보완대체의사소통 시스템 고도화를 위한 대형멀티모달모델 개발

김철희[1,2], 지승연[1], 김지원[1], 김채희[1], 임승찬[3], 한경림[1,4]*

[1]한국과학기술연구원 뇌과학연구소 뇌융합기술연구단, [2]고려대학교 컴퓨터학과, [3]에어패스,
[4]국가연구소대학교 KIST스쿨 바이오-메디컬 융합전공

setg1502@kist.re.kr, seungyeon0510@gmail.com, jiwon23@kist.re.kr, gimchaehui342@gmail.com,
vrsports@airpass.co.kr, *khan@kist.re.kr

Development of Large Multimodal Models for Advancing Augmentative Alternative Communication Systems

Cheolhee Kim[1,2], Seungyeon Ji[1], Jiwon Kim[1], Chaehee Kim[1], Sungchan Lim[3],
Kyungreem Han[1,4]*

[1]Center for Brain Technology, Brain Science Institute, Korea Institute of Science and Technology, Seoul 02792, Korea, [2]Department of Computer Science and Engineering, Korea University, Seoul 02841, Korea, [3]Airpass, Gyeonggi-do 12082, Korea, [4]Division of Bio-Medical Science & Technology, KIST School, Korea National University of Science and Technology, Seoul 02792, Korea

요 약

언어적 의사소통이 어려운 장애인은 보완대체의사소통(AAC) 시스템을 사용하여 제한적이나마 의사소통에 도움을 받을 수 있지만, 몸을 가누기조차 힘든 중증 장애인은 AAC을 직접 조작하기가 거의 불가능하다. 본 연구에서는 중증장애인 의사소통 보조를 위한 몸동작 및 음성 모달리티 기반 대형멀티모달모델(LMM)을 개발하였다. 개발한 LMM은 "네/아니요"로 대답하는 닫힌 질문에 대한 몸동작과 음성 반응을 딥러닝 모델로 분류한 후, 결과를 "네/아니요" 텍스트/음성으로 정확히 변환할 수 있다. 나아가 "네/아니요" 텍스트와 닫힌 질문을 한국어 대형언어모델(LLM)로 결합하여 최종 복합 의사 표현 텍스트를 변환할 수 있다. "네/아니요" 의미 분류 딥러닝 모델은 몸동작과 음성 모달리티에 대해 모두 90% 이상의 분류 정확도를 획득하였고, 올바른 복합 의사 표현 텍스트를 출력할 수 있었다. 비장애인의 몸동작과 음성 데이터 세트를 이용해 1차 전이 학습한 후, 중증 장애인 데이터를 사용하여 2차 전이학습을 수행하여 장애인 개별 맞춤 대형멀티모달모델 개발 플랫폼을 구축하였다. 본 연구에서 제시하는 의사소통 보조 LMM은 기존의 보완대체의사소통 시스템의 고도화를 통하여 중증 장애인과 비장애인 간의 의사소통 장벽을 낮출 중요한 수단을 제공할 것이다.

Ⅰ. 서 론

의사소통 보조를 위해 장애인 보완대체의사소통(Augmentative Alternative Communication, AAC) 시스템이 개발되고 있지만, 몸을 가누기 힘든 중증 장애인은 AAC 시스템을 직접 조작하기 어려워 큰 도움이 되지 않는다. 또한, 기존의 AAC는 그림 카드 등을 이용해 간접적으로 표현을 지정하는 방식으로, 개인의 언어 표현이나 몸동작과 같은 비언어적 모달리티를 직접 활용하지 못하는 한계가 있다.[1] 이 논문에서는 중증 장애인이 음성 및 몸동작을 통해 생활 필수 의사소통이 가능하도록 보조하는 대형멀티모달모델 (Large Multimodal Model, LMM) 기반 AAC 기술을 제안한다.

