Understanding Nonverbal Communication Cues of Human Personality Traits in Human-Robot Interaction

Zhihao Shen, Armagan Elibol, and Nak Young Chong, *Senior Member, IEEE*

***Abstract*—With the increasing presence of robots in our daily life, there is a strong need and demand for the strategies to acquire a high quality interaction between robots and users by enabling robots to understand users' mood, intention, and other aspects. During human-human interaction, personality traits have an important influence on human behavior, decision, mood, and many others. Therefore, we propose an efficient computational framework to endow the robot with the capability of understanding the user's personality traits based on the user's nonverbal communication cues represented by three visual features including the head motion, gaze, and body motion energy, and three vocal features including voice pitch, voice energy, and Mel-Frequency Cepstral Coefficient (MFCC). We used the Pepper robot in this study as a communication robot to interact with each participant by asking questions, and meanwhile, the robot extracts the nonverbal features from each participant's habitual behavior using its on-board sensors. On the other hand, each participant's personality traits are evaluated with a questionnaire. We then train the ridge regression and linear support vector machine (SVM) classifiers using the nonverbal features and personality trait labels from a questionnaire and evaluate the performance of the classifiers. We have verified the validity of the proposed models that showed promising binary classification performance on recognizing each of the Big Five personality traits of the participants based on individual differences in nonverbal communication cues.**

***Index Terms*—human-robot interaction, nonverbal communication cues, personality traits, machine learning.**

I. Introduction

With the population aging and sub-replacement fertility problems increasingly prominent, many countries have started promoting robotic technology for assisting people toward a better life. Various types of robotic solutions have been demonstrated to be useful in performing dangerous and repetitive tasks which humans are not able to do, or do not prefer to do. In relation to elderly care provision, assistive robots could replace and/or help human caregivers support the elderly socially in their home or residential care environments.

Researchers gradually realized that the interactions between a human user and a robot are far more than sending commands to the robot or reprogramming, as a new class of social robots are emerging in our daily life. It is now widely understood that not only the robot's appearance but also its behaviors are important for human-robot interaction [1] [2]. Therefore, synchronized verbal and nonverbal behaviors [3] were designed and applied to a wide variety of humanoid robots, like Pepper, NAO, ASIMO, and many others, to improve the user's engagement in human-robot interaction. For instance, the Honda ASIMO robot can perform various movements of arms and hands including metaphoric, iconic, and beat gestures [4]. Likewise, some researchers have designed such gestures using the SAIBA framework [5] for the virtual agents. The virtual agents were interfaced with the NAO robot to model and perform the combined synchronized verbal and nonverbal behavior. In [6], the authors tested the combined verbal and nonverbal gestures on a 3D virtual agent MAX to make the agent act like humans. Meanwhile, cultural factors are also considered to be crucial components in human-robot interaction [7]. In [8], the authors designed emotional bodily expressions for the Pepper robot and enabled the robot to learn the emotional behaviors from the interacting person. Further investigations on the influence of the robot's nonverbal behaviors on humans were conducted in [9]. These efforts were made to enable robots to act like humans. However, the synchronized behaviors are unilateral movements with which robots track the person's attention. Therefore, the authors in [10] claimed that social robots need to act or look like humans, but more importantly they will need to be capable of responding to the person with the synchronized verbal and nonverbal behavior based on his/her personality traits. Inspired by their insight in [10], we aim to develop a computational framework that allows robots to understand the user's personality traits through their habitual behavior. Eventually, it would be possible to design a robot that is able to adapt its combined verbal and nonverbal behavior toward enhancing the user's engagement with the robot.

A. Why are the personality traits important during the interaction?

In [11], the authors investigated how personality traits affect humans in their whole life. The personality traits encompass relatively enduring patterns of human feelings, thoughts, and behaviors, which make each different from one another. When the human-human conversational interaction is considered, the speaker's behavior is affected by the speaker's personality traits, and the listener's personality traits also affect their attitude toward the speaker. If their behaviors make each other

*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. The authors are also grateful for financial supports from the Air Force Office of Scientific Research (AFOSR-AOARD/FA2386-19-1-4015).

feel comfortable and satisfying, they would enjoy talking to each other. In social science research, there have been different views toward the importance of interpersonal similarity and attraction. Some people tend be attracted to other people with similar social skills, cultural background, personality, attitude, and several others [12] [13]. Interestingly, in [14], the authors addressed the complementary attraction that some people prefer to talking with other people whose personality traits are complementary to themselves. Therefore, we believe that if the robot is able to understand the user's coherent social cues, it would improve the quality of human-robot interaction, depending on the user's social behavior and personality.

In previous studies, the relationships between the user's personality traits and the robot's behavior were investigated. It was shown in [16] that humans are able to recognize the personality of the voice that was synthesized by the digital systems and computers. Also, a compelling question was explored to better understand the personality of people whether they are willing to trust a robot or not in an emergency scenario in [15]. Along the lines, a strong correlation between the personality traits of users and the social behavior of a virtual agent was presented in [17]. In [18], the authors designed the robot that have personalities to interact with a human, where significant correlation between human and robot personality traits were revealed. Their results showed how the participants' technological background affected the way they perceive the robot's personality traits. Also, the relationship between the profession and personality was investigated in [19]. The result conforms with our common sense such as that doctors and teachers tend to be more introverted, while managers and salespersons tend to be more extroverted. Furthermore, the authors investigated how humans think about the NAO robot with different personality traits (Introversion or Extroversion), when the robot plays different roles in human-robot interaction [20]. However, their results were not in accordance with our common sense. The robot seems smarter to the human when the robot acted as an introverted manager and extroverted teacher. On the contrary, the extroverted manager and introverted teacher robots were not perceived intelligent by the participants. These two results conflict with each other. This could be due to the fact that people treat and perceive robots differently than humans in the aforementioned settings. Another reason could be that the introverted manager robot looked like more deliberate, because it took more time to respond, while the extroverted teacher robot looked like more erudite, because it took less time to respond during the interaction. Even though these two studies found conflicting results, the results imply the importance of robot personality traits in designing professional roles for human-robot interaction.

In light of the previous studies on personality match in human-robot interaction, some of the findings are inconsistent with each other. The result that was shown in [21] indicated that the participants enjoyed interacting more with the AIBO robot when the robot has a complementary personality to the participants'. While the conclusions from [22] showed that the participant was more comfortable when they interacted with the robot with a similar personality to theirs. Similarly, the engagement and its relation to the personality traits were analyzed during human-robot interaction in [23], where the participants' personality traits played an important role in evaluating individual engagement. The best result was achieved when the participant and robot both were extroverted. Note that when both the participant and the robot were introverted, the performance was the worst. Although the complementary and similar attraction theory may need further exploration in the future, these studies clearly showed that how the personality traits are important in human-robot interaction.

On the other hand, the personality traits have been shown to have a strong connection with the human emotion. In [26], it was discussed that how the personality and mind model influence the human social behavior. A helpful analogy for explaining the relationship between personality and emotion is "personality is to emotion as the climate is to weather" [27]. Therefore, theoretically, once the robot is able to understand the user's personality traits, it would be very helpful for the robot to predict the user's emotion fluctuation.

Fig. 1 illustrates our final goal by integrating the proposed model of inferring human personality traits with the robot's speech and behavioral generation module. The robot will be able to adjust its voice volume, speed, and body movements to improve the quality of human-robot interaction.

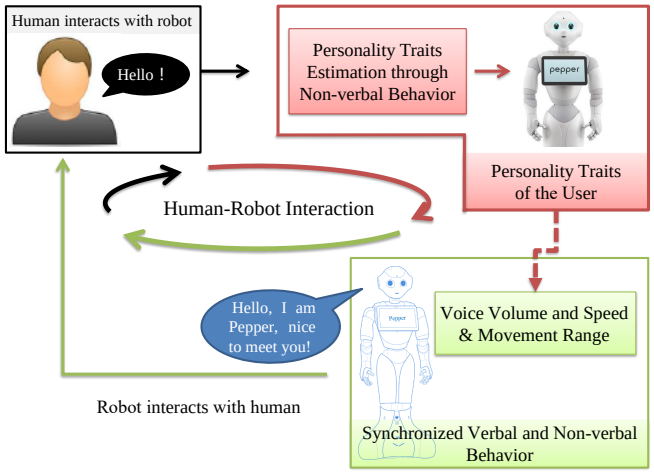


Fig. 1. Integrating Proposed Model of Inferring Human Personality Traits into Robot Behavior Generation

B. Architecture for Inferring Personality Traits in Human-Robot Interaction

In the sub-section above, the importance of the personality traits in human-human and human-robot social interactions is clearly stated. Here we propose our computational framework for enabling the robot to recognize the user's personality traits based on their visual and vocal nonverbal behavior cue. This paper is built upon our preliminary work in [32].

