Command Feedback System Based on Context Awareness for Minimizing Control Error in Human-Robot Interaction

1st Taein Yong
Department of Information and Communications Engineering
Sejong University
Seoul, Republic of Korea
k62570@gmail.com

2nd Pyeongjoo Kim
Department of Information and Communications Engineering
Sejong University
Seoul, Republic of Korea
pyongjoo.sejong@gmail.com

3rd Juyeon Weon
Department of Information and Communications Engineering
Sejong University
Seoul, Republic of Korea
juyeon.sejong@gmail.com

4th Yonghyun Kwon
Department of Information and Communications Engineering
Sejong University
Seoul, Republic of Korea
yonghyun.sejong@gmail.com

Jaeho Kim*
Department of Information and Communications Engineering
Sejong University
Seoul, Republic of Korea
kimjh@sejong.ac.kr

***Abstract*—The advancement in Artificial Intelligence has made it possible to control the robots through intuitive interfaces such as speech and gestures. A robotic control system for human-robot interaction has been developed, enabling robot control using human speech or hand gestures. Speech and gesture recognition technology has become a part of robotic control systems, acting as a bridge for interaction between humans and robots. However, simple robotic control systems process control commands independently, without considering the situation in which the robot is located. This leads to collision risks and a decline in control reliability. This paper introduces a Command Feedback System based on Context Awareness (CFS-CA) to minimize control errors by integrating real-time situational analysis with user commands. The system provides context-aware feedback, enabling more accurate and efficient robotic control. Experiments performed in a virtual environment with seven participants demonstrated that CFS-CA significantly reduces control errors and improves performance across key metrics, including the number of commands, driving time, collisions, and path efficiency, ensuring safer and more reliable robotic control.**

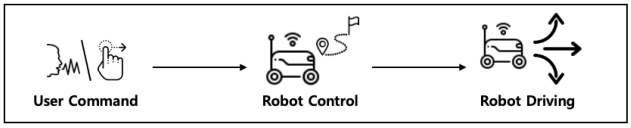
***Index Terms*—Human-robot interaction, Feedback, Context awareness, Speech recognition, Gesture recognition**

I. Introduction

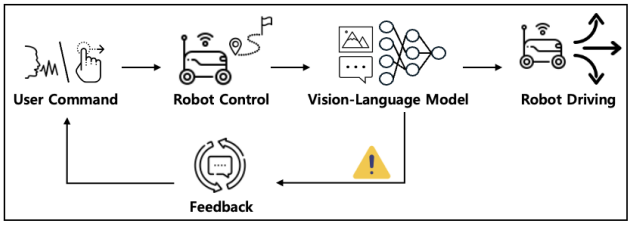
Recent advancements in Artificial Intelligence have been observed to bring significant attention to technologies that control robots through speech and gestures, leading to the emergence of various robotic control systems. Human-robot interaction (HRI) through speech and gestures enhances the intuitiveness and efficiency of robotic control systems, playing a crucial role in various applications [1], [2]. Speech and gesture-based robotic control systems provide user-friendly interfaces by leveraging intuitive human input and real-time responsiveness [3], [4]. In particular, continuous research on deep learning and computer vision technologies has significantly improved the accuracy of speech and gesture recognition [5]. As a result of these advancements and the intuitive nature of speech and gestures, they have been recognized as essential means of HRI, with their practicality gradually expanding [6], [7]. However, simple robotic control systems are designed to process control commands independently, focusing solely on basic robotic control. There is a limitation in that the situation in which the robot is located is not considered.

Conventional speech [8] and gesture-based robotic control systems [9], as shown in Fig. 1(a), are designed such that commands input by the operator are executed as robotic control commands if they exceed a certain accuracy threshold. However, if incorrect commands are input into the system and executed, it may lead to robot malfunction or damage, as well as harm to the surrounding environment. This risk becomes more pronounced in complex or visually constrained situations, where such issues could result in critical failures within robotic control systems in HRI. Additionally, there is a concern that commands incompatible with the robot's real-world context may be executed due to the operator's carelessness, compromising safety and reliability. This design does not account for critical aspects of HRI, including the recognition of the operator's intent and adaptability to situational contexts, ultimately reducing user convenience and operational efficiency.

