딥러닝을 활용한 실시간 Lip-Reading 시스템 설계

정도윤*, 오연재**, 김남호*
*호남대학교 컴퓨터공학과, **배재대학교 게임공학과
*rabbitsun2@gmail.com, **oksug10@naver.com, *nhkim@honam.ac.kr

Real-time Lip-Reading system design using deep learning

Doyoon Jung*, Yeonjae Oh**, Namho Kim*
*Dept. of Computer Engineering, Honam Univ., **Dept. of Game Engineering, PaiChai Univ.

요 약

본 연구는 청각 장애인, 소음이 심한 환경에서의 의사소통, 그리고 비밀스러운 대화가 필요한 상황 등을 대상으로 하는 독순술(Lip-reading) 시스템을 설계하였다. 이러한 시스템은 입술 독해를 통한 의사소통의 효율성과 정확성을 향상하는 것을 목표로 한다. 본 연구에서는 인간의 입술 움직임을 실시간으로 분석하고, 이를 통해 발화되는 문장이나 단어를 정확하게 인식할 수 있는 시스템을 개발하였다. 이를 위해 "Visual Speech Recognition for Multiple Languages"라는 오픈소스 프로젝트를 활용했을 때 기존의 텍스트 중심 처리 과정을 영상처리 과정을 포함할 수 있도록 개선하였다. 또한, MediaPipe와 같은 오픈소스 라이브러리를 이용하여 입술의 다양한 움직임을 정확하게 추적하였으며, 독순술 시스템이 실시간으로 처리할 때 지연시간으로 인한 동영상의 연속성에 대한 문제를 해결할 방법을 제시하고자 한다.

Ⅰ. 서 론

본 논문은 인공지능(AI) 기술을 활용하여 실시간 독순술(Lip-reading) 시스템을 설계하였다. 독순술은 사람의 입술 움직임을 관찰함으로써 발화되는 단어나 문장을 이해하는 과정을 말한다. 이러한 기술은 특히 청각 장애인이나 소음이 심한 환경에서의 의사소통, 비밀스러운 대화가 필요한 상황에서 유용하다. 그러나 기존의 독순술 기술은 주로 수동적인 관찰에 의존하며, 이에 따라 의사소통의 효율성과 정확성이 제한됐다.

최근 AI 기술의 발전은 비언어적 의사소통 방식을 자동화하고 개선할 새로운 기회를 제시하였다.

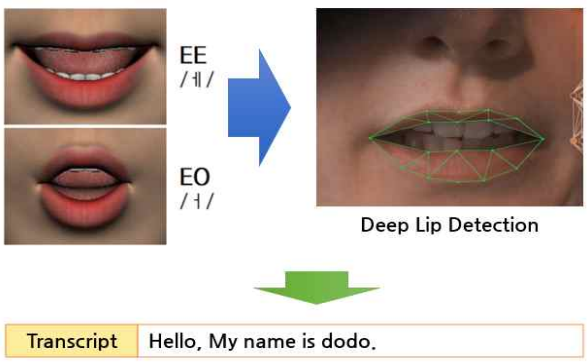

그림 1 독순술 시스템의 원리

본 연구에서는 AI 기반의 실시간 독순술 시스템을 통해 이러한 한계를 극복하고자 한다[1]. 특히, 딥러닝과 영상처리 기술을 활용하여 인간의 입술 움직임을 정확하게 추적하고, 이를 기반으로 발화된 단어나 문장을 실시간으로 인식하는 시스템을 선행 연구로서 공개된 오픈소스 프로젝트 "Visual Speech Recognition"를 활용하여 개발하였다[2][4][5].

본 연구는 청각 장애인의 의사소통 능력 향상뿐만 아니라, 일상생활 속 다양한 상황에서의 의사소통 효율성과 정확성을 높이는 데 기여할 것으로 기대된다.

Ⅱ. 본론

선행 연구에서는 그림 2와 같이 얼굴에 센서를 부착하여 sEMG 파형을 CNN으로 학습시켜 독순술을 구현한 사례가 있다[3]. sEMG(표면 근전도)란 근육의 움직임을 위한 전기 신호를 정량적으로 측정하기 위한 근전도 진단기기의 일종이라고 할 수 있다. 또한, 영상처리와 CNN 그리고 LSTM(RNN)을 활용해 독순술 일부를 구현한 연구가 있었다[1][9]. 그러나 이러한 연구의 한계점은 독순술 시스템을 구현할 때, 언어체계의 정확한 문법 구조, 어순, 문맥 등을 해석할 수 있는 언어전문가가 필요하다는 점이다.