In this study, the Pepper robot [25] equipped with two 2D cameras and four microphones interacts with each participant. In the previous research on the emergent LEAder corpus (ELEA) [33], when recording the video of a group meeting, the camera was set in the middle of the desk to capture

each participant's facial expression and upper body movement. [23] also used the external camera to record the individual and interpersonal activities for analyzing the engagement of human-robot interaction. However, we do not use any external devices for the two reasons; First, we attempt to make sure that all audio-visual features are captured from the first-person perspective, ensuring that the view from the robot is closely similar to that from the human. Secondly, if the position and pose of the external camera changes for some reasons, it would yield a significant difference between the visual features. Thus, we use the Pepper's forehead camera only.

Fig. 2 briefly illustrates our experimental protocol which consists of the nonverbal feature extraction and the machine learning model training: a) All participants recruited from the Japan Advanced Institute of Science and Technology were asked to communicate with the Pepper robot. The robot keeps asking questions related to the participant, and each participant answers the questions. The participants are supposed to reply to the robot's questions with their habitual behavior. Before or after each participant finished interacting with the robot, they were asked to fill out a questionnaire to evaluate their personality traits. The personality traits scores were binarized to perform the classification task. b) We extracted the participants' audio-video features that include the head motion, gaze, body motion energy, voice pitch, voice energy, and MFCC during the interaction. c) The nonverbal features and personality traits labels will be used to train and test our machine learning models.

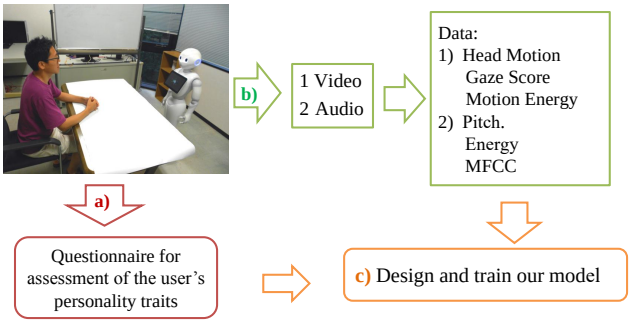


Fig. 2. Experimental Protocol for Inferring Human Personality Traits

To the best of our knowledge, this is the first work that shows how to extract the user's visual features from the robot's first-person perspective, as well as the prosodic features, in order to infer the user's personality traits during human-robot interaction. In [24], the non-verbal cues were extracted from the participant's first-person perspective and used to analyze the relationship between the participant and robot personalities. With our framework, the robot is endowed with the capability of understanding human personalities during face-to-face interaction. Without using any external devices, the proposed system can be conveniently applicable to any type of environment.

The rest of this paper is organized as follows. Section II explains the personality traits model used corresponding to Part a. Section III explains why we used the nonverbal features, and what nonverbal features were used for recognizing the participant personality traits corresponding to Part b. Section IV presents the technical details of our experiments. Section V is devoted to experimental results and analysis corresponding to Part c. Section VI draws conclusions.

II. Personality Traits

Based on the definition in [34] [35], personality traits have a strong long-term effect in generating the human's habitual behavior: "the pattern of collective character, behavioral, temperamental, emotional, and mental traits of an individual that has consistently over time and situations".

In the most of existing studies on personality traits, the researchers proposed many different personality models including Meyers-Briggs (Extroversion-Introversion, Judging-Perceiving, Thinking-Feeling, and Sensation-Intuition) [28]; Eysenck Model of Personality (PEN) (Psychoticism, Extroversion, and Neuroticism) [29]; and the Big-Five personality model (Extroversion, Openness, Emotional Stability, Conscientiousness, Agreeableness) [30] [31]. The Big-Five personality traits are the very common descriptor of human personality in psychology. In [36] [37], the authors investigated the relationship between the Big-Five personality traits model and nonverbal behaviors. We also use the Big-Five personality traits model in this study. Table I denotes the intuitive expressions for the Big-Five personality traits.

TABLE I
Big-Five Personality Traits

Big-Five	High on this trait	Low on this trait
Extroversion	Enjoy meeting new people Like being attention center Easy to make new friends Has a wide social circle	Prefer solitude Dislike being attention center Think things through Do not talk much
Agreeableness	Care about others Prefers to cooperate Enjoy helping others Kind and compassionate	Do not interest in others Manipulates others frequently Insult and belittle others Competitive and stubborn
Conscientiousness	Keep things in order Pay attention to details Enjoy having a schedule Goal- and detail-oriented	Make messes Do not take care of things Delay to finish tasks Less detail-oriented
Emotional Stability	Do not worry much Deal well with stress Rarely feel depressed Emotionally stable	Worry about many things Experience a lot of stress Get upset easily Appears anxious or irritable
Openness	Enjoy tackling challenges Like abstract concepts Open to trying new things	Do not enjoy new things Resist new ideas Not very imaginative

As the personality traits become more popular in the last few decades [38], various questionnaires were proposed in the literature for the assessment of human Big-Five personality traits. The most popular format of questionnaire is the Likert scale: Ten Item Personality Inventory (TIPI) which has 10-items and each question is on a 7 point scale [39]; The Revised NEO Personality Inventory (NEO PI-R) which contains 240 items [40]; the NEO Five-Factor Inventory (NEO-FFI), a shortened version of NEO PI-R, which comprises 60 items [41]; and

the International Personality Item Pool (IPIP) Big-Five Factor Markers which has been simplified to 50 questions [42]. We used the IPIP questionnaire in this paper, and all participants were asked to fill out the questionnaire to evaluate their Big-Five personality traits. The IPIP questionnaire is relatively easier to answer, and it does not need too much time to complete.

Specifically, the participants are asked to rate the extent to which they agree/disagree with the personality questionnaires on a five-point scale. A total of 50 questions are divided into ten questions for each of the Big-Five traits and the questions also include the reverse-scored and positive-scored items. For the reverse-scored items, Strongly Disagree equals 5 points, Neutral equals 3 points, and Strongly Agree equals 1 point; for the positive-scored items, Strongly Disagree equals 1 point, Neutral equals 3 points, and Strongly Agree equals 5 points. After the participants rate themselves for each question, each personality trait is represented by the mean score of 10 questions. We did not use the scale of 1-5 to represent the participant's personality traits. Instead, the personality traits are binarized using the mean score of all participants as a cut-off point to indicate whether the participant has a high or low level of each of the Big-Five traits. For instance, if a participant's trait of extroversion was rated 2 which is less than the average value 2.8, then, this participant is regarded as introvert and his/her trait score will be re-assigned 0. Then, we used the binary labels to train our machine learning models and evaluate the classification performance accordingly.

III. Feature Representation

It is known that the personality trait encompasses the human's feeling, thoughts, and behaviors. The question to be investigated then arises as "how can it be inferred human personality traits based on their verbal and nonverbal behaviors?"

A. Related Work on Verbal and Nonverbal Behaviors

The influences of personality traits on linguistic speech production have been addressed in previous works [43] [44]. The user's daily habits were investigated to ascertain whether they are related to the user's personality traits. The changes of facial expression were also used to infer the personality traits, which was proposed in [54]. In [49], the participants were asked to use the Electronically Activated Recorder (EAR) to record their daily activities, which included locations, moods, language, and many others, to verify the manifestations of personality. Moreover, the authors investigated how the writing language reflects the human personality style based on their daily writing diaries, assignments, and journal abstracts [45]. More specific details were presented in [46]. In that study, two corpora that contain $2,479$ essays and $15,269$ utterances more than 1.9 million words were categorized and used to analyze the relation to each participant's Big-Five personality traits. Although the participant's verbal information can be used to analyze their personality traits based on Pennebaker and King's work [45], it should be noted that categorizing so many words would be an arduous task. In [47], the authors addressed that the language differences could influence the annotator's impressions toward the participants. Therefore, they asked three annotators to watch the video that was recorded in the meeting without audio and to annotate the personality traits of each participant. Notably, the issue of conversational error was addressed in [48], where the error caused the loss of trust in the robot during human-robot interaction. In light of the aforementioned studies, the participants in our study were free to use any language to talk with the robot. It can generally be said that the nonverbal behavior would be a better choice in this study.