Ⅱ. 의사소통 보조 LMM 시스템

의사소통 보조 LMM 시스템의 전체 구조는 그림 1과 같다. LMM 시스템은 몸동작과 음성 형태의 모달리티에서 ResNet을 이용해 특징을 추출하고, 특징을 분석해 긍정/부정의 의미를 지니는지 분류한다. 그리고, 긍정/부정의 의미를 텍스트 "네/아니요"로 변환한 후 질문 텍스트와 함께 대형언어모델(Large Language Model, LLM)에 입력하여 자연스러운 답변으로 바꾸어 중증 장애인의 의사소통이 원활하게 이루어지도록 한다.

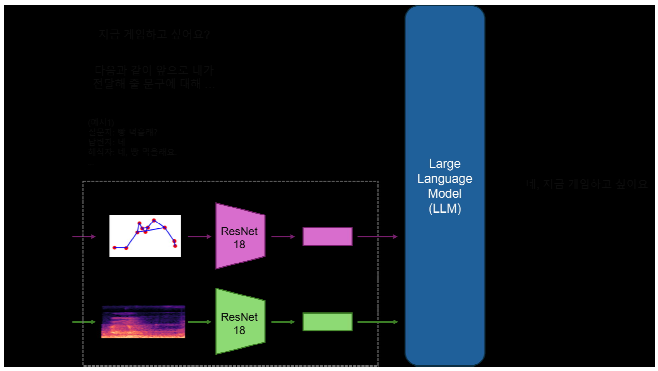

그림 1. 의사소통 보조 LMM의 구조

Ⅲ. 모달리티 특징 추출

학습 데이터는 사전학습을 위한 비장애인의 영상과 그림 2의 왼쪽 그림

과 같이 전이 학습을 위한 중증 장애인 영상을 수집해 사용하였다. 학습 데이터 영상은 “네/아니요”로 대답할 수 있는 질문에 대한 대답 영상으로, 답변 시의 몸동작과 음성 반응을 담고 있다. 질문의 경우, 그림 2의 오른쪽 그림과 같이 장애인이 생활하는 데 필요한 리스트를 작성하여 문답을 구성하였다. 수집한 영상으로부터 신체 관절점 및 음성 데이터를 추출하였다.

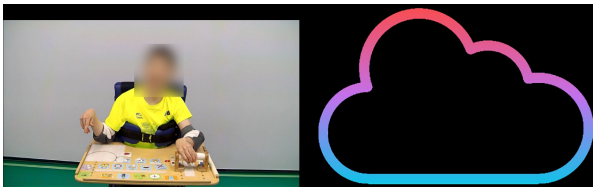

그림 2. 중증 장애인 데이터(왼쪽) 생활 필수 질문 리스트(오른쪽)

1. 모달리티 데이터 추출

신체 관절점 추출 YOLO-Pose 모델[2]을 사용해 인물의 신체 관절점 정보를 추출한다. 영상은 정면에서 촬영된 한 명의 인물을 대상으로 하며, COCO 데이터 세트 형식에 따라 각 프레임에서 17개의 관절점 x, y 위치 정보를 얻는다. 단, 책상이 촬영 대상의 하반신을 가리기 때문에 상반신의 11개 관절점만을 추출해 사용한다. 상반신 11개의 관절점은 $(x_1, y_1, x_2, y_2, \cdots, x_{11}, y_{11})$ 위치 정보가 영상 전체 프레임 수만큼 구성된 2차원 배열로 저장된다.

음성 데이터 추출 긍정과 부정 음성의 차이를 분석하기 위해 FFT(고속 푸리에 변환)를 사용해 음성 신호를 주파수 성분으로 변환한다. 주파수 분석 결과, 긍정 표현은 주파수 성분이 모여 있고, 부정 표현은 주파수 성분이 분산된 양상을 보인다. 긍정/부정 표현 간의 주파수 차이를 CNN 모델에서 음성 특징으로 추출하기 위해 음성 신호를 멜 스펙트로그램(Mel Spectrogram)으로 변환한다.

2. 모달리티 데이터 증강

신체 관절점 증강 데이터 사이의 레이블 불균형을 해소하고, 크게 진동하며 수렴하는 불안정한 학습 과정을 줄이기 위해 데이터 증강을 진행한다. 데이터 증강 방법에는 그림 3과 같이 가우시안 노이즈 및 가우시안 블러(Gaussian noise with Gaussian blur)과 무작위 각도로 회전(random rotation)을 순차적으로 적용한다.

