2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

A Gaze-Speech System in Mixed Reality for Human-Robot Interaction

John David Prieto Prada[1], Myung Ho Lee[1], and Cheol Song[1†]

Abstract— **Human-robot interaction (HRI) demands efficient time performance along the tasks. However, some interaction approaches may extend the time to complete such tasks. Thus, the time performance in HRI must be enhanced. This work presents an effective way to enhance the time performance in HRI tasks with a mixed reality (MR) method based on a gaze-speech system. In this paper, we design an MR world for pick-and-place tasks. The hardware system includes an MR headset, the Baxter robot, a table, and six cubes. In addition, the holographic MR scenario offers two modes of interaction: gesture mode (GM) and gaze-speech mode (GSM). The input actions during the GM and GSM methods are based on the pinch gesture and gaze with speech commands, respectively. The proposed GSM approach can improve the time performance in pick-and-place scenarios. The GSM system is 21.33 % faster than the traditional system, GM. Also, we evaluated the target-to-target time performance against a reference based on Fitts' law. Our findings show a promising method for time reduction in HRI tasks through MR environments.**

I. Introduction

The human-robot interaction (HRI) field has evolved over the years to help humans to resolve different tasks [1], [2]. The task completion time in HRI environments is a critical variable that requires attention [3], [4]. A highly effective HRI system can complete tasks more efficiently. Hence, an intuitive HRI task can lead to a highly effective system in terms of time performance [5], [6]. However, certain interactions with the robot can highly impact task-completion times [7]. Thus, the task completion time of collaborative robotic systems is a problem that requires extensive exploration [8], [9], [10].

Time performance in industrial processes is a challenging factor frequently influenced by how the user interacts with the robot [11]. An inefficient time performance in HRI tasks can bring less volume production to the industrial sector. Furthermore, a slow task completion time in HRI provokes mental and physical load on the users [12]. In general, the industrial sector is constantly seeking new approaches that effectively enhance the time performance in HRI [13]. For example, immersive methods have shown improved time performance during HRI tasks [14].

Regarding immersiveness, some studies have used mixed reality (MR) methods in HRI tasks. For instance, a hand gesture system enabled the control of an HRI task through a MR environment [15]. The study compared their proposed system against conventional methods. However, the standard method based on joysticks performed slightly better than the gesture system with average times of 20.07 and 27.61 s, respectively. In addition, robotic surgery based on MR was useful for novice doctors [16]. The subjects reduced their navigation time by 34.57 %. However, the system used fiducial markers and external tools for the interaction. Similarly, an MR-HRI system based on heading gestures outperformed a typical MR pointing gesture in terms of time performance [17]. The system provided users intuitive and natural robot control for selection and manipulation tasks. Also, one study presented an eye-gazing method for HRI in MR [18]. The task completion time was approximately 200 s faster than traditional methods. Hence, MR methods in HRI have promising potential and require more improvement in terms of time performance.

The contribution of this study is an evaluation of our proposed gaze-speech system that effectively enhances time performance in HRI tasks through MR. To the best of our knowledge, this is the first attempt to use an MR approach based on a gaze-speech system for HRI tasks. We designed an MR environment for pick-and-place tasks. The MR scenario had two distinct ways of interaction, gesture mode (GM) and gaze-speech mode (GSM). The MR world showed four targets: blue, pink, green, and orange. Moreover, the input mechanisms during the GM and GSM methods were the pinch gesture and gaze-with-speech commands, respectively. In addition, we evaluated each mode under two conditions: big targets (BT) and small targets (ST). We asked 20 subjects to conduct six pick-and-place MR tasks with the Baxter robot using the GM and GSM methods. The subjects were divided into two groups: A and B. Furthermore, all the individuals answered a survey regarding the usability of the proposed method. We compared the time performance of the experiments during the GM and GSM systems. Under the BT condition, participants using the GM and GSM systems had an average completion time of 178.3844 s and 133.8183 s, respectively. Also, we calculated the target-to-target time performance against a reference based on Fitts' law. The lowest root-mean-square error (RMSE) found was 1.5252 s using the GSM method under the BT condition. In addition, the survey results showed a high preference for the GSM system in terms of time efficiency, ease of use, and physical load.

