Composite Emotion Recognition and Feedback of Social Assistive Robot for Elderly People

Yegang Du[1], Kaiyuan Zhang[2], and Gabriele Trovato[3]

[1] Future Robotics Organization, Waseda University, Tokyo, Japan
[2] School of Creative Science and Engineering, Waseda University, Tokyo, Japan
[3] Innovative Global Program, Shibaura Institute of Technology, Tokyo, Japan

Abstract. As the world's population ages, the issue of medical care and daily care for the elderly population is becoming more and more critical. While there are many efforts to ensure the physical health of older adults, there is a lack of effective solutions to the psychological problems caused by the lack of companionship. To solve this problem, we built a social robot with an emotion recognition and feedback model. The emotion recognition module achieves composite emotion recognition by integrating voice emotion recognition, semantic emotion recognition and topic emotion recognition. The audio data and corresponding text data are mined using CNN and LSTM algorithms, and the emotional information they contain can be further refined. Meanwhile, the emotion feedback module provides the appropriate emotion feedback based on the recognised emotions. The emotional feedback is achieved by generating the appropriate robot facial expressions. The proposed emotion recognition and feedback model is validated by the corresponding database respectively. In addition, the real-world implementation validation and real scenario application of the model are also discussed in the paper.

Keywords: Emotion Recognition · Elderly Care · Social Robotic · Emotion Feedback.

1 Introduction

In recent years, the growth of the elderly population has brought the issue of medical and daily care to the forefront. As a result, advancements in the Internet of Things (IoT) and Artificial Intelligence (AI) have led to the creation of various home medical assistance applications for the elderly. These applications aim to monitor the physical well-being of the elderly through the use of sensors and smart homes[7][8]. However, despite their efforts to improve the physical health of the elderly, the emotional and psychological needs of this demographic are often left unfulfilled. And these needs have been clearly identified and analysed by scholars as early as the 1990s[12]. In recent years, social robots that integrate multiple technologies have been seen as an effective solution to this challenge[4][9].

Many researchers have come up with their own solutions for emotion recognition and feedback in robots. Aronsson argues that the introduction of emotional technologies into the care equation can lead to new relationships and ways of interacting[3]. Castillo et al. proposed a framework that can recognise emotions by studying physiological signals, facial expression and voice. Emotion regulation is enabled by tuning music, colour and light to the specific needs of the elderly[5]. Khosla et al. reported on the design and field trial of Matilda, a human-like assistive communication robot in Australian residential and community care settings[13]. Facial expression and voice analysis were used to identify and measure the emotional state of the participants during the interaction between elderly participants and the social robot Misa[6]. An empathic robot Ryan was created using a multimodal emotion recognition algorithm and a multimodal emotion expression system. Using different input modalities for emotion, i.e. facial expression and speech sentiment, the empathic Ryan detected the emotional state of the user and uses an affective dialogue manager to generate a response[1]. These studies provide compelling evidence of the potential for emotional technologies to revolutionize the care of the elderly. Through the integration of emotion recognition and feedback, robots have the potential to provide not only physical assistance, but also emotional support and companionship, thus improving the quality of life for the elderly population.

However, the above studies still suffer from several shortcomings in practical application scenarios. Firstly, some studies require the use of multiple sensor combinations, and deploying these complex sensor networks indoors would require extensive installation and commissioning. Second, emotion recognition systems based on facial expression recognition often rely on one or more cameras. These indoor cameras pose a significant privacy risk to the user. In addition, due to the relatively high real-time processing requirements, the computational process of emotion recognition and feedback can only be performed on the robot, making it difficult to apply algorithms that rely on server computing power in practice.

