A Scalable Robot-Agnostic Voice Control Framework for Multi-Robot Systems

Valentina Pericu[a], Federico Rollo[b] and Navvab Kashiri[c]
Robotics, Innovation Labs, Leonardo S.p.A., Genoa, Italy

Keywords: Multi-Robot Systems, Robot-Agnostic, Scalability, Voice Control.

Abstract: In recent years, Multi-Robot Systems (MRS) have gained increasing relevance in domains such as industrial automation, healthcare, and disaster response, offering effective solutions to manage complex and dynamic tasks. However, controlling such systems remains a challenge, particularly for users without expertise in robotics. A critical factor in addressing this challenge is developing intuitive and accessible Human-Robot Interaction (HRI) mechanisms that enable seamless communication between humans and robots. This paper introduces a scalable, robot-agnostic voice control framework designed to simplify interaction with MRS. The framework enables users to issue voice commands that are processed into actionable, robot-specific instructions through a centralized architecture. At its core, the framework features a centralized Control Management System (CMS) that is responsible for processing voice commands and interpreting them into robot-agnostic actions. System scalability is achieved through namespace management and a flexible structure, allowing new robots to be integrated and larger teams to be accommodated with minimal effort. By minimizing hardware requirements and leveraging voice commands as the primary interaction modality, the framework reduces technical barriers and provides an accessible, cost-effective solution for non-expert users. Experimental validation demonstrates its flexibility, scalability, and effectiveness in multi-robot scenarios. This work contributes to advancing HRI by offering a robust, intuitive, and adaptable solution for managing heterogeneous robot teams across dynamic environments.

1 INTRODUCTION

Human-Robot Interaction (HRI) serves as the fundamental bridge between human intent and robotic actions, enabling robots to perform tasks that align with human requirements and expectations. As robots become increasingly integrated into human-centric environments, HRI has emerged as a crucial area of research, aiming to develop intuitive, accessible, and efficient interaction methods that enhance usability and acceptance. The rapid growth of the Internet of Things (IoT) has further accelerated the integration of robots into everyday life, expanding the need for effective communication between humans and robotic systems (Su et al., 2023).

This need becomes particularly critical in Multi-Robot Systems (MRS), where the challenge of coordinating heterogeneous robotic devices requires interaction methods that are not only natural but also scalable and efficient, ensuring effective collaboration in complex and dynamic environments (Dahiya et al., 2023). MRS have revolutionized numerous fields such as disaster response, industrial automation, and healthcare by leveraging various robotic capabilities to address tasks that are beyond the scope of individual robots (Dahiya et al., 2023; Rizk et al., 2019; Stone and Veloso, 2000). The ability of MRS to dynamically distribute tasks based on individual robot capabilities further enhances their applicability in addressing real-world challenges, as demonstrated by Heppner et al. (Heppner et al., 2024), which present a decentralized approach using behaviour trees to dynamically allocate tasks based on robots' capabilities, leveraging runtime auctions for optimal assignments.

Many studies have explored different HRI modalities to facilitate effective interaction between humans and robots. Hang Su et al. (Su et al., 2023) review recent advances in multi-modal HRI, emphasizing the strengths and limitations of different interaction styles. The review highlights that existing HRI methods span multiple modalities, including audio-

[a] https://orcid.org/0009-0000-4661-8293
[b] https://orcid.org/0000-0001-5833-5506
[c] https://orcid.org/0000-0002-1219-2447

Pericu, V., Rollo, F. and Kashiri, N.
A Scalable Robot-Agnostic Voice Control Framework for Multi-Robot Systems.
DOI: 10.5220/0013672500003982

