A prototype of multi-modal interaction robot based on emotion estimation method using physiological signals

KAORU SUZUKI[†1], TAKUMI IGUCHI[†1], YURI NAKAGAWA[†1], and MIDORI SUGAYA[†1]

Abstract: In recent years, the type of robot that estimates emotions in real time has been proposed and is expected to be introduced into nursing care and home use. Among emotion estimation technologies, those incorporating methods using the physiological signals have satisfied the requirements of obtaining real-time emotional response of the human by using physiological signals such as EEG and HRV from the wearable sensors. However, the current real-time emotion estimation response robot only has a speech function or a facial expression function accordingly to emotional states of the human that is in front of the robot and does not use multiple modalities. With those limited modalities of output, we assume that it would be difficult to improve the emotional state of the person that uses the robot. And we would like to know which modalities are effective and what happens when modalities are combined. Therefore, in this research, we propose a multi-modal robot that combines not only speech but also facial expressions and body movements aiming the practical application of a robot that responds to the person's emotion in real time. This robot outputs facial expressions/speeches/body movements to improve or maintain the user's emotional state. As a result of the experiment, it is confirmed that this robot has a relaxing effect on the user when output includes speech.

Keywords: physiological signals, emotion estimation, robot, multimodal

1. Introduction

In recent years, the possibility and realization of a human-centric (anthropocentric) society, i. e. Society 5.0 that achieves both economic development and a resolution of social issues [1], have been explored. Various technologies and systems such as IoT, AI, and robots are to be utilized there.

As one of the utilizations of such systems, technology for human mental care is expected. The present day is said to be a stressful society, and to achieve QoL (Quality of Life), the care for people's emotional aspects has become one of issues to be mattered. To address the issue, understanding human emotions through the use of IoT, AI, and robot technology, as well as technology to care for human emotions, will be important [14]. By applying emotion understanding and corresponding care to the interaction between robots and humans, it is believed that robots will be able to exert greater effects on humans.

Various techniques have been proposed for estimating user's emotions [2,3]. Techniques for estimating emotions from user's facial expressions and voice have been proposed. They use specific patterns based on image processing and voice analysis [2, 13]. However, emotion estimation from images has a problem of high learning costs due to large differences in whetaher emotions are expressed as facial expressions or not [7, 8]. In addition, emotion estimation from voice also has a problem that emotion can be estimated only at the timing of speaking.

On the other hand, Ikeda et al. proposed a method to estimate emotion with physiological signals, specifically electroencephalograms (EEG) and heart rate variability (HRV) [3]. This method has the benefit of estimating emotions that users cannot realize by themselves. By using such method, it is thought that it will be possible to respond to more detailed emotions that even the person himself/herself cannot comprehend. As robots using emotion estimation technology via physiological signals, talking robots [4, 5] and facial expression robots [7, 8] have been proposed and their effectiveness has been reported. However, currently, talking robots control only speeches, and facial expression robots control only facial expressions accordingly to the user's emotions estimated from pysiological signals. The effects of multiple modalities, including body movements and their combinations, have not yet been clarified.

2. Purpose/Proposal

As we described, with the limited modality of the output, we assume that it would be difficult to improve the emotional state of the person that uses the robot. And we would like to know which modalities are effective and what happens when modalities are combined. Therefore, the purpose of this research is to achieve high emotional care by a robot using multiple modalities and to clarify that its effectiveness can improve the emotional state of the person. Here, high emotional care means providing appropriate interaction by performing detailed emotional estimation based on physiological signals. In this paper, we first introduced the emotion estimation method based on physiological signals. Then, we design and implement a multimodal robot that responds with facial expressions, body movements, and speeches accordingly to the estimated person's emotions. Lastly, assessment of the emotional care is given. In the evaluation, it is verified whether the user is in a pleasant state coming from an unpleasant state.

2.1 Russell's Circumplex Model of Affect

As we described previously, we will first explain about the technique for estimating emotions using physiological signals. In the emotion estimation method using physiological signals, Russell's circumplex model of affect is used as the base for emotional interpretation through psychological information [6].

