Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 59-10 (2025) 2274–2279

Bidirectional Verbal Communication in Human-Robot Collaboration

Rami Ojanen [*] Roel Pieters [*]

[*] Cognitive Robotics group, Unit of Automation Technology and Mechanical Engineering, Tampere University, 33720, Tampere, Finland (e-mail: firstname.surname@tuni.fi).

Abstract: To fully leverage the capabilities of both the robot and the human operator while working towards shared goals, effective communication and interaction between them are crucial. However, in many collaborative robot applications, interaction is often restricted, resulting in ineffective and unintuitive collaboration. Speech is a promising communication method for human-robot collaboration, as it is natural for the human operator and allows for two-way communication. In this work, we demonstrate a collaborative robotic system that integrates speech recognition, text-to-speech, object detection, and robot control. Object detection equips the robot with environmental awareness, while the speech-related functionalities enable bidirectional communication. The system is tested through several assembly-related test cases, demonstrating basic and advanced communication features such as task coordination, dialogue, and quality inspection.

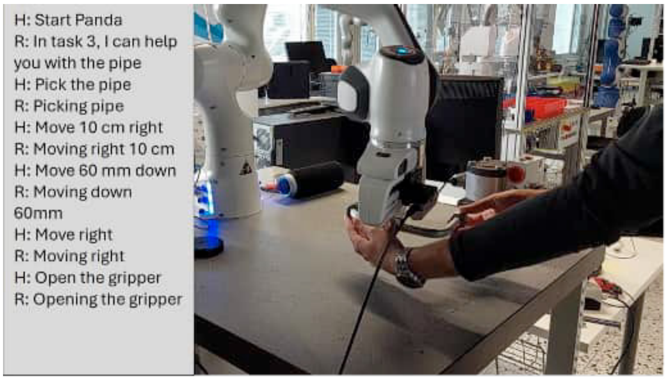
Keywords: Human-Robot Collaboration, Speech perception, Text-To-Speech, Sensor Integration, Human-in-the-Loop, Human-Robot Task Allocation

1. INTRODUCTION

In recent years, the integration of collaborative robots into different industrial tasks such as assembly and inspection has significantly transformed manufacturing processes. Collaborative robots are flexible, easily programmable, and adaptable to various tasks. These benefits have led to increased interest in human-robot collaboration (HRC). HRC enhances efficiency, productivity, and safety by combining human intelligence and adaptability with the precision and reliability of robots (Inkulu et al., 2022).

Effective human-robot interaction is key to further improving the fluency and safety of collaborative work. Speech interaction holds great promise, as it is a natural and intuitive method of communication for humans and enables bidirectional communication (Salehzadeh et al., 2022). Through automatic speech recognition (ASR), the robot can interpret human commands, while text-to-speech (TTS) enables it to provide instructions and information about its environment, intents, and goals. This kind of robot-to-human communication can reduce the cognitive load on human operators and enhance trust in the robot (Buisan et al., 2020; Zhu and Williams, 2020). Effective communication is especially valuable in assembly tasks, where the human operator and the robot must collaborate closely in shared work spaces (see Fig. 1).

Incorporating vision and object detection further enhances the robot's situational awareness. This makes it more flexible and robust as it can perform tasks based on visual detections rather than predefined locations. This capability is particularly important in dynamic assembly tasks and also enables more interactive and efficient robot-to-human communication.



H: Start Panda
R: In task 3, I can help you with the pipe
H: Pick the pipe
R: Picking pipe
H: Move 10 cm right
R: Moving right 10 cm
H: Move 60 mm down
R: Moving down 60mm
H: Move right
R: Moving right
H: Open the gripper
R: Opening the gripper

Fig. 1. Bidirectional verbal communication between human and robot can offer the human a better understanding of robot actions. This work presents such approach and enables speech and vision to be included for communication.