On the other hand, it is desirable that the robot can change its distances with the user depending on a variety of social factors leveraging a reinforcement learning technique in [50]. In [51], the author also used the changes in the distance between the robot and the participant as one of their features for predicting the participant's extroversion trait. Similarly, the authors proposed a model of automatic assessment of human personality traits by using body postures, head pose, body movements, proximity information, and facial expressions [52]. The results in [53] also revealed that the extrovert could accept people to come closer than the introvert. However, the proxemics feature was not considered in our study, as the human-robot distance remains unchanged during our communicative interaction settings.

In the related research on inferring human personality traits, a variety of fascinating multimodal features were proposed. In [36] [55], the authors used vocal features to infer personality traits. In [37], they used vocal and simple visual features to recognize the personality traits based on MS-2 corpus (Mission Survival 2). [47] [56] detailed how to infer personality traits in the group meeting. They used the ELEA corpus, and the participant's personality traits were annotated by the external observer. Meanwhile, the participant's vocal and visual features such as voice pitch, voice energy, head movement, body movement, and attentions were extracted from audio and videos. The similar features were used in [57] to infer the personality traits with Youtube video blogs. The convolutional neural networks were also applied to predict human personality traits based on an enormous database that contains video, audio, and text information from YouTube vlogs [58] [59]. In [60] [61], the authors explained a nonverbal feature extraction approach to identifying the emergent leaders. The nonverbal features that were used to infer the emergent leaders included prosodic speech feature (pitch and energy), visual features (head activity and body activity), and motion template-based features. In [77] [70], the frequently-used audio and visual nonverbal features in existing research were summarized for predicting the emergent leader or personality traits. Similarly, a method was proposed in [69] for identifying the human's confidence during human-robot interaction with the sound pressure, voice pitch, and head movement.

In the previous studies [47] [37] [60] [62], the authors used the statistical features and activity length features. Since the personality traits are long-term characteristics that affect people's behaviors, they believed that the statistical features can well represent the participants' behaviors. Similar nonverbal features were used in our study. However, we believe that the state transitions of the nonverbal behaviors or features are also

importance to understand the human's personality traits. The study in [56] proposed their co-occurrent features to indicate some movements of other participants that happened at the same time. Hence, in our study, the raw form and time-series based features of the visual and vocal nonverbal behavior were used to train the machine learning models.

B. Nonverbal Feature Representation

Taking into account the findings of the aforementioned studies, we intend to extract similar features from the participant's nonverbal behaviors. Nonverbal behaviors include vocal and visual behaviors. Table II shows the three visual features including the participant's head motion, gaze score, and upper body motion energy, as well as the three vocal features including the voice pitch, voice energy, and Mel-Frequency Cepstral Coefficient (MFCC).

In our basic human-robot interaction scenario, it is assumed that the participant talks to a robot using gestures the way a person talks to a person. Therefore, the participant's visual features can be extracted using the robot's on-board camera while the participant or the robot talks. Note that, in Table II, some of the visual features $HM2$, $GS2$, and $ME2$ are extracted when the participant listens to the robot asking four simple questions. The total time duration was too short to capture sufficient data enough to train our machine learning models. Therefore, we did not use these three features in our study.

TABLE II
NONVERBAL FEATURE REPRESENTATION

Activity	Abbreviation	Description
Head Motion	HM1	Users move head while they are talking
	$HM1_b$	Binarized HM1
	HM2	Users move head while pepper is talking
Gaze Score	GS1	Users' gaze score while they are talking
	$GS1_b$	Binarized GS1
	GS2	Users' gaze score while pepper is talking
Motion Energy	ME1	Users move body while they are talking
	$ME1_b$	Binarized ME1
	ME2	Users move body while pepper is talking
Pitch	Pn	Normalized pitch
	Pn_b	Binarized pitch
Energy	En	Normalized energy
	En_b	Binarized energy
MFCC	MFCCO	One of the 13 MFCC vectors
	$MFCCO_b$	Binarized MFCCO
	m_FCCO	The average vector of the 13 MFCC vectors

1) Head Motion: An approach to analyze the head activity was proposed in [60]. They applied the optical flow on the detected face area to decide whether the head was moving or not. Based on the head activity states, they were able to understand when and for how long the head moved. We followed the method that was proposed in [63]. First, every frame captured by the Pepper's forehead camera was used for scanning procedure to extract the sub-windows. The authors in [63] has trained 60 detectors based on left-right rotation-out-of-plane and rotation-in-plane angle, and each detector contains many layers that are able to estimate the head pose and detect a human face. Each sub-window was used as an input to each detector which was trained by a set of the face with a specific angle. The output would provide the 3D head pose (pitch, yaw, and roll) as shown in the left image of Fig. 3. In this study, the pitch angle covers [-90°, 90°], the roll angle covers [-45°, 45°], and yaw angles covers [-20°, 20°]. And then the Manhattan distance of every two adjacent head angle was used to represent the participant's head motion. Let α, β, and γ denote the pitch, yaw, and roll angles, respectively. Then the head motion (HM1) can be calculated by the following equation:

$$HM1_{(i)} = |\alpha_{(i)} - \alpha_{(i+1)}| + |\beta_{(i)} - \beta_{(i+1)}| + |\gamma_{(i)} - \gamma_{(i+1)}|, \quad (1)$$

where i and $i+1$ are two consecutive frames at 1 sec time interval.

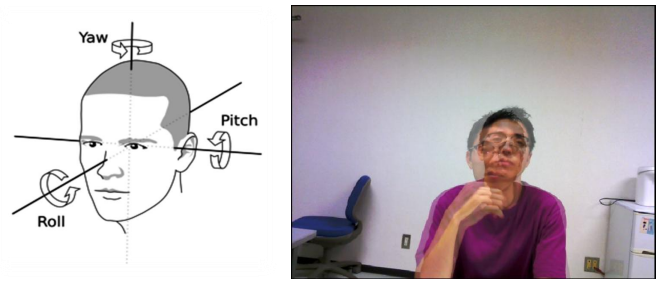


Fig. 3. Visual Features (The left image illustrates the 3D head angles, and the right image shows the different pixels by overlapping two consecutive frames)

2) Gaze Score: In [62], the influence of gaze in the small group human interaction was investigated. The previous studies used the visual focus of attention (VFOA) to represent the participant's gaze direction [61] in the group discussion. However, the high-resolution image is required for the analysis of the gaze direction, which will tremendously increase the computational cost. In our experiment, the participant sits at a table in front of the robot positioned 1.5 to 1.7 m away. In practice, the calculation of gaze direction might not be feasible, if we consider the image resolution and the distance, since the eye occupies only a few pixels in the image. As the head pose and gaze direction are highly related with each other [64], an efficient way of calculating the gaze direction was proposed based on the head orientation in [65]. Therefore, we used the head direction to estimate the gaze direction which is highly related to the head yaw and pitch angles. In the real experimental environment, we found that the face was hardly detected when the facial plane exceeds $\pm 20°$. When the participant faces the robot's forehead camera, the tilt/pan angle is $0°$. Therefore, we measure the Euclidean distance from the $0°$ to the head yaw and pitch angle. Then, the full range (distance) of tilt/pan angles [0°, 20°] is normalized to 0

to 1. Finally, the normalized score between 0 and 1 is used as the gaze score which indicates the confidence in the fact that the participant is looking at the robot. If we denote by α and β the head pitch and yaw angles, respectively, the gaze score of the frame i can be calculated by the following equation:

$$GS1_{(i)} = 1 - \sqrt{\frac{\alpha_{(i)}^2 + \beta_{(i)}^2}{{\alpha_{max}}^2 + {\beta_{max}}^2}}, \tag{2}$$

where α_{max} and β_{max} represent the maximum degree of the head pitch and yaw angle, respectively.

3) Motion Energy: The motion energy images [66] [67] were used in the previous studies to describe body motion. Their basic idea is to compute the number of different pixels of every two consecutive frames. We applied the same idea to calculate the ratio of the different pixels between every two frames. The right image of Fig. 3 shows an example of different pixels between two frames. This method is simple and effective. However, it requires the image to have stationary background and distance between the robot and each participant. Otherwise, the change of the background will be perceived as the participant's body movement, and the number of different pixels will increase if the participant sits closer to the robot. Now, all three visual features were calculated and normalized in the whole database, denoted by $HM1$, $GS1$, and $ME1$. The binary features $HM1_b$, $GS1_b$, and $ME1_b$ mentioned in Table II are the binarized $HM1$, $GS1$, and $ME1$ which were simply calculated by comparing whether the value is larger than 0 or not.