†1 Shibaura Institute of Technology

Russell showed 28 emotinal terms that are circularly distributed on a two-dimensional plane, with the horizontal axis representing pleasure/unpleasure and the vertical axis representing arousal/sleepiness, and the relative emotional terms are arranged on opposite sides of the origin. Each emotional term is positioned at a unique angle from the pleasure-unpleasure axis. By this method, the effectiveness of this model has been recognized in many documents, so we assume that would be suitable for adoption.

2.2 Physiological emotion estimation technology

In the emotion estimation method based on physiological signals, emotions are estimated by associating the pleasure-unpleasure values and arousal-sleepiness values in Russell's circumplex model of affect with values directly measured from EEG and HRV [3][16]. By associating objective physiological information with emotional terms, even unconscious emotions can be estimated from physiological signals. Fig. 1 shows the principle of emotion estimation on the circumplex model [15]. The oberved data is a vector determined by the pleasure-unpleasure value and arousal-sleepiness value in the figure, and the angle from the pleasure-unpleasure axis of the vector gives the emotion, while the length of the vector gives the intensity of the emotion.

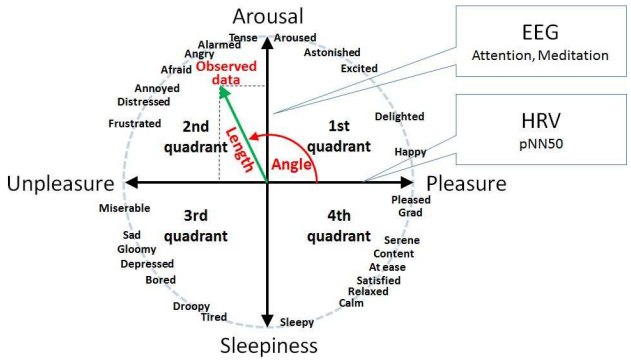

Fig. 1 Adaptation of Russell's Circumplex Model of Affect

2.3 Emotion estimation algorithm

In this study, we classified emotional states into four quadrants of the two-dimensional model that coordinates of pleasure-unpleasure and arousal-sleepiness [6]. That is, when observed data is obtained from the EEG sensor and the optical pulse sensor, angle and length in Fig. 1 are calculated, and the quadrant where the emotional state is classified, is defined by the angle.

The values obtained from the sensors used in this study are updated once per second. Emotion estimation also calculates the quadrant to which the emotion belongs and the intensity value once every second as the instantaneous emotional state. Furthermore, the emotion estimation extracts the leading emotion for each period by accumulating the intensities for 5 times (5 seconds), choosing the quadrant with the maximum accumulated intensity value as the estimated emotional state, and using the maximum accumulated intensity value as the intensity of the emotion [7, 8]. The robot's response, which will be described later, is determined once every 5 seconds based on this estimated emotional state.

2.4 Multimodal interaction robot

Fig. 2 shows the software configuration of a prototype robot system that estimates the user's emotional state in real time using the emotion estimation algorithm described above and outputs multimodal responses accordingly. Fig. 3 shows how this robot is operated.

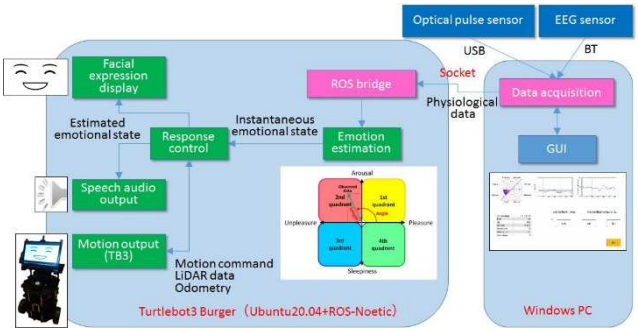

Fig. 2 Software configuration of the prototype robot system

Fig. 3 Operation of the prototype robot system

The robot body is built with Turtlebot3 Burger (hereinafter TB3). The SBC of TB3 is a RaspberryPi 3B+, and Ubuntu20.04 and ROS (ros-noetic) are run on this SBC to control the robot's body movements. In addition, TB3 is equipped with a 5-inch monitor to display facial expressions, and a USB active speaker for speech audio output. The optical pulse sensor and EEG sensor are connected to a Windows PC, which is separated from TB3, and this PC acquires physiological data and presents it visually to the experimenter, while passing the data to the ROS system of TB3 via socket communication. Emotion estimation from physiological data is performed in a dedicated ROS node (emotion estimation in Fig. 2) on TB3. Then, based on the estimated emotional state, subsequent facial expressions, body movements, and speech outputs are similarly executed by dedicated ROS nodes (response control, facial expression display, speech audio output, and motion output in Fig. 2).

