NO.1

Year 2023	**Summary of Thesis**

Student No.	Last name, First name
M2220200	Nitta, Yamato

(Title)
Research on a Machine Learning Model for Emotion Estimation Based on Multimodal Information from Voice and Text

Recently, AI technology for mimicking human sense and perception has made progress in the areas such as natural language and image processing. However, the AI generally handles a series of the processing as separated tasks without correlating each information for expressing the human emotion. This AI is referred to as 'singlemodal AI'. On the other hand, humans can recognize, combining various sensor information. This recognition style is referred to as 'multimodal'. The research on multimodal AI generally uses a machine learning approach in which voice and texts are input simultaneously [1]. The difficulty of the research is to reflect the output in emotion estimation similar to humans.

In this study, we assume multimodal in dialogue with AI. In general, when a human interacts with an AI, multimodal information such as voice, texts, facial expressions, and gestures are combined. Here, facial expressions and gestures are mainly affected by the emotions of the speaker and the other party. Therefore, we consider that the speaker's speech, sentences, and the speaker's emotion estimated from them are very important features for multimodal information. Usually, 8 emotions are used in emotion estimation by machine learning [2,3]. However, emotions that affect physical information in dialogue are assumed to be classified in a more coarse-grained manner and responded with facial expressions and gestures based on them, rather than in a fine-grained manner such as 8 emotions, because real-time processing is prioritized. We aimed to construct a multimodal machine learning modeling that can perform coarser emotion classification by integrating voice and texts, leveraging the conventional machine learning modeling of single-modal AI for emotion classification.

We extracted features from an online game voice chat corpus with emotion ratings (OGVC) using openSMILE. The features were trained on a Random Forest to estimate

Graduate School of Science and Technology, Chitose Institute of Science and Technology

NO.2

the emotional intensity of the voice. The results showed an average accuracy of 50 to 60% or better for all emotions. This study uses this emotion intensity estimation model. We re-trained BERT using the subjective and objective emotion analysis dataset (WRIME) to estimate the emotional intensity of texts. The results of emotion intensity estimation for texts showed that there was a good agreement for joy and anticipation, moderate agreement for disgust, fear, sadness and surprise, and poor agreement for anger and trust. This study uses this emotion intensity estimation model.

We tested NNs that can perform coarser emotion classification by integrating voice and texts. Specifically, we used OGVC to perform inference by inputting emotional features obtained by Random Forest and BERT into the NN. Precision was confirmed, since it is important that the predicted emotion is the actual emotion when considering a coarser emotion classification. Precision of the 8 emotions is 0.61 for trust, 0.07 for anger, 0.73 for anticipation, 0.12 for disgust, 0.19 for fear, 0.59 for joy, 0.21 for sadness, and 0.51 for surprise. Precision of the 5 emotions is 0.61 for trust/joy, 0.18 for anger/disgust, 0.73 for anticipation, 0.54 for fear/surprise, and 0.23 for sadness. Precision of the four emotions is 0.64 for trust/joy, 0.29 for anger/disgust/sadness, 0.73 for anticipation, and 0.48 for fear/surprise. These Precision results suggest that a coarser classification of emotions is possible when the four emotions are used. In addition, we believe that further integration of emotions cannot be achieved by responding with facial expressions and gestures.

Accordingly, in the construction of multimodal machine learning modeling that can perform coarser emotion classification by integrating voice and texts, it is shown that the parameter of 4 emotions is a coarser emotion classification that affects the physical information.

[1] G. Krishnamurthy, N. Majumder, S. Poria, E. Cambria, “A deep learning approach for multimodal deception detection”, ArXiv preprint arXiv:1803.00344

[2] W. Abe, D. Makabe, T. Kosaka, “Study on Training Data for Emotion Recognition in Spontaneous Dialogue Speech Using SVM”, IPSJ Tohoku Branch SIG Technical Report, 2016-7-A3-3

[3] H. Suzuki, K. Akiyama, T. Kajiwara, T. Ninomiya, N. Takemura, Y. Nakashima, H. Nagahara, “Emotional Intensity Estimation based on Writer’s Personality Information”, The 36th Annual Conference of the Japanese Society for Artificial Intelligence, 2022

Graduate School of Science and Technology, Chitose Institute of Science and Technology