In *Proceedings of the 22nd International Conference on Informatics in Control, Automation and Robotics (ICINCO 2025) - Volume 2*, pages 194-201
ISBN: 978-989-758-770-2; ISSN: 2184-2809

based, visual, tactile, and multi-modal approaches, each with distinct advantages and drawbacks. Audio-based interaction offers intuitive and hands-free communication; however, it struggles with ambient noise, ambiguous phrases, and variations in dialect, and it can be sensitive to the distance between speaker and receiver (Marin Vargas et al., 2021; Kumatani et al., 2012). Visual interaction, employing gestures and facial expression recognition, provides rich contextual information. Nevertheless, it is sensitive to environmental factors such as lighting and occlusions (Rollo et al., 2023b; Rollo et al., 2024b), and requires the appropriate installation of cameras and/or other visual sensors. Similarly, haptic interaction enhances intuitiveness through touch-based feedback (Pyo et al., 2021), which requires high precision and reliable sensors, as well as the corresponding control algorithms to ensure safe and effective interactions. Moreover, both cameras and haptic sensors require the proximity of the user to the robot, which is not always desirable in terms of safety and usability in MRS. Multi-modal systems aim to integrate these approaches to replicate human-like communication by combining their strengths (Muratore et al., 2023). However, their hardware requirements and integration complexity further limit their practicality, particularly in heterogeneous robot systems.

The growing interest in verbal communication within both industry and academia reflects its potential to enable more natural and intuitive interactions between humans and robots, as noted by Marin et al. (Marin Vargas et al., 2021). This communication approach offers significant advantages in various fields, including industrial applications (Papavasileiou et al., 2024) (Del Bianco et al., 2024), medical robotics (Rogowski, 2022), assistive technologies (Padmanabha et al., 2024), and education-focused robotics (Budiharto et al., 2017; Belpaeme et al., 2018). Verbal communication stands out by simplifying the interaction between advanced technological systems and individuals with limited technical knowledge, making it more accessible and user-friendly. Moreover, voice control offers several advantages over other interaction modalities, as it allows users to interact freely without being constrained by the field-of-view or proximity requirements of visual or tactile systems. Recent advances in speech recognition and natural language processing (Su et al., 2023; Amadio et al., 2024) have further improved the reliability and practicality of voice-based systems, making them well-suited for real-world applications.

The need for a scalable, robot-agnostic interaction framework for MRS has drawn significant attention to this field, aiming to minimize hardware dependencies while ensuring simplicity and accessibility. Carr et al. (Carr et al., 2023) propose a human-friendly verbal communication platform for MRS, which eliminates the dependency on network infrastructures by enabling robots to communicate between themselves through microphones and speakers. However, this approach relies on the availability of onboard audio hardware, which can be challenging to exploit in mobile robots such as drones and/or quadrupeds due to the potential interference from operational noise.

To overcome these limitations, this work proposes a centralized framework for voice control with hardware-minimal and robot-agnostic characteristics. Unlike visual or tactile interaction modalities that require specialized hardware, such as cameras or haptic sensors on each robot, adding complexity, cost, and reducing compatibility, this approach centralizes command processing within a Control Management System (CMS). Relying solely on a single microphone, the CMS eliminates the need for additional sensors or hardware modifications on the robots, enabling straightforward integration with diverse robotic platforms. The CMS translates voice commands into robot-agnostic instructions, ensuring adaptability across heterogeneous robot teams.

Moreover, the CMS features a capability-aware mechanism that ensures voice commands are directed only to robots capable of executing the requested tasks. This capability-handling feature improves reliability by filtering out invalid commands, enhancing the system's efficiency in managing diverse robot teams, and ensuring effective operation in heterogeneous multi-robot environments.

The following Section 2 outlines the methods, providing a detailed explanation of the proposed system architecture, with an emphasis on the CMS modules. Section 3 presents the experimental setup, describes two use case experiments, and discusses the obtained results to illustrate the system's functionality. Finally, Section 4 concludes the paper by summarizing key findings and suggesting directions for future work.

2 METHODS

This work focuses on developing a hardware-minimal, robot-agnostic, scalable approach leveraging voice control. Robot-agnostic refers to a system design that operates independently of the specific hardware or internal architecture of individual robots. Instead, it relies on generic interfaces and a shared communication protocol, allowing the same interaction and processing logic to be applied

seamlessly across heterogeneous platforms such as wheeled robots, quadrupeds, and drones. This section details the architecture and core functionalities of the framework.