This paper is organized as follows: Section II describes the study methods with a deep explanation of the hardware and software setup. Section III presents the experimental plan used in this study. Section IV gives a detailed analysis of the experimental results involved in this study and the respective

† Corresponding author
[1]Department of Robotics and Mechatronics Engineering, DGIST, Daegu 42988, South Korea prada.john, lmho777, csong@dgist.ac.kr
*This work was supported by the DGIST R&D Program of the Ministry of Science and ICT (22-RT-01).

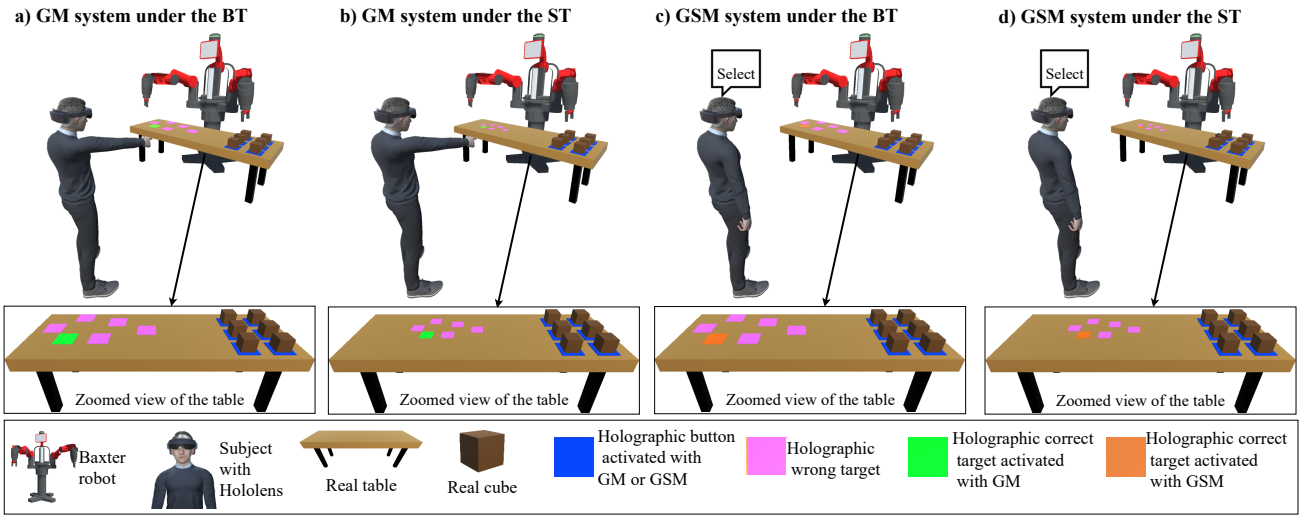

Fig. 1. General schematic of the MR interface for HRI. a) The MR system using the GM under BT condition. b) The MR interface using the GM under ST condition. c) The MR world with our proposed system, using the GSM under BT condition. d) The MR environment with our proposed method, using the GSM under ST condition. We assign distinct colors to the correct holographic targets for distinguishing purposes during both GM and GSM paradigms.

discussion. The conclusions are drawn in Section V.

II. METHODS

The system is composed of two different parts: hardware and software systems. Fig. 1 shows the general schematic of the system.

A. Hardware

This study used a mixed reality headset (MRH), Microsoft HoloLens 2. This MRH is a powerful device that can combine data from multiple built-in sensors, including speech recognition, and eye and hand tracking. In addition, we used an industrial robot, Baxter, for the HRI task with the subject, as shown in Fig. 1.

The MR environment offered two different ways to interact with the Baxter robot, GM and GSM, as shown in Fig. 1. The real environment is composed of a table and six cubes. The dimensions of the table and each cube were 1.6 m $\times$ 0.7 m $\times$ 0.75 m and 0.05 m $\times$ 0.05 m $\times$ 0.05 m, respectively. Both the subject and robot interact using the real cubes during pick-and-place experiments.