To address the above issues, a new solution has been proposed: the development of a social assistance robot for the elderly named DarumaTO. Fig. 1 illustrates the design of this innovative solution. The robot is designed with emotional interaction as a key component, with the aim of alleviating the loneliness and isolation commonly experienced by the elderly population. The robot is not equipped with a camera as shown in the Figure. Emotion recognition is performed by analysing speech data only, which avoids privacy issues and reduces the amount of data to be processed, allowing the algorithm to run on robots with relatively low computing power. The speech data was decomposed into audio data and textual data, and the voice emotion, semantic emotion and topic emotion embedded in them were analysed using convolutional neural networks (CNN) and long-short term memory (LSTM) algorithms respectively. The algorithm is first trained offline on an existing database and then deployed to the robot, reducing computation time while maintaining model performance. In addition, the robot's emotional feedback is achieved by dynamically generating appropriate robot facial expressions on the display.

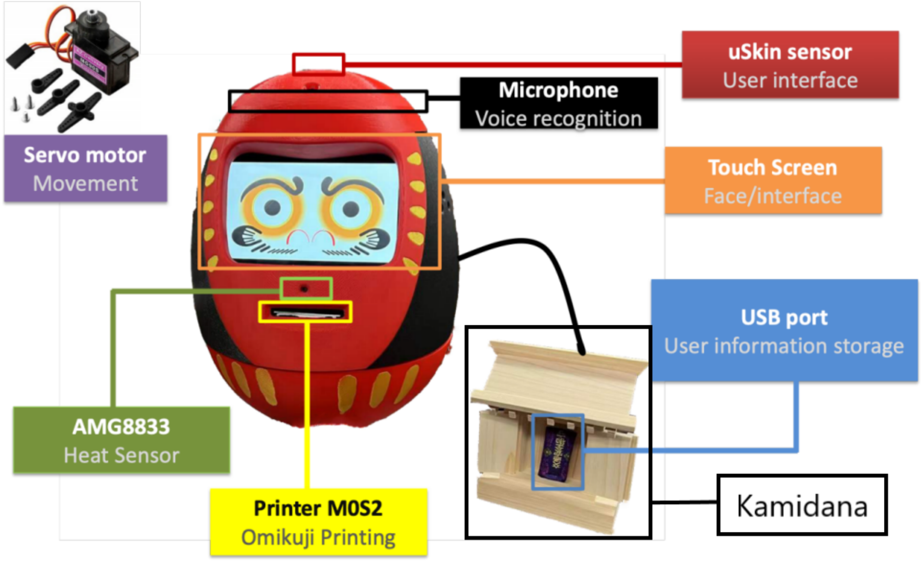


Fig. 1. The DarumaTO social assistive robot for elderly perple.

By incorporating emotional intelligence into the design of DarumaTO, the robot is able to provide not only physical assistance, but also companionship and emotional support to the elderly. The emotion recognition and feedback model introduced by this paper minimises the difficulty of operation for older users, while presenting the robot's feedback emotions in the most intuitive way. With its focus on emotional interaction, DarumaTO has the potential to revolutionize the way we provide care and support to the elderly.

2 DarumaTO Social Robot

The design of DarumaTO is specifically intended to evoke a sense of familiarity and comfort among Japanese elderly individuals, who may recognize the cultural significance of the Daruma doll. The robot is equipped with various sensors and actuators, allowing it to collect data from the environment and make decisions on the most effective way to interact with the user. In this sense, DarumaTO operates as a form of an IoT system. This enables the robot to respond to the needs and preferences of the user in real-time, providing a highly personalized and interactive experience. The design of DarumaTO highlights the importance of cultural awareness and sensitivity in the development of social robots for the elderly. By incorporating cultural elements into the design, these robots can provide elderly individuals with a sense of comfort and familiarity, and enhance the quality of the caregiving experience.