In this study, to minimize control command errors, a Command Feedback System based on Context-Awareness (CFS-CA) was designed by integrating context awareness into the robotic control system, as shown in Fig. 1(b). This system



(a) General robot control system



(b) Command Feedback System based on Context Awareness (CFS-CA)

Fig. 1. Concept of CFS-CA

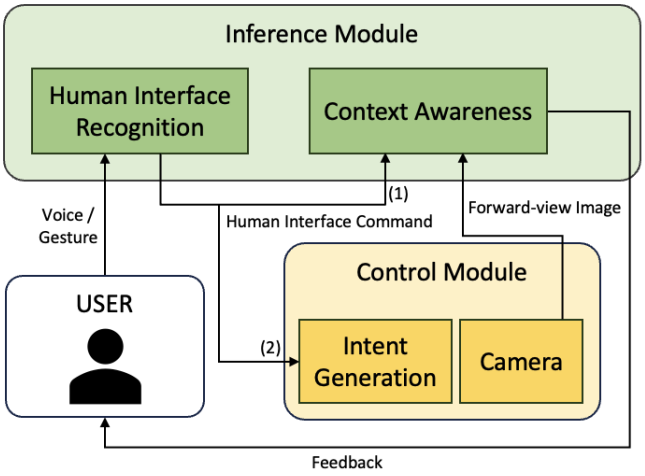


Fig. 2. System architecture of CFS-CA

simultaneously analyzes questions including the operator's command and the robot's surrounding environment in real time, providing command feedback based on situational analysis. By analyzing the robot's surrounding environment and recommending more accurate commands to the operator, the system aims to minimize control error and enable more stable, efficient, and contextually appropriate robotic control. To validate the efficiency of the proposed system, we conducted a robotic control experiment with seven participants in a virtual environment modeled after the 5th floor of the Daeyang AI Center at Sejong University.

This paper is organized as follows: Section II introduces the internal modules that consist the CFS-CA. Section III describes the experimental environment and the artificial intelligence technologies incorporated into the system. Section IV provides experimental results to verify whether the proposed system meets its intended objectives. Finally, Section V concludes with a summary of the study and a brief discussion of future work.

II. System Architecture of CFS-CA

The CFS-CA is designed to minimize command errors during robotic operation through feedback. The architecture of CFS-CA, as shown in Fig. 2, consists of User, Inference Module, and Control Module. User's speech and gestures are converted into Human Interface Command through Human Interface Recognition. First, a question containing the converted command and the image captured by the forward-view camera mounted on the robot are input into Context Awareness, where situational analysis and feedback generation are performed. Context Awareness determines whether the user's initial command is an appropriate control command and generates feedback accordingly. The user's final command, adjusted based on the feedback, is used to generate the robot control intent in Intent Generation through Human Interface Recognition.

A. Inference Module of CFS-CA

- **Human Interface Recognition:** Human Interface Recognition is part of Inference Module and processes the user's speech and gestures into actionable commands. The speech recognition model converts the user's speech into text, regarding the gesture recognition model, it generates control commands using the joint values and angles of both detected hands in the image. This module serves as a connecting intermediary that enables humman-robot interaction within the CFS-CA.
- **Context Awareness:** Context Awareness simultaneously receives questions, including human interface commands, in text form and the forward-view image as inputs. Additionally, it is multimodal, which is capable of processing text and images within a single model, and generates feedback based on whether the control command is executable in the robot's environment. The generated feedback is presented in text format and is displayed on the screen for the user's reference.