For emotion estimation, the optical pulse sensor (manufactured by pulsesensor.com [9]) attached to the user's finger is used to obtain pulse wave data using an Arduino UNO, and the pNN50 value is calculated. The pNN50 value is the heart rate variability (HRV) index and is used as a measurement of pleasure-unpleasure in the autonomic nervous system. A pleasure-unpleasure value is calculated from this pNN50 value, and the emotional state is classified as pleasant if the pNN50 value exceeds the threshold th, and unpleasant if it is below th. This

threshold th is set to 0.23, referring to the average of pNN50 [12].

In addition, the attention value (concentration level) and meditation value (relaxation level) are obtained from the EEG sensor (NeuroSky MindWave™ Mobile2 [10]) worn on the user's head. At this time, a arousal-sleepiness value is given as attention value minus meditation value, and the emotional state is classified as arousal side when the attention value exceeds the meditation value, and as sleepiness side when it is less than.

2.5 Mechanism to make the robot face the user

The motion of the robot is controlled depending on the odometry obtained by internal sensors. At first, even if the robot and the user are facing each other, there is a possibility that the face-to-face state would gradually collapse due to an accumulation of errors in the odometry.

Therefore, as shown in Fig. 4, the closest object point within ±45° from the front direction of the odometry is detected by LiDAR, and that direction is always used as the reference for operation.

This closest object point is expected to be a part of the user's body and should serve as a reference for making the robot face the user. Both the forward/backward movement of the body and the left/right turns, which will be described later, are executed with this reference direction as the front direction of the movement, so even if the odometry deviates from the actual one, the robot will be able to keep the face-to-face state with the user.

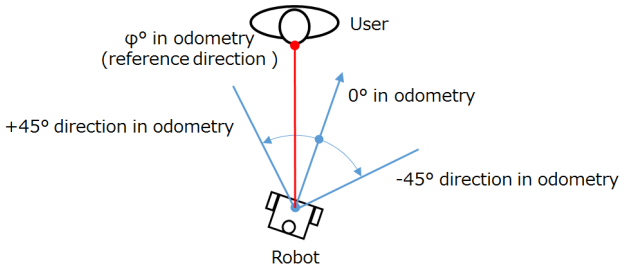

Fig. 4 Mechanism to make the robot face the user

2.6 Response selection rules

The mission of this system is to make transition from the user's unpleasant state (2nd and 3rd quadrants) to the pleasant state (1st and 4th quadrants). The robot present expressions that match the user's emotional state [7,8]. Specifically, a happy face is displayed in the first quadrant, a tense face is displayed in the second quadrant, a gloomy face is displayed in the third quadrant, and a smily face is displayed in the fourth quadrant.

To develop a synchronized nuance in the body movements, the first quadrant is a happy movement (forward and backward motion), the second quadrant is a tense movement (large and fast left and right rotations), and the third quadrant is a gloomy movement (small and slow left and right rotations), and the 4th quadrant expresses a relaxed state (stand still) (Fig. 5). The same operation is repeated if the estimated emotional state does not change.

We designed the different response speeches for pleasant and unpleasant states. As shown in Table 1, in the 1st and 4th quadrants, since there is no need to reverse the states, only synchronous utterances are output. In the second quadrant (tense/frustrated) and the third quadrant (bored/tired), three types of speech are output: synchronous speech, suggestive speech, and encouraging speech. We had to narrow down the emotional terms for each quadrant in order to make the appropriate utterances. Therefore, we decided on emotional terms for each quadrant as shown in table 1, assuming application in a relatively calm state in which strong emotions such as anger and sadness do not occur.