4) Voice Pitch and Energy: The vocal behavior is another important feature when humans express themselves. Pitch and energy are the two well-known vocal features and very commonly used in emotion recognition. Pitch, which is generated by the vibration of vocal cords, is perceived as $F0$ the fundamental voice frequency. There are many different methods to track the voice pitch. For instance, AMDF (Average Magnitude Difference Function [71]), SIFT (Simple Inverse Filter Tracking [72]), and ACF (Auto-correlation Function [73]) are the time domain approach, while HPS (Harmonic Product Spectrum [74]) is the frequency domain approach. We used the auto-correlation function denoted by $acf(\tau)$ given in Eq. 3 to calculate pitch:

$$acf(\tau) = \sum_{i=1}^{N-1-\tau} s(i)s(i+\tau), (0 \leq \tau < N), \tag{3}$$

where $s(i)$ is the audio signal of each frame, τ is the time delay, and N is the frame size.

Using the $acf(\tau)$ function, we divided the sampling frequency by the index number of the second peak to calculate the pitch of each frame. Generally, the audio signal of each frame that was used to extract vocal feature contains more than two periods, and the pitch range of a human's voice is higher than $50Hz$. For an audio file whose sampling frequency is $16,000Hz$, we can calculate the range of the frame size N based on the following equation.

$$\frac{16,000}{50} \leq \frac{N}{2}, \tag{4}$$

$$N = 16000 \times T. \tag{5}$$

In Eq. 5, T is the time duration of the audio signal in one frame. Since the frame size N used in this study is 800, the time duration T is 50 milliseconds.

Now the average of the short-term energy can be calculated by the following equation:

$$Energy = \frac{1}{N}\sum_{i=1}^{N} s(i)^2, \tag{6}$$

where $s(i)$ is the audio signal of each frame, and N is the frame size.

5) Mel-Frequency Cepstral Coefficient: Mel-Frequency Cepstral Coefficient (MFCC) [75] is a vocal feature well known for its good performance in speech recognition [76]. The procedures to calculate MFCC are highly related to the vocalism principle and also able to discard the redundant information that the voice carries, *e.g.*, the background noise, emotion, and many others. We intend to test this pure and essential feature which reflects how the sound was generated. We calculated the MFCC based on the following steps;

First, we calculate the power spectrum by calculating the Fast Fourier transform (FFT) of each frame. The motivating idea is from the concept of how our brain understands the sound. The cochlea in the ear converts sound waves, which caused the vibrations in different spots, to the electrical impulses to inform the brain that some frequencies are present. Usually, only 256 points were kept from 512 points in FFT.

Then, 20-40 (usually 26) triangular filters of the Mel-spaced filterbank were applied to the power spectrum. This step is to simulate how the cochlea perceives the sound frequencies. The human ear is less sensitive to the closely spaced frequencies, and it becomes even harder when the frequency is increasing. This is why the triangular filter becomes wider as the frequency increases.

Third, the logarithm was applied to the 26 filtered energies. This is also motivated by human hearing. We need to put 8 times more energy to double the loudness of the sound. Therefore, we used the logarithm to compress the features much closer to what humans actually hear.

Finally, we compute the discrete cosine transform (DCT) of the logarithmic energies. In the previous step, the filterbanks were partially overlapped, which provide high correlated filtered energies. The DCT was used to decorrelate the energies. Only 13 coefficients were kept as the final Mel Frequency Cepstral Coefficients.

Now we have 13 MFCC features vectors. Each feature was used to train and test our machine learning models. As shown in the Table II, $MFCCO$ is one of the MFCC vectors. And the m_MFCC is the average vector of all 13 MFCC vectors. The three vocal features were normalized among the whole dataset. As for the binarized pitch and energy, Pn_b and En_b were calculated by estimating the trend of Pb and Eb. *e.g.*, if the pitch of frame i is greater or equal to the value of

frame $i-1$, we assign 1. Otherwise, we assign 0. While the binarized feature $MFCCO$ is different, we again pay attention to whether the value is greater than 0 or not.

We normalized the features by the following equation:

$$X = \frac{F - E(F)}{Var(F)}, \tag{7}$$

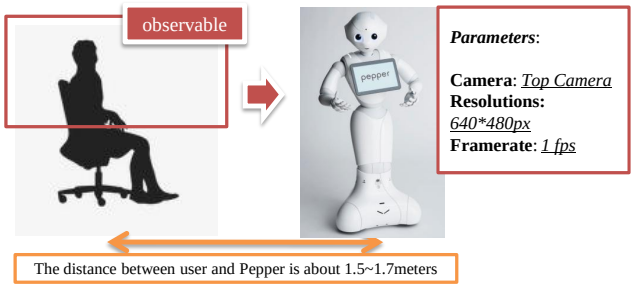
where X is the normalized feature vector, F is the raw form feature, $E(F)$ is the mean value of the raw form feature, and $Var(F)$ is the variance of the raw form feature.

IV. Experimental Design

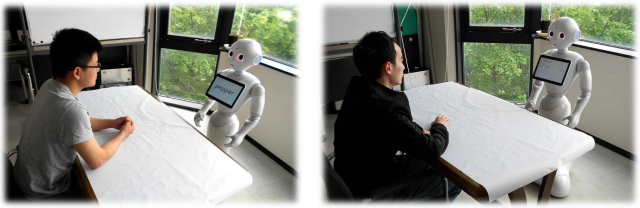
The experiment was designed in the scenario that the robot asks questions as the robot meets with the participant. In the following, we introduced the experimental environment and the machine learning methods used.

A. Experimental Setup

The relationship between people's professions and personality traits was investigated in [19]. In our study, all the participants were recruited from the Japan Advanced Institute of Science and Technology. Therefore, the relationship between professions and personality traits was not considered. On the other hand, the interactions between participants and the robot were assumed to be casual everyday conversations. Specifically, each participant sits at a table with his/her forearm resting on the tabletop and talks with the robot. The participants did not have any strenuous exercises before they were invited to the experiment.



(a) Illustrative Diagram of Experimental Setup



(b) Snapshots of Real Experiments

Fig. 4. Details of Experimental Setup

The experimental setup is shown in Fig. 4. Each participant was asked to sit at a table in front of the robot standing 1.5 to 1.7 m away in a separate room. Only the upper part of the participant's body was observable from the robot's onboard camera that extracts the visual features. The robot keeps asking questions one after another. The participant was asked to respond to each question using his/her habitual gesture. As mentioned in Section III, the participants were free to use any language (such as English, Italian, Chinese, and Vietnamese) to communicate with the robot.

Fig. 4 (a) shows some parameters used in the experiment. The top camera of the robot in the middle of the forehead was used, which makes the view of the robot very similar to that of humans. The resolution of the video camera is 640×480 pixels. If the frame rate is too high, it may provide too many subtle movements. On the contrary, if the frame rate is too low, the subtle movements will be hard to detect. We used one frame per second. Participants were aware of all the questions they will receive, such as "Hello, I am Pepper, nice to meet you. Can you introduce yourself?", and many others. Therefore, they would have time to prepare the answers. The audio was recorded with a microphone on the robot's head with 16,000 Hz. Before starting to extract the vocal features, the robot's fan noise was removed from the audio.

We recruited 15 participants in the study; however, 3 of the participants were too nervous during the experiment, and they looked at the experimenter frequently. Therefore, they were excluded, and our database contains the data of 12 participants with a total duration 2,000 sec. One of the convenient ways to infer personality traits is using the fixed time length. Once the robot has enough data, it would be able to infer the personality traits. Therefore, we divided the data into 30-sec long clips. The 30-sec clip may contain data from different sentences. When we divided the clips, each clip has 50% overlap with the previous one, and then we were able to generate more data generalized.