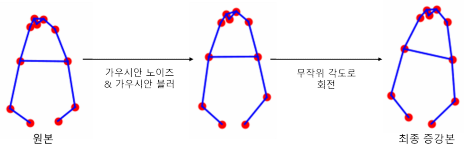

그림 3. 신체 관절점 데이터 증강

음성 데이터 증강 음성 데이터 또한 불안정한 학습을 방지하기 위해 데이터 증강을 적용한다. 증강 방법으로 속도 변경, 피치 변경, 화이트 노이즈 추가를 적용한다. 하나의 음성파일 당 5개의 증강 데이터를 생성하여 최종적으로 196개의 원본 음성 데이터와 980개의 증강 음성 데이터를 학습 데이터로 확보한다. 테스트 데이터 세트는 84개의 원본 음성 데이터를 사용한다.

3. 모달리티 분류 학습

영상 속 인물의 몸동작과 음성을 이용해 “네/아니요”의 긍정/부정 의사 표현을 구분하는 분류 학습을 진행한다.

신체 관절점 분류 학습 진행 전 관절점 데이터는 크기 조정 및 정규화 전처리 과정을 거친다. 분류 모델은 CNN 모델 중 ResNet18을 사용하며, 배치 크기는 16, 에포크는 150, 학습률은 1e-4로 학습을 진행한다. 학습을 수행한 결과, 그림 4와 같이 학습 초반에 손실과 정확도 모두 빠르게 수렴하며 테스트 데이터 세트에 대해 99%의 정확도를 기록한다.

음성 데이터 분류 학습 전 데이터 전처리를 통해 z-score 표준화를 적용하고 데이터의 feature 크기는 128로 설정하고, 데이터의 길이는 모든 음성 데이터의 시간 프레임 평균으로 맞추어 조정한다. 분류 모델은 ResNet18을 사용하며, 배치 크기는 16, 에포크는 150, 학습률은 1e-4로 학습을 진행한다. 학습 결과는 그림 5와 같다. 그림 4에서 확인할 수 있듯이 테스트 데이터 세트에 대해 정확도가 94%를 기록하며, 학습 초반부터 전체적으로 정확도가 매우 높은 것을 확인할 수 있다.

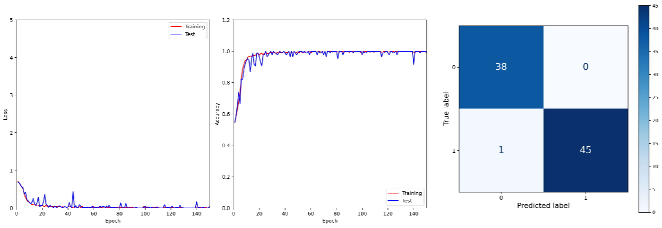

그림 4. 신체 관절점 분류 결과

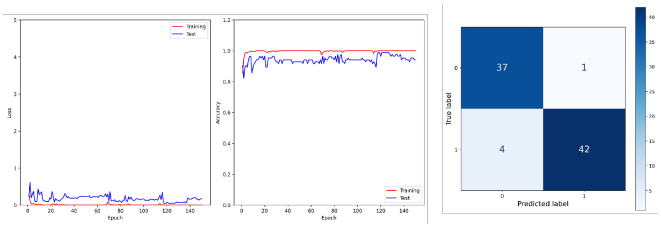

그림 5. 음성 분류 결과

중증 장애인 데이터 전이 학습 사전학습을 진행한 신체 관절점 분류 모델과 음성 분류 모델에 중증 장애인 데이터를 적용해 전이 학습을 진행한다. 중증 장애인의 몸동작 및 음성 데이터는 비장애인에 적용한 것과 같은 증강 방식을 적용한다. 증강한 데이터를 이용하여 신체 관절점 분류 모델과 음성 분류 모델 모두 전이 학습을 진행하며, 배치 크기 16, 에포크 150, 학습률 1e-4로 학습을 진행한다. 학습 결과는 다음 표 1과 같다. 음성 데이터의 경우, 재현율과 F1-Score가 정밀도보다 낮은 것을 확인할 수 있다. 낮은 재현율과 F1-Score의 원인은 “아니요” 응답 데이터의 부족으로 인해 레이블 간 균형이 맞지 않았기 때문이다. 또한, ‘네’ 응답은 또렷한 음성이 있지만, “아니요” 응답 데이터는 음성을 발현하지 않아 구분되는 특징을 얻기 어려워 재현율과 F1-Score가 낮게 나옴을 알 수 있다.