In quadrants with multiple response speeches, one is selected in turn at a time. Also, the response speech is outputted when the estimated emotional state is updated once every 5 seconds, however if the same estimated emotional state continues, it will be outputed once every two updates (i. e. once every 10 seconds). These response speeches are created with OpenJTalk's female voice using the voice synthesis software "Textalk " [11].

Table 1 Response selection rules for each quadrant

Emotional terms for each quadrant	Facial expressions	Body movements	Response speeches
1st quadrant (Pleasure, arousal) Delighted / Excited	Happy	move forward and back repeatedly	Synchronization: I'm happy too
2nd quadrant (Unpleasure, arousal) Tense / Frustrated	Tense	rotate large and fast repeatedly	Synchronization: I'm nervous too. Suggestion: Would you like to take a deep breath? (play one from above)
3rd quadrant (Unpleasure, sleepy) Bored / Tired	Gloomy	rotate small and slow repeatedly	Suggestion: Would you like to take a deep breath? Encouragement: Are you okay? Encouragement: Cheer up. Encouragement: Keep it up. (play one from above)
4th quadrant (Pleasure, sleepy) Relaxed / Satisfied	Smile	stand still	Synchronization: I am relaxed too.

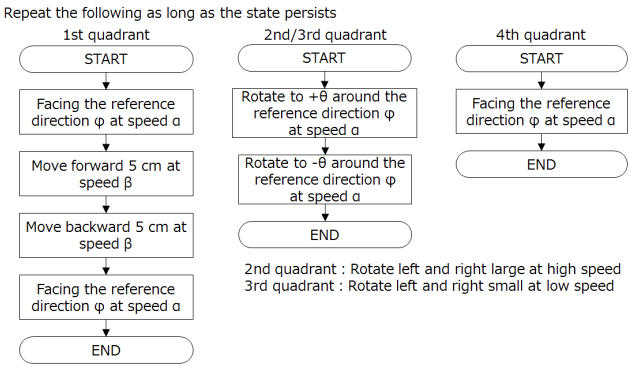

Fig. 5 Body movement algorithm

3. Evaluation

We measure and compare the pNN50 values of users of this system in single modal mode of facial expression, body movement, and speech, and in multimodal mode.

3.1 Experimental method

Experiment participants who play the role of users are asked to wear an optical pulse sensor and an EEG sensor and face the robot placed on the desk, and the pNN50, attention, and meditation values are measured in the order of experiments 1 to 4 in Table 2. In addition, an unstructured interview is conducted with the

participants after each experiment.

In each experiment, the robot is activated after 1 minute of rest. After about 1 minute and 20 seconds from the start of rest, the robot starts responding. The period from the start of rest to the start of response is defined as the rest period, and the 5 minutes from the start of response is defined as the interaction period. During the interaction period, the robot acts accordingly to the physiological data of the participants based on the response selection rules in Table 1.

The participants are three males in their twenties (named A, B, and C). A and B have no background knowledge of this robot. On the other hand, C already knew the robot's function before participating in the experiment. C has experienced multimodal, but has not experienced single modal, and knows that the robot responds to his emotions.

Table 2 Modalities for each experiment

Experiment 1	Single modal action with only facial expressions
Experiment 2	Single modal action with only body movements
Experiment 3	Single modal action with only speeches
Experiment 4	Multimodal action with facial expressions, body movements, and speeches

3.2 Interview result

Table 3 lists excerpts of comments from participants obtained through interviews after each experiment.

Table 3 Comments from the interview (excerpts)

AB common	・When only facial expressions or movements of the robot were performed, it was understood that the expressions were the robot's own emotional state. ・When only the speech was performed, we understood the intention of the robot.
A	・I became a little sleepy when there were only facial expressions and body movements. ・During multimodal, I was happy when the robot said, "I'm happy too."
B	・When only facial expressions were used, I felt uncomfortable with the expression of gloomy and tense. ・I was surprised at the movement of the robot when only body movements were performed. ・When only the speech was performed, I felt that the robot's "Cheer up" was full of emotion. ・When multimodal, the meaning of movement was understood by adding facial expressions and speech.
C	・When only facial expressions were used, I thought that the robot was expressing its own emotions with facial expressions. (I knew how it worked) ・During the movement only, only a pleasant state was shown, so I thought it was a program that simply repeated forward/ backward movements instead of reacting to emotions. ・During multimodal, I was worried that the robot might fall off the desk.

