In-Contact Manipulation and Voice-Controlled Interaction for Underwater Robotic Arms

Zachary Speiser
Oregon State University
Corvallis, OR, USA
Email: speiserz@oregonstate.edu

***Abstract*—This report presents the research conducted as part of the 2024 Distributed Research Experiences for Undergraduates (DREU) program, supported by the Office of Naval Research (ONR). The project, under the CHARISMA team, led by Dr. Heather Knight, focused on developing an in-contact manipulation dataset and voice-controlled interaction algorithms for robotic arms. The in-contact manipulation dataset will serve as a foundation for future research. At the same time, the voice-controlled system, with Text-to-Speech (TTS) feedback, allows intuitive user interaction with the robot. The dataset is designed to capture how user phrases—particularly adverbs—map to robotic actions, enabling robots to interpret instructions like "poke" or "reset your position." Future integration of the ROS 2 joint trajectory controller will further enhance system capabilities. These advancements aim to enable human-in-the-loop control of robots in underwater applications, such as cleaning boats, sampling sensors, or welding, with the hope that voice-based control will provide a more intuitive interface compared to screen-based systems.**

I. Introduction

A. Background

Robotic manipulation in underwater environments poses significant challenges due to fluid dynamics, pressure variations, and the need for precise control in delicate tasks. Supported by the Office of Naval Research (ONR) and conducted under the CHARISMA team led by Dr. Heather Knight, this research addresses these challenges by developing algorithms for in-contact manipulation tasks and integrating voice-controlled systems to enhance human-robot interaction. The primary goal is to enable robotic arms to perform stable manipulation in underwater conditions, with a focus on real-time motion control, force sensing, and operator feedback.

B. Problem Statement

This project encompasses two major areas of focus: (1) the development of an in-contact manipulation dataset that connects natural language user instructions—especially adverbs—to robotic actions, and (2) the creation of a voice-controlled algorithm with TTS feedback for real-time operator interaction with the robotic arm. Both components are critical for advancing the capabilities of robotic systems in challenging environments, such as underwater scenarios. The joint trajectory controller integration is ongoing, and once fully operational, it will enable more precise control of the robotic arm in real-time applications.

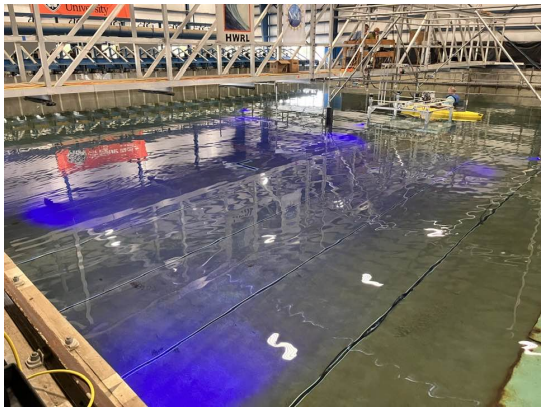

Fig. 1. Hinsdale Wave Research Lab: representative image from prior testing.

C. Objectives

The objectives of this project are to:

- Develop and validate algorithms for stable in-contact manipulation.
- Create an in-contact manipulation dataset that emphasizes the use of adverbs to connect user commands to robotic actions.
- Implement a voice-controlled algorithm for robotic arm interaction, with TTS feedback for operator guidance.
- Integrate the ROS 2 joint trajectory controller for enhanced real-time control.

II. Related Work

Robotic manipulation in dynamic environments has been the subject of extensive research, particularly regarding adaptive control and machine learning-based techniques. Prior work highlights the complexity of achieving precise in-contact manipulation, which this project seeks to address [1], [2]. Voice-controlled interaction is also gaining prominence, with studies demonstrating the potential for real-time feedback to improve user experience [3], [4]. The combination of tactile and voice input in robotics has been shown to enhance operator interaction, providing a more intuitive interface for controlling complex systems [4]. This research builds on these concepts by contributing an in-contact manipulation dataset that focuses on mapping adverbs like "quickly" or "precisely" to specific robot actions [5].

Fig. 2. Franka Research 3 robotic arm used for in-contact manipulation and voice-controlled interaction.

