An Emotion-based Korean Multimodal Empathetic Dialogue System

Minyoung Jung*[1], Yeongbeom Lim*[1,2], San Kim[1], Jin Yea Jang[1], Saim Shin[1], and Ki-Hoon Lee[2]

[1]AIRC, Korea Electronics Technology Institute, South Korea
[2]School of Computer and Information Engineering, Kwangwoon University, South Korea
{minyoung.jung, warf34, kimsan0622, jinyea.jang, sishin}@keti.re.kr
kihoonlee@kw.ac.kr

Abstract

We propose a Korean multimodal dialogue system targeting emotion-based empathetic dialogues because most research in this field has been conducted in a few languages such as English and Japanese and in certain circumstances. Our dialogue system consists of an emotion detector, an empathetic response generator, a monitoring interface, a voice activity detector, a speech recognizer, a speech synthesizer, a gesture classification, and several controllers to provide both multimodality and empathy during a conversation between a human and a machine. For comparisons across visual influence on users, our dialogue system contains two versions of the user interface, a cat face-based user interface and an avatar-based user interface. We evaluated our dialogue system by investigating the dialogues in text and the average mean opinion scores under three different visual conditions, no visual, the cat face-based, and the avatar-based expressions. The experimental results stand for the importance of adequate visual expressions according to user utterances.

1 Introduction

As dialogue systems for human-machine conversations have attracted attention from the public, various multimodal dialogue systems with the purpose of healthcare (Wada and Shibata, 2007), empathetic conversation (Ishii et al., 2021) or multiparty attentive listening (Inoue et al., 2021b) have been recently introduced because multimodality makes conversations more entertaining (Pollmann et al., 2020). Most research in this field has been conducted by few research groups in industry or university because of the complicated architecture inherent in multimodal dialogue systems to control multimodal recognition or representation. Consequently, most multimodal dialogue systems are limited to a few languages such as English and Japanese.

Empathy is also the main factor for more humanized conversation (Zech and Rimé, 2005) along with multimodality. Researches on empathetic dialogues (Lin et al., 2020; Zheng et al., 2021; Zhong et al., 2020; Li et al., 2021; Kim et al., 2021a; Sabour et al., 2022) are also focused on a few languages from a lack of empathetic dialogue datasets. Although a Korean empathetic dataset (Yang et al., 2020) and a Korean empathetic dialogue generation model (Jang et al., 2022) have been recently published, a Korean empathetic dialogue system supporting multimodality has not been studied.

This paper makes the following contributions:

1. We propose an emotion-based Korean multimodal empathetic dialogue system composed of an emotion detector, an empathetic response generator, a monitoring interface, a voice activity detector, a speech recognizer, a speech synthesizer, a gesture classification, and several controllers.

2. We provide three different visual-representing conditions to compare the user's behaviors and opinion scores. The three conditions include no visualization (a black screen), a cat face-based emotion expression, and an avatar-based gesture expression.

3. We evaluate our dialogue system with six participants collected for our experiments. The experiments are performed under three different visual-expressing conditions. We analyze the experimental results which are dialogues in text form and average mean opinion scores.

The remainder of this paper is formed as follows. We explain our emotion-based Korean multimodal empathetic dialogue system in Section 2. In Section 3, the experimental results of our dialogue system are discussed. Section 4 contains the related

*Equal contribution

Proceedings of the 2nd Workshop on When Creative AI Meets Conversational AI (CAI2), COLING 2022, pages 16–22
October 12–17, 2022.

work in multimodal dialogue systems and empathetic dialogues. Finally, we draw our conclusion in Section 5.

2 Empathetic Dialogue System

We illustrate the emotion-based Korean multimodal empathetic dialogue system. As shown in Fig. 1, the overall architecture of the dialogue system is composed of modules on a device and server(s). The device must be equipped with at least a microphone, a speaker, a display, and a computer for voice activity detection, speech recognition, speech synthesis, and visual expression. The visual expression is derived from either a cat face-based emotion expression (V1) or an avatar-based gesture expression (V2). The modules on server(s) are an emotion detector, an empathetic response generator, a monitoring service, and the main controller to receive inputs (user information and a user speech in text) from the device and to send outputs (a system response in text, a detected emotion class, and estimated probabilities of a user emotion and a system dialogue strategy) to the device. Those modules can operate on the device instead of server(s) if the computing and memory resources on the device afford them. Otherwise, they can be executed on a single server or several servers in consideration of the resources on the server(s).