3.3 Inter-experimental comparison and discussion of the pNN50 mean value during the interaction period

Fig. 6 shows the pNN50 mean values of each participant during the rest period and the interaction period for each experiment. From this figure, looking at the mean value of pNN50 during the interaction period, that of body movements only was the lowest for participants A and B, followed by facial expressions only, multimodal, and speeches only, in that order. On the other hand, for C, the facial expressions only was the lowest, followed by body movements only and speeches only, and multimodal was the highest. It seems that the reaction is different between those who have never experienced this type of experiment and those who have experienced multimodal.

When comparing speech only and multimodal, two-tailed, paired t-tests with a significance level of 5% showed that A's mean value was significantly higher in speech only, and C's was significantly higher in multimodal. For B, there was no significant difference. From these results, it is difficult to determine the superiority of speech only and multimodal.

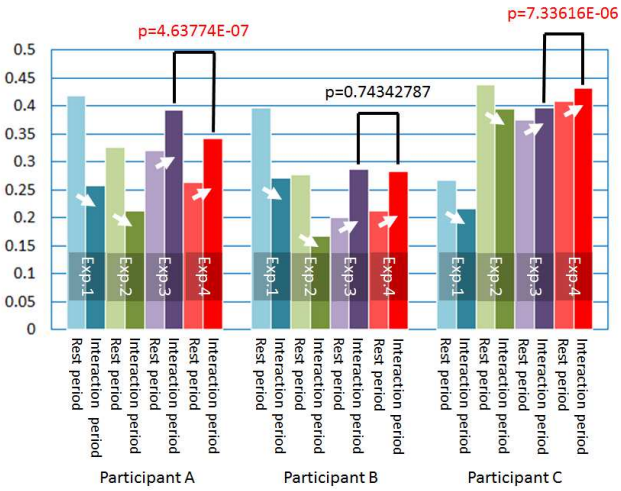

Fig. 6 Comparison of pNN50 mean value between rest period and interaction period

3.4 Comparison and discussion of pNN50 mean values at rest/interaction

In addition, from Fig. 6, the pNN50 mean values of all participants decreased when the rest period transitioned to the interaction period for facial expressions only and body movements only. An increase in pNN50 can be interpreted as transition to a relaxed state, and a decrease as transition to a tense state. From the interview, it seems that A, B could not grasp the robot's intentions only by facial expressions and body movements alone. Therefore, it is conceivable that this tense state is caused by participants' efforts trying to understand the robot's intentions, or by being frustrated and bored because they do not understand the relationship with the robot.

On the other hand, during speech only and multimodal, the pNN50 mean value increased in all participants when the rest period transitioned to the interaction period. By including speech in the response, it is thought that the intention of the robot and the relationship with it became clear, therefore the tension was released.

3.5 Effect verification of suggestion-type speech "Deep breath"

In all cases, pNN50 increased when participants took deep breaths at the suggestion from the robot" Would you like to take a deep breath ?". Fig. 7 shows a time-series graph of pNN50 observed during multimodal of participant A. At the time indicated by the red line in the figure, the robot suggested, and when A took a deep breath accordingly, it was confirmed that the pNN50 value increased. In this example, during the interaction period the participant was suggested only once to take a deep breath.

In this experiment, only "deep breathing" was added, but if there were any suggestions that would affect emotions, they could be added.