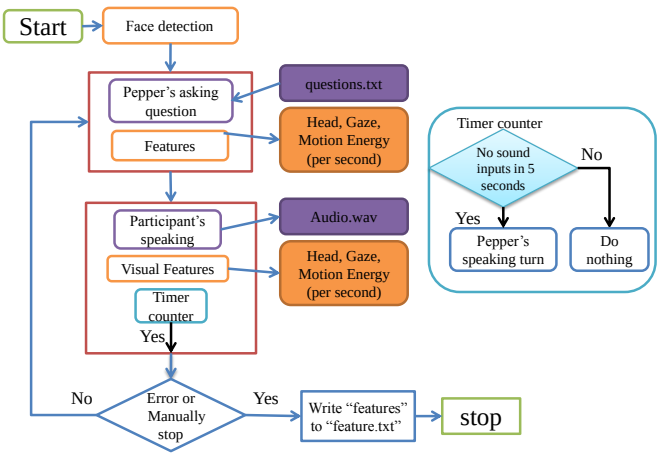


Fig. 5. The Pipeline for Feature Extraction

The flowchart in Fig. 5 shows the architecture for extracting features. The robot first detects whether there is a person to talk to. Then the robot would sequentially select a question from a file "questions.txt" and use the speech synthesizers to start the conversation. Meanwhile, the robot also extracted the visual features every second. Even after the robot finished asking its question, the vocal and visual features extraction would be continued while the participant was responding to the question. The participant was instructed that they were

expected to stop talking for 5 seconds for letting the robot know that it may ask the next question.

B. Classification Model

In [77], the authors summarized different methods used for the prediction of the leadership style such as the Logistic regression [47] [56] [78], Rule-based [79], Gaussian mixture model [80], and Support vector machine [47] [56] [81]. The ridge regression and linear SVM were both used in [47] [56]. We opted to apply the same methods in our study to make a simple comparison. The cross-validation was used to find the optimal regression parameters. The following formulas were used to calculate the regression parameters:

$$\omega = (X^T X + \gamma I)^{-1} X^T y, \tag{8}$$

where X is the feature matrix, I is an identity matrix, y is the binarized label of the personality traits, and γ is the ridge parameter calculated using the following equation:

$$\gamma = e^{i-10} (i \in [0, 29], i \in \mathbb{N}). \tag{9}$$

In Eq. 9, i is an integer indicating that each regression model is executed for 30 times for optimizing the regression parameter ω. As we used the regression model to perform a classification task, we used the accuracy rather than the mean squared error, which would give more meaningful results.

SVM is used to perform the linear or nonlinear classification task by using different types of kernel functions. It requires a longer time to train an SVM classifier than ridge regression. From Table III, it can be noticed that the binary features did not present their advantages in ridge regression. Therefore, the binary features were discarded in SVM. Then, we trained an SVM classifier with the linear kernel with the penalty parameter of the error term which was chosen from [0.1, 0.4, 0.7, 1]. Therefore, each SVM classifier was trained for 4 times based on the equation that was mentioned in [82] which is shown in the following equation.

$$y(x) = \sum_{m=1}^{M} a_m y_m \mathcal{K}(x, x_m) + b, \tag{10}$$

where $y(x)$ is the predicted sign of the testing sample x, and a_m is a set of Lagrange multipliers. This model was trained by the training data x_m with the corresponding label y_m. b is the bias parameter, and $\mathcal{K}(x, x_m)$ is the linear kernel function. The linear kernel function can be described as the following equation:

$$\mathcal{K}(x_i, x_j) = x_i^T x_j, \tag{11}$$

where x_i and x_j are two data samples.

The leave-one-out method was used to evaluate the performance of ridge regression and linear SVM. The results of the linear SVM was presented in Table IV.

V. Experimental Results

A. Classification Results

Tables III and IV showed the results of the ridge regression and linear SVM classifier that were trained by each feature corresponding to each personality trait, respectively. The best result of each trait was shown in bold. It is obvious that the linear SVM classifier outperforms the ridge regression on four traits, except the results of Extroversion. As shown in Table III, the highest accuracies of five traits were acquired by $GS1$, En, and the 6-th $MFCC$. These three features were concatenated as the combined feature and used to train five more ridge regression models for the five traits. We used the abbreviation FRR to represent the combined feature for ridge regression. Similarly, we trained five more SVM classifiers for each trait by concatenating $HM1$, $GS1$, $ME1$, Pn, and the 6-th $MFCC$ features. The abbreviation $FSVM$ was used to represent the combined feature for SVM. The details about the classification results were given in the following two tables.

TABLE III
Averaged Accuracies for Big Five Personality Traits (Ridge Regression Classifier)

Trait / Feature	**Extro-version**	**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
FRR	0.6243	0.8641	0.6082	**0.8320**	0.6165

TABLE IV
Averaged Accuracies for Big Five Personality Traits (Linear SVM Classifier)

Trait / Feature	**Extro-version**	**Agreeableness**	**Conscientiousness**	**Emotional Stability**	**Openness**
HM1	0.5068	0.6689	0.7364	**0.7635**	**0.7162**
GS1	**0.5946**	0.5878	0.6149	0.5472	0.4459
ME1	0.4122	0.7973	0.6014	0.6622	**0.7162**
Pn	0.5581	0.5814	**0.8682**	0.5194	0.6202
En	0.5349	0.5193	0.8527	0.5891	0.6976
MFCCO	0.5113	**0.8915**	0.8527	0.7209	0.5504
m_MFCC	0.4806	0.5736	0.7984	0.6899	0.5349
FSVM	**0.6401**	0.8411	0.8645	0.6963	0.5761

In ridge regression, the combined feature FRR achieved higher accuracy of Emotional Stability. And the accuracies of

Extroversion and Agreeableness were close to the best result. However, the results of the other two traits, Conscientiousness and Openness, remained comparatively less accurate. On the other hand, in Table IV, the feature $FSVM$ increased the accuracy of Extroversion. And the accuracy of Conscientiousness reached, as closely as possibly, the best result that was acquired by Pn. The combined feature $FSVM$ did not improve the results of the other three traits Agreeableness, Emotional Stability, and Openness.

The highest accuracy of Extroversion that was achieved by using gaze score $GS1$ with ridge regression is 0.6363, which, however, is the lowest among the five traits. This problem seems to be caused by the following two reasons: the relationship between a human and the robot is pretty confusing, which is hard for the participant to define whether the robot as a friend or a machine. The experimental environment also has some limitations for the participants expressing themselves. The participants sat at a table in front of them, and their movements could be restricted due to the experimental settings.

We acquired the highest accuracy of Agreeableness 0.8915 by using MFCC with the linear SVM classifier, which is the highest accuracy among all traits. We tested all 13 MFCC feature vectors with ridge regression and SVM. In the ridge regression, we calculated all accuracies of 13 MFCC features on each trait. Only the results on each trait that was obtained by using the sixth MFCC feature vector were better than the average accuracy of 13 MFCC features. Therefore, the results of $MFCCO$ as shown in Table IV also were acquired by using the sixth MFCC feature vector. Agreeableness appeared to be the trait that could be most easily recognized.

The highest result of Conscientiousness 0.8682 was obtained by using pitch Pn with the linear SVM. Based on the results, the features Pn, En, and $MFCCO$ provided such promising results, which shows a strong correlation between Conscientiousness and the human's voice features.

The best result of Emotional Stability is 0.7635 that was obtained by using the Head Motion $HM1$ with linear SVM. Moreover, the two best results of Openness also were achieved by $HM1$ and $ME1$ with 0.7162. The visual features provided better results when we used the SVM classifier than the ridge regression.

Even the feature head motion $HM1$ and gaze score $GS1$ were calculated based on the head angle, they still provided very different results. Also, the MFCC showed its promising performance for recognizing the personality traits. If we compare our results with the results of previous works [47] [56], except the results of Extroversion, our experiment outperformed all of the previous competing methods.

B. Regression

For the ridge regression model, we used the average personality trait scores that were calculated from the questionnaire ranging from 1 to 5. For evaluating the regression model, we calculated MSE (Mean Squared Error) values and R^2 which is known as the coefficient of determination used to evaluate the goodness of fit of the regression model [47]. The maximum R^2 values of Conscientiousness and Openness are smaller than 0.1. Therefore, we only presented the R^2 values of Extroversion, Agreeableness, and Emotional Stability in Table V. We calculated R^2 based on the following equation:

$$R^2 = 1 - \frac{\sum_{n=1}^{M} \left(Y_n - \hat{Y}_n\right)^2}{\sum_{n=1}^{M} \left(Y_n - \overline{Y}_n\right)^2}, \tag{12}$$

where M is the total number of the data samples, Y_n is the personality trait score of the sample n, $\hat{Y}_n$ is the regression personality trait score of the sample n, and $\overline{Y}_n$ is the average score of the trait.

