CS639 Project Proposal
Voice-to-Action: Multimodal Language-Driven Control for a 6-DOF Robotic Arm

Project Team:
Yuanpei Zhang(yzhang2939@wisc.edu) Bin Xiao (bxiao@wisc.edu)

Project Setting:

This project aims to develop a natural-language interface that allows intuitive human control of a Hiwonder 6-DOF robot arm(The end-effector is a grasper). The robot interprets spoken instructions (e.g., "Pick up the red cube and place it on the left") and autonomously executes the corresponding motion. The system bridges voice commands, visual perception, and motion control through a Large Language Model (LLM)-based task planner. The implementation will use the ROS 2 framework to integrate speech recognition, object detection, and motion execution nodes. This task is chosen to explore how multimodal understanding—combining language, perception, and kinematics—can enable more natural and adaptive robot behavior beyond rigid programming.

Techniques:

The system architecture consists of four major modules:

1. **Speech Recognition (Perception I)** – Real-time transcription of voice commands using Whisper or Google Speech-to-Text.
2. **Task Parsing via LLM (Planning / High-Level Reasoning)** – Converts natural language commands into structured subtasks (e.g., detect, pick, place) using a predefined schema.
3. **Visual Perception (Perception II)** – Using a depth camera to get the location of our target cubes and put them into proper locations based on the voice command.
4. **Motion Execution (Kinematics Control)** – Motion planning via MoveIt 2, implementing primitive functions such as `move(x,y,z)`, `grasp()`, and `ungrasp()`.

This system integrates multiple core robotics topics—including perception, planning, and control—covered in CS639.

Evaluation Plan:

1. **Experiment 1 – Command Understanding Accuracy:** Measure how accurately natural-language commands are mapped to correct subtask sequences. 10 spoken commands (simple to compositional) will be tested. *Metric:* Parsing accuracy (%).
2. **Experiment 2 – Task Execution Robustness:** Evaluate success rate and precision of pick-and-place execution under different object positions. *Metrics:* Completion rate, positional error (cm), and latency (s).

Clarity and Scope:

The project is modular and feasible within the 6-week timeline. The final demo will show the 6-DOF robotic arm performing real-time pick-and-place tasks in response to spoken commands. This system demonstrates an end-to-end robotic pipeline integrating speech, vision, reasoning, and motion—a full-stack embodiment of the CS639 learning objectives.