B. The MR environment configuration

The MR environment is programmed using three languages and two operating systems (OS). Hence, the configuration parts of the MR environment are as follows:

1) Windows 10 Holographic OS: We created an MR environment based on the C# language and Unity software. Fig. 1 shows the MR world. In addition, the MR world has two holographic interaction modes, GM and GSM. Fig. 1 a) and b) and Fig. 1 c) and d) show the subject using the GM and GSM, respectively. Furthermore, each interaction mode presents two different conditions: BT and ST. The MR world is deployed, installed, and run on the MRH based on Windows 10 holographic OS.

2) GM paradigm: Fig. 1 a) and b) represent the typical MR interaction in HRI, using the GM. The MR world displays three styles of holographic buttons: blue, pink, and green. The blue holographic buttons represent a fixed location of each real cube. The pink and green holographic buttons mean the wrong and correct targets. Furthermore, the MR software places the targets randomly on the experimental table.

The size of the blue holographic buttons is always 0.1 m^2. In contrast, the size of the pink and green targets varies depending on the current condition. Fig. 1 a) shows the GM method under the BT condition, where the size of the pink and green targets is 0.1 m^2. Under the ST conditions, the targets' size is 0.05 m^2, as shown in Fig. 1 b).

The software tracks and recognizes the subject's hand and pinched gestures, respectively. Furthermore, the subject uses hand motions as a pointer to navigate the MR world. Also, the pinch gestures aid the subject in selecting the holographic buttons and targets.

3) GSM paradigm: Fig. 1 c) and d) show our proposed system, GSM. Similar to the GM paradigm, the MR environment shows three interactive buttons: blue, pink, and orange. All buttons' conditions and sizes are the same as in GM mode. However, the interaction process is different.

The software recognizes the subject's gaze and a speech command called "select." The individual uses the gaze as a pointer to navigate the MR world. In addition, the gaze system is head-assisted. Furthermore, the speech command helps the subject to select the desired button while focusing on it.

4) Linux OS: We use Python and ROS to operate the Baxter robot. In addition, the Python program has the TCP/IP protocol to process the MRH data. The subject's actions in the MR world can trigger Baxter's motions. Moreover, the robot receives two position variables: the location of the

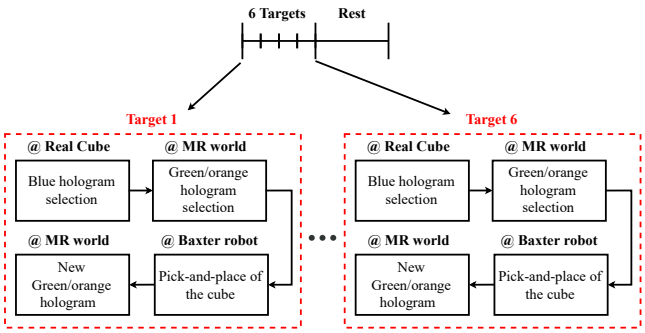

Fig. 2. Experimental design for pick-and-place task for HRI in MR.

picked cube and the selected correct target. However, the robot is motionless whenever the subject selects the wrong target information. Thus, we only consider the correct target data during the experiments.

C. TCP/IP protocol for HRI in MR

Our system consists of two separate workstations, the MRH and Linux computer. Thus, the MR world and the Baxter control use different programming languages. Hence, we introduce the TCP/IP protocol, a common communication between two workstations.

III. EXPERIMENTS

We recruited 20 individuals (17 males and 3 females) between the ages of 24–36 (M = 28.90, SD = 3.33) years old. In addition, 80 % of the participants stated to have low experience with industrial robots. Furthermore, we split the participants into two groups: A and B. The participants in group A started the experiments using the GM system, followed by the GSM method. In contrast, the individuals in group B performed the experiments using the GSM paradigm, followed by the GM system. DGIST Institutional Review Board approved the study with the research management number (DGIST-20210608-HR-118–01).