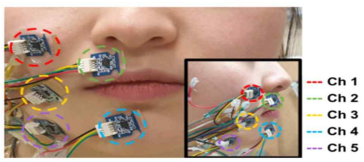

그림 2 sEMG electrodes를 활용한 독순술 구현 사례

본 연구에서 그림 3은 독순술 시스템을 구현하는 데 있어서 표준화되지 않은 개발 방법을 대신하여 기존에 지속해서 연구 중인 프로젝트를 활용하여 실시간 독순술 시스템을 설계하였다. 인공지능 모델 개발의 시간을 줄이고 데이터 세트와 다국어가 지원될 수 있는지 가능성에 초점을 두고 실험을 진행하였다.

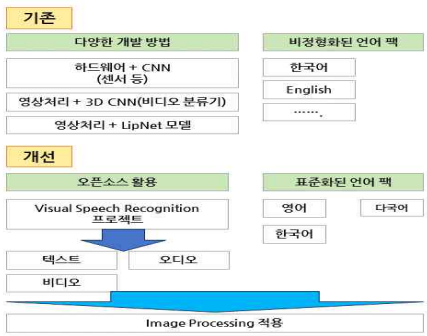

그림 3 실시간 독순술 시스템 설계 - 프로그램 구조

본 연구에서 적용한 LRS3(Lip-Reading-Sentences 3) 모델은 옥스퍼드 대학(Oxford University)에서 만든 언어 모델로서 테드(TED)와 테드엑스(TEDx) 강연에서 수집한 음성을 모아둔 데이터 세트이다.[4].

본 연구에서는 기존 프로젝트를 OpenCV와 연동하여 동영상을 실험하였다. 그림 4는 동영상 클립을 통해 독순술 시스템으로 스크립트를 추출한 모습이다.

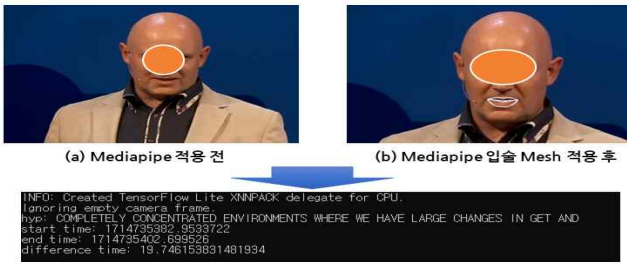

그림 4 독순술 시스템 구현 - 단일 영상 동작 모습

그림 4의 start time과 end time의 차이 시간을 살펴보면, 약 19.74초의 지연 시간이 발생한다는 것을 알 수 있다.

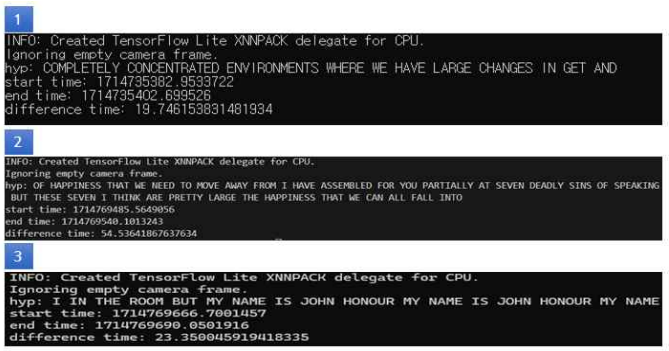

그림 5 독순술 시스템 구현 - 클립 3개 동작 모습

그림 5는 독순술 시스템에 클립 3개를 각각 돌렸을 때 발생한 지연 시간이다. 1번은 영상 길이가 7초, 2번은 15초, 3번은 10초이며 학습을 시도한 모습이다. TED 관련 영상은 1번과 2번이다. 3번은 웹 카메라로 직접 촬영하였다. 지연 시간은 각각 19.74초, 54.53초, 23.35초가 나왔다.

