Fuzzy Systems and Data Mining X
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA241453

A Mobile Robot Arm for Providing Daily Support to Cantonese Speaking Elderly Persons

Chi-Yat Lau [a] , Man-Ching Yuen [b, 1] , Chi-Wai Yung [c] ,
Tsz-Fung Wong [a], Ka-Keung Chan [a], Check-Hang Chow [a], Tsz-Him Ho [a],
Hong-San Wong [d]
[a] *Department of Information Technology, Vocational Training Council, China*
[b] *Department of Digital Innovation and Technology, Technological and Higher Education Institute of Hong Kong, China*
[c] *Linux Foundation (Asia Pacific)*
[d] *R2C2 Limited*

Abstract. Many elderly need help to perform tasks around their home, such as grabbing their clothes and care takers can provide an assistant. The objective of this project is to utilize human language for controlling robot arm operations. Additionally, the robot will engage in interactions with humans through a large language model (LLM), enhancing its responses to be more human-like. Moreover, the robot arm will have the capability to discern various syntaxes in human language, allowing it to respond to commands in a manner akin to human speech. The project focuses on the robot arm's ability to respond to Cantonese (廣東話/粵語) input and execute basic robotic movements based on Cantonese commands. This prototype can help to develop a robot arm with more functions based on Cantonese commands.

Keywords. Large language model, LLM, robot

1. Introduction

1.1. Background and Motivation

Recently, elderly population increased globally, and more care takers are needed to provide daily supports to elderly. With the help of robot arm, elderly can perform a lot of tasks without others' help, such as grabbing objects. It is a good idea to have a robot arm which can understand elderly's verbal commands and perform tasks for them.

In recent years, products and software related to generative artificial intelligence have seen rapid development. Examples include ChatGPT and Midjourney. These large language models offer people significant convenience and access to a wealth of rich content, with very low threshold requirements. Furthermore, we found that large language models can act as a bridge between robots and humans that allow humans to

[1] Corresponding Author, Man-Ching YUEN. Technological and Higher Education Institute of Hong Kong, Hong Kong, China; E-mail: connieyuen@thei.edu.hk.

command robots by language, and robots able to understand our verbal command correspondingly. This finding brings innovative ideas to robot remoting aspects.

This project seeks an innovative approach to robotic remoting. We found that there are a series of issues with typical robotic controlling approaches like the long learning path to the robot controller, and the limited dexterity and precision of the human. Therefore, we tried to develop a robot-understandable translator to execute our command. To achieve this, we developed Integrate Advanced Natural Language Platform to translate our verbal language into robotic commands. Also, we developed a Seamless Voice-to-Text Conversion and communication platform to allow the controller to control the robot by speaking or texting respectively. We believe that this innovative approach can reduce the learning cost of robotic remoting and also be the most effective and the simplest way to control different types of robots.

1.2. Our Contributions

Our contributions in this project are shown as follows:

- **Integrate Advanced Natural Language Processing (NLP) with Robotics:** To combine cutting-edge NLP technologies, including OpenAI's Whisper and Breeze model, with robotic systems to create a user-friendly, voice-controlled robotic assistant.
- **Develop Seamless Voice-to-Text Conversion:** Utilize Whisper technology for efficient and accurate conversion of voice commands into text, ensuring the system can understand and process user inputs with high precision.
- **Enable Natural Language Understanding and Response Generation:** Leverage the capabilities of the Mistral model to ensure the robotic assistant can understand complex language structures and respond in a human-like manner, enhancing the interaction experience.
- **Create an Intuitive User Interface:** Design an interface that is easy to navigate and accessible for a wide range of users, regardless of their technical expertise or experience with robotic systems.
- **Ensure High Accuracy and Responsiveness:** Aim for a high degree of accuracy in voice recognition and response generation to minimize errors and misunderstandings in human-robot interaction.
- **Promote Accessibility and Usability:** Make the robotic assistant accessible to a diverse user base, including those with disabilities or limited technical knowledge, enhancing the inclusivity of the technology.