Despite the advancements in collaborative robotics, current systems often suffer from limited and inefficient communication between human operators and robots. Additionally, in cases where communication exists, it is often one-way, usually only from human to robot, leading to sub-optimal collaboration. Bidirectional verbal interaction presents a promising solution to bridge this gap by offering a more intuitive and seamless form of communication (see Fig. 1). The goal of this research is to develop and implement a collaborative assembly system in which the

10.1016/j.ifacol.2025.09.382

human communicates with the robot through speech, and to showcase the possibilities of bidirectional verbal communication in collaborative industrial assembly tasks. The contributions of our work are:

- A system combining object detection, speech recognition and text-to-speech for human-robot collaboration
- An approach for bidirectional verbal communication between human and robot
- Experimental demonstration of the system in various industrial collaborative assembly tasks

2. RELATED WORK

2.1 Human-Robot Collaboration

Human-robot collaboration aims to advance the productivity, safety and efficiency of factories (Arents et al., 2021). The synergy of robots and humans can be utilized to accomplish diverse tasks and minimize the total production time and cost. By combining the characteristics of humans, such as intelligence and flexibility, with accuracy and consistency of robots, manufacturing companies can ensure the adaptability of the production setup to make varying products.

To ensure effective, flexible and safe collaboration between human and robot, interaction and communication between them are essential. In addition to verbal communication, different non-verbal and multimodal methods can also be utilized for communication. Non-verbal methods include, for instance, gestures (Liu and Wang, 2018), haptics (Villani et al., 2018) and gaze (Palinko et al., 2016). With multimodal methods, the idea is to combine the characteristics and advantages of different methods by using them simultaneously. Common combinations include, e.g., speech and gestures (Halim et al., 2022; Ekrekli et al., 2023) or speech and haptics (Gustavsson et al., 2017).

2.2 Verbal Communication

Speech is one of the main methods for direct communication between humans (Rocci and Saussure, 2016). Verbal communication is highly intuitive and the most direct form of explicit communication in human-robot collaboration. By using speech recognition and TTS functionalities, human and robot can communicate through verbal commands in both directions: human commanding the robot and the robot responding to the human operator (Salehzadeh et al., 2022). It allows the human to work uninterruptedly and with both hands (Villani et al., 2018). Moreover, the amount of information and different commands transferred is usually greater compared to other methods (Ekrekli et al., 2023). Thus, it has been successfully applied in various cases and applications particularly in assembly tasks (Angleraud et al., 2021) but also in welding (Niculescu et al., 2014).

The level of recognition capabilities varies in different applications and implementations of speech recognition in the context of communication between human and robot. Some of them recognize single words, specific commands or phrases (Angleraud et al., 2021; Ekrekli et al., 2023; Telkes et al., 2024) and some natural language (Niculescu et al., 2014; Williams et al., 2015; Thierauf et al., 2024).

While human-robot communication is more common, there are also systems focusing on robot-human communication (Buisan et al., 2020; Zhu and Williams, 2020). Other systems include bidirectional communication (Ferrari et al., 2022; Thierauf et al., 2024), however, these are not very common, especially in industrial scenarios. Different kinds of virtual assistants, social and service robots often have advanced capabilities for both natural and bidirectional communication, but their industrial applications are limited, being mostly utilized in healthcare or entertainment (Gunson et al., 2022; Li et al., 2023).

The different approaches of utilizing speech in robotics can be divided into lexical grounding and learning-based methods. In lexical grounding, the words and commands are directly connected to robot actions or targets. Meanwhile, learning-based methods can use high-level natural language instructions and generalize input commands to desired outcomes with the help of Large Language Models (LLMs). Lexical grounding is usually faster and more reliable in connecting the commands and actions, while learning based methods can produce more natural and versatile communication (Telkes et al., 2024). When it comes to this categorization, the implemented system represents lexical grounding.

3. SYSTEM

3.1 Visual Perception

Detectron2 (Wu et al., 2019) is utilized for object detection, which gives information to monitor the work space and grasp objects. A custom dataset was created by collecting 216 images of 12 object and target classes with an Intel Realsense D-435 camera attached to the robot. These images were annotated with segmentation polygons and bounding boxes using open-source data labelling tool Label Studio (HumanSignal Inc., 2024). Following, this data was augmented using python library Albumentations to include variations in noise and lighting conditions. The augmentation increased the number of images to around 20,000 to form the whole dataset. The model was trained by using Faster R-CNN R50-FPN, which is a basic bounding box detector extendable to Mask R-CNN, from Detectron2. Object detection served as a direct visual input for object grasping and for the monitoring of the work space. This includes quality monitoring of object assemblies by checking if all required parts are present on the assembly.

