# 보완대체 의사소통 시스템 고도화를 위한 발화인식모델 개발

지승연[1], 김철희[1,2], 임승찬[3], 한경림[1,4]*

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

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

# Development of a speech recognition model for advancing augmentative and alternative communication systems

Seungyeon Ji[1], Cheolhee Kim[1,2], 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

대부분의 중증장애인 의사소통 보조기기는 음성 출력을 위해 장애인이 미리 저장된 전형적 이미지나 텍스트를 선택하는 방식으로 작동되기 때문에 정확한 의미 전달이 어렵고 시간적 지연이 발생한다. 음성인식 모델이 탑재된 인공지능 스피커를 이용하는 경우에도 발화가 불완전한 장애인의 음성인식 성공률은 현저히 떨어진다. 본 연구에서는 구음장애 환자의 음성 데이터를 활용하여 발화 예측 모델을 최적화하고, 이를 기반으로 중증장애인 보완대체 의사소통 시스템 고도화를 위한 발화 인식 모델을 개발하였다. 구음장애 음성인식 데이터를 대상으로 한 실험에서, MFCC와 VGG16을 함께 사용하였을 때 구음장애 환자의 발화가 93.41%로 가장 정확하게 분류됨을 확인하였고, 이러한 MFCC-VGG16를 기반으로 중증장애인 발화 인식 모델을 개발하였다. 중증장애인의 일상생활과 삶의 질에 필수적인 12가지 상황별 대규모 데이터를 수집/정제하고 이를 모델 개발과 검증에 활용하였다. 이러한 중증장애인 발화 인식 모델의 개발은 환자별 맞춤 발화 인식 모델 개발을 통하여 중증장애인의 의사소통 문제 해결에 크게 기여할 것이다.

## Ⅰ. 서 론

현재 보건복지부에서 지원하는 중증장애인을 위한 대면 의사소통 보조기기의 사용자 인터페이스는 대부분 사전에 환자의 말뭉치를 수집하여 제작된 애플리케이션에서 텍스트나 이미지를 손으로 선택하는 방식으로 동작한다. 그러나, 단순히 이미 저장된 몇 가지 전형적 텍스트나 이미지를 선택하는 방식은 정확한 의사 표현이 어렵고, 다른 사람과 의사소통하는 데 시간적 지연이 발생하며, 손이나 눈을 사용할 수 없는 상황에서는 사용이 불가능하다. 본 연구에서는 구음장애 환자의 발화를 예측하는 최적의 모델을 도출하여 음성을 통해 환자가 발화하고자 하는 내용을 정확하게 분류한다. 나아가, 구음장애 발화 인식 최적화 모델을 기반으로 중증장애인 보완대체 의사소통 시스템 고도화를 위한 장애인 개인 맞춤 발화 인식 모델을 개발하고자 한다.

## Ⅱ. 본론

본 연구에서는 구음장애 환자의 발화 분류 모델을 비교하여 최적의 모델을 도출하고 이를 활용하여 중증장애인 의사소통 보조를 위한 발화 인식 모델을 개발한다.

### 1. 데이터 세트

구음장애 환자의 음성 데이터로 AI-Hub에서 제공하는 뇌신경 장애 환자 216명의 702시간 음성 기록 데이터를 사용하였다.[1] 본 연구에서는 주로 언어+뇌신경 장애가 있는 한 명이 발화한 20개의 문장을 다루었다 (표 1).

| | | | |
|---|---|---|---|
| 1. 문 열어. | 6. 집에 아무도 없겠지. | 11. 날씨가 더워요. | 16. 내일 모레 날짜 알려줘. |
| 2. 내일 만나. | 7. 거기 갈래. | 12. 관련 뉴스 알려줘. | 17. 뉴스 기사 읽어줘. |
| 3. 머리 말려. | 8. 그렇게 될 리가 없지. | 13. 기본 알람음으로 변경해 줘. | 18. 메모 읽어줘. |
| 4. 여기 놔요. | 9. 전화 왔어요. | 14. 날씨 알려줘. | 19. 미국 국가번호 알려줘. |
| 5. 밥 먹었죠. | 10. 방이 추운 거 같애. | 15. 나비야 동요 들려 줘. | 20. 오늘 일정 삭제해줘. |