As shown in the Fig. 1, the Jetson Nano is responsible for managing the various inputs from the sensors and executing the algorithms that drive the

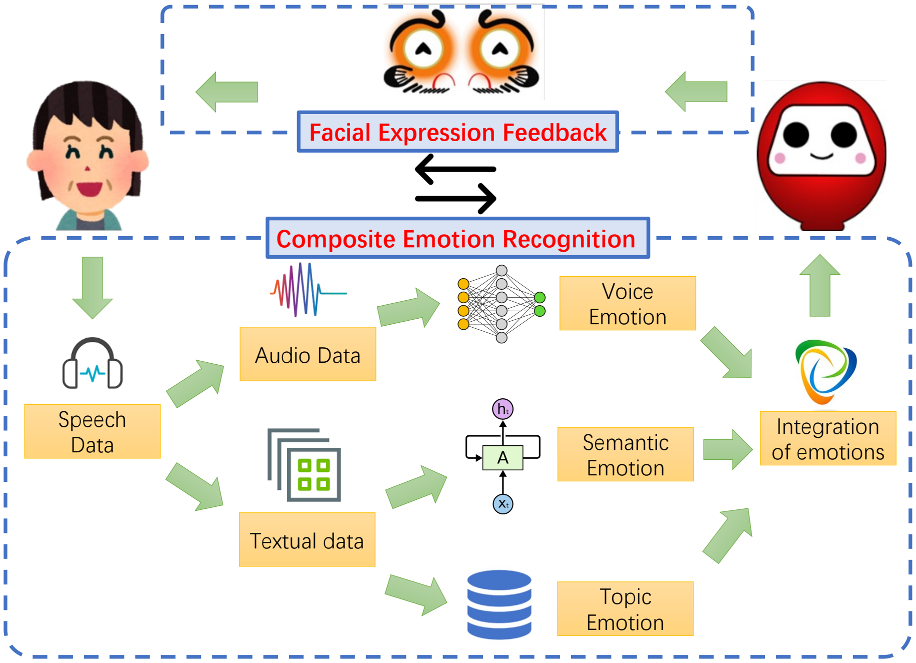


Fig. 2. Overall architecture of composite emotion recognition and feedback model.

robot's behavior. The sensors include a heat sensor, touch sensor, and microphone, which provide the robot with real-time information about the environment and the user's physical and emotional state. The touch screen interface allows the elderly to interact directly with the robot, providing a simple and intuitive means of communication. Additionally, the robot is equipped with two servo motors that control its Yaw and Pitch movements, allowing it to move in response to the user's actions and gestures. The printer feature enables DarumaTO to provide users with Omikuji (Japanese fortune-telling), advice, and reminders, adding a fun and engaging element to the overall experience. Finally, the external cabling to the Kamidana allows DarumaTO to read and transfer personal information, providing the robot with a more complete understanding of the user's preferences and needs.

3 Design of Composite Emotion Recognition

DarumaTO is able to communicate with people with simple dialogue as well as the facial expressions. Because of the privacy issue, DarumaTO does not equip with a camera. And this poses a challenge to its emotional interaction with older people. Different from the widely used recognition of user's emotion through visual information, this paper proposes a fused emotion recognition model through speech data. Under the premise of using only speech data, the user's emotion

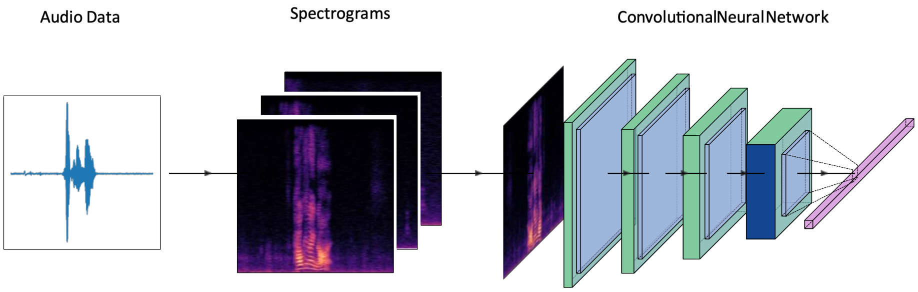


Fig. 3. Schematic visualization of the structure of the voice emotion recognition.

recognition is achieved by integrating the analysis of voice, semantic and topic. In addition, an emotional feedback model was designed to allow the robot to respond to the user's emotions by facial expressions appropriately. The above two parts constitute the emotional interaction model of DarumaTO, as shown in Fig. 2.

