청각장애인의 의사소통 접근성 향상을 위하여 AI 를 활용한 애플리케이션 개발

양세진 [1], 오소연 [1], 전보경 [1], 하원호 [1], 이지유 [2]
[1] 한국폴리텍대학교 성남캠퍼스 인공지능소프트웨어과 학생
[2] 한국폴리텍대학교 성남캠퍼스 인공지능소프트웨어과 교수

iamjiyoo@kopo.ac.kr

Improving Communication Access for the Hearing Impaired through AI

SeJin Yang [1], SoYeon Oh[1], BoKyeong Jeon[1], WonHo Ha[1], JiYoo Lee[2]
[1]Dept. of, Artificial Intelligence Software, Korea Polytechnics
[21]Dept. of, Artificial Intelligence Software, Korea Polytechnics Professor

요 약

본 논문은 청각장애인들이 음성 언어 중심의 사회에서 겪는 의사소통 제한을 해소하기 위해 제안되었다. 이를 위해 수어 인식, 입모양 탐지, 음성 인식 모델을 통해 추출된 데이터를 텍스트로 변환하여 접근성을 향상시키는 시스템을 개발하였다. 이 시스템은 청각장애인의 원활한 의사소통을 지원하고 사회적 참여를 증진하는 것을 목표로 한다.

1. 서론

의사소통에 자유롭고 원활하게 참여할 수 있는 권리를 누리는 것은 한 개인의 사회 통합에 필수불가결한 전제조건이다[1]. 그러나 많은 수의 청각장애인이 청력의 기능적 손상으로 인해 음성 언어 중심의 사회에서 의사소통에 제한을 경험하고 일상생활, 교육, 사법, 의료, 문화, 여가 등 다양한 영역에서 차별을 경험한다. 따라서 청각장애인들의 의사소통에 대한 접근성과 편의성을 향상시키고 이들이 사회적으로 더 많은 참여와 활동을 할 수 있는 환경을 조성하는데 기여하기 위해 다양한 의사소통 방법을 제시하고자 한다.

2. 모델연구

본 논문에서는 개인의 선호에 따라 의사소통 방법을 선택할 수 있도록 음성 인식, 입모양 탐지, 수어 번역의 세 가지 주요 기능을 제안한다. 각 기능 구현에 사용된 기술과 모델, 데이터세트는 다음과 같다.

① 음성 인식

음성을 텍스트로 변환하는 데에는 OpenAI 의 Whisper 모델을 사용하였다. Whisper 는 68 만 시간 이상의 다언어 음성 데이터를 기반으로 학습된 개방형 음성 인식 모델로, 음성 인식뿐만 아니라 음성 번역, 음성 활동 감지(VAD), 타임스탬프 예측, 언어 식별 등 다양한 작업을 동시에 처리할 수 있는 다중 작업(multitask) 및 다국어(multilingual) 지원 기능을 갖추고 있다[2]. Whisper 의 제로샷(Zero-shot) 성능을 통해 다양한 언어의 음성 데이터를 효과적으로 처리하며, 음성을 텍스트로 변환하고 자막 생성 등의 작업을 효율적으로 수행한다.

② 입모양 탐지

입모양 탐지를 위한 딥러닝 아키텍처는 LSTM, 완전 합성곱(Fully Convolutional), 트랜스포머(transformer) 세 가지 모델을 비교 분석하였다[3]. 트랜스포머 모델이 긴 문장 해석에 유리한 성능을 보인다는 연구 결과에 따라 이를 선택하였다. 입모양 탐지에 필요한 학습 데이터세트는 AI HUB 의 립리딩 음성인식 데이터세트(OLKAVS)를 사용하여 입모양을 효과적으로 추출하고 이를 텍스트로 변환하였다[4].