A. Pick-and-place experiments

We asked 20 participants to perform pick-and-place tasks in MR with the Baxter robot. Each participant was 1.5 meters away from the table, as shown in Fig. 1. Our MR experimental scenario is based on a Fitts' experiment environment for 2D pointer performance [19]. Every participant had 1-2 minutes to familiarize themselves with the MR world using both GM and GSM systems.

Fig. 2 shows the experimental design of this study. The experiment began with the interaction of the subjects and the real cube. The participants were instructed to press the blue holographic button corresponding to the current cube. The individuals were then required to locate and select the correct target from among the incorrect targets. In the MR world, all correct and wrong targets were randomly arranged. The Baxter robot picked the current cube and placed it in the correct chosen target.

The task ended once the robot successfully performed the pick-and-place exercise. We asked the subjects to perform the pick-and-place task for each of the six cubes on the table. The experiment finished after the robot successfully placed all cubes. Furthermore, every participant rested for 2 minutes at the end of the experiments. Moreover, the individuals conducted their experiments in two distinct modes:

- GM system: The subjects trigger the holographic buttons with a pinch gesture, as shown in Fig. 1 a) and b). The MRH tracks and detects the index finger of the users. A guiding ray emerges parallel and coincident to the finger. Furthermore, the guiding ray can help the individuals as a moving cursor. In addition, the users are required a pinch gesture to activate the buttons, after the selection with the guiding ray occurs.
- GSM system: As our proposed method, the subjects are required to do the gaze cursor and voice command to trigger the holographic buttons, as shown in Fig. 1 c) and d). The gaze cursor guides the subject to choose the buttons in the MR environment. The voice command and gaze cursor activate the button event. Furthermore, the voice command is "select." In addition, the gaze cursor is assisted with head motions.

We evaluated each experimental mode under two conditions, BT and ST. Under the BT condition, the MR system shows pink, green, and orange holograms with a size of 0.1 m^2. Under the ST condition, however, the MR world displays the pink, green, and orange buttons with a size of 0.05 m^2. In addition, the size of the blue holograms is always 0.1 m^2 for all the cases.

We calculated the total completion time of each participant during each experiment. In addition, we compared each subject's target-to-target completion time with a reference, which is based on the Fitts' law equation for 2D pointing methods [20]. The Fitts' law equation is expressed as:

$$MT = 230 + 166 \cdot (\frac{2D}{W}) \quad (1)$$

where MT represents the average time to complete a target-to-target selection, D is the euclidean distance from the starting target to the final target, and W is the width of the final target.

B. Survey

All participants answered four survey questions at the end of each experiment. Table I shows the questions of the survey. Furthermore, each question was based on a 7-point

TABLE I
SURVEY QUESTIONS.

Q1: Did you feel safe during the experiments?
Q2: Were you satisfied with the time performance during the experiments?
Q3: Was the system easy to use during the experiments?
Q4: Was the system physically demanding during the experiments?

Likert scale. In addition, the survey aimed to determine a usability score for our proposed method in terms of safety perception, time efficiency, ease of use, and physical load.

C. Hypotheses

This study proposes a gaze-speech interface in MR for HRI tasks. Hence, we expect that our GSM will outperform the GM in terms of time performance. Also, we hypothesize that the participants will similarly perform the tasks regardless of the group. Furthermore, we predict that the GSM will result in a higher sense of safety and time performance satisfaction than the GM. In addition, the GSM may be more intuitive and less physically demanding than GM. Overall, we propose the next hypotheses:

- H1: Participants using GSM will require less time to perform the experimental tasks compared to participants using GM.
- H2: The subjects' performance will be statistically similar regardless of their group.
- H3: The participants will complete the experiments under the BT condition faster than they will complete the experiments under the ST condition.
- H4: The GSM system will have the highest safety perception.
- H5: The GSM system will be rated significantly higher than the GM system in terms of time performance satisfaction.
- H6: The GSM system will be the easiest system to use.
- H7: The GSM system will be the least physically demanding to use.