In the composite emotion recognition model, the speech data is first recorded and converted into audio data and textual data respectively. And the time-series audio data is transformed into time-frequency resolved spectrograms after audio decomposition. The tokenization and text irrelevant word elimination of textual data needs to be conducted before further analysis.

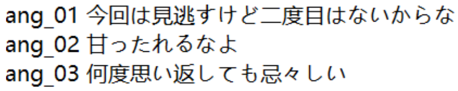
3.1 Voice Emotion

In order to recognize and respond to emotional cues in the voice, DarumaTO employs advanced signal processing techniques that analyze various acoustic parameters of speech. These parameters are known to reflect physiological changes that occur in the speech production system during emotional states, such as changes in respiration, vocal fold frequency, and articulation[11]. The voice emotional detection model in DarumaTO categorizes a sample of audio data into one of several emotional classes, such as joy, anger, or sadness. This is achieved through the analysis of the parameters of the acoustic waveform, including fundamental frequency, jitter, pitch, speech rate, speech pauses, syllable rates, intensity, energy, and relative energy in different frequency bands. The use of these parameters allows DarumaTO to accurately detect and respond to the emotional state of the user, providing a more engaging and personalized experience. By combining speech analysis with other modalities, DarumaTO provides a comprehensive solution for emotional recognition and feedback in social robots.

In practice, the voice emotion detection is based on the signal decomposition of time-series audio wave signals into time-frequency resolved spectrograms. Then, the spectrograms generated above are passed through a CNN with partially frozen weights, which allows for transfer learning in parts of the network. Some of these weights were frozen because the Japanese language database was too small to train the model. We first pre-trained the model using other language databases, and later used the Japanese database for transfer learning. Note that the new Japanese variant is currently based on the OGVC[2], MULTEXT-J[14], and KEIO-ESD[15]. And the emotion labels are placed at the output of the network, as shown in Fig. 3.

When we want to use the model to recognise emotion, the model expects clear audio input from the end user's speech. The input data should be in wave format and last for several seconds to provide a reasonable basis for estimation. The data must be sampled at a minimum of 16kHz sampling frequency and 16 bit depth (i.e. 256 kbit/s bit rate). The model works with speech samples as short as one second. However, detection will be more reliable if longer samples are provided.

3.2 Semantic Emotion



ang_01 今回は見逃すけど二度目はないからな
ang_02 甘ったれるなよ
ang_03 何度思い返しても忌々しい

Fig. 4. Examples of semantic emotion dataset.

The so-called semantic model is to judge emotion by analysing the meaning of the linguistic information in the conversation, using the method as the Recurrent Neural Network (RNN) in the popular NLP deep learning. The sound signal received from the microphone is converted into text that can be used for semantic analysis, which can be simply understood as "dictation". To create the training set for the model, a 600-character text dataset was created with the labels "angry", "happy", "neutral" and "sad" to generally cover the emotions and attitudes of human conversations in life, as shown in Fig. 4. However, after experimental validation, we found that this database was too small to train the proposed model, so we created the necessary Japanese database by translating an English database[16].

Language is a communication tool, which is essentially a system of signs. Semantic is the information meaning expressed by language, which can be the form of the objective world or subjective thoughts and feelings. And we assume the dialogue is sequence of words that contains the consciousness and emotions of the speaker. In this paper, LSTM network is used for semantic emotion recognition because of its ability to process sequential data. The textual data is converted into word vector available to the neural network after word embedding by word2vec. The LSTM network is constructed as shown in Fig. 5 and trained by the dataset we made.