표 1. 구음장애 발화 인식 모델 개발에 활용한 음성인식 데이터

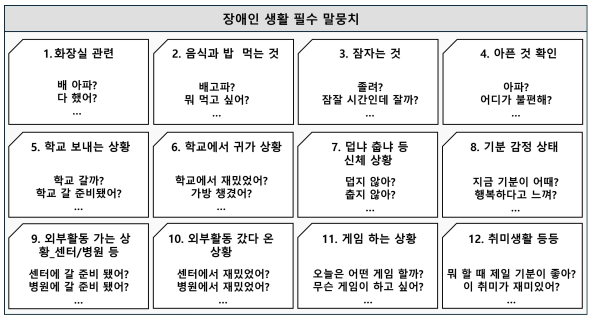


그림 1. 자체 수집한 장애인 생활 필수 말뭉치 모식도

중증장애인의 일상생활에 필수적인 12가지 상황별 대규모 데이터를 수집/

정제하고 이를 모델 개발과 검증에 활용하였다 (그림 1).

### 2. 음성 데이터 증강

구음장애 음성인식 데이터의 경우 각 문장에 대해 하나의 음성 데이터가 존재하기 때문에 원활한 학습을 위하여 데이터 증강이 필요하다. 또한, 발화가 담긴 음성 데이터를 활용하기 때문에 시간적 변형이 발생하지 않는 피치 조절, 속도 조절, 음량 조절, 화이트 노이즈 추가, 생활 환경 노이즈 추가 기법을 사용하였다.[2] 기존의 증강 방식은 다양성 측면에서 한계가 있기 때문에 생활 환경 노이즈를 추가하여 데이터 세트의 다양성을 높였다. 생활 환경 노이즈는 Freesound에서 강아지 소리, 발소리, 음악 소리, 거리 소리, 청소기 소리를 선정하였다.[3] 피치 조절은 -4, 속도 조절은 0.3, 음량 조절은 -20, 화이트 노이즈 추가는 0.015로 설정하였다. 이런 증강 기법을 최대 3개 조합하여 최종적으로 하나의 원본 데이터당 25개의 증강 데이터를 추출하였다.

### 3. 음성 특징 추출

Mel spectrogram은 음성 신호를 시간 및 주파수 영역에서 시각적으로 표현한 것이다. 일반적인 스펙트로그램의 주파수 축을 인간의 청각 특성에 맞춰 멜 스케일로 변환한 것으로 음성 특징을 효과적으로 추출한다.

MFCC(Mel-Frequency Cepstral Coefficients)는 음성 신호의 멜 스펙트럼에서 로그 변환과 이산 코사인 변환을 거쳐 추출된 특징으로 음성 특징을 저차원으로 압축하여 효율적인 분류가 가능하다.

### 4. 기본 발화 분류 모델

ResNet18은 잔차연결을 활용하여 기울기 소실 문제를 해결한 모델로, 깊지 않은 네트워크에서도 효과적으로 학습할 수 있으며 음성 데이터에서 다양한 주파수 특징을 탐지할 수 있다.

VGG16은 3x3 필터를 사용하는 합성곱 레이어를 쌓은 모델로, 음성의 세부적인 패턴과 변화를 포착할 수 있다.

Efficientnet은 모든 차원의 깊이, 너비, 해상도를 균형 있게 확장한 모델로 적은 파라미터로 높은 성능을 보인다.[4] 계산 비용을 줄여 복잡한 음성 데이터를 처리할 수 있다.

