires. We use the following equations to calculate the ridge parameter γ:

$$\gamma = e^{i-10} (i \in [1, 30], i \in N) \tag{4}$$

and the following equation embodied how we normalize all of the aforementioned features:

$$X = [F - E(F)]/Var(F) \tag{5}$$

where F is a non-verbal feature matrix or vector, *E(F)* is the mean of the matrix or vector F, and *Var(F)* is the variance of the matrix or vector F. We will evaluate the inferring performance by testing each non-verbal feature.

TABLE III

THE AVERAGED ACCURACIES OF BIG-FIVE PERSONALITY TRAITS

Average Accuracy	Extroversion	Agreeableness	Conscientiousness	Emotional Stability	Openness
HM1	0.5601	0.5739	0.5282	0.5693	0.5280
HM1$_b$	0.5289	0.5483	0.5455	0.5399	0.5335
GS1	**0.6363**	0.5087	0.5087	0.5298	0.5601
GS1$_b$	0.5951	0.5629	0.6364	0.5418	0.6391
ME1	0.5252	0.6961	0.6658	0.5554	0.5923
ME1$_b$	0.5151	0.6126	0.5695	0.5262	0.5455
Pn	0.5703	0.5142	0.5189	0.6033	0.5592
P$_b$	0.5400	0.5776	0.6446	0.5280	0.6281
En	0.5363	0.5611	0.6979	0.6612	**0.7053**
E$_b$	0.5473	0.5868	0.6446	0.5409	0.6281
MFCCO	0.5225	**0.8696**	**0.7594**	**0.7456**	0.6079
MFCCO$_b$	0.5629	0.6171	0.6694	0.5666	0.6574
m_MFCC	0.5751	0.6430	0.6141	0.6588	0.6219

IV. EXPERIMENTAL RESULTS

The performance of inferring user personality traits was evaluated by the classification accuracy of the test data on each personality trait as shown in Table III. The accuracy is presented by calculating the average of the accuracy over all folds. The first to the sixth rows showed the classification results of the six non-verbal features separately. Since we did not let Pepper talk very much, the classification results of the *HM2*, *GS2*, and *ME2* were excluded. We used thirteen MFCC vectors, the seventh row of Table III shows the mean accuracy of thirteen normalized MFCC vectors rather than binary MFCC vectors. The results of the sixth vector were used as *MFCCO*, and the fifth vector were used as *MFCCO*$_b$. The results of *MFCCO* provide the three highest accuracies of agreeableness, conscientiousness, and emotional stability out of thirteen MFCC vectors, and the accuracies of three traits are the highest among all non-verbal features. The results of the rest of the vectors are omitted due to their comparatively poor results.

We highlighted the highest accuracies of each trait with a bold font. From Table III, we can see that the visual feature of head motion and the vocal non-verbal feature pitch did not provide as accurate as the other features. The feature *ME1* works well in inferring Agreeableness and achieves the second best accuracy. The vocal feature *En* achieved the highest accuracy on Openness and the second highest accuracy on Conscientiousness. From Table III, it can be seen that the feature *MFCCO* achieved three highest accuracies on Conscientiousness, Emotional-Stability, and Openness. We noticed that the accuracies of all non-verbal features for inferring Extroversion are not as high as other traits. The highest accuracy achieved by *GS1* is 0.6363 which is regarded as low compared to the other traits. From the experimental results, it can be concluded that binary features do not help to improve the accuracy.

Extroverted people use lots of gestures and an energetic voice when they talk. During the experiment, all participants were asked to sit very close to the table, which might limit the movement of each participant. On the other hand, the engagement is crucial in HRI. The first interactions of participants and robot may not be very degage. Meanwhile, the feature motion energy also shows that few movements of each participant were detected. In such case, the visual features provide limited information of each participant, which could explain why the inferring performance of three visual non-verbal features is not very good. During the experiment, there is no movement of Pepper. Without movement, Pepper is considered as an introverted robot. According to [5], the engagement of people with an introverted robot is worse than that of people with an extroverted robot. Extroverted people may lose their interest in talking to Pepper.

V. Conclusions and Future Works

The proposed framework in this paper is used to infer the user's personality traits during human-robot face to face interactions. The user's non-verbal features were efficiently extracted from the robot's perceptual functions during social interaction. Summarizing the results of our extensive experiments, we are convinced that the proposed framework provides promising results toward a real-time implementation. Notably, the proposed usage of MFCC feature has proven its advantage in inferring personality traits. It can be said that Agreeableness is the easiest trait that can be inferred in HRI. On the contrary, Extroversion is a little bit harder to be inferred. The visual features did not always provide good results partly due to the limitation of the experimental environment. Moreover, it should be emphasized that the conversation topics and the robot's bodily behavior also need to be interesting. If the participants lose interest in talking to Pepper, their reactions may be unauthentic. For this reason, there is a need for a way to analyze the quality of engagement in HRI.

It should be also noted that personality traits can be accurately inferred from information obtained through long-term interactions. This work was not intended to learn and tell the user's personality traits on-the-fly. However, the work enables the robot to update its knowledge about the user's personality every time it interacts with the user. Therefore, it is expected that the accuracy of inferring user's personality traits tends to converge gradually through long-term interaction. There remains a technical problem with different types of features and/or poor accuracy. The current results were obtained from a different single non-verbal feature that we separately used. We plan to improve the accuracy further by fusing multi-modal features.