B. Control Module of CFS-CA

- **Camera:** Camera in Control Module transmits real-time images captured by the robot's forward-view camera to the context awareness model. The model analyzes the robot's surrounding environment using forward-view images and evaluates the appropriateness of the command.
- **Intent Generation:** Intent Generation takes Human Interface Command, refined by feedback, as input and generates robot control intents containing speed, direction, and angular velocity. The generated control intent is transmitted to the robot's hardware to perform the actual actions.

III. Experimental Setup

We conducted experiments to verify that the CFS-CA system enables more accurate and efficient robot control by providing context-aware feedback. To achieve this, we

created a virtual environment modeled after the 5th floor of the Daeyang AI Center at Sejong University. The virtual environment was constructed using Isaac Sim, and control experiments were performed with a four-wheeled robot involving seven participants. Human Interface Recognition in the CFS-CA Inference Module includes a speech recognition model and a gesture recognition model. For the speech recognition model, the Whisper model [10] was fine-tuned to improve the accuracy of speech-based control commands. For the gesture recognition model, gesture-based control commands were generated by training CNN-LSTM network on joint values extracted through MediaPipe [11]. Context Awareness in the CFS-CA Inference Module utilized the Vision-Language model, LLaVA [12].

A. Simulation Environment

We selected the four-wheeled robot as the Unmanned Ground Vehicle (UGV) and conducted experiments in the virtual environment we generated. The indoor space was scanned using a LiDAR sensor to create a 3D map. Additionally, rendering techniques were applied to add textures, visualizing the indoor environment. Fig. 3 shows the virtual environment we constructed. Fig. 3(a) shows the LiDAR-scanned indoor virtual environment with textures added through rendering. Fig. 3(b) illustrates the virtual box and path designed for the control experiments. Fig. 3(c) indicates the starting point of the four-wheeled robot and Fig. 3(d) shows its target endpoint.

(a) LiDAR-scanned indoor virtual environment

(b) Control experiment environment

(c) Starting point of robot driving

(d) Target endpoint of robot driving

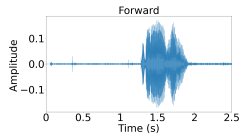
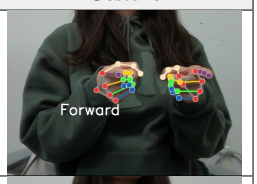
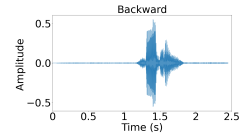
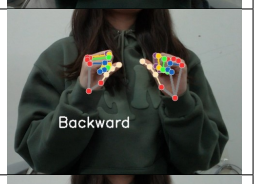
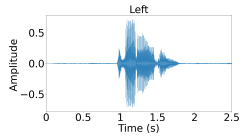
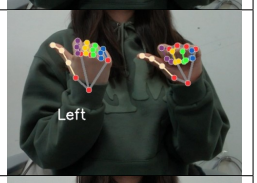
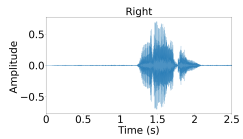
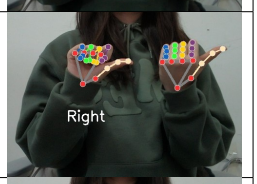
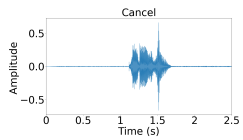
Fig. 3. Simulation environment

B. Speech and Gesture Models for Robot Control

We fine-tuned the Whisper-small model to serve as the speech recognition component of our system. The fine-tuning process involved 13 participants constructing 12 instruction sets, and fine-tuning the model using the dataset consisting of 2,087 instruction sentences. Among the 12 commands, five were selected as the most suitable for controlling the four-wheeled robot: forward, backward, left, right, and cancel. In addition, parameters such as the robot's linear velocity and angular velocity can be specified by speech input.