### 5. 구음장애 환자 발화 분류 모델 최적화

본 연구에서는 구음장애 환자의 발화를 분류하는 최적의 모델을 도출하기 위해 음성 특징 2가지와 분류 모델 3가지를 조합한 6가지 경우를 비교하였다. 모든 모델에서 에포크는 200, 학습률은 0.00001, 배치 사이즈는 16으로 20번 반복하여 평균 정확도를 추출했다. MFCC와 VGG16을 사용하는 경우, MFCC의 특성상 13개의 특징이 추출되어 입력 데이터의 크기가 작기 때문에 패딩을 추가하였다. 그림 2와 같이 MFCC와 VGG16을 사용하였을 때 평균 정확도가 93.41%로 가장 높게 나타났다. 표2는 MFCC+VGG16의 F1-score를 출력한 결과이며, 전체 6가지 경우의 모델 추론 정확도는 표3과 같다. 모델 추론 정확도란 테스트 데이터인 원본 데이터 104개에 대한 예측 정확도로, 모델의 성능 평가 지표로 활용하였다.

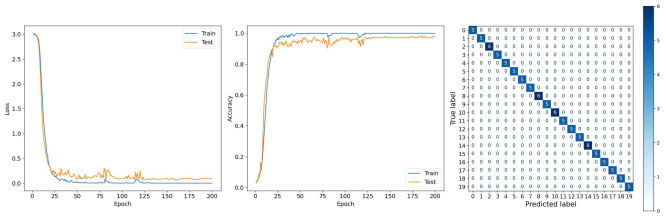


그림 2. MFCC+VGG16의 손실-정확도 그래프 및 혼동행렬

| 정밀도(Precision) | 0.9917 |
|---|---|
| 재현율(Recall) | 0.9900 |
| F1 score | 0.9899 |

표 2. MFCC+VGG16의 F1 score

| | Mel spectrogram | MFCC |
|---|---|---|
| ResNet18 | 78.32±4.11% | 70.91±4.43% |
| VGG16 | 87.16±4.24% | 93.41±3.07% |
| Efficientnet | 84.90±3.10% | 65.48±5.44% |

표 3. 모델 추론 정확도

### 6. 중증장애인 발화 인식 프로토타입 모델

구음장애 환자의 음성 데이터를 통해 도출한 최적의 모델인 MFCC+VGG16을 활용하여 중증장애인 발화 인식 모델을 개발하였다. 전체 구조는 그림3과 같다.

맞춤형 VGG16은 생활 필수 말뭉치를 발화한 중증장애인의 음성데이터를 통해 학습 및 파인튜닝된 모델로, 특정인의 맞춤형 음성인식에 효과적이다. 중증장애인의 발화를 통해서 내용을 정확하게 이해할 수 없는 경우 맞춤형 VGG16을 이용하여 발화 내용을 인식할 수 있다.

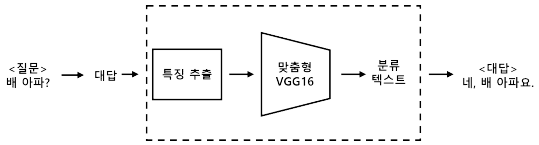


그림 3. 중증장애인 발화 인식 모델 구조

## Ⅲ. 결론

본 연구에서는 구음장애 환자의 발화 분류를 위한 최적의 모델인 MFCC+VGG16을 도출하고 중증장애인 맞춤형 발화 인식 모델의 프로토타입을 개발하였다. 이는 중증장애인이 음성을 통해 다른 사람들과 의사소통 할 수 있는 생활 필수 보완대체 의사소통 시스템 고도화에 기여한다. 나아가 컴퓨터 게임 등 개인의 취미 생활을 보조할 수 있는 모델로 발전시켜 중증장애인의 삶의 질 향상을 이루고자 한다.

## ACKNOWLEDGMENT

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

## 참 고 문 헌

[1] 이 연구는 과학기술정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 구축된 구음장애 음성인식 데이터을 활용하여 수행된 연구입니다. 본 연구에 활용된 데이터는 AI 허브(aihub.or.kr)에서 다운로드 받으실 수 있습니다.

[2] Ferreira-Paiva, L. et al, "A survey of data augmentation for audio classification," XXIV Brazilian Congress of Automatics (CBA), 2022.

[3] Font, F. et al, "Freesound technical demo," in Proc. ACM Multimed., 2013.

[4] Tan, M. et al, "Efficientnet: Rethinking model scaling for convolutional neural networks." International conference on machine learning. PMLR, 2019.