인간-로봇 상호작용을 위한 안전 강화형 음성 인식 로봇 제어 시스템 설계 및 구현

홍준기, 정하륜, 김재호*

세종대학교

{joongi.sejong, haryun.sejong}@gmail.com, *kimjh@sejong.ac.kr

Safety-Enhanced Speech Recognition Robot Control System for Human-Robot Interaction

Jungi Hong, Haryun Jeong, Jaeho Kim*

Sejong University

요 약

본 논문에서는 음성 인식과 언어 모델을 결합한 대화 기반 로봇팔 제어 시스템을 구현하고, 호출어 감지, 발화 구간 검출, 프롬프트 설계, 실행 단계 제약을 통해 안전성을 강화하였다. 시스템은 사용자 음성을 텍스트로 변환한 뒤 언어 모델로 의도와 대상 물체를 추론하고, 대상 물체에 대응하는 사전 정의 좌표를 ROS 토픽(topic)으로 전송하여 로봇팔 동작을 수행하며, UI를 통해 호출어 대기부터 추론 및 실행까지의 진행 상태를 시각화하여 사용성을 향상했다.

Ⅰ. 서 론

최근 대규모 언어 모델(LLM)의 비약적인 발전으로 자연어를 이해하고 대화 맥락을 반영하여 사용자의 의도를 해석하는 연구가 활발히 진행되고 있다. LLM은 발화의 문맥을 고려해 의미를 추론함으로써 기존 명령 체계의 경직성을 완화할 수 있다. 따라서 같은 작업 의도라도 표현이 달라질 수 있는 상황에서 명령을 일관되게 해석하여 동일한 목표 수행으로 연결할 수 있어 활용 가치가 높다. 반면 기존의 규칙 기반 인터페이스는 구현 특성상 사전에 정의된 제한적인 명령어에 의존해야 하며 이로 인해 표현 다양성과 예외 상황을 포괄하기 어렵다. 이러한 한계를 보완하기 위해 다양한 시스템과 응용 분야에서 LLM의 적용이 확장되고 있다.

특히 인간-로봇 상호작용(Human-Robot Interaction, HRI) 분야에서는 사용자와 로봇이 상호작용할 수 있도록 LLM을 로봇 제어에 접목하려는 시도가 증가하고 있다. 그러나 실제 환경에서는 배경 소음에 의한 음성 인식 오류 및 발화의 모호성으로 인해 의도 해석의 신뢰성이 저하될 수 있으며, 이는 로봇의 오작동이나 충돌과 같은 안전사고로 이어질 위험이 있다. 따라서 음성 입력과 LLM의 추론 결과를 실제 로봇 동작으로 연결하는 과정에서 불확실성을 완화하고 안전성을 보장할 수 있는 제어 구조가 필수적이다.

본 논문에서는 LLM 추론의 오류를 최소화하면서 화자의 의도에 맞게 로봇팔이 움직이도록 하고 안전 규칙을 추가한 시스템을 설계하였으며 이를 실제 데모를 통해서 검증하였다. 제안된 시스템은 상호작용 과정에서의 오류를 줄이고 의도하지 않은 동작의 가능성을 완화한다.

Ⅱ. 본론

2-1.시스템 설계

본 연구에서 제안하는 시스템은 그림 1과 같이 음성 인식 인터페이스와 언어 모델 추론 및 좌표 송신으로 구분된다. 시스템은 사용자의 발화를 텍스트로 변환한 뒤 LLM이 대화 문맥을 기반으로 의도를 추출하고 대상 물체를 선정하며 이를 로봇팔 제어 명령으로 변환하여 작업 수행으로 연결하도록 설계하였다.