2.1 Emotion Classification Model

For generating more empathetic responses, utilization of user emotions is essential. Therefore we need an emotion classification model recognizing the user's emotion from the current user utterance among happy, sad, fear/anxiety, angry, surprise, disgust, and neutral in accordance with Ekman's six basic emotions (Ekman, 1992). The text emotion classification model (Lim et al., 2021) on the basis of Korean-English T5 (KE-T5) (Kim et al., 2021b), a T5 (Raffel et al., 2020)-based pre-trained model for both English and Korean, is adopted as the emotion detection model in our architecture. And the emotion detection model is re-trained on the extended version of the Korean empathetic conversation corpus (Yang et al., 2020) because the dataset used in (Lim et al., 2021) is on the basis of eight emotions.

2.2 Dialogue Generation Model

The dialogue generation model aims to automatically generate system responses in an empathetic manner, based on the latest three user utterances by utilizing the user emotion and the system's dialogue strategy. The user emotion is decided among the seven emotions as defined in Section 2.1, and the system dialogue strategy is determined among clarification, back-channel, facilitation, approval, disapproval, surprise, encouragement, evaluation, echoic, greeting, opinion, suggestion, and persona according to the extended version of the Korean empathetic conversation corpus (Yang et al., 2020). The KE-T5-based empathetic dialogue model (Jang et al., 2022) is employed as the empathetic response generation model in our architecture after the model is re-trained on the extended version of the Korean empathetic conversation corpus (Yang et al., 2020) because the persona class is added to the strategy classes.

2.3 User Interface

For human-machine multimodal interaction, we provide two versions of a user interface which are a cat face-based and an avatar-based user interface. Whenever our empathetic dialogue system starts, either of them can be chosen to deliver adequate visual-representation to the system responses. Both versions receive user information such as a user ID and user voice in speech. Once the user voice is detected, the speech recognition (speech to text) of the Web Speech API transforms the voice into the text so that emotion detection and empathetic response generation modules can obtain and process the text through the main controller on a server. After the emotion detection and empathetic response generation modules produce the recognized user emotion and the system response in the form of text, their outputs are sent to the chosen version of the user interface for the motion expression and the speech synthesis (text to speech).

2.3.1 Cat Face-Based User Interface

The first version (V1) of the user interface, a cat face-based Web user interface, receives the generated system response in text form and the detected user emotion for the speech synthesis and the emotion expression respectively. According to the emotion types in Section 2.1, seven different cat face-based motions are designed to express the user's emotion as shown in Fig. 2. The device can therefore provide the audio and visual interaction simultaneously to the user, through the audio controller and the emotion expression controller.