③ 수어 번역

수어 번역은 seq2seq 모델을 사용하여 구현하였다[5]. KETI 수어 데이터셋을 활용하였으며, 이 데이터셋은 11578 개의 고해상도 영상으로 구성되어 있으며 419 개의 단어와 105 개의 문장을 포함하고 있다. 수어 영상에서 얼굴, 손 등 신체부위의 키포인트를 OpenPose 를 통해 추출하고, 이를 객체 2D

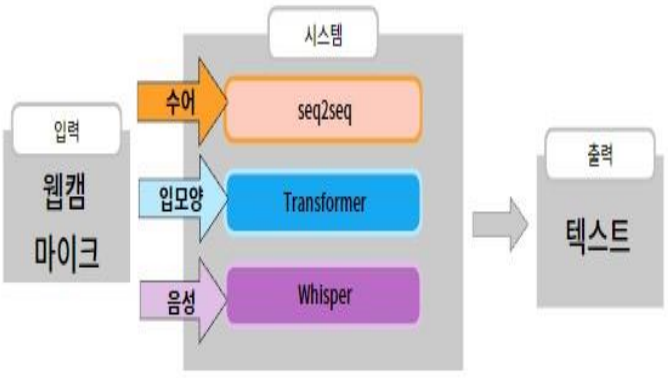
정규화하여 수어를 자연어 문장으로 번역하는 과정을 수행하였다.

3. 시스템 구성

첫번째, AI-HUB 에서 제공하는 한국어 수어 데이터셋을 이용하여 수어 인식과 텍스트 변환을 한다. 사용자가 수어 버튼을 클릭 시 수어 화면으로 이동한다. 웹캠, 모바일 카메라를 통해 수어 손동작을 인식하고 럽리딩 후 실시간으로 해당 단어가 무엇인지 텍스트 자막으로 출력한다.

두번째, 동영상에서 입모양 부분을 추출하기 위해서 동영상의 프레임을 하나씩 이미지로 변환하여 화면에서 얼굴을 찾아 좌표를 배열에 저장한다. 얼굴에 랜드마크를 표시할 모델을 적용하여 입 영역을 추출한다. 이후 해당 동영상을 저장하는 방식으로 전처리를 하여 동영상에서 입모양만을 추출한 뒤 해당 입모양을 번역하고 텍스트로 변환하여 출력한다.

세번째, 사용자가 구현된 애플리케이션에 마이크로 음성을 입력하면 Whisper 모델을 사용하여 해당 음성을 텍스트로 변환 후, 이를 화면에 출력한다.



<그림1> 시스템 구조

<그림 2> Open Pose 를 사용하여 수어 인식

4. 결론

본 논문은 청각장애인들이 음성 언어 중심 사회에서 겪는 의사소통의 제약과 차별을 해결하기 위해 수어 인식, 입모양 탐지, 음성 인식을 통합한 솔루션을 제안한다. 이를 통해 청각장애인들이 일상생활, 교육, 의료 등 다양한 사회적 영역에서 더욱 자유롭고 원활하게 소통할 수 있는 환경을 제공하는 것을 목표로 한다. 현재 수어 데이터의 부족으로 실사용에 제약이 있지만, 우리는 이를 극복하기 위해 직접 데이터셋을 구축하여 시스템을 고도화할 계획이다. 향후 시스템 성능을 지속적으로 개선하고 사용자 경험을 향상시킴으로써 청각장애인의 사회적 참여와 통합을 확대하는 데 기여할 것이다. 궁극적으로, 청각장애인이 평등하게 소통할 수 있는 사회적 기반을 마련하는 데 중요한 역할을 할 것으로 기대된다.

참고문헌

[1] 정은. (2002). 중증장애인의 의사소통에 대한 이해. 특수교육학연구, 37(3), 75-94

[2] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision”. arXiv:2212.04356, 2022

[3] Triantafyllos Afouras, Joon Son Chung, Andrew Zisserman, "Deep Lip Reading: a comparison of models and an online application", Visual Geometry Group, Department of Engineering Science, University of Oxford, UK, arXiv:1806.06053, 2018

[4] Jeongkyun Park, Jung-Wook Hwang, Kwanghee Choi, Seung-Hyeon Lee, Jun Hwan Ahn, Rae-Hong Park, Hyung-Min Park, "OLKAVS: An Open Large-Scale Korean Audio-Visual Speech Dataset", ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 10.1109/ICASSP48485.2024.10446901, 2024

[5] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, Choongsang Cho, “Neural Sign Language Translation based on Human Keypoint Estimation”: Korea Electronics Technology Institute 22 Dawangpangyo-ro 712 beon-gil, Seongnam-Si, Gyeonggi-do 13488, South Korea

※ 본 논문은 과학기술정보통신부 대학디지털교육 역량강화 사업의 지원을 통해 수행한 ICT 멘토링 프로젝트 결과물입니다.