For our training data set, most of the training samples are composed of one or two long and short sentences, which is very suitable for the characteristics of the LSTM model and can well capture the information and interaction relationship

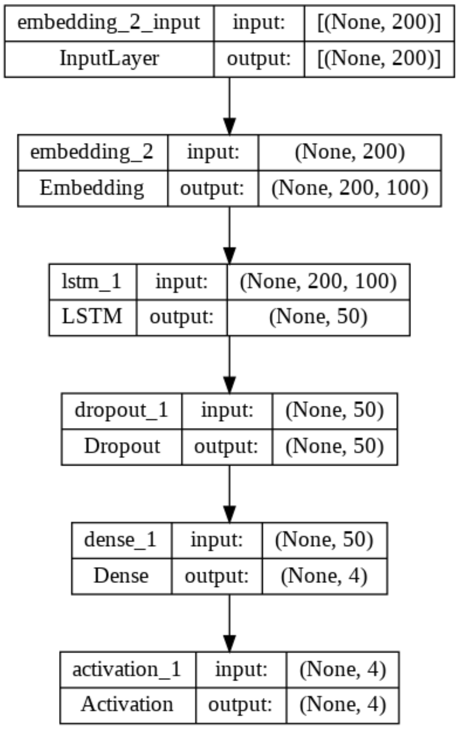

embedding_2_input | input: | [(None, 200)]
InputLayer | output: | [(None, 200)]
embedding_2 | input: | (None, 200)
Embedding | output: | (None, 200, 100)
lstm_1 | input: | (None, 200, 100)
LSTM | output: | (None, 50)
dropout_1 | input: | (None, 50)
Dropout | output: | (None, 50)
dense_1 | input: | (None, 50)
Dense | output: | (None, 4)
activation_1 | input: | (None, 4)
Activation | output: | (None, 4)

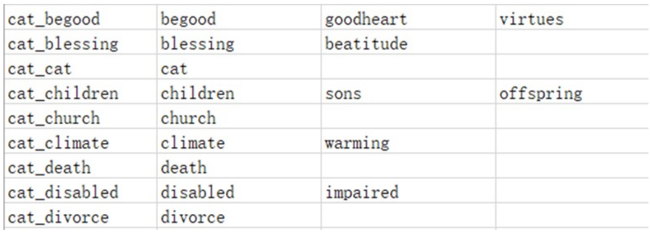
Fig. 5. The structure of LSTM for semantic emotion recognition.

of word combination in sentences. Therefore, the core algorithm of the model uses the LSTM algorithm to fit the data set.

3.3 Topic Emotion

The proposed topic analysis method in DarumaTO is a novel approach to speech recognition that captures keywords and emotions in daily conversations. The method involves the creation of a set of topic keywords that reflect the emotions frequently expressed in recent years. These keywords, such as "COVID-19," "headache," "stress," and "birthday," play an important role in determining the emotional direction of conversations and avoiding misjudgments due to data noise. Therefore, we constructed a database of topics and mapped the topics in the database to emotion tags, as shown in Fig. 6.

The topic analysis method enhances DarumaTO's ability to analyze and understand speech, as it allows for error correction and judgment in ambiguous situations. This approach represents a significant improvement over existing speech recognition systems, as it emulates the human ability to understand statements and emotional nuances in conversation.



cat_begood	begood	goodheart	virtues
cat_blessing	blessing	beatitude	
cat_cat	cat		
cat_children	children	sons	offspring
cat_church	church		
cat_climate	climate	warming	
cat_death	death		
cat_disabled	disabled	impaired	
cat_divorce	divorce		

Fig. 6. Examples of topic dataset.