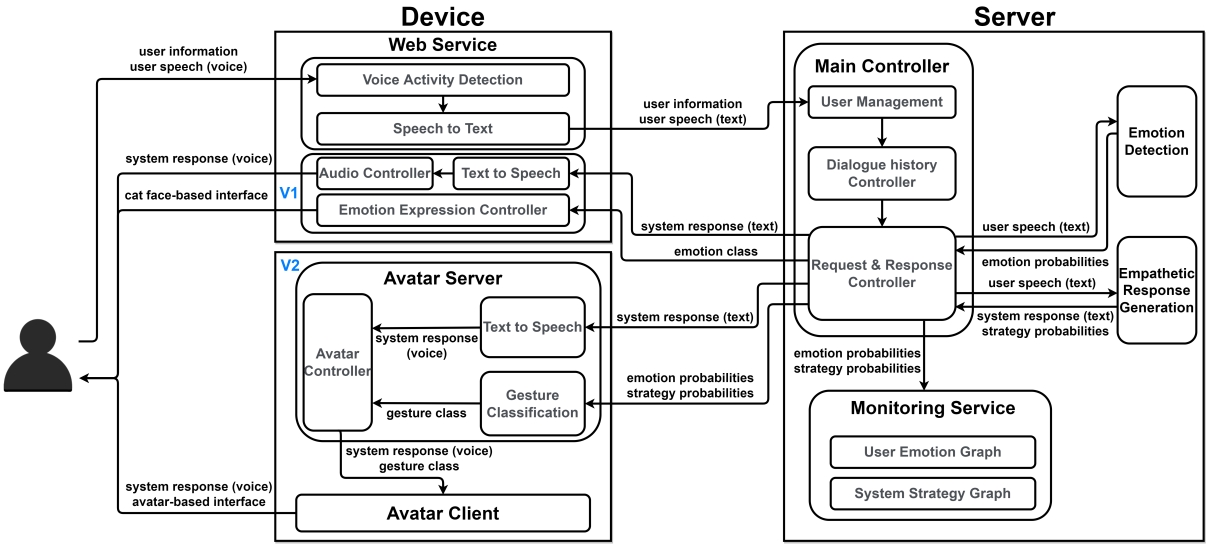

Figure 1: Overall architecture of our emotion-based Korean multimodal empathetic dialogue system

(a) Happy (b) Sad (c) Fear/Anxiety (d) Angry (e) Surprise (f) Disgust (g) Neutral

Figure 2: Seven different cat face-based motions

2.3.2 Avatar-Based User Interface

The second version (V2) of the user interface, an avatar-based Unity user interface, receives the generated system response in the form of text, the detected user emotion, and the suggested system dialogue strategy for speech synthesis and gesture expression. The current gesture classification module randomly selects a gesture from the seven different general-purpose avatar gestures as depicted in Fig. 3. The gestures include holding out one hand (A) or both hands (D), tilting (B) or nodding (E) the head, crossing the arms (C), and putting one hand (F) or both hands (G) on the chest. If some specific-purpose gestures are added afterward, the gesture classification module can utilize the given user emotion and system strategy to choose a more appropriate gesture for future work. The synthesized system voice in speech and the chosen gesture class are transmitted to the avatar controller so that the avatar server can send both information to the avatar client. Then the avatar client on the device can play the voice and gesture motion concurrently.

2.3.3 Monitoring Interface

The monitoring web interface is provided for participants so that they can check their current and some recent past emotions, and the current system dialogue strategy, as illustrated in Fig. 4. The x-axis and y-axis of the user emotion graph represent the time when the emotion is detected and the estimated emotion probabilities. And the system dialogue strategy probabilities are presented in the radial graph.

3 Experiments

For evaluating our emotion-based Korean multimodal empathetic dialogue system, we analyze the dialogue logs and the averaged mean opinion scores (MOS) achieved by six participants. MOS is commonly used to assess the dialogue system since no existing automatic evaluation metrics correctly measure the performance of the dialogue generation task. Our dialogue system was also evaluated in three different visual-representing conditions which are no visual (a black screen), the cat face-based, and the avatar-based expression methods.

3.1 Experimental Settings

A 160 cm kiosk built in a microphone, a speaker, a display, and a computer is employed for all our experiments conducted with six participants and

(a) Gesture A (b) Gesture B (c) Gesture C (d) Gesture D (e) Gesture E (f) Gesture F (g) Gesture G

Figure 3: Seven different general-purpose avatar gestures

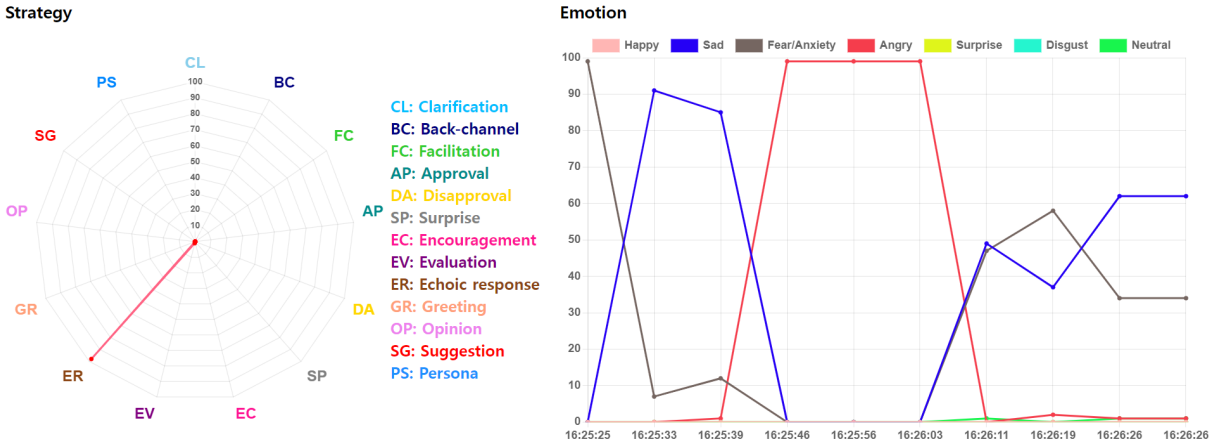

Figure 4: Monitoring interface drawing estimated probabilities of a current system strategy and latest user emotions