For gesture recognition, 42 joint values of both hands were extracted using MediaPipe. These joint values were processed by CNN to maximize feature extraction and further analyzed by LSTM to recognize dynamic gestures, forming CNN-LSTM model [13]. A total of 19 different commands were collected from 17 participants, with 500 samples per command. From this dataset, only five commands that were identical to the speech commands were used. The linear velocity of the robot in response to gestures is set to 2 m/s, while the angular velocity is 1 rad/s. Speech and gesture commands are shown in Table. I.

TABLE I
COMMAND OF SPEECH AND GESTURE

Command	Speech	Gesture
Forward		
Backward		
Left		
Right		
Cancel		
Linear Velocity & Angular Velocity	Speech Parameters	2 m/s & 1 rad/s

C. Vision-Language Model for Generate Feedback

The Vision-Language model, LLaVA was used to process a question containing a human interface command along with the front-view image. By processing these two inputs within a unified model, the Vision-Language model generates efficient feedback that accounts for environmental factors.

IV. Experimental Results and Discussions

We conducted experiments with seven participants controlling the four-wheeled robot in the indoor virtual environment. To evaluate the efficiency of the system, control experiments

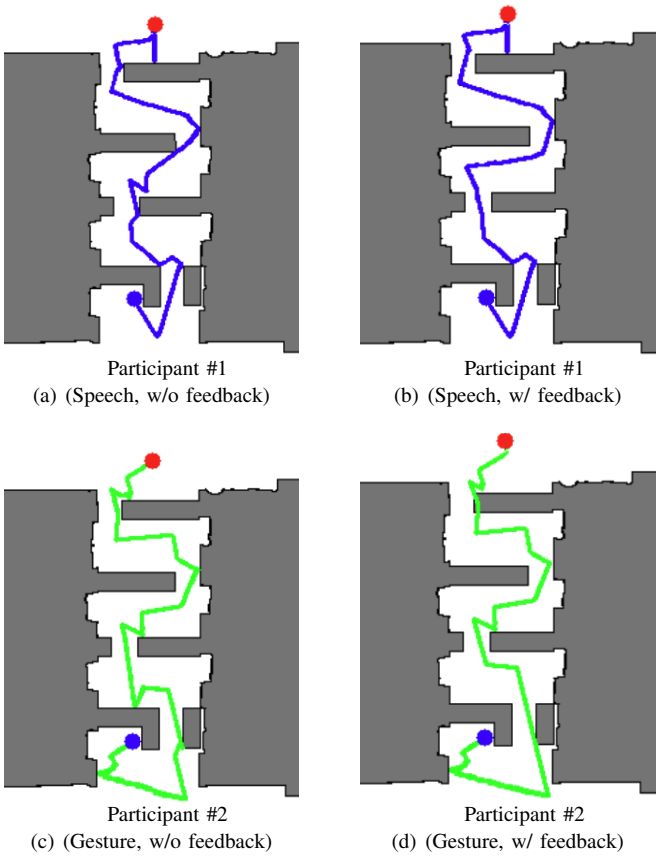
TABLE II
PERFORMANCE METRICS OF CFS-CA

Evaluation Metrics		Speech								Gesture							
		#1	#2	#3	#4	#5	#6	#7	Avg.	#1	#2	#3	#4	#5	#6	#7	Avg.
Number of Commands [Times]	w/o Feedback	40	35	46	46	55	49	39	43	40	61	52	42	**35**	50	53	48
	w/ Feedback	**35**	**31**	**35**	**32**	**42**	**37**	**34**	**35**	**35**	**50**	**46**	**35**	37	**48**	**40**	**41**
Driving Time [Seconds]	w/o Feedback	527	450	560	552	620	598	482	541	285	363	319	245	228	267	343	292
	w/ Feedback	**467**	**408**	**507**	**434**	**473**	**486**	**421**	**456**	**271**	**311**	**301**	**228**	**226**	**242**	**293**	**267**
Number of Collisions [Times]	w/o Feedback	6	3	7	6	4	5	4	5.0	**2**	6	4	5	3	4	6	4.2
	w/ Feedback	**4**	**3**	**4**	**3**	**3**	**3**	**3**	**3.2**	**2**	**4**	**3**	**3**	**2**	**3**	**3**	**2.8**
Path Efficiency [Meter]	w/o Feedback	21.5	20.8	21.6	22.2	22.9	23.7	21.8	22.0	18.8	26.2	**23.6**	20.0	19.2	22.2	25.3	22.1
	w/ Feedback	**18.7**	**20.5**	**19.1**	**18.4**	**20.8**	**20.1**	**18.3**	**19.4**	**17.5**	**22.4**	24.7	**18.5**	**18.6**	**21.7**	**20.5**	**20.5**