TABLE V
THE MAXIMUM VALUES OF R^2 OF THE REGRESSION RESULTS FOR EXTROVERSION, AGREEABLENESS, AND EMOTIONAL STABILITY

Trait \ Feature	HM1	GS1	ME1	Pn	En	MFCC6	FRR
Extroversion	0.05	**0.30**	0.01	0.01	0.11	0.01	0.15
Agreeableness	0.11	0.01	0.12	0.01	0.01	**0.28**	0.18
Emotional Stability	0.17	0.01	0.05	0.01	0.09	0.12	**0.31**

In Table V, the best classification result of three personality traits were inferred by the features with the highest R^2 values marked in bold.

The MSE values were given in Figs. 6 to 10. In order to show the changes of the MSE values clearer, we only revealed the i (the parameter for calculating the ridge parameter γ from Eq. 9) from 0 to 16. The variables that were shown in Fig. 6 to 10 were represented by using two capital letters of the abbreviation of personality trait and the feature name (refer to Table II, $mfcc6$ is the 6-th $MFCC$ vector). Figs. 6, 7, and 9 of Extroversion, Agreeableness, and Emotional Stability also showed that the feature with the smallest MSE value acquired the best classification result. The differences of the other two traits Conscientiousness and Openness were not very obvious compared to the aforementioned three traits.

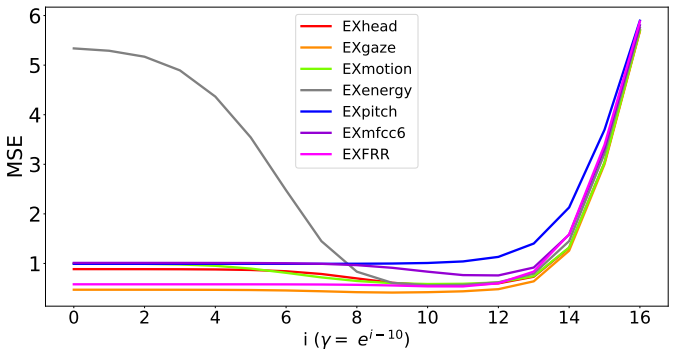


Fig. 6. MSE Values of the Ridge Regression for Inferring Extroversion

VI. CONCLUSION AND THE FUTURE WORKS

In this paper, we have proposed a new computational framework to enable a social robot to assess the personality

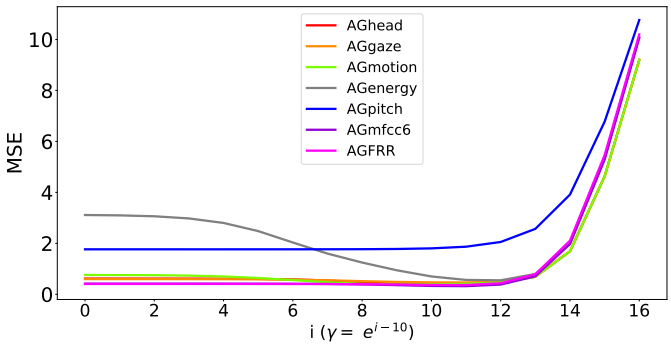


Fig. 7. MSE Values of the Ridge Regression for Inferring Agreeableness

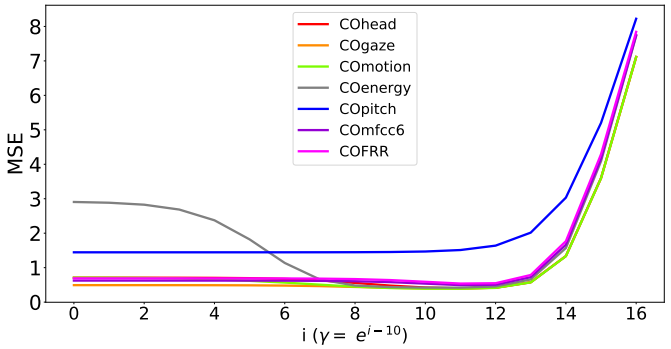


Fig. 8. MSE Values of the Ridge Regression for Inferring Conscientiousness

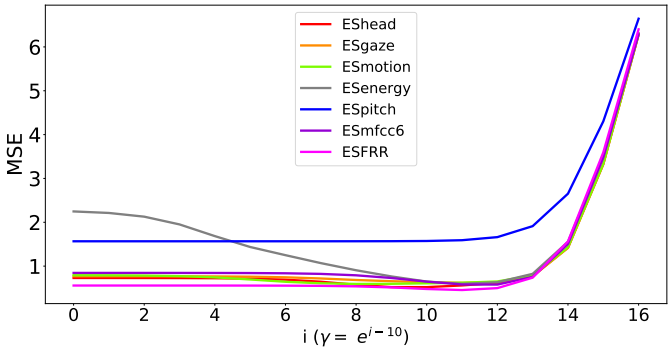


Fig. 9. MSE Values of the Ridge Regression for Inferring Emotional Stability

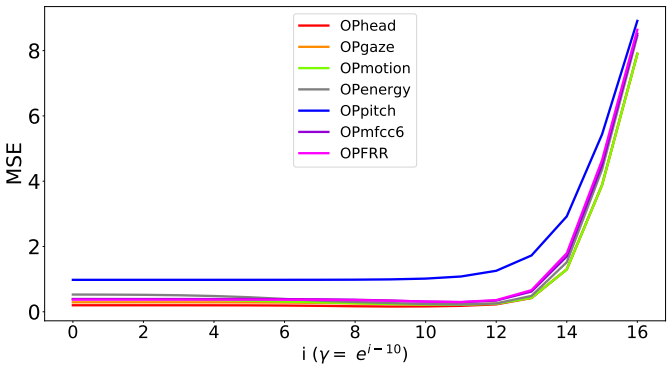


Fig. 10. MSE Values of the Ridge Regression for Inferring Openness

traits of the user it is interacting with. In the beginning, the user's nonverbal features were defined as easily-obtainable as possible and extracted from video and audio collected with the robot on-board camera and microphone. By doing so, we have decreased the computational cost in the feature extraction stage, yet the features provided promising results in the estimation of the Big Five personality traits. Moreover, the proposed framework is generic and applicable to a wide range of off-the-shelf social robot platforms. To the best of our knowledge, this is the first study to show how the visual features can be extracted in the first-person perspective, which could be the reason that our system outperformed the previous studies. Notably, the MFCC feature was beneficial to assessing each of the Big Five personality traits. We also found that, apparently, extroversion appeared to be the hardest trait. One reason could be the current experimental settings, where the participants sat at a table with their forearms resting on the tabletop that limited their body movements. Another reason could be the confusing relationship between the participants and the robot, which made the participants hesitate to express themselves naturally in the way they do in everyday situations.

Each feature showed its advantage in a different aspect. However, there is not a standard way of drawing the conclusion that declares the user's personality traits. Therefore, one of the future works is to find an efficient way to fuse the multi-modal features. On the other hand, the personality traits can be better understood through frequent and long-term interaction. This means that the system should be able to update its understandings of the user's personality traits whenever the robot interacts with its user. It is also needed to evaluate the engagement between a human and a robot, and attitude of the human toward the robot, since the user's behaviors can be precarious when the user loses interest in interacting with the robot. Finally, in order to achieve the best possible classification performance, more sophisticated machine learning models need to be incorporated.

REFERENCES

[1] T. Minato, M. Shimada, H. Ishiguro, and S. Itakura. "Development of an android robot for studying human-robot interaction," Lecture Notes in Computer Science, vol. 3029, pp. 424-434, 2004.

[2] S. Woods, K. Dautenhahn, and J. Schulz. "The design space of robots: Investigating children's views," Proc. IEEE International Workshop on Robot and Human Communication, pp. 47-52, 2004.

[3] V. Ng-Thow-Hing, P. Luo, and S. Okita, "Synchronized gesture and speech production for humanoid robots," Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4617-4624, 2010.

[4] D. McNeill, Language and gesture, Cambridge University Press, 2000.

[5] S. Kopp, B. Krenn, S. Marsella, A. Marshall, C. Pelachaud, H. Pirker, K. Thrisson, and H. Vilhjlmsson, "Towards a common framework for multimodal generation: The behavior markup language," Intelligent Virtual Agents, pp. 205-217, 2006.