3.4 Integration

Afterwards, the emotion labels identified by the three approaches are mapped to the same emotional space, which in turn generates a composite emotion vector. Thus, the emotions of the elderly can be recognized by means of dialogues with the DarumaTO.

$$\widehat{Y} = \frac{1}{N}\sum_{i=1}^{N}\omega_i y_i \tag{1}$$

As shown in the Equtation. 1, the results of each sentiment recognition model were mapped to the same dimensional space and a different result vector y was obtained separately. Correspondingly, the results of each model will have a confidence factor ω, which we use as a weight for their results. After a simple weighted average calculation, the final fusion result $\widehat{Y}$ can be obtained. In this paper, N is equal to 3, but offers the possibility of incorporating a wider range of emotion recognition models at a later stage.

4 Emotion Feedback with Facial Expression

The research shows that the facial display of a robot has a significant influence on the user experience of human-computer interaction[10]. And the screen mounted on top of the robot makes it possible to display a wide range of facial expressions. We therefore chose to use facial expressions as a channel for emotional feedback from the robot. DarumaTO's facial expressions consist of the eyes, eyebrows, nose, and beard. These organs are programmed to produce a certain degree of deformation and displacement to create different types of facial expressions.

Emotional feedback is not simply the reproduction of emotions or the feedback of opposing emotions to suppress negative emotions. Some people want to vent their feelings when they are angry or sad, or get empathy and understanding from others, so as to release the negative emotions in their hearts. Of course, when most people are faced with negative emotions, they can reconcile negative energy by receiving positive emotions, so as to achieve the function of comfort. Therefore, in this paper, we establish emotional feedback rules in such a way that older adults always receive positive emotional facilitation, as shown in Table. 1.

Table 1. Human emotion and corresponding facial expression feedback.

Human Emotion	DarumaTO's Reaction
Angry	Fear
	Shocked
	Angry
Joy	Smiling
	Laughing
Neutral	Smiling
	Neutral
	Laughing
Sad	Smiling
	Sad
	Crying

5 Experiments and Analysis

In this section, we separately validate the sentiment recognition models proposed in this paper. As topic-based emotion recognition is performed by means of keyword retrieval, its performance is not discussed in this section.

5.1 Voice Emotion

The results in Fig. 7 indicate that the performance of the DarumaTO's voice emotion detection system was tested on each independent validation set, and the performance was evaluated using various metrics. In total, nine emotions were identified. Compared to a baseline zeroR classifier, the weighted average recall (WAR) of the system improved slightly from 0.26 to 0.30, while the unweighted average recall increased from 0.11 to 0.27. The weighted F1-score also improved from 0.10 to 0.30.

5.2 Semantic Emotion

As mentioned earlier, we trained the model by translating the English database due to the insufficient size of the Japanese database. However, this approach does have some limitations. One such limitation is the potential bias in Google Translate's understanding of English to Japanese translations, which can lead to significant cultural differences in the dataset. To overcome this issue, the authors used the English translation dataset as the training data and the original Japanese speech text dataset as the test data to verify the model's ability to understand and analyze real Japanese cultural contexts and dialogue situations.

Despite these limitations, the preprocessing of the dataset and the algorithmic model remained unchanged, and the previously mentioned method was still used. After 100 times of complete learning and training of the whole data set, we obtained the accuracy curve and loss function curve respectively on the training