using speech and gesture inputs were conducted separately, with and without context-awareness. Additionally, the evaluation metrics included Number of Commands, Driving Time, Number of Collisions, and Path Efficiency.

Table. II presents the experimental results. In the speech-based experiments, the system demonstrated improved performance across all evaluation metrics when feedback was enabled. Notably, the number of commands decreased by up to 14, and the number of collisions was reduced by up to 3. Additionally, the driving time was reduced by over 100 seconds, and the path efficiency improved by approximately 3 meters. In experiments using gestures, feedback resulted in better performance on average across all evaluation metrics. Specifically, the number of commands improved by 7, the driving time was reduced by 30 seconds, the number of collisions decreased by 1, and the path efficiency improved by 2 meters. However, the performance varied depending on the skill level of the operator. The number of collisions remained unchanged for the first participant, and the path efficiency metric increased for the third participant when feedback was used. The results indicate that CFS-CA significantly reduces control errors and improves key metrics such as the number of commands, driving time, collision frequency, and path efficiency. This enhancement ensures safer and more reliable robot control.

Fig. 4 shows the trajectories and collision points observed during experiments conducted in the Daeyang AI Center indoor virtual environment. Fig. 4(a) shows the results of participant #1 performing the task without feedback, resulting in 6 collisions. In contrast, Fig. 4(b) presents the results with feedback, showing that the number of collisions was reduced to 4. Similarly, Fig. 4(c) and Fig. 4(d) describes the results of participant #2's gestures, showing 6 and 4 collisions, respectively. These results demonstrate that integrating feedback can effectively reduce the number of collisions.



Participant #1
(a) (Speech, w/o feedback)

Participant #1
(b) (Speech, w/ feedback)

Participant #2
(c) (Gesture, w/o feedback)

Participant #2
(d) (Gesture, w/ feedback)

Fig. 4. Number of collisions without and with feedback

Table. III shows the results of feedback generated by the context-awareness model by combining the questions with the operator's command and the forward-view image. In the first and third images, we can see that the response is generated based on accurate facts. However, in the second image, although the response correctly states that there is an obstacle

when turning left, it also generates incorrect information that there is a box instead of a wall in the additional explanation. This shows that the context-aware model generally reflects feedback well, but has limitations in certain situations.

TABLE III
FEEDBACK RESULTS

Human Interface Command	Forward-view Image	Feedback
Is there an obstacle if you go to the left in the given picture?		No, from this perspective there does not appear to be any obstacle if one were to turn left at the door. The area to the left looks clear and unobstructed by any objects or physical barriers.
Is there an obstacle if you go to the left in the given picture?		Yes, if you were to turn left at the door, you would encounter a stack of boxes or bins, which could be considered an obstacle.
Is there an obstacle if you go forward in the given picture?		No, if you were to move forward in the direction of the boxes, there is no apparent obstacle.

V. CONCLUSION

In this study, we proposed the Command Feedback System based on Context-Awareness (CFS-CA) to address the limitations of traditional speech and gesture-based robot control systems. By integrating real-time situation analysis with user commands, the system generates context-aware feedback, enabling more accurate, reliable, and efficient robot control.