3.2 Speech Perception and TTS

Creoir EdgeVUI SDK provides tools for ASR and TTS. It utilizes Speech Signal Enhancement (SSE) and other techniques for enhancing the speech recognition and clarity of voice commands (Creoir, 2024). It recognizes whole sentences, so called utterances, which are connected to intents. The intents are short keywords, which define what should happen when a utterance is recognized. Internally, it uses MQTT protocol for handling of data and is connected to the robot through ROS topics. When an intent is activated, a message with corresponding command is published to a ROS topic, which then leads to corresponding

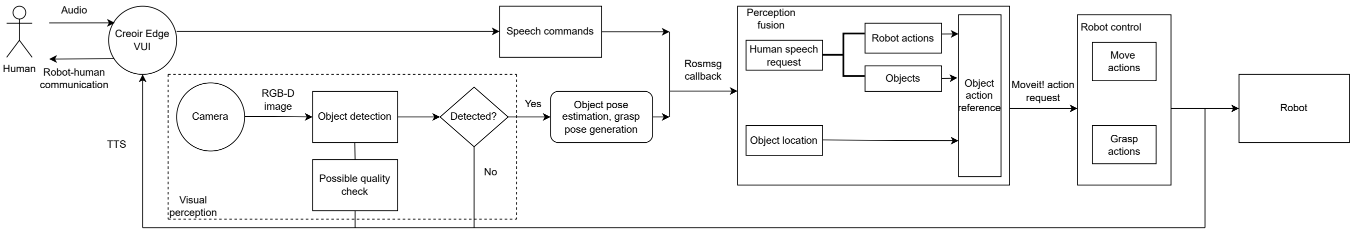


Fig. 2. Data architecture of the bidirectional verbal communication approach between human and robot. Visual and speech perception are utilized to support the communication and generate robot actions and TTS commands.

robot actions. Similarly, the robot sends ROS messages to the TTS functionality to communicate relevant information from robot to human.

3.3 Robot Control Architecture and Perception Fusion

The data architecture of the system is depicted in Fig. 2, consisting of visual and speech perception, the fusion of perception and the generation of robot actions and TTS commands. Perception results, in the form of visual perception and speech recognition are combined in the perception fusion block, where both the recognized speech commands and the object poses have their respective callback functions. From speech commands, robot actions are selected and referred to objects as perceived from the object detection module. The TTS functionality is called every time the robot will execute an action, a human speech request requires a response or when perception requires information to be communicated to the human. From a robot control perspective, robot actions are either move, grasp, pick-and-place or hand-over actions, or a combination of those. These actions are then sent to the robot (Franka Panda) to be performed. Robot control is implemented based on MoveIt and OpenDR (Passalis et al., 2022) libraries. ROS Noetic is utilized for both robot programming and communication between different components of the system. Our implementation can be found open-source at *https://github.com/ramblam/Bidir_HRC*, including a video of all experiments.

4. RESULTS

Different test-cases were designed, by utilizing a hydraulic pump assembly set, with the aim to test the system and demonstrate its functionalities. This includes the perception results and the results for communication from human to robot (see Fig. 4), robot to human (see Fig. 5) and bidirectional (see Fig. 6). Following, we explain these results in more detail.

4.1 Perception Results

Results of visual object detection are demonstrated by the ability to successfully grasp objects from visual input, as can be seen in the corresponding video. In our use cases the precision and recall of all 12 classes is over 90%, on average (see Fig. 3). Similarly, speech perception results are demonstrated by the successful experiments. As the tool used (Creoir EdgeVUI) is a commercial product, quantitative results were not deemed necessary for evaluation.

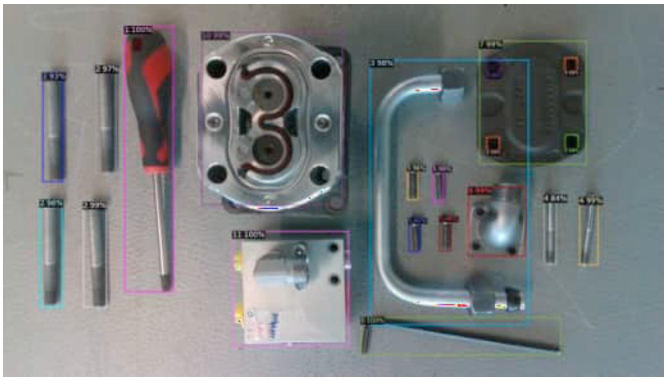


Fig. 3. Detection results for object detection (bounding box) with class number and confidence score.