이는 TensorFlow Lite를 이용하여 CNN과 RNN 같은 딥러닝 처리를 수행할 때 동영상 내용을 처리하는 데 계산하는 시간이 필요하다[4][9]. 본 연구와 관련하여 한국어에 대한 시각 음성인식(VSR) 연구로는 ETRI에서 개발한 KMSAV 모델과 ICASSP 2024 회의에서 발표된 OLKAVS 모델이 있다. KMSAV 모델은 150시간 분량의 수동 전사 데이터를 기반으로 하여, 자동음성인식(ASR)과 시각 음성인식(AVSR)에서 각각 11.1%, 18.9%의 오류율을 기록했다. 여기서 ASR은 음성신호를 텍스트로 변환하는 기술이며, AVSR은 영상 속 입모양을 통해 음성을 인식하는 기술을 의미한다. OLKAVS 모델은 1,150시간의 녹음 데이터와 1,107명의 한국어 사용자로부터 수집한 오디오 및 다양한 소음 환경에서의 음성 데이터를 포함하는 데이터 세트를 구축했다[6][7].

OLKAVS 모델의 경우, AI-HUB에 립리딩(입모양) 음성인식 데이터를 19.72TB의 데이터 세트를 사용하였다[7][8]. 딥러닝 기반의 립리딩 시스템을 사용하려면 현재에는 고사양의 GPU와 저장공간을 가진 시스템이 요구된다. 또한 한국어 데이터 세트 학습 모델을 구축하는 데 있어서 mp4 파일과 녹음 파일 등을 학습하는 데 많은 컴퓨팅 성능이 필요하다는 것을 알 수 있다.

Ⅲ. 결론

본 논문에서는 독립적인 립리딩 시스템 개발뿐만 아니라, 오픈소스를 활용한 립리딩 시스템의 다양한 가능성을 탐구하였다. 고유한 인공지능 학습 모델 개발의 기회와 더불어 언어 전문가의 필요성과 같은 일부 과제들도 발견되었다. 또한, 본 연구를 통해 실시간 동영상 처리와 TensorFlow Lite를 사용시 지연 시간 등의 문제도 발견되었다.

한국어 립리딩 데이터 세트인 KMSAV와 OLKAVS 모델 분석을 통해 이 두 모델이 표준화되지 않았음을 확인할 수 있었고, 학습 모델을 생성하는데 있어서 상당한 컴퓨팅 자원을 요구함을 알 수 있었다. 이에 따라, 앞으로 VSR 분야에 대한 한국어 자연어 처리와 동영상 기반 식별 시스템에 대한 깊이 있는 연구를 진행할 계획이다.

ACKNOWLEDGMENT

"본 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 지역지능화혁신인재양성사업(IITP-2024-00156287, 100%)의 연구결과로 수행되었음".

참 고 문 헌

[1] 임대영, 김선광, & 정길도, "영어발음 향상을 위한 실시간 인공지능 입모양 인식 프로그램 개발," 제어로봇시스템학회 논문지, 24(4), 327-333, 2018.

[2] Zhu, X., & Ramanan, D. "Face detection, pose estimation, and landmark localization in the wild," In 2012 IEEE conference on computer vision and pattern recognition pp. 2879-2886, June 2012.

[3] Kwon, J., Nam, H., Chae, Y., Lee, S., Kim, I. Y., & Im, C. H, "Novel three-axis accelerometer-based silent speech interface using deep neural network," Engineering Applications of Artificial Intelligence, 120, 105909, 2023.

[4] Ma, P., Petridis, S., & Pantic, M, "Visual speech recognition for multiple languages in the wild," Nature Machine Intelligence, 4(11), 930-939, 2022.

[5] Ma, P., "lip reader," 2023 (https://mpc001.github.io/lipreader.html).

[6] Park, K., Oh, C., & Dong, S, "KMSAV: Korean multi speaker 8spontaneous audiovisual dataset," ETRI Journal, 46(1), 71-81, 2024.

[7] Park, J., Hwang, J. W., Choi, K., Lee, S. H., Ahn, J. H., Park, R. H., & Park, H. M, "OLKAVS: an open large-scale Korean audio-visual speech dataset," In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6385-6389), April 2024.

[8] AI-HUB, "립리딩(입모양) 음성인식 데이터," 2022 (https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=538).

[9] Xing, G., Han, L., Zheng, Y., & Zhao, M., "Application of deep learning in Mandarin Chinese lip-reading recognition," EURASIP Journal on Wireless Communications and Networking, 2023(1), 90, 2023.