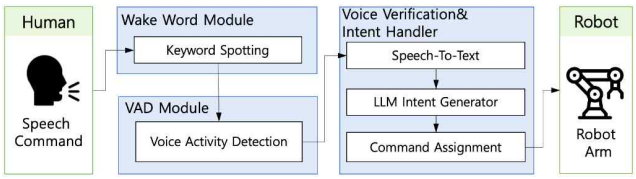

그림 1. 시스템 구성도

2-2. 음성인식 모듈

음성 인식 모듈은 불필요한 연산을 줄이고 시스템 효율을 높이기 위해 호출어(Wake Word) 감지 방식을 채택하였다. 대기 상태에서 Porcupine Wake Word[1] 기반의 Keyword Spotting이 항시 작동하며 사용자가 특정 호출어를 발화했을 시 메인 시스템이 활성화된다. 시스템이 호출되면 자동 음성 감지 기능인 Voice Activity Detection(VAD) 모듈이 작동하고 입력 오디오로부터 발화 구간을 자동으로 검출한다. Silero VAD[2] 기반의 VAD 모듈은 오디오 신호에 대한 발화 확률을 실시간으로 계산하고 해당 확률이 설정된 임계값을 연속적으로 초과하는 구간을 발화 시작으로 판단함으로써 배경 소음과 사용자 발화를 구분한다. 또한 확률이 일정 시간 이상 임계값을 연속적으로 넘지 못하는 경우 발화 종료로 판단하여 후속 처리를 위한 음성 구간을 확정한다. 이러한 구조를 통해 사용자는 버튼 조작 없이 호출어만으로 시스템과 상호작용할 수 있으며 배경 소음으로 인한 음성 인식 오류를 방지하여 로봇팔의 의도하지 않은 동작 가능성을 완화한다. 발화 종료를 감지하면 Whisper-large-v3[3] 기반으로 개발된 Speech-To-Text가 오디오를 텍스트로 변환하고, LLM Intent Generator가 변환된 텍스트를 기반으로 사용자의 의도 추론을 수행하며 LLM 모델로는 Llama-3.1-8B[4]가 사용되었다.

2-3. 언어 모델 추론 및 좌표 송신

언어 모델은 사용자의 발화로부터 의도를 분석하여 필요한 물체를 식별한다. 이후 시스템은 식별된 물체에 대응하는 사전 정의 좌표를 조회하고 ROS(Robot Operating System)의 토픽(topic) 통신을 통해 로봇팔 제어 노드에 목표 정보를 전달함으로써 실제 동작 수행으로 연결한다.

표 1.언어 모델 프롬프트

규칙 및 절차: step 1. 현재 발화에 요구(필요, 욕구, 요청)이 있는지 판단 step 2. 판단한 발화가 단순 언급, 관찰, 나열은 아닌지 다시 한 번 유의 step 3. 요구가 있다는 것이 확인될 경우 각 요구를 다음과 같이 연결 #1 갈증→Water, #2 허기→Snack, #3 놀이→Toy #4 기타→none 예시(Few-shot; none 포함): - "목말라."→ Water/ "간식 먹고 싶어."→ Snack / "심심해."→ Toy "물 필요 없어."→ none / "테이블에 물 있네." → none.....

본 시스템의 안전성을 높이기 위해 언어 모델 프롬프트의 경우 표 1에서와 같이 예시 기반 프롬프팅(Few Shot Prompting)[5]을 통해 사용자의 의도가 불분명하거나 시스템이 지원하지 않는 물체를 요구할 경우 물체를 선정하지 않도록 유도하는 부정 예시들을 프롬프트에 포함하였다. 또한 생각의 절차 기법(Chain of Thought)[6]을 통해 해당 물체의 언급과 실제 필요에 의한 요청을 구분하는 절차를 포함하여 의도 파악의 정확도를 높였다. 또한 추론 성능 보완을 위해 발화를 영어로 번역하여 입력으로 사용하였다. 로봇팔은 수신한 목표 정보를 기반으로 물체가 위치한 방향으로 이동하며 작업을 수행한다. 이때 로봇팔이 동작하는 동안에는 추가 음성 명령을 처리하지 않도록 ROS의 서비스(service)를 통해 명령 수용을 일시적으로 비활성화하고 동작 완료 이후에 다시 음성 인식을 수행하도록 제어 흐름을 구성하였다.