4.2 Human-to-Robot Communication

Communication from human to robot serves to command robot actions or request specific information regarding the shared task of collaborative assembly. Depending on the human speech request, visual information or task plan information is then queried and used to take follow-up actions or reply via TTS. Fig. 4 shows examples of such human speech requests: requesting the hand-over of a bolt (Fig. 4a), requesting the quality check of the object assembly (Fig. 4b) and requesting what actions should be done next in the shared task (Fig. 4c). Resulting robot actions are respectively: robot detecting the bolt, picking it and handing it over to the human, robot detecting whether all required objects are on the assembly, and robot detecting which objects are in the scene and querying which actions can be done by robot or human.

4.3 Robot-to-Human Communication

Communication from robot to human is mainly in the form of verbal commands by TTS and is triggered either by a human command or as a result of perception and robot actions that are planned to be executed. Fig. 5 shows examples of such robot to human communication and relate directly to the previous Fig. 4: explaining that the next action of the robot is to hand over a bolt (Fig. 5a), explaining that the current state of the assembly has quality errors and that a check should be done next (Fig. 5b) and explaining what actions can be done by the robot and what actions by the human (Fig. 5c), based on the current state. As such, these communication actions are open-ended and do not require a direct action or response from the human.

4.4 Bidirectional Communication

Communication between human and robot that is bidirectional allows for dialogue, question-answering or resolving conflicts in the collaboration. To enable this, both verbal commands and visual information are used and lead to resulting robot actions and TTS commands. Fig. 6 shows an example of such bidirectional communication: the human requests the Allen key to be handed over (Fig. 6a), the robot visually detects only a screwdriver in the scene and asks if the human meant this tool (Fig. 6b) and the person confirms (Fig. 6c). These closed-ended questions therefore expect a response from the human before any follow-up actions or command will be taken.

4.5 Collaborative Assembly Results

Besides these individual tests, also human-robot collaborative assembly experiments were performed, where all mentioned capabilities were included (see Fig. 7, Table 1 and the corresponding video). The tests show that the system is able to help the human operator by performing different actions, such as hand-overs of tools and parts, assembling some of the parts directly (Fig. 7a and 7b) and helping to place objects in situations where the human operator has both hands on the work piece (Fig. 7c). In addition, the system has functionalities for simple quality checking, stopping and continuing, giving instructions and simple dialog, among others. The system's ability to recognize sentences, instead of single words, combined with text-to-speech functionality leads to rather natural and comfortable communication. The latency of the system is also rather small (less than one second, on average), enabling a responsive system for fluent collaboration.

5. DISCUSSION

When comparing our system with related work, a few main differences can be noticed. Considering the level of recognized speech commands, our approach stands between the extremes of single words, specific commands or phrases (Angleraud et al., 2021; Ekrekli et al., 2023; Telkes et al., 2024), and fully natural language (Niculescu et al., 2014; Williams et al., 2015; Thierauf et al., 2024). Our system recognizes a wide range of natural language commands and command sentences, but they still need to be predefined, meaning it does not achieve full natural language flexibility.

In many of the current applications of collaborative robots, the amount of actual collaboration is limited. Meanwhile, our system allows close collaboration, as shown in Fig. 7c, with both human and robot physically handling the same object. This could be utilized, for example, in co-manipulation of a part that is difficult to handle alone due to size or shape. The system also shows proactivity by suggesting future actions, instead of only the current state, without an explicit human request.

To make the system viable in real-world industrial applications, additional testing and improvement in safety, reliability and robustness of the system are required. With the current implementation, although the system allows recognition of sentences rather than isolated words and accommodates various phrasings, the commands still need to be predefined. Achieving full natural language flexibility remains a challenge to be solved. Artificial intelligence (AI), in the form of generative AI and LLMs, for example, could help in overcoming this. The scope and complexity of dialogue between human and robot could also be extended. This would require the ability to handle pauses and mid-task changes from the robot, posing a challenge from a robot control point of view. The robot could keep track of the completed and pending tasks and thus give more exact instructions. Incorporating other communication methods, such as gestures and haptics, thus moving towards multi-modal communication, would also enhance the communication and could make it more fluent and robust.