	신체 관절점 분류	음성 데이터 분류
정밀도	88%	90%
재현율	80%	53%
F1-Score	84%	50%

표 1. 중증 장애인 데이터 전이 학습 결과

Ⅳ. LLM 중심 모달리티 통합

LLM을 이용하여 몸동작과 음성 모달리티를 통합해 질문에 맞는 대답 문장을 생성할 수 있도록 한다. 몸동작과 음성 모달리티 정보는 분류 학습을 통해 확인한 긍정/부정 결과를 “네/아니요” 텍스트로 변환해 이용한다. LLM은 LLaMA 3의 8B instructed 버전을 한국어로 파인튜닝을 진행한

Bllossom-8B 모델[3]을 사용한다.

그림 6과 같이 “네/아니요”로 대답하는 닫힌 질문 텍스트, 의사 표현 텍스트를 Bllossom-8B 모델의 입력으로 주어 대답 문장을 생성할 수 있도록 한다. 또한, 퓨샷(few-shot) 데이터를 함께 입력하여 인컨텍스트러닝(in-context learning)을 진행한다. 모델의 출력 텍스트를 분석한 결과, 파라미터의 미세조정과 다량의 학습 데이터 없이도 질문에 맞는 대답 문장을 구성할 수 있음을 확인하였다.

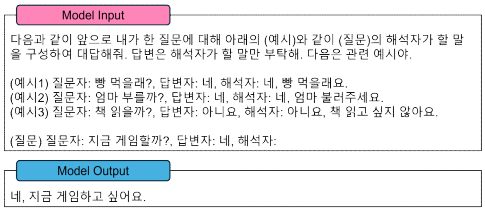

그림 6. Bllossom-8B 입력 프롬프트와 출력 문장

V. 요약 및 결론

본 연구에서는 비언어적인 몸동작 및 음성 모달리티와 LLM을 이용한 LMM 기반 장애인 개별 맞춤 대형멀티모달모델 개발 플랫폼을 구축하였다. 몸동작과 음성은 ResNet 기반의 모델로 1차 분석한 후, 몸동작과 음성 모달리티마다 분류된 “네/아니요” 대답 텍스트와 닫힌 질문 텍스트를 LLM에 입력하여 자연스러운 복합 텍스트를 생성하는 방법을 제시하였다. 본 연구에서 제시하는 의사소통 보조 LMM은 기존의 보완대체의사소통 시스템의 고도화를 통하여 중증 장애인이 겪는 의사소통 어려움 해소에 크게 기여할 수 있다. 후속 연구에서는 정확한 생활 필수 의사표현 기능에 더하여 개인의 취미, 예술, 사회활동을 보조할 수 있는 감정 표현 및 인식이 가능한 시스템으로 발전시키고자 한다.

ACKNOWLEDGMENT

과학기술정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 개발 중인 '장애인 소통 지원 초거대 AI 멀티모달 기반 서비스 개발'을 활용하여 수행한 연구입니다.

참 고 문 헌

[1] Light, J. et al., “Challenges and opportunities in augmentative and alternative communication: Research and technology development to enhance communication and participation for individuals with complex communication needs,” Augmentative and Alternative Communication, 35(1), 1-12, Jan. 2019.

[2] Maji, D. et al., “Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2637-2646, 2022.

[3] Choi, C., Jeong, Y., Park, S. et al., “Optimizing Language Augmentation for Multilingual Large Language Models: A Case Study on Korean,” arXiv preprint, arXiv:2403.10882., Mar. 2024.