2. System Architecture of Voice-Controlled Robotic Assistant with Natural Language Interaction (VCRNI) Project

2.1. Overview of the Solution

The data product implementation for the Voice-Controlled Robotic Assistant with Natural Language Interaction (VCRNI) revolves around creating a smooth integrated system that inscoterprets natural language commands through large language models and translated them into actionable tasks for a robotic arm. Leveraging technologies like Whisper for speech recognition, GGUF-Mistral (GPT-Generated Unified Format-

Mistral) for natural language understanding, the solution aims to offer an intuitive and interactive experience. The system will be presented through a user-friendly web interface, providing real-time insights into LLM conversations, historical interactions, and a live feed of the robotic arm's actions.

In the context of the Voice-Controlled Robotic Assistant with Natural Language Interaction (VCRNI) project, the presentation and application of data involve several key components, such as voice input processing, text understanding, text generation, and real-time camera feed.

The project employs a series of sophisticated algorithms to translate human interaction into robotic actions seamlessly. The Voice-to-Text Conversion Algorithm is the first critical step in this chain. It involves capturing audio inputs, preprocessing them to reduce noise, and feeding the cleaned audio into the Whisper model for accurate transcription. The output is a refined text version of the user's spoken commands, ready for further processing. Next, it is the Command Processing Algorithm, which is central to the system's intelligence. It takes the transcribed text and utilizes Mistral to analyze and extract the user's intent. This algorithm maps the understood intent to specific robotic actions or responses, effectively translating human language into machine-understandable commands. Finally, the Robot Control Algorithm acts as the executor of these commands. It receives instructions from the natural language processing module and translates them into actionable messages or directives. This algorithm is responsible for the direct control of the robotic hardware, ensuring that the robot responds accurately to the processed commands. It also includes mechanisms for real-time feedback and action adjustment based on sensor data, ensuring a responsive and adaptive robotic assistant. Together, these data models and algorithms form the backbone of the VCRNI project, enabling it to deliver an unparalleled user experience in human-robot interaction through advanced natural language processing and robotic control technologies.

2.2. Real-time Camera Feed

Integrating real-time camera feed of the robotic arm into the web interface adds a visual element to the user experience. Users can observe the physical actions and movements of the robotic arm as it responds to voice commands. This can enhance the transparency and understanding of how the system interprets and executes commands in the physical environment. Figure 1 shows how the real-time camera feeds. First, we proceed to preprocess the chessboard to ensure that the robot arm can move within a legal range. This step involves marking out 126 positions on the chessboard, representing the areas where the robot arm can legally reach. While marking these positions, we also determine the necessary joint and angle information for each position for the subsequent control.

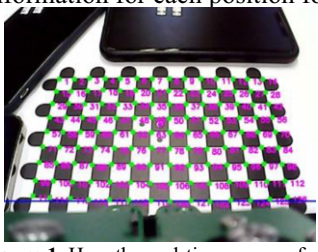

Figure 1. How the real-time camera feeds

Object modelling

Afterward, YOLO (You Only Look Once) v5 is utilized to train the target object, such as screws, tape and label the data through Roboflow. To enhance the flexibility and adaptability of the robot arm system, the combination of YOLO v5 and Roboflow enables quick and accurate detection of target objects in images.

Calculating Target Object Center Coordinates and Corresponding Joint Angles

The center coordinates of the target objects within the real-time camera are computed. Subsequently, distances between each target object and every point on the preprocessed chessboard are calculated using the Euclidean distance formula to determine the closest point to each object. This method ensures that identification of the most suitable points, accurately positioning the target objects on the image, with the corresponding joint and angles for the movement of the robot arm.

$$\text{Distance} = \sqrt{((x2 - x1)^2 + (y2 - y1)^2)}$$

2.3. Data Analytical Framework

Data Acquisition and Understanding

In data acquisition, the process of the LLM model for integration with a robotic arm begins with a comprehensive approach to data acquisition. This critical phase involves gathering a diverse array of textual data, including conversations, technical manuals, and user commands. Ensuring the inclusion of various linguistic styles, technical jargon, and everyday language is paramount. Additionally, the collection of voice data is crucial, especially for understanding nuances in spoken language. This may include recordings of user commands and dialogues. To customize the system for a robotic arm, it is essential to acquire contextual and environmental data, such as details about physical environments, types of objects interacted with, and typical user interaction scenarios. For data understanding, a thorough analysis of both text and voice data is conducted to unravel common language patterns, terminologies, and user intentions, essential for effective Natural Language Processing (NLP).