2.1 System Overview

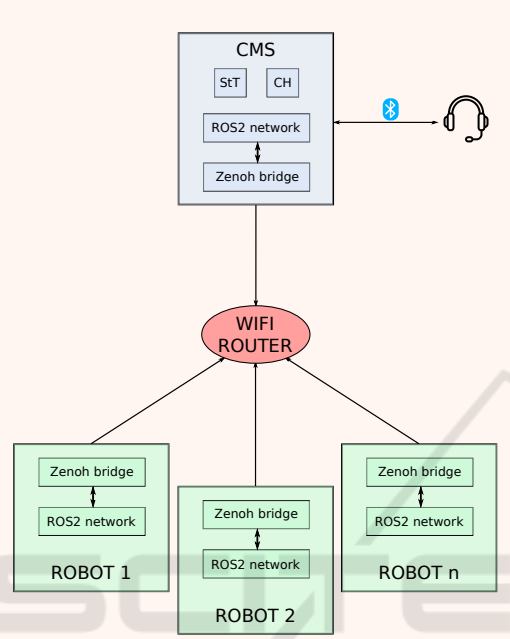

Figure 1: The system architecture of the voice control framework for MRS. The CMS runs the StT and CH, and also acts as a Zenoh router, connecting to Zenoh clients on each robot via a centralized Wi-Fi network. ROS2 manages internal operations on the CMS and robots, while Zenoh bridges communication between them.

Figure 1 illustrates the architecture of the scalable robot-agnostic voice control framework. The framework exploits the Robot Operating System 2 (ROS2) middleware, version Humble[1], and the Zenoh protocol[2]. ROS2 is a set of software libraries and tools for building robot applications, relying on DDS as its communication middleware. Zenoh acts as a bridge for ROS2, enabling efficient data exchange across networks by mirroring DDS topics, reducing communication overhead in distributed or wireless environments. Additionally, Zenoh simplifies topic diversification by allowing flexible namespace management, which facilitates the integration of MRS. As indicated in the figure, the voice command processing is carried out within the CMS as a centralized unit, to allow for a simple and hardware-minimal solution, relying only on a single microphone connected to the CMS unit. The system is built on a Wi-Fi network, enabling inter-system connectivity through Wi-Fi routers. In this setup, the CMS operates as a Zenoh router, while each robot functions as a Zenoh client. These components are interconnected and mutually visible. This shared network ensures direct and efficient communication, eliminating the need for intermediate configurations or additional networking infrastructure. The CMS manages human-to-robot interaction through two core modules: the Speech-to-Text (StT) module, which processes voice commands, and the Command Handler (CH) module, which interprets and dispatches commands to the appropriate robots in a robot-agnostic and scalable manner, allowing the integration of new robots with minimal reconfiguration. These modules operate independently, promoting modularity and adaptability within the system. Their specific roles and functionalities are detailed in the following subsections.

2.2 Speech-to-Text Module

The StT module converts verbal commands into textual commands and publishes them on ROS2. This module is implemented as a ROS2 node that exploits the Python `speech_recognition` library[3] for speech-to-text conversion. To enhance usability and robustness in noisy environments, the node dynamically adjusts for ambient noise by analyzing the environment to distinguish background sounds from actual speech before processing. This ensures reliable transcription even in environments with varying noise levels, improving the accuracy of recognized commands.

A Bluetooth microphone connected to the pilot PC further reduces interference and enables remote voice input, offering greater user mobility, especially beneficial in collaborative or vision-based tasks.

2.3 Command Handler Module

The CH module is implemented as a ROS2 node that listens to commands provided by the StT module. Upon receiving a command, it performs parsing and validation. The handler relies on predefined lists of *robots names*, *actions*, and action-specific *information* to ensure the commands are valid. Commands must follow the format *"robot name"* + *"action"* + *"info"*; where *"info"* is optional, depending on the action type. The robot name is used to specify the target robot for the action and is incorporated as a namespace in ROS2 to route commands to the appropriate robot. If a command is invalid, an error is

[1]https://docs.ros.org/en/humble/index.html
[2]https://zenoh.io/
[3]https://pypi.org/project/SpeechRecognition/

logged, and no action is performed. Valid commands are transformed into robot-specific actions and published to the ROS2 topic with the appropriate robot namespace. For example, the command *"robot1 rotate left"* is converted by the CH into a `Twist` message and published on the `/robot1/cmd_vel` topic. A key strength of the CH lies in its scalability, enabled by namespace management through Zenoh. Integrating additional robots into the system simply requires assigning unique namespaces to each robot and updating the predefined *robots names* list, making the system adaptable to larger teams without significant reconfiguration. Similarly, additional actions can be easily integrated by expanding the list of supported commands, ensuring the framework remains flexible and extensible to evolving operational requirements.