three visual expressing conditions. A participant starts a conversation with the kiosk given the condition, finishes the conversation when the participant wants, has a pause while other participants have a conversation with the kiosk, starts another conversation with the kiosk under another condition different from the first condition, and iterates the same steps until the participant tests all three visual conditions. The order of conditions given to each participant is randomly shuffled so that the evaluation results are not affected by the order.

For the speech synthesis, the Kakao text-to-speech API is selected because it provides a calm female voice in Korean, which sounds proper for most empathetic dialogues.

3.2 Experimental Results

For observing the changes in terms of participants' behavior, the dialogue logs were recorded individually depending on the participant and the visual condition. The numbers of user utterances per dialogue and words per user utterance are calculated on average, as shown in Table 1. The average number of words per user utterance for all three conditions is almost the same, whereas the users tend to talk less with the cat face and more with the avatar.

The participants graded each evaluation item on a 5-point scale from 1 to 5. A participant considers an evaluation item very bad if the participant scores 1 for the item, whereas scoring 5 means very good. The questionnaire was given to the participants before the experiment and contained the questions as described in Table 2. Except for Q4, all participants gave a mark for each conversation under a given visual condition. Question Q4 was only rated when no black screen was provided. We observed that the participants gave higher MOS with the cat face although we utilize the same emotion detector and the empathetic dialogue generator for all conditions. In case of question Q4, the participants considered that the emotion-based cat face expression was more proper than the random general purpose gesture-based avatar expression. The overall satisfaction scores (Q5) showed that the participants were the most satisfied with the cat face and the least satisfied with the avatar. The result that the avatar-based representation achieved lower MOS than the black screen implies the importance

Evaluation item	None	Cat face	Avatar
The average number of user utterances per dialogue	17.0	16.3	**18.3**
The average number of words per user utterance	3.8	4.0	4.0

Table 1: Average numbers of user utterances per dialogue and words per user utterance under three visual conditions

	Evaluation item	None	Cat face	Avatar
Q1	The recognized emotion was correct	**4.2**	**4.2**	**4.2**
Q2	The system strategy was appropriate	3.8	**4.0**	3.7
Q3	The system response was appropriate	**4.0**	**4.0**	3.3
Q4	The cat face or avatar gesture matched with the system response	n/a	**3.8**	2.8
Q5	The overall dialogue satisfied me	4.0	**4.2**	3.3

Table 2: Average mean opinion scores under three visual conditions

of providing appropriate visual-representation by understanding given user utterances.

4 Related Work

Several social robots providing multimodal interaction have been introduced for different purposes. The baby seal-shaped robot PARO was developed by the National Institute of Advanced Industrial Science and Technology in Japan for robot therapy (Wada and Shibata, 2007). And the PARO robot was utilized for examining whether the robot can support family caregivers caring for older persons with dementia (Inoue et al., 2021a). The Pepper robot, a wheeled humanoid robot produced by SoftBank Robotics, was initially designed for business-to-business in SoftBank stores and has been utilized for a variety of applications for business-to-consumer, business-to-academics, and business-to-developers (Pandey and Gelin, 2018). (Glas et al., 2016) created the ERICA robot, one of the most humanlike android robots, whose functionalities include conversation, advanced sensing, and speech synthesis. And the abilities of the ERICA robot extended into one-on-one attentive listening (Inoue et al., 2020) and multi-party attentive listening (Inoue et al., 2021b). The ERICA robot was also utilized for empathetic conversation during the Covid-19 quarantine (Ishii et al., 2021).