IV. Results and Discussion

In this study, we asked 20 participants to perform interactive HRI tasks with the Baxter robot while wearing the MRH. We designed two distinct MR interaction methods, GM and GSM. All participants were divided into groups A (10) and B (10) to avoid the learning effect on the subjects. For group A, the participants began the experiments using the GM, followed by the GSM. In group B, the subjects performed the tasks in the opposite order. In addition, every individual answered a survey after the experiments.

A. Pick-and-place tasks results

All participants performed pick-and-place tasks with the Baxter robot using the MRH. We made two different interaction paradigms, GM and GSM. Furthermore, we evaluated each paradigm under two conditions: BT and ST.

Fig. 3 shows the total completion time of the participants during the pick-and-place experiment. Subjects under the influence of our GSM system performed the experiments in less time than with the GM system. Table II shows the mean completion time of both A and B groups during the pick-and-place experiments. The GSM system outperformed the GM method under BT and ST conditions in both groups. The A and B groups, using the GSM system under the BT condition, completed experiments 45.3180 and 43.8140 s faster than when using the GM, respectively.

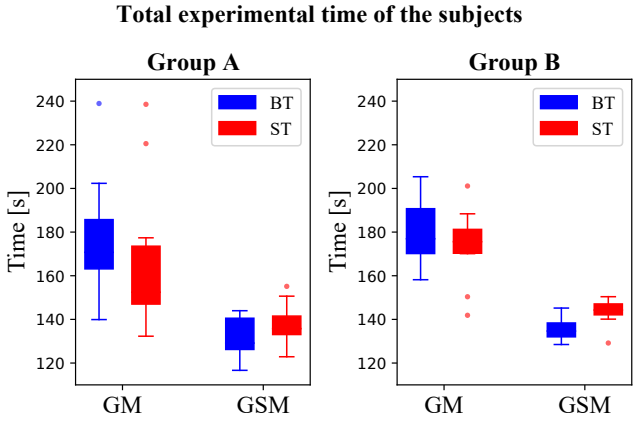

Fig. 3. Total completion time during the experiments.

TABLE II
AVERAGE TIME OF THE GROUPS DURING THE EXPERIMENTS.

Mode	Condition	Group average time (s)	
		A	B
GM	BT	177.2610	179.5078
	ST	167.4790	173.5972
GSM	BT	131.9430	135.6937
	ST	137.6651	143.6782

In addition, we tested the H1-H3 using a t-test with a significance level of $p < 0.05$ between each paradigm and its corresponding condition, as shown in Table III. The participants' performance using GSM, either under BT or ST conditions showed significant values of $p < 0.001$ in most cases in comparison to the GM system. Hence, H1 was confirmed. In addition, regardless of the group, all individuals performed statistically the same with a p-value of 1.1516, validating the H2.

Furthermore, the A and B groups using our GSM paradigm under the BT condition performed better than under the ST condition, with p-values of 0.0332 and 0.0011, respectively. Thus, H3 was confirmed. However, the subjects' performance of both groups using the GM method under the BT condition lacks significance compared to the ST condition. The p-values for groups A and B were 0.1716 and 0.2136, respectively. In this case, H3 was rejected.

TABLE III
P-VALUES FROM THE T-TEST RESULTS OF THE PICK-AND-PLACE EXPERIMENTS.

Paired T-test	Group p-value	
	A	B
GM BT vs GM ST	0.1716	0.2136
GM BT vs GSM BT	<0.001	<0.001
GM BT vs GSM ST	0.0020	<0.001
GM ST vs GSM BT	0.0045	<0.001
GM ST vs GSM ST	0.0088	<0.001
GSM BT vs GSM ST	0.0332	0.0011