[6] S. Kopp, K. Bergmann, and I. Wachsmuth, "Multimodal communication from multimodal thinking - towards an integrated model of speech and gesture production," Semantic Computing, vol. 2, no. 1, pp. 115-136, 2008.

[7] B. Bruno, N. Y. Chong, H. Kamide, S. Kanoria, J. Lee, Y. Lim, A. K. Pandey, C. Papadopoulos, I. Papadopoulos, F. Pecora, A. Saffiotti, and A. Sgorbissa, "Paving the Way for Culturally Competent Robots: a Position Paper," Proc. IEEE International Symposium on Robot and Human Interactive Communication, pp. 553-560, 2017.

[8] Nguyen T. V. T., S. Jeong, and N. Y. Chong, "Emotional Bodily Expressions for Culturally Competent Robots through Long Term Human-Robot Interaction," Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2008-2013, 2018.

[9] S. Saunderson and G. Nejat, "How robots influence humans: A survey of nonverbal communication in social humanrobot interaction," International Journal of Social Robotics, pp. 1-34, 2019.

[10] A. Aly and A. Tapus, "A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction," Proc. ACM/IEEE Int. Conf. Human-Robot Interaction, pp. 325-332, 2013.

[11] R. Hogan, J. A. Johnson, A., and S. R. Briggs (Eds.), Handbook of personality psychology, Academic Press, 1997.

[12] J. E. Lydon, D. W. Jamieson, and M. P. Zanna, "Interpersonal similarity and the social and intellectual dimensions of first impressions," Social Cognition, vol. 6, no. 4, pp. 269-286, 1988.

[13] C. A. Reid, J. D., Green, and J. L. Davis, "Attitude alignment increases trust,respect, and perceived reasoning ability to produce attraction," Personal Relationship, vol. 25, pp. 171189, 2018.

[14] K. Isbister and C. Nass. "Consistency of personality in interactive characters: Verbal cues, nonverbal cues, and user characteristics," Human-Computer Studies, vol. 53, pp. 251-267, 2000.

[15] A. Rossi, K. Dautenhahn, K. Koay, and M. L. Walters. "The impact of peoples personal dispositions and personalities on their trust of robots in an emergency scenario," vol. 9, no. 1, pp. 137-154, 2018.

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

[17] J. Cassell and J. Bickmore, "Negotiated collusion: Modeling social language and its relationship effects in intelligent agents," User Modeling and User-Adapted Interaction, vol. 13, pp. 89-132, 2003.

[18] 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, vol. 8, no. 2, pp. 281-305, 2007.

[19] M. R. Barrick and M. K. Mount, "The big five personality dimensions and job performance: A meta-analysis," Personnel Pshychology, vol. 44, pp. 1-26, 1991.

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

[21] K. M. Lee, W. Peng, S-A. Jin, and C. Yan. "Can robots manifest personality? An empirical test of personality recognition, social responses, and social presence in human robot interaction," Communication, vol. 56, pp. 754-772, 2006.

[22] E. Park, D. Jin, and A. P. del Pobil, "The law of attraction in human-robot interaction," International Journal of Advanced Robotic Systems, vol. 9:35, pp. 1-7, 2012.

[23] H. Salam, O. Celiktutan, I. Hupont, H. Gunes, and M. Chetouani, "Fully automatic analysis of engagement and its relationship to personality in human-robot interactions," IEEE Access, vol. 5, pp. 705-721, 2017.

[24] O. Celiktutan and H. Gunes, "Computational analysis of human-robot interactions through first-person vision: Personality and interaction experience," Proc. IEEE International Symposium on Robot Human Interactive Communication, pp. 815-820, 2015.

[25] NaoQi documentation, http://doc.aldebaran.com/2-5/home_pepper.html. Last visit: July 29, 2019

[26] H. Nakajima, S. B. C. Nass, R. Yamada, Y. Morishima, and S. Kawaji, "The functionality of human-machine collaboration systems-mind model and social behavior," Proc. IEEE Conference on Systems, Man, and Cybernetics, pp. 2381-2387, 2003.

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

[28] I. Myers-Briggs and P. B. Myers, "Gifts differing: Understanding personality type," Davies-Black Publishing, 1980.

[29] H. J. Eysenck, "Dimensions of personality: 16, 5 or 3? Criteria for a taxonomic paradigm," Personality and Individual Differences, vol. 12, pp. 773-790, 1991.

[30] L. R. Goldberg, "An alternative description of personality: The big-five factor structure," Personality and Social Psychology, vol. 59, pp. 1216-1229, 1990

[31] L. R. Goldberg, "A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models," Personality Psychology in Europe, vol. 7, pp. 7-28, 1999.

[32] Z. Shen, A. Elibol, and N. Y. Chong, "Inferring human personality traits in human-robot social interaction, Proc. ACM/IEEE International Conference on Human Robot Interaction, pp. 578-579, 2019.

[33] J. Kickul and G. Neuman, "Emergent leadership behaviours: The function of personality and cognitive ability in determining teamwork performance and KSAs," Journal of Business and Psychology, vol. 15, no. 1, pp. 27-51, 2000.

[34] L. W. Morris, "Extraversion and introversion: An interactional perspective," Hemisphere Publishing Corporation, 1979.

[35] A. Tapus and M. J. Mataric, "Socially assistive robots: The link between personality, empathy, physiological signals, and task performance," Proc. AAAI Spring Symposium on Emotion, Personality and Social Behavior, pp. 133-140, 2008.

[36] G. Mohammadi, M. Mortillaro, and A. Vinciarelli, "The voice of personality: Mapping nonverbal vocal behavior into trait attributions," Proc. International Workshop on the Social Signal Processing, pp. 17-20, 2010.

[37] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," Proc. International Conference on Multimodal Interfaces, pp. 53-60, 2008.

[38] A. Vinciarelli and G. Mohammadi, "A survey of personality computing, IEEE Transactions on Affective Computing, vol. 5, no. 3, pp. 273-291, 2014.

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

[40] P. T. Costa and R. R. MacCrae, "Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual," Psychological Assessment Resources, 1992.

[41] R. R. McCrae and P. T. Costa, "A contemplated revision of the neo five-factor inventory," Personality and Individual Differences, vol. 36, no. 3, pp. 587-596, 2004.

[42] L. R. Goldberg, "The development of markers for the big-five factor structure," Psychological Assessment, vol. 4, no. 1, pp. 26-42, 1992.

[43] A. Furnham, Language and personality, H. Giles and W. Robinson (Eds.), Handbook of Language and Social Psychology, Wiley, 1990.

[44] J. M. Dewaele and A. Furnham, "Extraversion: The unloved variable in applied linguistic research," Language Learning, vol. 49, no. 3, pp.

509-544, 1999.

[45] J. W. Pennebaker and L. A. King, "Linguistic styles: Language use as an individual difference," Personality and Social Psychology, vol. 77, pp. 1296-1312, 1999.

[46] F. Mairesse and M. Walker, "Words mark the nerds: Computational models of personality recognition through language," Proc. Annual Conference of the Cognitive Science Society, pp. 543-548, 2006.

[47] O. Aran and D. Gatica-Perez, "One of a kind: Inferring personality impressions in meetings," Proc. ACM International Conference on Multimodal Interaction, pp. 11-18. 2013.

[48] G. M. Lucas, J. Boberg, D. Traum, R. Artstein, J. Gratch, A. Gainer, E. Johnson, A. Leuski, and M. Nakano, "Getting to Know each other: The role of social dialogue in recovery from errors in social robots," Proc. ACM/IEEE International Conference on Human-Robot Interaction, pp. 344-351, 2018.

[49] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker, "Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life," Personality and Social Psychology, vol. 90, pp. 862-877, 2006.

[50] P. Patompak, S. Jeong, I. Nilkhamhang, and N. Y. Chong, "Learning proxemics for personalized human-robot social interaction," International Journal of Social Robotics, 2019.

[51] S. M. Anzalone, G. Varni, S. Ivaldi, and M. Chetouani, "Automated prediction of extraversion during humanhumanoid interaction," International Journal of Social Robotics, vol. 9, no. 3, pp. 385-399, 2017.

[52] Z. Zafar, S. H. Paplu, and K. Berns, "Automatic assessment of human personality traits: A step towards intelligent human-robot interaction," Proc. IEEE-RAS International Conference on Humanoid Robots, pp. 670-675, 2018.