2.3.1 Robot States and Capabilities

The CH module also monitors the states and capabilities associated with each robot:

- **Robot States:** each robot is associated with two boolean state indicators: `is_active` and `is_in_action`. The `is_active` state indicates whether the specified robot is turned on and connected to the Wi-Fi router, as verified by a keep-alive check performed every second. This ensures that commands are processed only for active robots, while those targeting inactive robots are discarded. The `is_in_action` state tracks whether the robot is currently executing a given action, supporting effective task management and safe handling of stop commands. Together, these states are crucial for coordinating robots in dynamic environments and preventing unintended actions.
- **Robot Capabilities:** Robot capabilities are predefined for each robot, representing its functional abilities. This enables the system to automatically reject commands that exceed a robot's functionality. For instance, a quadruped robot can execute actions like "stand-up" and "sit-down", while a wheeled robot can't, as they fall outside its operational scope. This approach is particularly beneficial in heterogeneous multi-robot setups, such as those involving quadrupeds, wheeled robots, drones, and manipulators, as it allows the system to manage diverse robots seamlessly.

2.3.2 Implemented Actions

The CH module supports a range of actions for each robot, to be triggered by specific voice commands:

- **Exploration:** triggered by the command *"explore"*, this action does not require any additional information; however, it can potentially accept additional information for inspecting the environment, i.e. *"explore and inspect"*. The CH interprets the command and instructs the robot to initiate an exploration routine, enabling it to map the surrounding environment.
- **Rotation:** this action is initiated with the command *"rotate"* and requires additional information, i.e *"left"/"right"*. The CH processes the command and sends the appropriate instruction to the robot, enabling it to rotate in the indicated direction.
- **Translation:** the action is activated by the command *"move"* and also requires additional information, i.e *"forward"/"backward"*. The CH handles the command and directs the robot to move in the specified direction.
- **Navigation to Goal:** triggered by the instruction *"go to"*, the action requires additional information specifying the name associated with the target position. The target position and the corresponding names may be available a priori, or can be constructed within the *"explore"* action, e.g. *"robot1 go to desk"*. The CH processes the command and provides the robot with the navigation instructions necessary to autonomously reach the destination.
- **Stop:** the command *"stop"* halts all ongoing actions of the specified robot. This action does not require additional input, and the CH ensures that the robot immediately stops its current tasks.
- **Standing** and **Sitting:** triggered by the commands *"stand up"* and *"sit down"*, respectively, these actions require no additional information. The CH invokes the appropriate service to execute the action for the specified robot, provided that the robot supports the requested functionality.

3 EXPERIMENTS

To validate the proposed framework, we conducted a set of real-world experiments involving two heterogeneous robots. Section 3.1 presents the experimental setup used to test how the system can interpret and execute voice commands across multiple robots. Section 3.2 describes two representative experiments, chosen from the broader set performed. In the valid sample experiment, presented in Section 3.2.1, all voice commands were correctly executed, targeting

Table 1: Detailed breakdown of the command handling pipeline, from speech recognition to execution, across valid and invalid scenarios. The table shows how voice commands are processed, validated, and dispatched to heterogeneous robots, detailing the StT output, target robot, requested action, optional information, feasibility check, generated ROS 2 topic/service, and execution result.

	CH					
StT	**Robot**	**Action**	**Info**	**Validity**	**Topic/Service**	**Result**
”Max explore”	Max	exploration	-	valid	`max/cmd_vel`	success
”Max go to home”	Max	navigation to goal	home	valid	`max/goal_pose`	success
”Bob go to chair”	Bob	navigation to goal	chair	valid	`bob/goal_pose`	success
”Max go to box”	Max	navigation to goal	box	valid	`max/goal_pose`	success
”Max go to home”	Max	navigation to goal	home	valid	`bob/goal_pose`	success
”Bob go to home”	Bob	navigation to goal	home	valid	`bob/goal_pose`	success
”Max go to chair”	Max	navigation to goal	chair	valid	`max/goal_pose`	success
”Bob stand up”	Bob	standing	-	invalid	-	success
”Rob move forward”	Rob	translation	forward	invalid	-	success
”Max sit down”	Max	sitting	-	valid	`max/sit_down`	success
”Bob rotate left”	Bob	rotation	left	valid	`bob/cmd_vel`	success