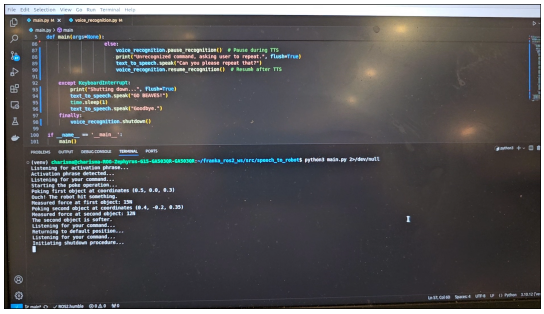

Fig. 3. Screenshot of the prototype video showing the software running on the laptop while controlling the robotic arm.

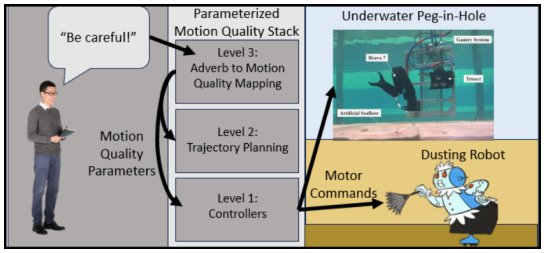

Fig. 4. Visualization showing motion quality stack, with levels of control from adverbs to motor commands.

III. In-Contact Manipulation Dataset: Motivation, Methods, and Status

A. Motivation

In-contact manipulation involves tasks where a robotic arm physically interacts with objects, often requiring precise force and motion control. The dataset developed in this project emphasizes connecting natural language instructions, particularly adverbs, to corresponding robotic actions. This approach aims to create a system where commands like "poke the object" or "reset your position" can be interpreted by the robot, ensuring that the robot's actions reflect the intended behavior described by the user.

The dataset is intended to support future research into improving the accuracy and stability of robotic manipulation, particularly in underwater environments where fluid dynamics add complexity. By linking adverbs to specific robot actions, researchers will be able to analyze how different manipulation strategies perform under varying conditions, contributing to the development of more robust control algorithms [1].

B. Methods

The dataset includes a wide range of in-contact manipulation tasks performed by the robotic arm, annotated with corresponding force data and voice commands. Each task involves the robot interacting with objects of different shapes, sizes, and materials. A particular emphasis is placed on capturing adverbs that modify robot behavior, such as "quickly" or "precisely," and mapping these linguistic cues to the robot's actions. Sensor feedback and motion data are collected during these interactions to enable detailed analysis of performance metrics such as force consistency, manipulation precision, and task completion time.

The dataset will provide insights into how verbal instructions, especially those using adverbs, can be reliably converted into actions by robotic systems. Different tasks are modeled, such as scraping, wiping, and poking, with varying levels of complexity, surface conditions (rigid vs. soft), and clutter (e.g., coral or seaweed obstructions). These diverse setups are crucial for creating a robust dataset that supports future algorithm development.

IV. Voice for Operator-in-Loop In-Contact Manipulation: Concept, Technology, and Status

A. Motivation

Voice-controlled interaction offers an intuitive way for operators to guide robotic systems, particularly in environments where hands-free control is advantageous, such as underwater. This project developed a voice-controlled algorithm to allow real-time interaction with the robotic arm, supported by TTS feedback to enhance operator understanding and communication. This approach simplifies the human-robot interface and enables operators to issue commands without the need for physical controls, focusing on how adverbs modify robot behavior [2].

B. Technology and Design

The voice-controlled system leverages the **Vosk speech recognition engine**, using the **vosk-model-en-us-0.22** model, which provides an accurate representation of US English [6]. Its performance has been tested across various domains, achieving 5.69% accuracy on the **librispeech test-clean** dataset, 6.05% on **TEDLIUM**, and 29.78% on **call center data**. This model was chosen for its versatility, but issues such as misrecognition of commands and difficulties in distinguishing similar phrases like "shutdown" versus