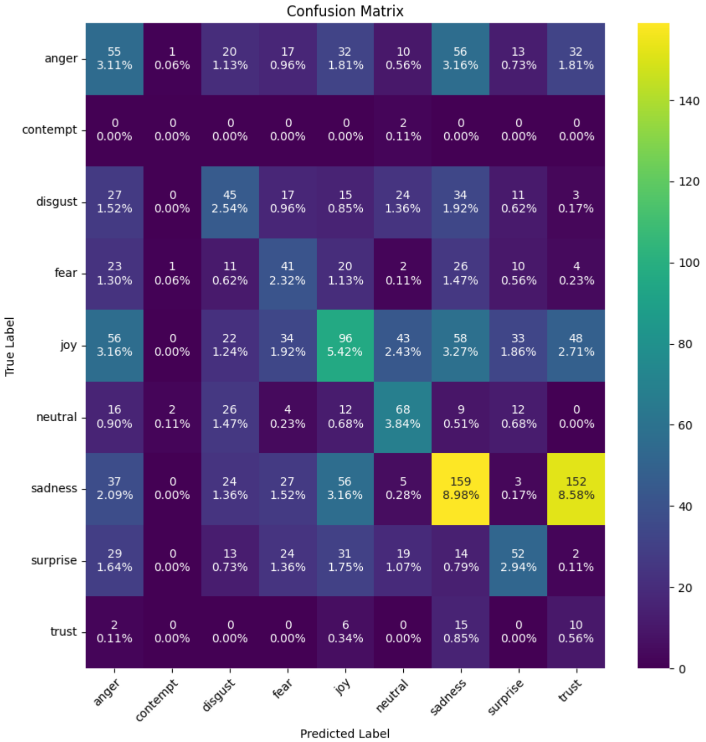


Fig. 7. Confusion matrix for the performance of voice emotion recognition in Japanese variant.

set and verification set. The accuracy of the new model was found to be 74% on the test dataset, as shown in Fig. 8. The loss was 0.52 and 0.69 on the training and validation sets, respectively.

6 Conclusion and Future Work

This study proposes a composite emotion recognition and feedback model for DarumaTO, a social robot for elderly users. The model performs emotion recognition in terms of voice, semantics and subject matter only through the analysis of the audio information acquired by the robot from the user. At the same time, the robot is made to make appropriate facial expressions based on the fused recognition results, which in turn enhances the interaction experience of the robot.

Although acceptable performance was achieved in the semantic emotion recognition section, the experimental results were poor in the voice emotion recognition. As the model was pre-trained in the English speech database and then transferred to the Japanese speech database, it was difficult to obtain good

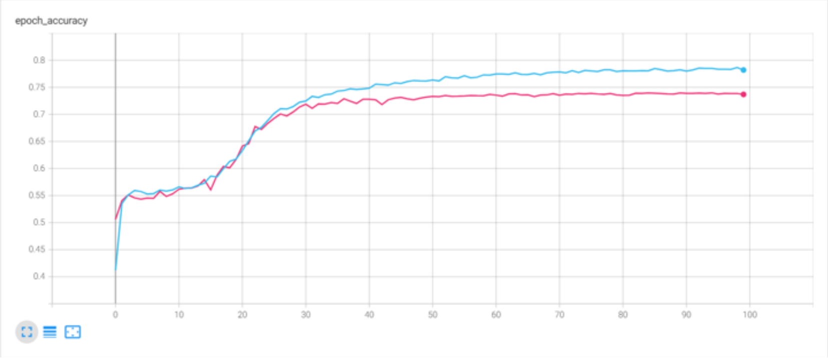


Fig. 8. Accuracy of semantic emotion recognition model.

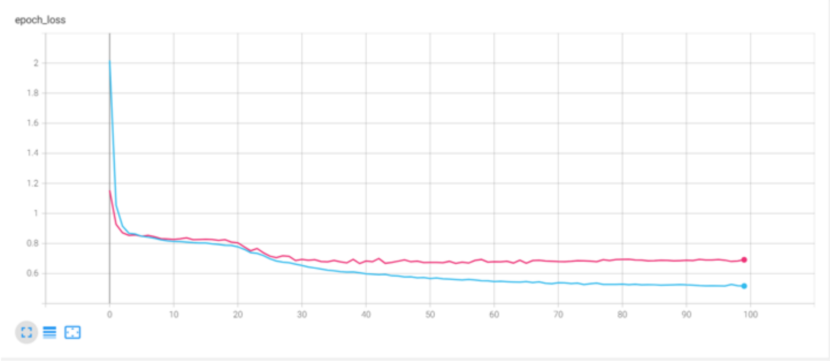


Fig. 9. Loss of semantic emotion recognition model.

training results with the small amount of data in the Japanese speech database. This reflects the fact that there are significant differences in pronunciation styles between languages and that audio features cannot be transferred between languages, especially when there is an unequal level of data.