active robots and requiring them to perform actions within their capabilities. In the second sample experiment, described in Section 3.2.2, we intentionally issued invalid commands, some directed at inactive robots, and others requesting actions that the robots do not support. Finally, Sections 3.3 and 3.4 present the results and discussion, further analyzing the framework’s scalability and robustness.

3.1 Experimental Setup

The experiments involve two robots operating within a shared environment: the Unitree B1 quadruped robot, called *Max*, and the Clearpath Husky wheeled robot, called *Bob*, both equipped with 16-channel 3D LiDAR sensors. The robots utilize SLAM (Simultaneous Localization and Mapping) to create a detailed map of their surroundings. RViz, a ROS visualization tool, is used to visualize the environment, the robots’ states, and to provide real-time feedback on their positions, paths, and actions, facilitating efficient monitoring of their operations.

Figure 2 shows the RViz visualization used during the experiments. A box and a chair represent the robot’s target locations, while yellow and green flags indicate the *home* positions of *Max* and *Bob*, respectively. The grey area represents the constructed map of the environment. To enable a common reference frame, a transformation was established between the shared *earth* frame and the individual maps of each robot. The relationships between frames are illustrated in Figure 2.

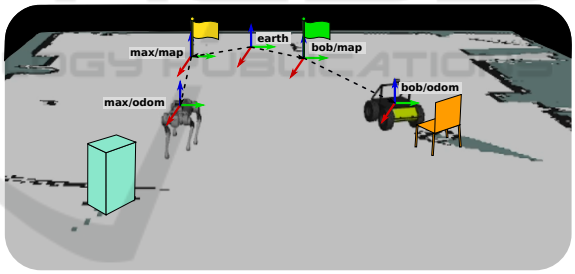

Figure 2: The RViz visualization of the experimental setup illustrates the environment where two robots, *Max* and *Bob*, operate. Their corresponding *home* positions are indicated by yellow and green flags, respectively. The target positions of the two robots, the box in cyan and the chair in orange, are highlighted. The visualization also shows the relationships between frames, including the shared *earth* frame as well as the individual map and odometry frames of each robot.

3.2 Experiments

Multiple experiments were conducted to evaluate the system under various conditions. Two illustrative experiments are presented below.