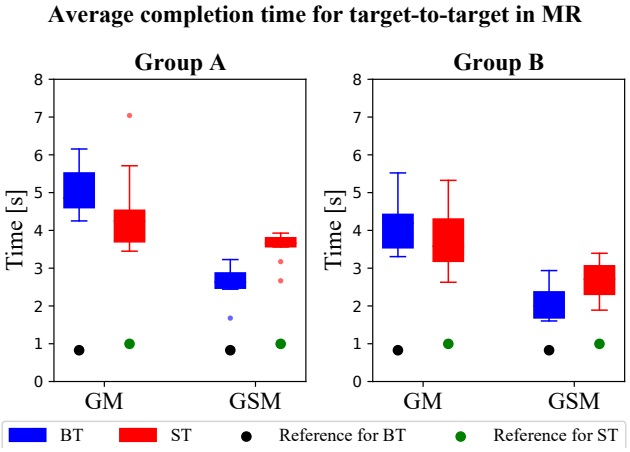

Fig. 4. Average completion time for target-to-target using the GM and GSM systems. The blue color represents the target-to-target average time performance under the BT condition. The red color shows the target-to-target mean time under the ST condition. Black and green colors are the average references for both BT and ST conditions, respectively.

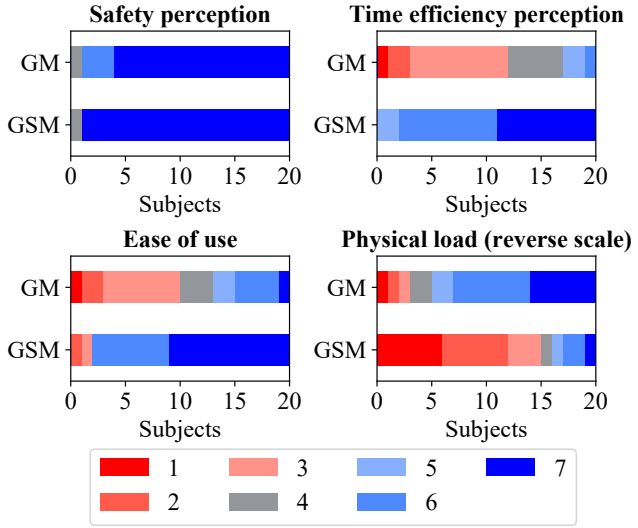

Fig. 5. Bar plots of the survey responses from the 20 participants on a 7-point Likert scale. The survey used statements, strongly disagree (1), disagree (2), slightly disagree (3), neutral (4), slightly agree (5), agree (6), and strongly agree (7).

B. Target-to-target performance results

The participants in this study conducted six pick-and-place tasks using the GM and GSM systems. Both methods used pointers to navigate and interact with the holographic targets. Hence, we evaluated the target-to-target time performance of the subjects during the experiments. Fig. 4 shows each subject's average target-to-target time performance in groups A and B. In addition, we calculated the average reference points based on (1) for all the tasks.

Furthermore, we determined all the RMSE values of the GM and GSM systems against the reference based on (1). Table IV shows the subjects' target-to-target average time and RMSE results. Regardless of their group, all participants demonstrated higher performance while using the GSM system than the GM method compared to the reference. The average RMSE of the subjects using the GSM system was 1.5252 and 2.1433 s under the BT and ST conditions, respectively. In addition, the GM system had the highest RMSE with values of 3.7867 and 3.1247 under the BT and ST conditions, respectively. Hence, the GM approach was slower in terms of time performance than our proposed system, GSM.

We made a t-test between the target-to-target results of all the subjects with a significance level of $p < 0.05$. The participants significantly improved the performance of the experiments while using our GSM system compared to the GM method. In addition, we found a statistical difference between the performance of the individuals using the GSM and GM with a p-value < 0.001.

C. Survey results

In Table I, the study participants answered a survey at the end of the tasks. Fig. 5 shows each participant's responses to the survey on a 7-point Likert scale based on the performance of the methods, GM and GSM. Each individual filled out the survey regarding safety perception, time efficiency, ease of use, and physically demanding.

We tested the H4-H7 by using an ANOVA with a significance level of $p < 0.05$, as shown in Table V. The safety perception (Q1) lacked significance when the subjects performed the experiments either using the GM or GSM.