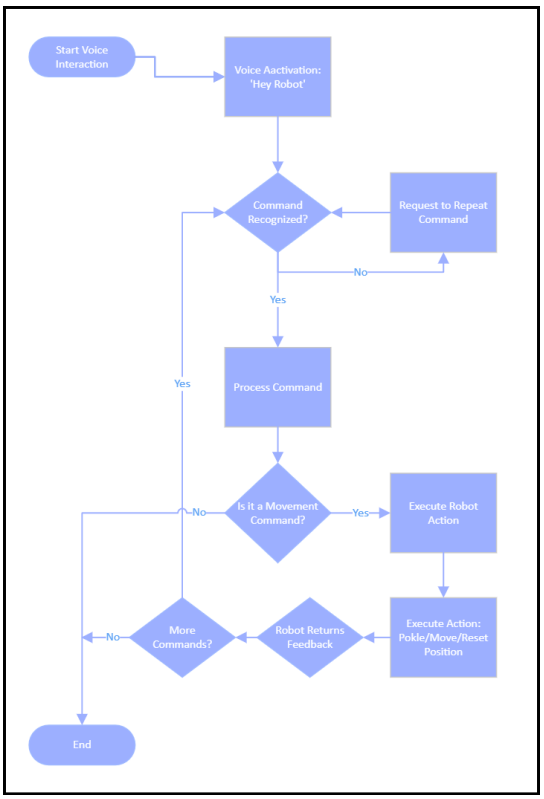

Fig. 5. Flowchart representing the software flow of voice-controlled interaction for the robotic arm, from activation to execution of commands.

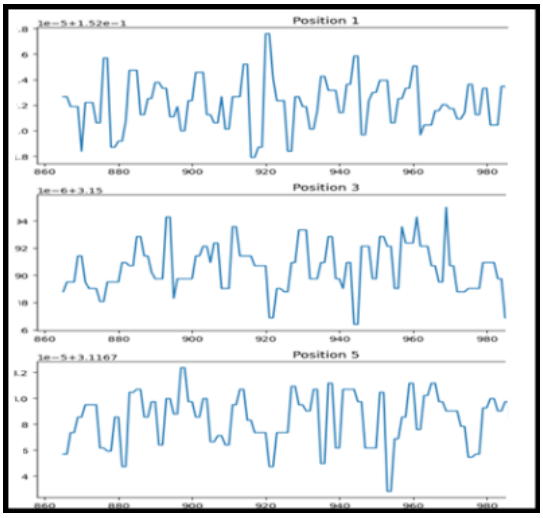

Fig. 6. Graphs showing data from the robotic arm during in-contact manipulation tasks.

”shut down” have led to less-than-ideal performance. These problems have prompted considerations of switching to an alternative voice recognition system in future iterations.

The system also incorporates Text-to-Speech (TTS) feedback through the **pyttsx3** library, which provides real-time verbal responses to the operator, ensuring that the robot’s actions are communicated effectively. The activation phrase is ”Hey robot,” followed by a verbal confirmation that the robot is ready to take commands. For example, an operator might say, ”Can you please poke those two objects over there and tell me which one is softer.” The robot would then execute the command, return a response regarding which object is softer, and wait for further instructions.

C. Status

The voice-controlled system is functional but needs improvement. Voice recognition accuracy is inconsistent, particularly in noisy environments, and adverb-based commands have not yet been fully integrated. The current system struggles with distinguishing between closely related phrases such as ”shutdown” versus ”shut down.” Future work will focus on enhancing the recognition system and integrating adverbs directly into the command structure. Additionally, exploring alternative voice recognition software may be necessary to improve reliability and performance.

V. Speech Recognition Evaluation and Challenges

The current implementation of the Vosk model for voice recognition has shown varying levels of success. While the system performs reasonably well in quiet environments, background noise, and reverberation have negatively impacted command recognition accuracy. The decision to use Vosk was based on its general versatility, but as the project evolved, limitations with Vosk’s ability to handle more nuanced command structures, such as adverbs, became apparent. This has prompted an ongoing evaluation of alternative voice recognition engines, such as Google’s Speech-to-Text API, to improve command recognition in future iterations.

VI. Future Applications and Use Cases

The voice-controlled and in-contact manipulation system developed in this research has broad potential for underwater applications beyond the current test environment. Future research teams can leverage the dataset for their own exploration of voice-to-robot interactions. Potential future use cases include:

- **Marine biology research**: Enabling robots to collect samples, probe delicate organisms, or inspect underwater habitats with high precision.
- **Ship maintenance**: The system could be adapted to clean hulls, scrape barnacles, and perform maintenance tasks autonomously while responding to voice commands from divers or operators.
- **Underwater welding**: Robots equipped with this technology could be directed by voice to perform welding operations in challenging underwater environments, where manual human intervention is dangerous or inefficient.
- **Environmental monitoring**: Robots could be used to collect data on ocean health, such as water quality,

temperature, and salinity, using voice-controlled actions to direct their sampling process.