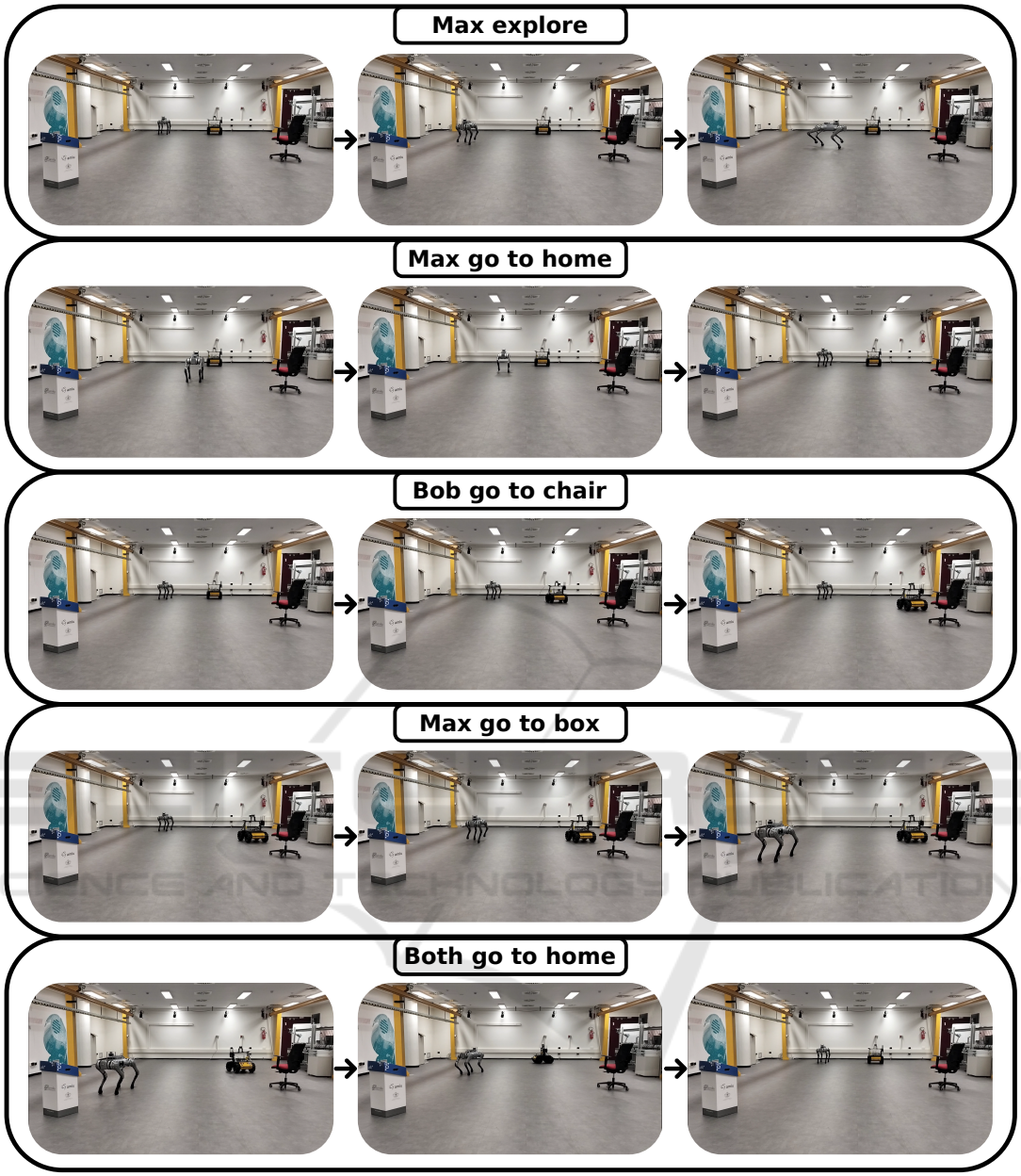

Figure 3: Sequence of actions performed during Experiment 1 (3.2.1) involving the two heterogeneous robots, *Max* and *Bob*. The figure illustrates key moments as the robots execute voice commands, showcasing the system's ability to manage multiple robots and diverse tasks within a shared environment.

3.2.1 Experiment 1: Valid Commands

In the first experiment, we evaluated the system's ability to process valid voice commands targeting active robots and requesting actions within their capabilities. The sequence of commands was as follows:

- *"Max explore": Max* starts an exploration routine and builds the map of the environment;
- *"Max go to home": Max* navigates to its *home*;
- *"Bob go to chair": Bob* navigates to the chair;
- *"Max go to box": Max* navigates to the box;
- *"Max go to home": Max* navigates to its *home*;
- *"Bob go to home": Bob* navigates to its *home*;

Figure 3 captures key moments from this experiment[4]. During the experiment, all voice commands

[4]A video demonstrating this experiment can be found at: https://youtu.be/Be6-bKC970M

were successfully processed and forwarded to the corresponding target robots. The robots correctly carried out the requested actions, operating within the shared environment without interruptions.

3.2.2 Experiment 2: Invalid Commands

The second experiment tested the system's handling of invalid inputs. Voice commands were intentionally issued to inactive robots or involved actions not supported by the robot. The sequence included:

- *"Max go to chair": Max* navigates to the chair;
- *"Bob stand up":* the command is refused since the *"stand-up"* ability does not belong to the wheeled robot;
- *"Rob move forward":* the command is refused since *Rob* is not an active robot;
- *"Max sit down": Max* sits down;
- *"Bob rotate left": Bob* performs a leftward rotation;

These outcomes confirm the effectiveness of robot availability and capability validation mechanisms.