TABLE IV
AVERAGE TARGET-TO-TARGET TIME PERFORMANCE AND RMSE RESULTS OF THE EXPERIMENTS.

Mode \ Condition	Average target-to-target time [s]			Average RMSE [s]
	Group		Reference	
	A	B		
GM \ BT	5.0453	4.1909	0.828	3.7867
GM \ ST	4.4941	3.7708	0.997	3.1247
GSM \ BT	2.6429	2.0580	0.828	1.5252
GSM \ ST	3.5670	2.6902	0.997	2.1433

TABLE V
ANOVA TEST RESULTS FROM THE SURVEY QUESTIONS.

Hypothesis	Subjective variables	Likert scale average		p-value
		GM	GSM	
H4	Safety perception	6.7000	6.8500	0.5035
H5	Time efficiency perception	3.4000	6.7000	<0.001
H6	Ease of use	3.9500	6.2000	<0.001
H7	Physical load	5.4000	2.7500	<0.001

Hence, H4 was rejected. However, the Likert scale average showed a slightly higher value in the GSM of 6.8500 than the GM of 6.7000 during the safety perception. We believe that the increment in the number of participants might affirm H4.

Subjects rated the GSM higher than the GM in terms of time efficiency perception (Q2), ease of use (Q3), and physical load (Q4), confirming the hypotheses H5, H6 and H7, respectively. The Likert scale average during the GSM outperformed the GM in Q2-Q4. The p-values found were < 0.001 for all Q2-Q4. The participants showed high preferences for the GSM than the GM.

D. Discussion

Our GSM approach effectively reduced the time to perform pick-and-place tasks. The participants who operated with the GSM system performed faster than those who operated the same tasks with the GM system.

Participants found the GM approach difficult to use due to the hand tracking and gesture recognition problems. Consequently, all participants required more time to complete tasks using the GM method. In contrast, subjects highly preferred our GSM system due to the fast response and ease of use. Hence, our findings suggest that subjects should avoid the GM paradigm for better time efficiency.

V. CONCLUSIONS

This study presented an effective method for reducing the time to perform HRI tasks by using MR technology based on a gaze-speech approach. The proposed system drastically reduced the completion time of pick-and-place experiments compared with the traditional GM method.

The time performance in HRI applications is affected by the interaction method. We used immersive MR technology for robot interaction to complete pick-and-place tasks. The MR environment, based on holographic targets, used two different interactive methods, GM and GSM. A pinch gesture was the primary input during the GM paradigm. In the GSM system, we set a voice command as the main input for the interaction. In addition, we evaluated every method under two different conditions, BT and ST. Furthermore, we asked 20 subjects to perform six tasks using both GM and GSM systems. The subjects were divided into two groups: A and B. After the experiments, subjects answered a survey regarding the usability of our proposed approach.

In the experimental results, we compared the time performance of tasks using the GM system against the GSM approach. The participants using GSM were 21.33 % faster than participants using GM. Also, we obtained the target-to-target time performance of the subjects for comparison to a reference based on Fitts' law equation. The lowest RMSE found was 1.5252 s when using the GSM method under the BT condition. In addition, the survey results showed that the subjects preferred the GSM system in terms of time efficiency, ease of use, and physical load.

Our findings proved an efficient way to improve the time performance in HRI with MR. We showed a new alternative for time reduction based on the GSM paradigm. Our proposed method can be implemented in scenarios where the operator has occupied hands and require interacting with robots. In future works, we will analyze the effect of color and wrong target selection during the experiments. In addition, we will explore other input interactions, such as VR handlers and haptic devices. Moreover, we will improve the GM system performance by implementing sensor fusion techniques.

REFERENCES

[1] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, vol. 55, pp. 248–266, 2018.

[2] R. Galin and R. Meshcheryakov, "Review on human–robot interaction during collaboration in a shared workspace," in *International Conference on Interactive Collaborative Robotics*. Springer, 2019, pp. 63–74.

[3] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, vol. 55, pp. 248–266, 2018.

