궤도로봇의 작업편의 향상을 위한 인공지능 인터페이스

Orbital Robot AI Interface for Improving User Usability

김 태 현[1] · 정 재 준[2] · 남 서 용[2] · 유 용 재†
Taehyeon Kim[1], Jaejun Jung[2], Seoyong Nam[2], Yongjae Yoo†

Abstract: The large infrastructures as like tunnels need to be inspected periodically. There are many efforts to adopt industrial robots, however, it is hard to apply because of the sound noise and low light. For this paper, we tried to implement the AI system to make easy to control industrial robot using Human-Robot Interaction. We adopted the speech recognition system, gesture recognition system, and multimodal system to understand the person's intention.

Keywords: Human-Robot Interaction, Command Recognition, Speech Recognition, Gesture Recognition, Multimodal, AI Interface

1. 서 론

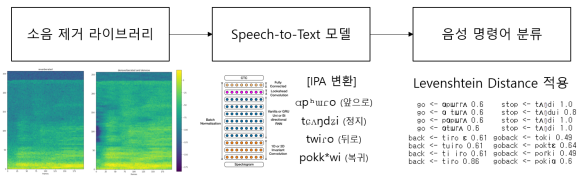
사회기반 시설물 중 터널과 같은 대규모 토목 건축물은 지속적인 점검이 필요하다. 현장 노동자를 투입하던 이전과 달리 기술의 발전으로 산업용 로봇의 도입이 계속되고 있지만, 주변 소음과 저조도의 환경에서는 적용이 제한되었다. 본 논문에서는 이러한 상황에서도 작업 환경에서의 인부가 Human-Robot Interaction을 바탕으로 산업용 로봇을 쉽게 조작할 수 있도록 인공지능 기술을 적용하였다. 음성인식 시스템과 제스처 인식 시스템을 바탕으로 인부의 의도를 파악하고, 멀티모달 시스템을 바탕으로 최종 결정을 내리는 알고리즘을 적용하는 방식으로 설계하였다.

2. 로봇 명령어 인공지능 인식 시스템

2.1 음성인식 시스템

2.1.1 구조 설명

소음 환경에서의 명령어 인식을 위해 소음 제거 라이브러리를 적용하고, 한국어에 대한 STT(Speech-to-Text) 모델을 바탕으로 IPA(International Phonetic Alphabet)를 반환한다. 그 후에 발음 유사도를 기반으로 하여 명령어를 인식한다.



[Fig. 1] 음성인식 시스템 구성도

2.1.2 소음 제거 라이브러리 적용

데이터 전처리 과정에 NARA-WPE 와 noisereduce 라이브러리를 바탕으로 소음 및 잔향 제거 과정을 진행하였다.

2.1.3 STT(Speech-to-Text) 모델

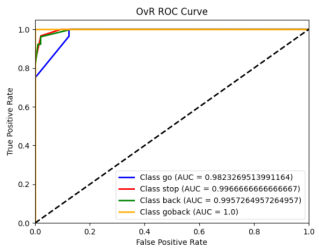
STT 모델은 Baidu 에서 개발한 Deep Speech 2 모델이 구현된 kospeech 툴킷을 바탕으로 학습을 진행하였다. AIHub 의 '극한 소음음성 인식 데이터[1]' 데이터셋에는 각종 소음 환경에서 녹음된 음성 파일과 라벨값으로 구성되어 있다. 따라서, 제시된 환경에서의 적용이 가능하다고 판단하여 데이터의 국문 라벨값을 IPA로 재구성[3]하여 학습할 수 있었다. 학습된 모델의 Validation Dataset CER(Character Error Rate) 결과는 약 0.12 이었으며 Test Dataset CER 은 약 0.18이었다.

2.1.4 음성 명령어 분류

※ This project was cooperated with (주)미래시티글로벌.
1. Principal Researcher, Hanyang University ERICA, Ansan, Korea (dev.taehyeon.kim@gmail.com)
2. Undergraduate Student, Hanyang University ERICA, Ansan, Korea (anayana9988@hanyang.ac.kr)
3. Undergraduate Student, Hanyang University ERICA, Ansan, Korea (namsyong0508@hanyang.ac.kr)
† Associate Professor, Corresponding author: Department of Artificial Intelligence, Hanyang University ERICA, Ansan, Korea (yongjaeyoo@hanyang.ac.kr)