Data Models Training

The data flow diagram shown in Figure 2 illustrates a comprehensive process in the speech recognition system for controlling a robot arm. Start with human speech input, captured through devices like microphones. The recognized speech was recorded and resulting in the creation of an audio file. Subsequently, the audio file is subject to the prowess of the Speech-to-Text Conversion Model, specifically the Faster Whisper generating textual data [1, 2]. Whisper is an automatic speech recognition (ASR) system trained on 680,000 hours of multilingual and multitask supervised data collected from the web. Complementing this is the Natural Language Understanding and Generation Model, powered by Mistral. Anthony Alford [3] said that Mistral AI's Open-Source Mixtral 8x7B Outperforms GPT-3.5. The Mistral model's prowess in understanding and generating human-like language responses is key to creating an intuitive user experience. The pre-training of Mixtral is conducted using data sourced from the open web, incorporating a simultaneous training mechanism for both the experts and the router networks, thereby optimizing its performance across diverse tasks.

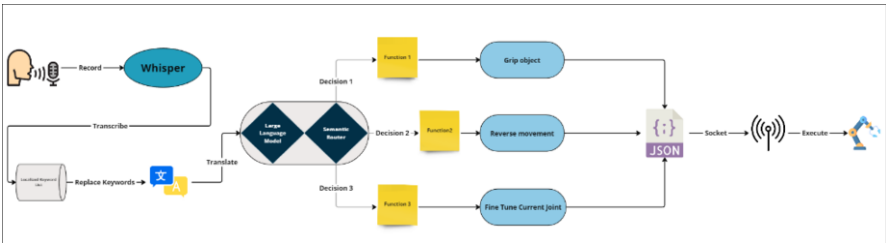

Figure 2. Speech-to-Command Processing for Robot Arm Control

Data Testing

Data testing focuses on evaluating the model's ability to accurately interpret and execute natural language commands in real-world scenarios. This would involve rigorous testing with diverse language inputs to ensure the robot's responses are appropriate and accurate. The paper may describe scenarios where the model's understanding of commands is tested, along with its ability to convert these commands into actionable tasks for the robot, emphasizing the importance of precision and reliability in human-robot interactions.

3. Hardware and Software Requirements

3.1. Hardware Requirements

Robotic arm as shown in Figure 3 is essential for executing precise movements and actions based on interpreted natural language commands. The arm should be compatible with integration into a networked system for seamless communication. Computer with GPU, the system requires a computer with GPU support for running the Mixtral-8x7B-Instruct Large Language Model (LLM). The GPU accelerates the model's learning process and ensures efficient real-time interpretation of natural language commands. Ideally with at least 12 GB of VRAM. Camera System which integrated cameras on the robotic arm and potentially on the computer facilitate real-time visual feedback. These cameras aid in monitoring the environment, contributing to precise robotic movements. This robotic arm is an ideal pilot test object.

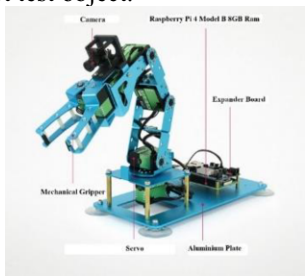

Figure 3. Robotic Arm

3.2. Software Requirements

Python serves as the primary programming language for the development of various components, including the integration of the robotic arm, implementation of the LLM model, and overall system development. LangChain Framework, as the language processing framework, plays a crucial role in interpreting and understanding user commands. It forms the backbone for processing natural language inputs and generating appropriate responses. Faster Whisper Speech-to-Text model is employed for speech-to-text conversion. This technology aids in accurately transcribing voice inputs, facilitating seamless communication between users and the robotic assistant. Socket programming is utilized for communication between the robotic arm, computer. This includes the transmission of commands, data, and feedback, ensuring a cohesive and responsive interaction.