By integrating adverb-based controls, the system could perform tasks with varying degrees of intensity or speed, allowing for fine-tuned control in highly variable environments such as the ocean floor. Future iterations of the dataset will include more complex manipulation tasks and explore how adverbs influence the execution of actions in real-time underwater conditions.

VII. Future Work

Moving forward, this project will focus on refining the voice recognition system and integrating it more seamlessly into the robotic control software. Specifically:

- **Adapting voice recognition**: Improving the system's ability to interpret nuanced commands, especially those involving adverbs.
- **Robot action integration**: Ensuring that the robotic arm can consistently execute the actions dictated by voice commands with greater precision and reliability.
- **Underwater testing**: The CHARISMA team will eventually test this system on an underwater robot, as the Franka Research 3 is not designed for underwater use.

VIII. Real-World Integration Challenges

Deploying this voice-controlled system for underwater robotic manipulation presents several real-world challenges. One major hurdle is the integration of the voice recognition system with the environmental constraints of an underwater setting. Acoustic properties in water, such as sound propagation and interference from surrounding elements, could negatively affect voice command recognition. Additionally, the robot's sensors must be capable of providing accurate real-time feedback despite these harsh conditions.

Hardware limitations, such as battery life and processing power on underwater robotic systems, could also restrict the duration and complexity of tasks. Ensuring reliable wireless communication between the operator and the robot in an underwater environment poses another technical challenge.

Future iterations of this system will focus on improving the robustness of sensor data transmission and voice command processing, to deploy this technology for long-duration underwater tasks without compromising performance.

IX. Conclusion

This research successfully developed and tested methods for in-contact manipulation and voice-controlled interaction with a robotic arm. The in-contact manipulation dataset connects user phrases—particularly adverbs like "quickly" and "precisely"—to robot actions, providing a foundation for future research. While the voice-controlled system has been implemented, it still requires improvements, particularly in terms of speech recognition accuracy and robustness in noisy environments. Future work will focus on integrating adverbs into the voice recognition system and improving overall system reliability, including switching to a different voice recognition engine if necessary. The joint trajectory controller is in the process of being integrated, which will further enhance the system's precision and reliability in executing real-time actions for underwater applications.

References

[1] R. Challa, H. Brown, C. Jain, Z. Speiser, and H. Knight, "Scrape that barnacle: Commanding underwater robot in-contact manipulation tasks with intuitive spatial-temporal-force features," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. Paris, France: IEEE, 2024. [Online]. Available: https://github.com/akzspeiser/akzspeiser.github.io/blob/master/files/Scrape_That_Barnacle.pdf

[2] H. W. Ka, D. Ding, and R. Cooper, "Aroma-v2: Assistive robotic manipulation assistance with computer vision and voice recognition," in *Proceedings of the 9th Conference on Rehabilitation Engineering and Assistive Technology Society of Korea*. Gyeonggi, Korea: Rehabilitation Engineering and Assistive Technology Society of Korea, 2015. [Online]. Available: https://d-scholarship.pitt.edu/26361/1/Ka_RESKO_2015.pdf

[3] B. House, J. Malkin, and J. Bilmes, "The voicebot: A voice controlled robot arm," in *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI)*. Boston, Massachusetts, USA: ACM, 2009. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=c788b209140c538ff4181a1ee30a89b45f0d5e49

[4] T. Sawabe, S. Honda, W. Sato, T. Ishikura, M. Kanbara, S. Yoshikawa, Y. Fujimoto, and H. Kato, "Robot touch with speech boosts positive emotions," *Scientific Reports*, vol. 12, p. 6884, 2022. [Online]. Available: https://www.nature.com/articles/s41598-022-10503-6

[5] J. Swaminathan, C. Jain, M. Miller, and H. Knight, "A semi-automated multi-robot comedy performance system with gesture," in *Proceedings of the International Conference on Social Robotics*. Collaborative Robotics and Intelligent Systems (CoRIS) Institute, Oregon State University, 2022. [Online]. Available: https://ir.library.oregonstate.edu/concern/graduate_thesis_or_dissertations/tx31qr60r

[6] A. C. Inc., "Vosk speech recognition," *GitHub Repository*, 2024. [Online]. Available: https://github.com/alphacep/vosk-api