As empathy plays a crucial role in communication, there have been several attempts to generate more empathetic system responses in text-based conversations. An end-to-end empathetic chatbot CAiRE (Lin et al., 2020) recognizes user emotions and generates responses in an empathetic manner, based on the Generative Pre-trained Transformer (Radford et al., 2018). (Zheng et al., 2021) proposed a multi-factor hierarchical framework for empathetic response generation, which consists of communication mechanism, dialog act, and emotion. (Zhong et al., 2020) suggested a novel large-scale dataset (PEC) and a BERT (Devlin et al., 2019)-based response selection model for persona-based empathetic conversations. (Li et al., 2021) and (Kim et al., 2021a) focused on emotion causes for generating empathetic responses. (Sabour et al., 2022) leveraged commonsense to achieve additional information such as user's situations and feelings. And the information was utilized for the enhancement of empathetic response generation.

5 Conclusion

This paper proposes an emotion-based Korean multimodal empathetic dialogue system whose submodules include an emotion detector, an empathetic response generator, a monitoring interface, a web interface, and a unity interface. We evaluated our dialogue system by analyzing the dialogues in text and the average mean opinion scores under the three different visual-representing conditions and observed the significance of proper visual expressions. For future research, gesture classification with more specific-purpose gestures and system emotion expression corresponding to the system response will be considered.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00320, Artificial intelligence research about cross-modal dialogue modeling for one-on-one multi-modal interactions)

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Dylan F. Glas, Takashi Minato, Carlos T. Ishi, Tatsuya Kawahara, and Hiroshi Ishiguro. 2016. Erica: The erato intelligent conversational android. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 22–29.

Kaoru Inoue, Kazuyoshi Wada, and Takanori Shibata. 2021a. Exploring the applicability of the robotic seal paro to support caring for older persons with dementia within the home context. *Palliative Care and Social Practice*, 15:26323524211030285. PMID: 34350398.

Koji Inoue, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2020. An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 118–127, 1st virtual meeting. Association for Computational Linguistics.

Koji Inoue, Hiromi Sakamoto, Kenta Yamamoto, Divesh Lala, and Tatsuya Kawahara. 2021b. A multi-party attentive listening robot which stimulates involvement from side participants. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 261–264, Singapore and Online. Association for Computational Linguistics.

Etsuko Ishii, Genta Indra Winata, Samuel Cahyawijaya, Divesh Lala, Tatsuya Kawahara, and Pascale Fung. 2021. ERICA: An empathetic android companion for covid-19 quarantine. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 257–260, Singapore and Online. Association for Computational Linguistics.

Jin Yea Jang, San Kim, Minyoung Jung, and Saim Shin. 2022. Utilization of emotions and strategies in generating system response of the empathetic dialogue models. In *ICEIC*.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021a. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

San Kim, Jin Yea Jang, Minyoung Jung, and Saim Shin. 2021b. A model of cross-lingual knowledge-grounded response generation for open-domain dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 352–365, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yanran Li, Ke Li, Hongke Ning, Xiaoqiang Xia, Yalong Guo, Chen Wei, Jianwei Cui, and Bin Wang. 2021. Towards an online empathetic chatbot with emotion causes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 2041–2045, New York, NY, USA. Association for Computing Machinery.

Yeongbeom Lim, San Kim, Jin Yea Jang, Saim Shin, and Minyoung Jung. 2021. Ke-t5-based text emotion classification in korean conversations. In *HCLT*, pages 496–497.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. Caire: An end-to-end empathetic chatbot. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13622–13623.

Amit Kumar Pandey and Rodolphe Gelin. 2018. A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine*, 25(3):40–48.

Kathrin Pollmann, Christopher Ruff, Kevin Vetter, and Gottfried Zimmermann. 2020. Robot vs. voice assistant: Is playing with pepper more fun than playing with alexa? In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, pages 395–397, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11229–11237.

Kazuyoshi Wada and Takanori Shibata. 2007. Living with seal robots—its sociopsychological and physiological influences on the elderly at a care house. *IEEE Transactions on Robotics*, 23(5):972–980.

Jae Hee Yang, Jin Yea Jang, Minyoung Jung, and Saim Shin. 2020. Establishing a corpus for an ai-based empathic response system. In *ICONI*.

Emmanuelle Zech and Bernard Rimé. 2005. Is talking about an emotional experience helpful? effects on emotional recovery and perceived benefits. *Clinical Psychology & Psychotherapy*, 12(4):270–287.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. CoMAE: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824, Online. Association for Computational Linguistics.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.