2-4. 시스템 구현

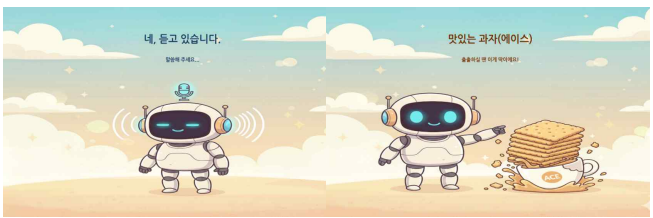

그림 2. 사용자 인터페이스

사용자 인터페이스(UI)를 통해 사용자는 현재 진행 상황을 확인한다. 사용자가 발화할때 시스템은 호출어 인식, 음성 감지, 음성-텍스트 변환 및 언어 모델 추론, 물체 선정의 순서로 동작하며 UI는 그림 2와 같이 시스템 상태가 변경될 때마다 ROS 토픽(topic)으로 전달되는 상태 메시지를 구독(subscribe)하고 수신한 상태 유형에 따라 화면에 표시되는 이미지를 전환한다.

시스템 노드 로그를 통해 절차를 더 자세히 확인할 수 있다. 그림 3과 같이 사용자의 발화가 끝났을 시 언어 모델이 사용자의 의도를 추론하고 그에 따른 물체 선정 및 해당 물체에 대응하는 목표 좌표를 ROS 토픽으로 송신하는 과정을 거치면 그림 4와 같이 로봇팔이 물체를 향해 이동하며 발화는 잠시 차단된다.

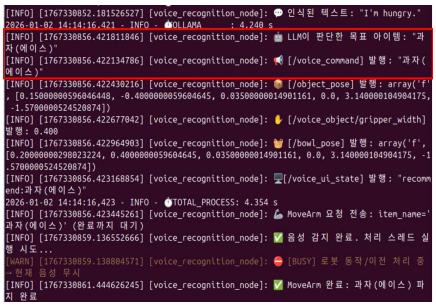

그림 3. '과자' 선정에 대한 시스템 작동 과정

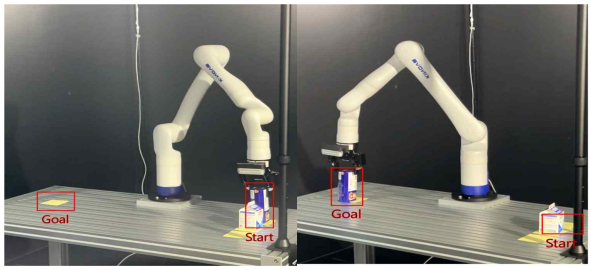

그림 4. 좌표 송신에 따른 로봇팔 작동 과정

Ⅲ. 결론

본 논문에서는 음성 기반 상호작용의 편의성을 유지하면서도 물리적 동작을 수행하는 로봇팔 시스템에 요구되는 안전성을 확보하기 위해 음성 인식과 언어 모델을 결합한 대화 기반 로봇팔 제어 시스템을 설계하고 구현하였다. 제안 시스템은 사용자의 음성을 음성-텍스트 변환 모듈로 텍스트화한 뒤 언어 모델 추론을 통해 의도를 도출하고 이를 좌표 송신을 통한 로봇팔 제어 명령으로 변환하여 실행 단계까지 연결한다. 또한 배경 소음 환경에서의 오동작을 줄이기 위해 자동 음성 감지 기반 모듈을 적용하고, 안전 지향 프롬프트 설계 및 제어 규칙을 통합하여 안정적인 동작을 지원한다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학 ICT연구센터(ITRC)의 지원(IITP-2026-RS-2021-II211816)과 산업통산자원부 및 산업기술평가관리원(KEIT)의 지원(RS-2022-00154678)과 한국연구재단 및 무인이동체원천기술개발사업단의 지원을 받아 무인이동체원천기술개발사업을 통해 수행되었음(RS-2020-NR117734).

참 고 문 헌

[1] Porcupine Wake Word [Website]. (2026, January 5). https://picovoice.ai/docs/porcupine/

[2] Silero-vad [Website]. (2026 January 5). https://github.com/snakers4/silero-vad

[3] Whisper-large-v3 [Website]. (2026, January 5). https://huggingface.co/openai/whisper-large-v3

[4] A. Grattafiori et a., "The Llama 3 Herd of Models," arXiv preprint arXiv:2407.21783, 2024.

[5] T. B. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, 33: pp. 1877 - 1901, 2020.

[6] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," In Advances in Neural Information Processing Systems(NeurIPS) 35: pp. 24824 - 24837, 2022.