4. Performance Evaluation

4.1. Object Detection and Joint Angle Retrieval System

To verify the functionality and accuracy of the object detection and joint angle retrieval system under various scenarios. In Figure 4, rotate the robotic arm to the left and verify if it accurately received command in natural language. Command is "向左轉" (Turn to the left). In Figure 5, rotate the robotic arm to the other size from the current position (Left) and verify if it accurately received command in natural language Command is "轉向另一個方向" (Turn to the other size). In Figure 6, test the system's ability to retrieve joint angles for gripping objects (Screw). Command is "握住最左邊的螺絲" (Grip the leftmost screw).

Figure 4. Turn Left Test

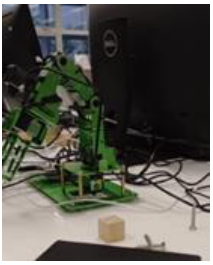

Figure 5. Turn to the other side Test

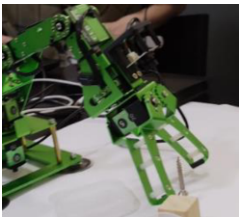

Figure 6. Object Gripping Test

4.2. Accuracy of Semantic Understanding facilitated by Semantic Router

The LLM will determine which tools and developed function will be utilized and extract parameters of target function with semantic router from the user instruction, enabling seamless integration and execution of the specified tasks [4]. Figure 7 shows the improvement of system robustness with semantic router. In Figure 8, the semantic router also represented a wise synonym understanding for comprehending the intent behind user's instructions.

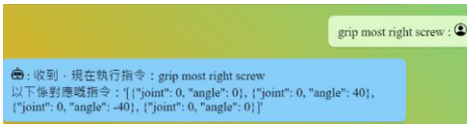

Figure 7. System Robustness in Syntactic Level

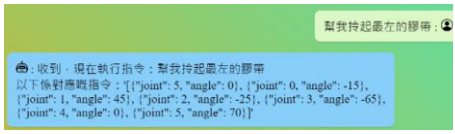

Figure 8. A wise synonym understanding for comprehending the intent behind user's instructions

4.3. Processing Time Comparison

Figure 9 clearly illustrates that the Faster Whisper Model showcases a significantly shorter processing time, completing its tasks in approximately 5 seconds. This remarkable reduction in processing time is a key performance metric. The performance metrics emphasize significant speed improvement, reduces processing duration with the Faster Whisper Model.

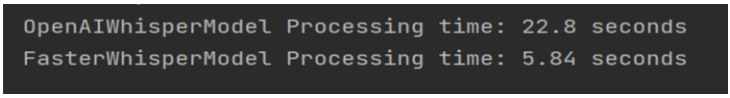

Figure 9. The quality versus inference budget tradeoff.

5. Conclusion

Throughout the implementation of the project, the process initiates with human language input and progresses to utilizing a large language model to determine the appropriate function for execution. Our system specifically targets Cantonese users, using Traditional Chinese as the main language. This specialization introduces substantial complexities in both the integration and operation of the large language model. A key aspect of the system is its ability to comprehend the semantic meanings of words and sentences, thereby avoiding misunderstandings and necessitating precise, strictly structured inputs. One target user group is Cantonese-speaking elderly, they can control the robot to complete some tasks for them in daily lives by verbal instructions. The cost of the robot arm is low and the robot is easy to set up and use. The robot is suitable for Cantonese-speaking elderly to use. Since there are many people who speak different dialects in Mainland China, we will build up a data model which can support other dialects and thus serve more users of different dialects in the future.

References

[1] Guillaume Klein et al. (2020, July). *Efficient and High-Quality Neural Machine Translation with OpenNMT* https://aclanthology.org/2020.ngt-1.25.pdf
[2] Guillaume Klein et al. (2023, Jan.). *SYSTRAN/faster-whisper: Faster Whisper transcription with CTranslate2.* GitHub. https://github.com/SYSTRAN/faster-whisper
[3] Anthony Alford. (2023, Jan 23). *Mistral AI's Open-Source Mixtral 8x7B Outperforms GPT-3.5.* InfoQ https://www.infoq.com/news/2024/01/mistral-ai-mixtral/
[4] Simonas Jakubonis et al. (2023, Nov 9). *aurelio-labs/semantic-router.* GitHub. https://github.com/aurelio-labs/semantic-router