STT 모델의 추론 결과인 IPA를 바탕으로 설정한 명령어와의 발음 유사도를 계산하여 명령어를 분류한다. 레벤슈타인 거리를 활용하여 자음 조음 위치 및 자음 조음 방법에 대한 Deletion 연산과 이외의 조음 위치, 조음 방법, 조음 강도, 유성음 여부, 입술 모양에 대하여 Substitution 연산을 진행한다. 그 이후에, 식 (1)을 적용하여 유사도를 계산한다. 해당 유사도에 대하여 Threshold를 기준으로 명령어에 대한 분류를 진행할 수 있다. 테스트 데이터셋에 대하여 임의의 Threshold(=0.3)를 설정하여 실험을 진행한 결과, Accuracy는 약 90%이었고, F1-Score는 약 90%이었다. ROC Curve 그래프는 아래 [Fig. 2]와 같았다.

$$(Similarity) = 1 - (Levenshtein\ Distance\ /\ len(cmd)) \quad (1)$$



[Fig. 2] 음성 명령어 분류에 대한 OvR ROC Curve 그래프

2.2 제스처 인식 시스템

2.2.1 라이브러리(mmaction2) 및 사용 모델(TSN) 소개

mmaction2 는 OpenMMLab 팀의 라이브러리이다. 해당 라이브러리는 Action Recognition에 특화되고, DDP(Distributed Data Parallel)로 구현되어있으며, 모듈화가 잘 되어있다는 점에 적용하였다. 또한, 지원되는 모델 중에서 TSN(Temporal Segment Networks) 모델을 이용하였다.

2.2.1 제스처 인식 모델

TSN 모델을 바탕으로 AIHub 의 '교통 수신호 데이터셋[2]'을 학습하였다. '교통 수신호 데이터셋'은 교통 수신호 중 6가지 동작에 대하여 자율주행차량의 실제 기상환경 및 시간대를 고려하여 다양한 촬영환경을 조성하고 촬영방법을 달리하여 구축한 교통 수신호 패턴 영상 데이터이다. 수신호를 이용하고, 로봇과 관리자 사이의 거리를 고려하였을 때 블랙박스 정도의 거리라는 점에서 학습 데이터로 적절하였다. 따라서, [Table 1]과 같은 라벨링 형식으로 구성한 뒤에 학습을 진행하였다. 그 결과, 정확도가 약 80%이었다.

2.3 멀티모달 시스템

[Table 1] 교통 수신호 데이터셋 라벨링 구성

우측에서 좌측으로	go (오른손들기)
좌측에서 우측으로	back (왼손들기)
전방정지	stop (전방으로 손들기)
후방정지	no operation
좌우측방 동시정지	goback (양손들기)
전후방 동시정지	no operation

2.3.1 알고리즘 설명

멀티모달 시스템에서는 앞서 나온 음성인식 시스템과 제스처 인식 시스템의 결과를 취합하여 최종 결정을 내리게 된다. 하지만, 상황에 따라 각 모델의 결과가 다르게 나올 수 있기에 환경에 따른 Threshold 값 설정이 가능하도록 Adaptive Threshold를 적용하였다. 각 모델에 대한 값(식 (2), 식 (3))을 큐(Queue)를 통해 받고, 나온 결과(식 (4))가 Adaptive Threshold(식 (5))보다 높으면 로봇에게 ROS를 통해 명령을 내리는 방식으로 구현을 하였다. 또한, 음성인식 추론 시간과 제스처 인식 추론 시간이 다른 경우, 이전에 추론된 데이터를 바탕으로 판단할 수 있게 되어 해당 경우에 decay factor(=0.7)를 곱해주는 방식으로 해결하였다.

$$(speech) = (go, stop, back, goback) \quad (2)$$

$$(gesture) = (go, stop, back, goback) \quad (3)$$

$$(prob) = (speech * decay + gesture * decay) / 2 \quad (4)$$

$$T = Mean(prob[-60:]) + Stdev(prob[-60:])$$

$$(Adaptive\ Threshold) = max(min(T, 0.8), 0.2) \quad (5)$$

3. 결론

3.1 인공지능 인터페이스 적용 효과

위에서 제시한 방법을 바탕으로 주변 소음 및 저조도 환경에서의 인공지능 인터페이스 적용에 대한 가능성을 보여줄 수 있었다.

References

[1] AIHub, "극한 소음 음성인식 데이터," [Online], https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71417, Accessed: June 26, 2023.

[2] AIHub, "교통 수신호 데이터셋," [Online], https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=496, Accessed: June 26, 2023.

[3] stannam, "hangul_to_ipa," [Online], https://github.com/stannam/hangul_to_ipa, Accessed: Jul 2, 2023.