In the future, we will implement the emotion recognition and feedback model proposed in this paper into DarumaTO and hand it over to the elderly in nursing homes for evaluation. We will verify the validity of the model proposed in this paper by means of group comparison experiments. The DarumaTO with immobile facial expressions will be used as the control group and the DarumaTO fitted with an emotion recognition and feedback model will be used as the experimental group. Questionnaires and interviews will be used to obtain the satisfaction of the elderly with the two groups of robots respectively.

Acknowledgements This work was supported by Ministry of Internal Affairs and Communications (MIC) of Japan (Grant no. JPJ000595).

References

1. Abdollahi, H., Mahoor, M., Zandie, R., Sewierski, J., Qualls, S.: Artificial emotional intelligence in socially assistive robots for older adults: a pilot study. IEEE Transactions on Affective Computing (2022)
2. Arimoto, Y., Kawatsu, H., Ohno, S., Iida, H.: Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment. Acoustical science and technology **33**(6), 359–369 (2012)
3. Aronsson, A.S.: Social robots in elder care the turn toward emotional machines in contemporary japan. Japanese review of cultural anthropology **21**(1), 421–455 (2020)
4. Broekens, J., Heerink, M., Rosendal, H., et al.: Assistive social robots in elderly care: a review. Gerontechnology **8**(2), 94–103 (2009)
5. Castillo, J.C., Fernández-Caballero, A., Castro-González, Á., Salichs, M.A., López, M.T.: A framework for recognizing and regulating emotions in the elderly. In: Ambient Assisted Living and Daily Activities: 6th International Work-Conference, IWAAL 2014, Belfast, UK, December 2-5, 2014. Proceedings 6. pp. 320–327. Springer (2014)
6. Demaeght, A., Miclau, C., Hartmann, J., Markwardt, J., Korn, O.: Multimodal emotion analysis of robotic assistance in elderly care. In: Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments. pp. 230–236 (2022)
7. Du, Y., Lim, Y., Tan, Y.: A novel human activity recognition and prediction in smart home based on interaction. Sensors **19**(20), 4474 (2019)
8. Du, Y., Lim, Y., Tan, Y.: Rf-arp: Rfid-based activity recognition and prediction in smart home. In: 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS). pp. 618–624. IEEE (2019)
9. Getson, C., Nejat, G.: Socially assistive robots helping older adults through the pandemic and life after covid-19. Robotics **10**(3), 106 (2021)
10. Hashimoto, T., Hitramatsu, S., Tsuji, T., Kobayashi, H.: Development of the face robot saya for rich facial expressions. In: 2006 SICE-ICASE International Joint Conference. pp. 5423–5428. IEEE (2006)
11. Johnstone, T., Scherer, K.R.: Vocal communication of emotion. Handbook of emotions **2**, 220–235 (2000)
12. Kasser, V.G., Ryan, R.M.: The relation of psychological needs for autonomy and relatedness to vitality, well-being, and mortality in a nursing home 1. Journal of Applied Social Psychology **29**(5), 935–954 (1999)
13. Khosla, R., Chu, M.T., Nguyen, K.: Enhancing emotional well being of elderly using assistive social robots in australia. In: 2013 International Conference on Biometrics and Kansei Engineering. pp. 41–46. IEEE (2013)
14. Kitazawa, S.: "prosody and speech processing" japanese multext prosodic corpus (multext-j) (2010)
15. Mori, S., Moriyama, T., Ozawa, S.: Emotional speech synthesis using subspace constraints in prosody. In: 2006 IEEE International Conference on Multimedia and Expo. pp. 1093–1096. IEEE (2006)
16. Mower, E., Matarić, M.J., Narayanan, S.: A framework for automatic human emotion classification using emotion profiles. IEEE Transactions on Audio, Speech, and Language Processing **19**(5), 1057–1070 (2010)