The proposed system was evaluated through experiments conducted in the virtual environment, focusing on key metrics such as the number of commands, driving time, number of collisions, and path efficiency. The results showed that CFS-CA significantly reduces command errors, improves safety, and increases overall performance compared to systems without feedback mechanisms. Specifically, the integration of context-awareness led to a noticeable reduction in number of commands, driving time, and number of collisons for both speech and gesture-based controls. Although this system generally showed improved performance across all evaluation metrics, its performance was limited by the skill level of individual participants.

Future research will focus on evaluating the feedback generation performance of state-of-the-art (SOTA) VLM models across various tasks. The ultimate objective is to develop a robust context-aware model for dynamic environments and robot control, enabling effective HRI in complex scenarios. Furthermore, the system will be extended into a more general framework by incorporating a wider range of commands beyond the existing ones.

ACKNOWLEDGMENT

This work was supported by the Technology Innovation Program (RS-2022-00154678, Development of Intelligent Sensor Platform Technology for Connected Sensor) funded by the Ministry of Trade, Industry & Energy(MOTIE, Korea) and by Unmanned Vehicles Core Technology Research and Development Program through the National Research Foundation of Korea (NRF) and Unmanned Vehicle Advanced Research Center (UVARC) funded by the Ministry of Science and ICT, the Republic of Korea (NRF-2023M3C1C1A01098414).

REFERENCES

[1] Kateryna Zinchenko, Chien-Yu Wu, and Kai-Tai Song. A study on speech recognition control for a surgical robot. *IEEE Transactions on Industrial Informatics*, 13(2):607–615, 2017.

[2] Wen Qi, Salih Ertug Ovur, Zhijun Li, Aldo Marzullo, and Rong Song. Multi-sensor guided hand gesture recognition for a teleoperated robot using a recurrent neural network. *IEEE Robotics and Automation Letters*, 6(3):6039–6045, 2021.

[3] Christian Deuerlein, Moritz Langer, Julian Seßner, Peter Heß, and Jörg Franke. Human-robot-interaction using cloud-based speech recognition systems. *Procedia Cirp*, 97:130–135, 2021.

[4] M Meghana, Ch Usha Kumari, J Sthuthi Priya, P Mrinal, K Abhinav Venkat Sai, S Prashanth Reddy, K Vikranth, T Santosh Kumar, and Asisa Kumar Panigrahy. Hand gesture recognition and voice controlled robot. *Materials Today: Proceedings*, 33:4121–4123, 2020.

[5] Sai Nikhilesh Reddy Karna, Jai Surya Kode, Suneel Nadipalli, and Sudha Yadav. American sign language static gesture recognition using deep learning and computer vision. In *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, pages 1432–1437, 2021.

[6] Wojciech Kaczmarek, Jarosław Panasiuk, Szymon Borys, and Patryk Banach. Industrial robot control by means of gestures and voice commands in off-line and on-line mode. *Sensors*, 20(21):6358, 2020.

[7] Xin Huang Deng Yongda, Li Fang. Research on multimodal human-robot interaction based on speech and gesture. *Computers Electrical Engineering*, 72:443–454, 2018.

[8] Meenu Gupta, Rakesh Kumar, Raju Kumar Chaudhary, and Jayshree Kumari. Iot based voice controlled autonomous robotic vehicle through google assistant. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pages 713–717, 2021.

[9] Xin Wang, Dharmaraj Veeramani, and Zhenhua Zhu. Wearable sensors-based hand gesture recognition for human-robot collaboration in construction. *IEEE Sensors Journal*, 23(1):495–505, 2022.

[10] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[11] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[13] Juyeon Weon, Taein Yong, and Jaeho Kim. Camera-based virtual drone control system using two-handed gestures. In *2024 IEEE International Conference on Metaverse Computing, Networking, and Applications (MetaCom)*, pages 291–296. IEEE, 2024.