3.3 Results

The proposed framework was validated through voice-controlled experiments involving two heterogeneous robots operating in a shared environment. Voice commands were issued via a Bluetooth microphone under realistic conditions, and their execution was monitored to evaluate system performance. Table 1 provides a detailed summary of the commands processed by the system during Experiment 1 (Section 3.2.1) and Experiment 2 (Section 3.2.2), including the output of the StT module, the target robot, the action type, the optional information, the feasibility validation, the ROS2 topic/service generation, and the execution results. Although multiple experiments were conducted, only these two are reported in the table for clarity and representativeness.

The CH consistently parsed each command, verified its feasibility by checking robot availability and capability, and routed it to the appropriate ROS2 topic or service via Zenoh. All valid commands were successfully executed. For example, the command *"Max sit down"* was correctly interpreted and dispatched, as the quadruped robot supports this capability. In contrast, commands that were syntactically incorrect, targeted undefined or inactive robots (e.g., *"Rob move forward"*), or requested unsupported actions (e.g., *"Bob stand up"*) were successfully rejected.

Throughout the experiments, the system demonstrated stable and repeatable behavior. No execution failures were observed, and latency between voice input and robot action remained consistently low, well under two seconds. These results confirm the framework's effectiveness in handling both valid and invalid input and its robustness in coordinating task execution within a heterogeneous multi-robot setup.

3.4 Discussion

The experiments highlight the framework's effectiveness in simplifying HRI for MRS. The centralized architecture and hardware-minimal design ensure adaptability across heterogeneous robot teams and ensure ease of deployment and scalability. The voice control modality removes the need for additional sensors on individual robots, reducing costs and complexity while enabling intuitive interaction for non-expert operators. The framework's ability to interpret and execute voice commands accurately highlights its robustness and practicality in real-world applications. The framework's flexibility is evident in its capacity to accommodate additional robots and dynamic task sequencing, addressing the challenges of managing diverse robot teams in shared operational environments.

Nevertheless, the current reliance on predefined voice commands limits the naturalness and adaptability, potentially reducing effectiveness in more complex or unstructured scenarios. Advancing towards natural language understanding techniques would enhance user experience and make the system more adaptable to a larger range of operators and contexts.

Although the system currently uses inter-robot communication infrastructure, it has not yet been fully leveraged to facilitate direct data sharing among robots. Enabling this functionality could allow the exchange of maps, environmental data, and task progress, thereby enhancing collaboration and enabling decentralized coordination. These improvements would significantly boost operational efficiency and the scalability of the system.

4 CONCLUSION

This paper introduces a scalable robot-agnostic voice control framework for MRS, offering an accessible hardware-minimal solution for HRI. By centralizing command processing and filtering tasks based on robot capabilities, the system improves efficiency while maintaining robustness across heterogeneous teams. The experimental results confirm the framework's practicality and highlight its potential for deployment in dynamic, real-world scenarios.

To address current limitations, including the reliance on predefined voice commands and the lack of environmental awareness, future work will focus on integrating semantic mapping frameworks (Rollo et al., 2023a) to enable contextual understanding and support advanced loco-manipulation skills (Rollo et al., 2024a). In addition, incorporating natural language understanding techniques is expected to enhance communication flexibility and user intuitiveness. These developments aim to evolve the framework into a more scalable, autonomous, and comprehensive HRI solution for multi-robot collaboration in complex and dynamic environments.

REFERENCES

Amadio, F., Donoso, C., Totsila, D., Lorenzo, R., Rouxel, Q., Rochel, O., Hoffman, E. M., Mouret, J.-B., and Ivaldi, S. (2024). From vocal instructions to household tasks: The inria tiago++ in the eurobin service robots coopetition. *arXiv preprint arXiv:2412.17861*.

Belpaeme, T., Vogt, P., Van den Berghe, R., Bergmann, K., Göksun, T., De Haas, M., Kanero, J., Kennedy, J., Küntay, A. C., Oudgenoeg-Paz, O., et al. (2018). Guidelines for designing social robots as second language tutors. *International Journal of Social Robotics*, 10:325–341.

Budiharto, W., Cahyani, A. D., Rumondor, P. C., and Suhartono, D. (2017). Edurobot: intelligent humanoid robot with natural interaction for education and entertainment. *Procedia computer science*, 116:564–570.