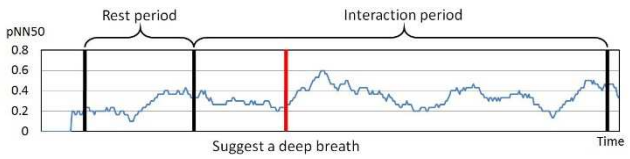

Fig. 7 Change in pNN50 during multimodal for participant A

3.6 Mysterious effect of gloomy face

Although the reason has not yet been clarified, when only facial expressions were used, sometimes pNN50 of participants B and C increased after a gloomy face. This is shown in Fig. 8. The vertical lines in the figure represent the timing of the gloomy face. The arrows in the figure indicate the points where pNN50 increased afterward.

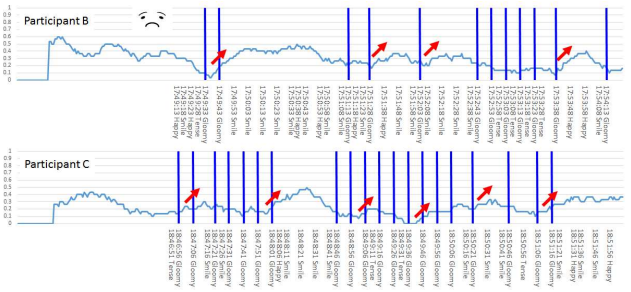

Fig. 8 Transition of pNN50 when B and C have only facial expressions (Vertical lines indicate the timing of the gloomy face)

On the other hand, it was also revealed that tense faces had no effect on raising pNN50. This is shown in Fig. 9. The vertical lines in the figure represent the timing of the gloomy face (blue lines) and the tense face (red lines). In particular, the positions of the tense face are indicated by circles in the figure, and there is no increase in pNN50.

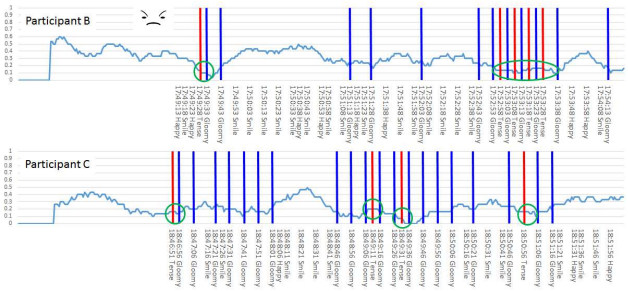

Fig. 9 Transition of pNN50 when B and C have only facial expressions (Circles indicate the positions of the tense face)

4. Conclusion

In this paper, we report on the multimodal prototype robot that estimates the user's emotional state in real time from the physiological signals and outputs responses via facial expressions, body movements, and speeches to improve the emotional state of the user. And the evaluation results are also reported.

The mission of this robot is transition the user's emotional state from unpleasant to pleasant (or to maintain the pleasant state). User's emotional states are classified into four quadrants on two-dimensional coordinates of pleasure-unpleasure and arousal-sleepiness, and the robot outputs responses accordingly to the response selection rules set for each quadrant.

The pNN50 mean value by period is used as an index of the pleasant state, we evaluated the effectiveness of the response selection rules and the characteristics of the modalities by changing the output response to facial expressions only, body movements only, speeches only, or multimodal.

As a result of the experiment with three participants, the effectiveness of the response output including speech was confirmed. In addition, the superiority of speech-only and multimodal was not determined. Regarding the response selection rules, we confirmed the effectiveness of the suggestive speech prompting deep breathing.

Although the reason is unknown, we also confirmed the phenomenon that pNN50 is increased by gloomy face when only facial expressions are used. On the other hand, the tense face has no effect on increasing pNN50.

There were only three participants in the experiments. It is necessary to continue the evaluation with more participants in the future.

In addition, the current response selection rules output facial expressions, body movements, and speeches accordingly to the user's emotional state. However, in another experiment, when participants intended their emotional state to be pleasant, this stressed them and prevented their emotional state from becoming pleasant. Current response selection rules do not assume that users will intentionally try to change their emotional state. However, the robot needs to change its response when such a situation occurs.

Furthermore, in this system, pleasure-unpleasure values were calculated from pNN50, arousal-sleepiness values were calculated as the difference between attention and meditation, and these values were applied to Russell's Circumplex Model of

Affect to estimate emotional states [3].