6. CONCLUSION

This work proposed how both human-robot, robot-human and bidirectional verbal communication can be utilized to enhance human-robot collaboration in an industrial assembly task. The implemented system combines speech perception and text-to-speech (TTS) with visual perception and robot control. Visual perception provides the robot with information about its environment, while speech perception and TTS enable bidirectional communication. The system was tested in various assembly-related test-cases, where functionalities such as quality inspection, dialog and the robot providing instructions were presented, in addition to common actions such as picking, placing and handing-over of objects. The results demonstrate the feasibility of bidirectional verbal communication in collaborative assembly tasks. Future work will focus on exploring the potential of Large Language Models (LLMs) to further improve the communication capabilities of the system.

ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101135708 (JARVIS).

REFERENCES

Angleraud, A., Mehman Sefat, A., Netzev, M., and Pieters, R. (2021). Coordinating shared tasks in human-robot collaboration by commands. *Frontiers in Robotics and AI*, 8, 734548.

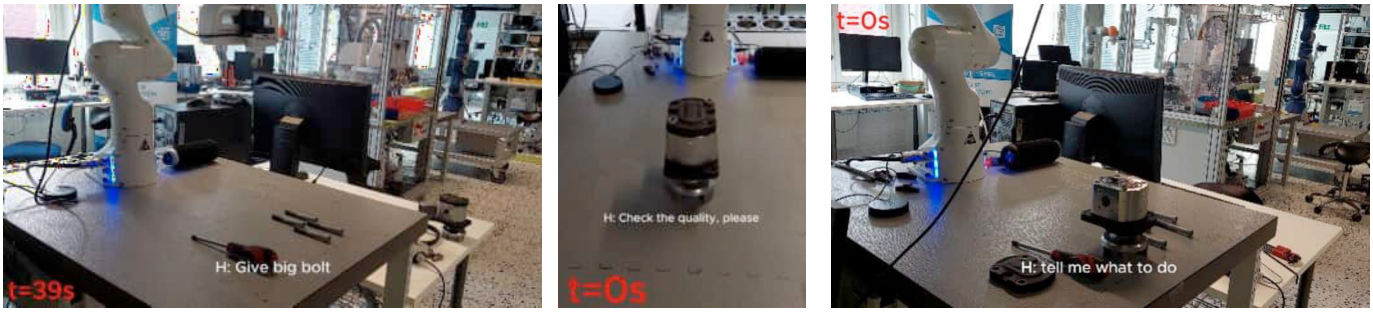
Arents, J., Abolins, V., Judvaitis, J., Vismanis, O., Oraby, A., and Ozols, K. (2021). Human–robot collaboration trends and safety aspects: A systematic review. *Journal of Sensor and Actuator Networks*, 10, 48.

Buisan, G., Sarthou, G., and Alami, R. (2020). Human aware task planning using verbal communication feasibility and costs. In *International Conference on Social Robotics*, 554–565.

Creoir (2024). Edgevui. https://creoir.com/edgevui. Accessed: 2024-07-19.

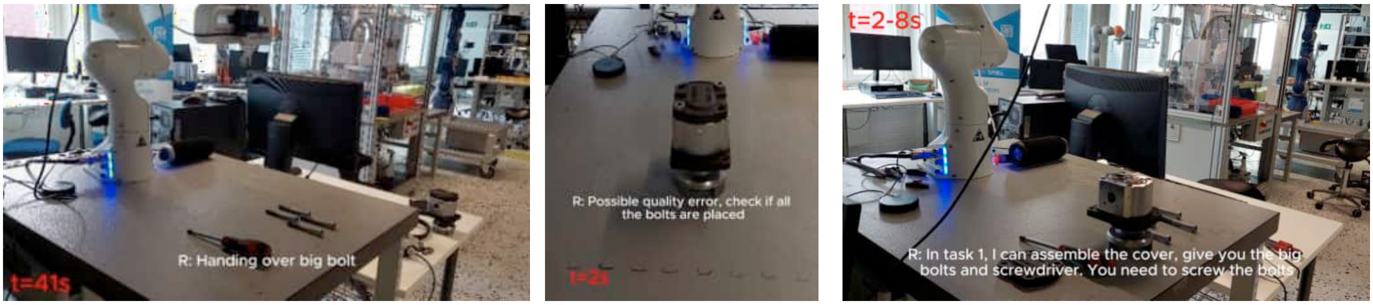
Ekrekli, A., Angleraud, A., Sharma, G., and Pieters, R. (2023). Co-Speech Gestures for Human-Robot Collaboration. In *IEEE International Conference on Robotic Computing (IRC)*, 110–114.