References

[1] Roberts, B. W. Back to the future: Personality and Assessment and personality development. Journal of Research in Personality, 43(2), 137-145, 2009.

[2] S.N. Woods, K. Dautenhahn, C. Kaouri, R.T. Boekhorst, K.L. Koay, and M.L. Walters. Are robots like people? Relationships between participant and robot personality traits in human-robot interaction studies. Interaction Studies, 8(2):281-305, 2007. 2

[3] C. Nass and M.K. Lee. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency attraction. Experimental Psychology: Applied, 7:171-181, 2001. 2, 4

[4] D. Windhouwer. The effects of the task context on the perceived personality of a Nao robot. In Proceedings of the 16th Twente Student Conference on IT, Enschede, The Netherlands, 2012.

[5] H. salam, O. Celiktutan, I. Hupont, H. Gunes, M. Chetouani, Fully Automatic Analysis of Engagement and Its Relationship to Personality in Human-Robot Interactions, IEEE Access 5(2016)705-721, https://doi.org/10.1109/ACCESS.2016.2614525

[6] Digman JM (1990). "Personality structure: Emergence of the five-factor model". Annual Review of Psychology. 41: 417440.

[7] A. Aly and A. Tapus, A Model for Synthesizing a Combined Verbal and Non-Verbal Behavior Based on Personality Traits in Human-Robot Interaction, in Proc. ACM/IEEE Int. Conf. Human-Robot Interaction, Mar. 2013, pp. 325-332.

[8] K. Isbister and C. Nass. Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. Human-Computer Studies, 53:251-267, 2000.

[9] T.F. Leary. The Interpersonal diagnosis of personality. Ronald Press, New York, USA, 1957.

[10] W. Revelle, K. R. Scherer. Personality, and Emotion. In: Oxford Companion to the Affective Sciences. Oxford University Press, Oxford, pp. 1-4, 2009.

[11] Celiktutan Oya, Sariyanidi Evangelos, Gunes Hatice. (2018). Computational Analysis of Affect, Personality, and Engagement in Human?Robot Interactions. University of Cambridge, United Kingdom. Computer Vision for Assistive Healthcare. 283-318. 10.1016/B978-0-12-813445-0.00010-1.

[12] J. W. Davis and A. F. Bobick. The representation and recognition of human movement using temporal templates. In Proc. of IEEE CVPR, pages 928-934. IEEE, 1997.

[13] D. Sanchez-Cortes, O. Aran, D. B. Jayagopi, M. S. Mast, and D. Gatica-Perez. Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. Journal on Multimodal User Interfaces, 7(1-2):39-53, 2013.

[14] A. Vinciarelli and G. Mohammadi, A survey of personality computing, IEEE TAC, 2014.

[15] D. Sanchez-Cortes, O. Aran, M. Mast, and D. Gatica-Perez. A non-verbal behavior approach to identify emergent leaders in small groups. Multimedia, IEEE Transactions on, vol. 14, No. 3, pp.816-832, 2012.

[16] S. D. Gosling, P. J. Rentfrow, and W. B. Swann. A very brief measure of the big-five personality domains. Journal of Research in Personality, vol.37, pp. 504-528, 2003.

[17] Goldberg LR (January 1993). "The structure of phenotypic personality traits". The American Psychologist. 48 (1): 26-34.

[18] Goldberg, Lewis R. The development of markers for the Big-Five factor structure. Psychological Assessment 4.1 (1992): 26.

[19] M. L. Knapp, J. A. Hall, and T. G. Horgan. Nonverbal Communication in Human Interaction. 8th ed. Boston: Wadsworth, Cengage Learning, 2008.

[20] B. Wu, H. Ai, C. Huang and S. Lao, Fast Rotation Invariant Multi-View Face Detection Based on Real Adaboost, IEEE Conf. on Automatic Face and Gesture Recognition, pp. 79- 84, 2004 .

[21] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. in Proc. of the Institute of Phonetic Sciences, vol. 17, no. 1193. Amsterdam, 1993, pp. 97-110.

[22] O. Aran and D. Gatica-Perez. One of a4913 kind: Inferring Personality Impressions in Meetings. In Proc.ACM International Conference on Multimodal Interaction (ICMI). pp. 11-18. 2013.

[23] S. Okada, O. Aran, and D. Gatica-Perez. Personality Trait Classification via Co-Occurrent Multiparty Multimodal Event Discovery. In Proc. ACM International Conference on Multimodal Interaction (ICMI),pp. 15-22. 2015.

[24] Min Xu; et al. (2004). "HMM-based audio keyword generation". In Kiyoharu Aizawa; Yuichi Nakamura; Shin'ichi Satoh. Advances in Multimedia Information Processing-PCM 2004: 5th Pacific Rim Conference on Multimedia (PDF). Springer. ISBN 3-540-23985-5. Archived from the original (PDF) on 2007-05-10.

[25] M. Sahidullah and G. Saha. Design, Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition. Speech Communication, Vol. 54, No. 4, 2012, pp. 543-565.

[26] Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. Prediction of the Leadership Style of an Emergent Leader Using Audio and Visual Nonverbal Features. IEEE Transactions on Multimedia 2017.