[53] J. T. Webb, "Interview synchrony: An investigation of two speech rate measures," Studies in Dyadic Communication, A. W. Siegman and B. Pope (Eds.), pp. 115-133, Pergamon Press, 1972.

[54] R. Hassin and Y. Trope, "Facing faces: Studies on the cognitive aspects of physiognomy," Personality and Social Psychology, vol. 78, pp. 837-852, 2000.

[55] A. Guidi, C. Gentili, E. P. Scilingo, and N. Vanello, "Analysis of speech features and personality traits," Biomedical Signal Processing and Control, vol. 51, pp. 1-7, 2019.

[56] S. Okada, O. Aran, and D. Gatica-Perez, "Personality trait classification via co-Occurrent multiparty multimodal event discovery," Proc. ACM International Conference on Multimodal Interaction, pp. 15-22. 2015.

[57] D. Gatica-Perez, D. Sanchez-Cortes, T. M. T. Do, D. B. Jayagopi, and K. Otsuka, "Vlogging over time: Longitudinal impressions and behavior in YouTube," Proc. International Conference on Mobile and Ubiquitous Multimedia, Slim Abdennadher and Florian Alt (Eds.), pp. 37-46, 2018.

[58] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung., "Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction," Proc. Annual Meeting of the Association for Computational Linguistics, vol. 2, pp. 606611, 2018.

[59] R. D. P. Principi, C. Palmero, J. C. S. Jacques Jr., and S. Escalera, "On the effect of observed subject biases in apparent personality analysis from audio-visual signals," IEEE Transactions on Affective Computing, 2019.

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

[61] 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, vol. 7, no. 1-2, pp. 39-53, 2013.

[62] M. L. Knapp, J. A. Hall, and T. G. Horgan, Nonverbal communication in human interaction, 8th Ed., Cengage Learning, 2009.

[63] B. Wu, H. Ai, C. Huang, and S. Lao, "Fast rotation invariant multi-view face detection based on Real AdaBoost," Proc. IEEE International Conference on Automatic Face and Gesture Recognition, pp. 79-84, 2004.

[64] R. Stiefelhagen and J. Zhu J, "Head orientation and gaze direction in meetings," CHI02 Extended Abstracts on Human Factors in Computing Systems, pp. 858-859, 2002.

[65] Ricci E, Odobez J., "Learning large margin likelihoods for realtime head pose tracking," International Conference on Image Processing, 2009.

[66] J. W. Davis and A. F. Bobick. "The representation and recognition of human movement using temporal templates," In Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 928-934. IEEE, 1997.

[67] A. F. Bobick and J. W. Davis. "The recognition of human movement using temporal templates," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257-267, 2001.

[68] I. McCowan, M. Krishna, D. Gatica-Perez, D. Moore, and S. Ba, "Speech acquisition in meetings with an audio-visual sensor array," Proc. IEEE International Conference on Multimedia and Expo, pp. 1382-1385, 2005.

[69] W.-F. Hsieh, Y. Li, E. Kasano, E.-S. Simokawara, and T. Yamaguchi, "Confidence identification based on the combination of verbal and non-verbal factors in human robot interaction," Proc. International Joint Conference on Neural Networks, pp. 1-7, 2019.

[70] J. C. S. Jacques Jr., Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, Marcel A. J. v. Gerven, R. v. Lier, S. Escalera, "First impressions: A survey on computer vision-based apparent personality trait analysis," IEEE Transactions on Affective Computing, 2019

[71] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, H. Manley, "Average magnitude difference function pitch extractor," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-22, no. 5, pp. 353-362, 1974.

[72] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," IEEE Transactions on Audio and Electroacoustics, vol. AU-20, no. 5, pp. 367-377, 1972.

[73] X.-D. Mei, J. Pan, and S.-H. Sun, "Efficient algorithms for speech pitch estimation," Proc. International Symposium on Intelligent Multimedia, Video and Speech Processing, pp. 421-424, 2001.

[74] M. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate," Proc. Symposium on Computer Processing Communications, pp 779-797, 1969.

[75] M. Xu, L. Y. Duan, J. Cai, L. T. Chia, and C. Xu, "HMM-based audio keyword generation," Advances in Multimedia Information Processing, K. Aizawa, Y. Nakamura, and S. Satoh (Eds.), pp. 566-574, 2004.

[76] 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, pp. 543-565. 2012.

[77] C. Beyan, F. Capozzi, C. Becchio, and V. Murino, "Prediction of the leadership style of an emergent leader using audio and visual nonverbal features," IEEE Transactions on Multimedia, vol. 20, no. 2, pp. 441-456, 2018.

[78] S. Feese, B. Arnrich, G. Trster, B. Meyer, and K. Jonas, "Quantifying behavioral mimicry by automatic detection of nonverbal cues from body motion," Proc. ASE/IEEE International Conference on Social Computing, pp. 520-525, 2012.

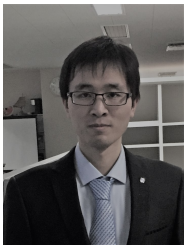
[79] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small

groups," IEEE Transactions on Multimedia, vol. 14, no. 3, pp. 816-832, 2012.

[80] D. Sanchez-Cortes, D. Jayagopi, and D. Gatica-Perez, "Predicting remote versus collocated group interactions using nonverbal cues," International Conference on Multimodal Interfaces, Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing, 2009.

[81] C. Beyan, F. Capozzi, C. Becchio, and V. Murino. "Identification of emergent leaders in a meeting scenario using multiple kernel learning," International Conference on Multimodal Interaction, Workshop on Advancements in Social Signal Processing for Multimodal Interaction, pp. 3-10, 2016.

[82] C. M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, 2006



Zhihao Shen is a Ph.D. student in the School of Information Science at the Japan Advanced Institute of Science and Technology. He received M.S. degree from Japan Advanced Institute of Science and Technology and B.S. degree from Tianjin University of Science and Technology. His research interests are in the area of human-robot interaction and machine learning.

Armagan Elibol is a full-time faculty member (as an assistant professor) of Robotics Lab. at School of Information Science of Japan Advanced Institute of Science Technology (JAIST) since October 2018. Prior to that, he was an assistant professor at Department of Mathematical Engineering of Yildiz Technical University (YTU) and was an adjunct teaching faculty member of Computer Engineering Department of FMV Isik University in Turkey. He obtained a Ph.D. in Computer Science (2011) from the Computer Vision and Robotics Group of University of Girona, Spain, both B.Sc. and M.Sc. in Mathematical Engineering from YTU, respectively 2002 and 2004. He was a DAAD-supported visiting researcher at the Pattern Recognition and Bioinformatics Group of Martin-Luther-University Halle-Wittenberg in Germany during summer 2012. Afterward, He was with the Ocean Robotics and Intelligence Lab (ORIN). of Korea Advanced Institute of Science and Technology (KAIST) in Daejeon for one year as a postdoctoral researcher. During 2015, he was with the School of Integrated Technology of Yonsei University at Yonsei International Campus in Songdo-Incheon, Rep. of Korea. His research interests include the wide area of Computer Vision and Robotics. Lately, he integrates his experience on optical mapping to the humanoid research using different sensors and intelligent methodologies.

Nak Young Chong received the B.S., M.S., and Ph.D. degrees from Hanyang University, Seoul, Korea, in 1987, 1989, and 1994, respectively. From 1994 to 2003, he was with Daewoo Heavy Industries, KIST, MEL, and AIST. In 2003, he joined the faculty of JAIST, where he currently is a Professor of Information Science. He also served as a Councilor and Director of the Center for Intelligent Robotics. He was a Visiting Scholar at Northwestern University, Georgia Tech, University of Genoa, and CMU, and served as an Associate Faculty at UNLV and Kyung Hee University. He serves as Editor of the IEEE Robotics and Automation Letters, International Journal of Advanced Robotic Systems, and Journal of Intelligent Service Robotics, and served as Editor of IEEE ICRA, IEEE CASE, IEEE Ro-Man, and UR, and Associate Editor of the IEEE Transactions on Robotics. He served as Program (Co)-Chair for JCK Robotics 2009, ICAM 2010, IEEE Ro-Man 2011/2013, IEEE CASE 2012, URAI 2013/2014, DARS 2014, ICCAS 2016, IEEE ARM 2019. He was a General Co-Chair of URAI 2017 and serves as a General Chair of UR 2020. He also served as Co-Chair for IEEE RAS Networked Robots TC from 2004 to 2006, and Fujitsu Scientific System WG from 2004 to 2008.