Attention and meditation are considered to indicate a person's degree of arousal and calmness, and are normalized from 0 to 100, respectively. Therefore, we decided to use the difference between attention and meditation as arousal-sleepiness for convenience. The pNN50 indicates the degree of tension of the autonomic nerves, and the smaller the value, the more stressed/unpleasant the person is, and the higher the pNN50, the more normal/plesant the person is[16].

However, this is not yet an established method, and the concordance rate between estimated emotion and subjective evaluation is 40-50% [17]. We are still searching for improvements.

Acknowledgments This research was supported by JST, CREST, and JPMJCR19K1. We would like to express our gratitude.

Reference

[1] Japanese Cabinet Office, "Society 5.0 Explanatory materials for New Society Pioneered by Science and Technology Innovation", https://www8.cao.go.jp/cstp/society5_0/society5_0.pdf (in Japanese)

[2] Omron, "OKAO Vision", https://components.omron.com/jp-ja/products/sensors/human-image-solution/software-library/software-library (in Japanese)

[3] Yuhei Ikeda, Ryota Horie, and Midori Sugaya, "Estimate Emotion with Biological Information for Robot Interaction", KES-2017, Vol.112, pp.1589-1600. (2017)

[4] Teppei Ito, Reiji Yoshida, Yoshito Tobe, and Midori Sugaya, "Supportive Voice-Casting Robots using Bio-Estimated Emotion for Rehabilitation", Intelligent Environments 2019. (2019)

[5] Kodai Matsumoto, Reiji Yoshida, Feng Chen, and Midori Sugaya, "Emotion Aware Voice-Casting Robot for Rehabilitation Evaluated with Bio-signal Index", Human Computer Interaction International 2019 (HCII 2019), LNCS. (2019)

[6] James A. Russell, "A Circumplex Model of Affect", Journal of Personality and Social Psychology, Vol.39, No.6, pp.1161-1178. (1980)

[7] Peeraya Sripian, Muhammad Nur Adilin Mohd Anuardi, Yushun Kajihara, and Midori Sugaya, "Empathetic robot evaluation through emotion estimation analysis and facial expression synchronization from biological information.", Artificial Life and Robotics 26.4, pp. 379-389. (2021), https://doi.org/10.1007/s10015-021-00696-w

[8] Peeraya Sripian, Muhammad Nur Adilin Mohd Anuardi, Yushun Kajihara, and Midori Sugaya, "Empathetic Robot Evaluation through Emotion Estimation Analysis & Facial Expression Synchronization from Biological Information", Artificial Life and Robotics, 2021.(2021)

[9] "Pulse Sensor", https://pulsesensor.com/

[10] NeuroSky, "MindWave™ Mobile2", https://www.neurosky.jp/mindwave-mobile2/

[11] "Textalk", https://gui.jp.net/textalk/ (in Japanese)

[12] Michael Trimmel, "Relationship of Heart Rate Variability (HRV) Parameters Including pNNxx With the Subjective Experience of Stress, Depression, Well-Being, and Every-Day Trait Moods (TRIM-T): A Pilot Study", The Ergonomics Open Journal, 2015, 8: pp.32-37. (2015), https://benthamopen.com/ABSTRACT/TOERGJ-8-32

[13] User Local, "Recognize emotion in speech", https://emotion-voice-ai.userlocal.jp/ (in Japanese)

[14] Takanori Shibata and Joseph F. Coughlin, "Trends of Robot Therapy with Neurological Therapeutic Seal Robot, PARO", Journal of Robotics and Mechatronics Vol.26 No.4, pp.418-425. (2014)

[15] "How emotions can be classified", https://note.com/celestia1212/n/n169242440f5c (in Japanese)

[16] Reiji Yoshida and Midori Sugaya, "Emotion Estimation using Biometric Information and its Application", IPSJ SIG Technical Reports, Vol.2019-MBL-90, No.13, pp.1-5. (2019) (in Japanese)

[17] Ikuya Kurono, Peeraya Sripian, Reiji Yoshida, and Midori Sugaya, "A study of methods to realize empathic robots using emotional attunement", IPSJ Interaction 2019, 3P-84. (2019) (in Japanese)