Ferrari, D., Benzi, F., and Secchi, C. (2022). Bidirectional communication control for human-robot collaboration. In *International Conference on Robotics and Automation (ICRA)*, 7430–7436.



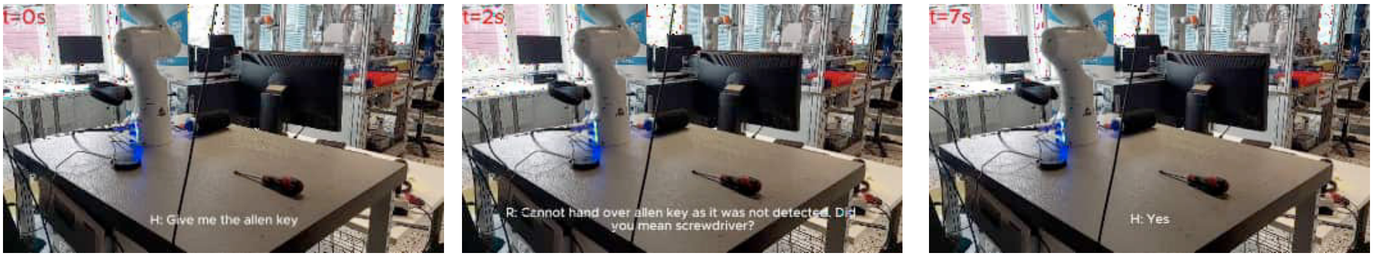
(a) (b) (c)

Fig. 4. Snapshots of human (H) to robot (R) communication: human requests a big bolt to be handed over (a), human requests the quality of the assembly to be checked (b) and human asks what actions can be done next (c).


t=41s
R: Handing over big bolt
R: Possible quality error, check if all the bolts are placed
t=2s
t=2-8s
R: In task 1, I can assemble the cover, give you the big bolts and screwdriver. You need to screw the bolts

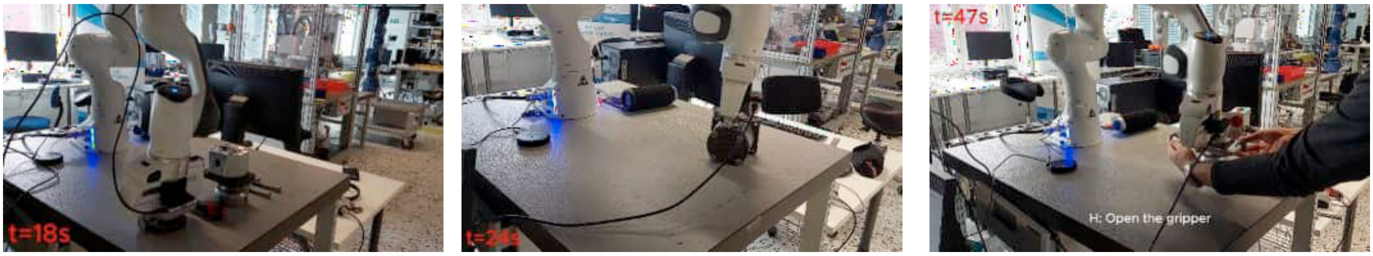
(a) (b) (c)

Fig. 5. Snapshots of robot (R) to human (H) communication: robot explains what it will do next; handing over a big bolt (a), robot explains that the assembly has a possible quality error, and what steps should be taken next; check if all bolts are placed (b) and robot explains what robot and human actions can be done next (c).


t=0s
H: Give me the allen key
t=2s
R: Cannot hand over allen key as it was not detected. Did you mean screwdriver?
t=7s
H: Yes

(a) (b) (c)

Fig. 6. Snapshots of bidirectional communication: human requests an Allen key (a), robot responds with the current status and a suggested solution (b) and human accepts the suggested solution (c).



(a) (b) (c)

Fig. 7. Snapshots of the collaborative assembly task, divided in three subtasks: cover assembly (a), elbow assembly (b) and pipe assembly (c). All commands utilized by human and robot are listed in Table 1.