Carr, C., Wang, P., and Wang, S. (2023). A human-friendly verbal communication platform for multi-robot systems: Design and principles. In *UK Workshop on Computational Intelligence*, pages 580–594. Springer.

Dahiya, A., Aroyo, A. M., Dautenhahn, K., and Smith, S. L. (2023). A survey of multi-agent human–robot interaction systems. *Robotics and Autonomous Systems*, 161:104335.

Del Bianco, E., Torielli, D., Rollo, F., Gasperini, D., Laurenzi, A., Baccelliere, L., Muratore, L., Roveri, M., and Tsagarakis, N. G. (2024). A high-force gripper with embedded multimodal sensing for powerful and perception driven grasping. In *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*, pages 149–156. IEEE.

Heppner, G., Oberacker, D., Roennau, A., and Dillmann, R. (2024). Behavior tree capabilities for dynamic multi-robot task allocation with heterogeneous robot teams. *arXiv preprint arXiv:2402.02833*.

Kumatani, K., Arakawa, T., Yamamoto, K., McDonough, J., Raj, B., Singh, R., and Tashev, I. (2012). Microphone array processing for distant speech recognition: Towards real-world deployment. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–10. IEEE.

Marin Vargas, A., Cominelli, L., Dell'Orletta, F., and Scilingo, E. P. (2021). Verbal communication in robotics: A study on salient terms, research fields and trends in the last decades based on a computational linguistic analysis. *Frontiers in Computer Science*, 2:591164.

Muratore, L., Laurenzi, A., De Luca, A., Bertoni, L., Torielli, D., Baccelliere, L., Del Bianco, E., and Tsagarakis, N. G. (2023). A unified multimodal interface for the relax high-payload collaborative robot. *Sensors*, 23(18):7735.

Padmanabha, A., Yuan, J., Gupta, J., Karachiwalla, Z., Majidi, C., Admoni, H., and Erickson, Z. (2024). Voicepilot: Harnessing llms as speech interfaces for physically assistive robots. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Papavasileiou, A., Nikoladakis, S., Basamakis, F. P., Aivaliotis, S., Michalos, G., and Makris, S. (2024). A voice-enabled ros2 framework for human–robot collaborative inspection. *Applied Sciences*, 14(10):4138.

Pyo, S., Lee, J., Bae, K., Sim, S., and Kim, J. (2021). Recent progress in flexible tactile sensors for human-interactive systems: from sensors to advanced applications. *Advanced Materials*, 33(47):2005902.

Rizk, Y., Awad, M., and Tunstel, E. W. (2019). Cooperative heterogeneous multi-robot systems: A survey. *ACM Computing Surveys (CSUR)*, 52(2):1–31.

Rogowski, A. (2022). Scenario-based programming of voice-controlled medical robotic systems. *Sensors*, 22(23):9520.

Rollo, F., Raiola, G., Tsagarakis, N., Roveri, M., Hoffman, E. M., and Ajoudani, A. (2024a). Semantic-based loco-manipulation for human-robot collaboration in industrial environments. In *European Robotics Forum 2024*, pages 55–59. Springer Nature Switzerland.

Rollo, F., Raiola, G., Zunino, A., Tsagarakis, N., and Ajoudani, A. (2023a). Artifacts mapping: Multi-modal semantic mapping for object detection and 3d localization. In *2023 European Conference on Mobile Robots (ECMR)*, pages 1–8. IEEE.

Rollo, F., Zunino, A., Raiola, G., Amadio, F., Ajoudani, A., and Tsagarakis, N. (2023b). Followme: a robust person following framework based on visual re-identification and gestures. In *2023 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*, pages 84–89. IEEE.

Rollo, F., Zunino, A., Tsagarakis, N., Hoffman, E. M., and Ajoudani, A. (2024b). Continuous adaptation in person re-identification for robotic assistance. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 425–431. IEEE.

Stone, P. and Veloso, M. (2000). Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots*, 8:345–383.

Su, H., Qi, W., Chen, J., Yang, C., Sandoval, J., and Laribi, M. A. (2023). Recent advancements in multimodal human–robot interaction. *Frontiers in Neurorobotics*, 17:1084000.