[4] R. R. Galin and R. V. Meshcheryakov, "Human-robot interaction efficiency and human-robot collaboration," in *Robotics: Industry 4.0 Issues & New Intelligent Control Paradigms*. Springer, 2020, pp. 55–63.

[5] A. Seth, J. M. Vance, and J. H. Oliver, "Virtual reality for assembly methods prototyping: a review," *Virtual reality*, vol. 15, no. 1, pp. 5–20, 2011.

[6] S. Stadler, K. Kain, M. Giuliani, N. Mirnig, G. Stollnberger, and M. Tscheligi, "Augmented reality for industrial robot programmers: Workload analysis for task-based, augmented reality-supported robot control," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 179–184.

[7] J. Abou Saleh and F. Karray, "Towards generalized performance metrics for human-robot interaction," in *2010 International Conference on Autonomous and Intelligent Systems, AIS 2010*. IEEE, 2010, pp. 1–6.

[8] Y. Mizuchi and T. Inamura, "Estimation of subjective evaluation of hri performance based on objective behaviors of human and robots," in *Robot World Cup*. Springer, 2019, pp. 201–212.

[9] W. Xiong, C. Pan, Y. Qiao, N. Wu, M. Chen, and P. Hsieh, "Reducing wafer delay time by robot idle time regulation for single-arm cluster tools," *IEEE Transactions on Automation Science and Engineering*, vol. 18, no. 4, pp. 1653–1667, 2020.

[10] A. Mayima, A. Clodic, and R. Alami, "Towards robots able to measure in real-time the quality of interaction in hri contexts," *International Journal of Social Robotics*, pp. 1–19, 2021.

[11] S. Kumar, C. Savur, and F. Sahin, "Survey of human–robot collaboration in industrial settings: Awareness, intelligence, and compliance," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 280–297, 2020.

[12] C. J. Lin and R. P. Lukodono, "Sustainable human–robot collaboration based on human intention classification," *Sustainability*, vol. 13, no. 11, p. 5990, 2021.

[13] A. Weiss, A.-K. Wortmeier, and B. Kubicek, "Cobots in industry 4.0: A roadmap for future practice studies on human–robot collaboration," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 4, pp. 335–345, 2021.

[14] J. Delmerico, R. Poranne, F. Bogo, H. Oleynikova, E. Vollenweider, S. Coros, J. Nieto, and M. Pollefeys, "Spatial computing and intuitive interaction: Bringing mixed reality and robotics together," *IEEE Robotics & Automation Magazine*, vol. 29, no. 1, pp. 45–57, 2022.

[15] E. Nazarova, O. Sautenkov, M. A. Cabrera, J. Tirado, V. Serpiva, V. Rakhmatulin, and D. Tsetserukou, "Cobotar: Interaction with robots using omnidirectionally projected image and dnn-based gesture recognition," in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2021, pp. 2590–2595.

[16] L. Qian, A. Deguet, Z. Wang, Y.-H. Liu, and P. Kazanzides, "Augmented reality assisted instrument insertion and tool manipulation for the first assistant in robotic surgery," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5173–5179.

[17] D. Krupke, F. Steinicke, P. Lubos, Y. Jonetzko, M. Görner, and J. Zhang, "Comparison of multimodal heading and pointing gestures for co-located mixed reality human-robot interaction," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–9.

[18] K.-B. Park, S. H. Choi, J. Y. Lee, Y. Ghasemi, M. Mohammed, and H. Jeong, "Hands-free human–robot interaction using multimodal gestures and deep learning in wearable mixed reality," *IEEE Access*, vol. 9, pp. 55 448–55 464, 2021.

[19] S. Ljubic, V. Glavinic, and M. Kukec, "Finger-based pointing performance on mobile touchscreen devices: Fitts' law fits," in *International Conference on Universal Access in Human-Computer Interaction*. Springer, 2015, pp. 318–329.

[20] I. S. MacKenzie, "Movement time prediction in human-computer interfaces," in *Readings in human–computer interaction*. Elsevier, 1995, pp. 483–493.