Table 1. List of commands from human (H) and robot (R) for the collaborative assembly task. TTS denotes text-to-speech.

Communication command	Human speech	TTS	Associated robot action
Task 1: cover assembly			
Hello, let's start by assembling cover		R	
Assemble the cover	H		
Assembling cover and then moving back home, picking it from predefined location		R	picking and placing actions
Give big bolt (4 times)	H		
Handing over big bolt (4 times)		R	picking and hand-over actions
Give me the screwdriver	H		
Handing over screwdriver		R	picking and hand-over actions
Task 2: Elbow assembly			
Hello. In task 2, let's assemble the elbow		R	
Assemble the elbow	H		
Assembling elbow and then moving back home		R	picking and placing actions
Task 3: Pipe assembly			
Hello. In task 3, I can help you with the pipe		R	
Pick the pipe	H		
Picking pipe		R	Picking action
Move 10 cm right, 60 mm down, right	H		
Moving 10 cm right, 60 mm down, right 50 mm		R	Moving action
Open the gripper	H		
Opening the gripper		R	Gripper action
Move up	H		
Moving up 50 mm		R	Moving action

Gunson, N., Garcia, D.H., Sieinska, W., Dondrup, C., and Lemon, O. (2022). Developing a social conversational robot for the hospital waiting room. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1352–1357.

Gustavsson, P., Syberfeldt, A., Brewster, R., and Wang, L. (2017). Human-robot collaboration demonstrator combining speech recognition and haptic control. *Procedia CIRP*, 63, 396–401.

Halim, J., Eichler, P., Krusche, S., Bdiwi, M., and Ihlenfeldt, S. (2022). No-code robotic programming for agile production: A new markerless-approach for multimodal natural interaction in a human-robot collaboration context. *Frontiers in Robotics and AI*, 9, 1001955.

HumanSignal Inc. (2024). Label studio. `https://labelstud.io/`. Accessed: 2024-07-19.

Inkulu, A.K., Bahubalendruni, M.R., Dara, A., and K., S. (2022). Challenges and opportunities in human robot collaboration context of Industry 4.0 - a state of the art review. *Industrial Robot: the international journal of robotics research and application*, 49(2), 226–239.

Li, C., Chrysostomou, D., and Yang, H. (2023). A speech-enabled virtual assistant for efficient human–robot interaction in industrial environments. *Journal of Systems and Software*, 205, 111818.

Liu, H. and Wang, L. (2018). Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics*, 68, 355–367.

Niculescu, A.I., Banchs, R.E., and Li, H. (2014). Why industrial robots should become more social: On the design of a natural language interface for an interactive robot welder. In *International Conference on Social Robotics*, 276–278.

Palinko, O., Rea, F., Sandini, G., and Sciutti, A. (2016). Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5048–5054.

Passalis, N. et al. (2022). OpenDR: An open toolkit for enabling high performance, low footprint deep learning for robotics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 12479–12484.

Rocci, A. and Saussure, L.D. (eds.) (2016). *Verbal Communication.* De Gruyter.

Salehzadeh, R., Gong, J., and Jalili, N. (2022). Purposeful communication in human–robot collaboration: A review of modern approaches in manufacturing. *IEEE Access*, 10, 129344–129361.

Telkes, P., Angleraud, A., and Pieters, R. (2024). Instructing hierarchical tasks to robots by verbal commands. In *IEEE/SICE International Symposium on System Integration (SII)*, 1139–1145.

Thierauf, C., Thielstrom, R., Oosterveld, B., Becker, W., and Scheutz, M. (2024). 'Do this instead' - Robots that adequately respond to corrected instructions. *ACM Transactions on Human-Robot Interaction*, 13, 1–23.

Villani, V., Pini, F., Leali, F., and Secchi, C. (2018). Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics*, 55, 248–266.

Williams, T., Briggs, G., Oosterveld, B., and Scheutz, M. (2015). Going beyond literal command-based instructions: extending robotic natural language interaction capabilities. In *AAAI Conference on Artificial Intelligence*, 1387–1393.

Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., and Girshick, R. (2019). Detectron2. `https://github.com/facebookresearch/detectron2`.

Zhu, L. and Williams, T. (2020). Effects of proactive explanations by robots on human-robot trust. In *International Conference on Social Robotics*, 85–95.