조음 장애인들의 음성인식을 위한 AI 어플리케이션

전재민 ○ 이혁 황성수
한동대학교 전산전자공학부
rmk1075@gmail.com, 21400611@handong.edu, sshwang@handong.edu

An AI Speech Recognition Application for the Articulation Disorder

Jaemin Jeon○, Hyuk Lee, Sungsoo Hwang
School of Computer Science and Electrical Engineering Handong Global University

요 약

음성인식 기술을 이용한 다양한 상품들이 대중화되고 있지만 조음장애인들은 이를 사용하는 데에 많은 어려움을 느낀다. 발음의 부정확함으로 기기가 음성을 제대로 인식하지 못한다. 이에 본 논문은 조음장애인이 일반인과 같은 수준으로 음성인식 인터페이스를 사용할 수 있도록 이들의 음성을 변환해주는 방식 및 이를 이용한 어플리케이션을 제안한다. 조음 장애인들의 음성은 일정한 특징을 가지고 있지 않기 때문에 일반적인 방법의 학습으로는 음성 인식이 쉽지 않다. 따라서 이 논문에서는 특정한 명령어를 미리 선정하고 학습을 통해 명령어를 구분(classification) 하는 방식을 제안한다.

1. 서 론

음성 인식은 음소 분석과 음성 모델의 정확성이 동시에 요구되는 시스템이다. 현재 스마트 홈 시스템, 스마트 기기의 음성 비서 시스템 등이 활발히 활용되고 있으며, 활용 분야와 음성 인식 시장이 계속 확대되고 있다. 이러한 음성인식 시스템들은 수행하는 구체적인 활용 결과는 서로 다르지만, 음성을 정확히 인식해야 한다는 공통점이 있다. 일반적인 음성 인식의 경우 발화자의 음성을 음소 단위로 분해하고 미리 학습시켜 놓은 학습망에 음성 데이터를 넣어 다음에 올 수 있는 말들의 확률을 구해 음성을 인식한다. 또한 음성 인식 학습망 제작시 많은 양의 데이터를 사용하여 음성 인식의 정확도를 높이고 사람들의 일반적인 발화 방식을 분석하여 인식할 수 있도록 한다.

하지만 조음 장애인의 경우 다른 방식으로 음성 인식을 해야 한다. 조음 장애란 구강 안에서 말소리를 만드는 데에 어려움이 있어 발음이 제대로 되지 않는 언어 상태인 경우, 즉, 발음이 부정확한 상태를 말한다. 그런데 조음 장애인의 경우 발화의 공통적인 특징이 없다. 따라서 위에 제시한 모델처럼 음소 단위로 분석하고 일반적인 발화 방식을 통해 음성 인식을 하는 모델을 적용하기 어렵다. 이처럼 사용자들의 발화에 공통된 특징이 없을 때, 사용자들의 음성을 인식하고 그에 따라 올바른 언어로 출력하는 것이 이 연구의 목표이다. 발음의 특성으로 인해 일반적인 방법인 음성인식의 방법이 아닌 고립단어 인식을 통해 화자에 알맞는 음성인식 프로그램을 제안한다. [1]

본 논문에서는 모바일 어플리케이션과 기계학습의 신경망을 통하여서 화자의 음성을 인식하고 이에 알맞는 명령어를 출력해주고자 한다. 이를 위해서 화자의 음성을 학습하고 학습된 신경망을 포함한 백엔드와 인식하고자 하는 명령어를 인식할 어플리케이션을 프론트 엔드로 구성한다. 이때 프론트 엔드는 모바일 기기에 맞게 안드로이드 기반의 어플리케이션으로 제작한다. 해당 어플리케이션을 통하여 입력받은 음성은 MFCC 방식을 통해서 특징이 추출되어지고, 추출된 특성은 신경망을 통해서 정답 class 를 출력하게 된다.[2] 출력된 결과값은 어플리케이션으로 전달되어 전자 음성으로 출력된다.

관련된 연구로는 '도어락에 적용한 화자 종속형 음성인식 시스템 구현에 관한 연구'와 'IDMLP 신경회로망을 이용한 화자종속 고립단어 인식 시스템'이 있다. 첫 번째 연구에서는 음성 데이터 입력시 설정된 절대 에너지보다 클 경우에만 실제 음성으로 간주하고 음성프레임 길이를 100 으로 설정하여 이 음성프레임이 6 개 이상이 되면 메모리에 저장한다. 또한 설정된 절대 에너지보다 작은 음성프레임이 20 개 이상이 되면 15 개의 음성프레임은 삭제하고 나머지 5 개의 음성프레임만을 음성영역으로 사용했다. 이 연구에서는 음성 전처리 과정을 단순화하여 빠른 결과값 출력을 추구한다.[3] 두 번째 연구에서는 IDMLP 신경망을 통한 음성인식 시스템을 고안하였다. 해당 논문에서 제안한 IDMLP (Input Driven Multiplelayer Perceptron) 는 각 층에서 최초의 입력값과 이전 층의 출력값, 이 두가지를 입력받아 학습을 진행하는 신경망이다. 해당 논문은 한 사람에 의해 30 번 발음된 0~9 까지의 숫자음을 입력받아서 학습을 진행하고 이를 통해 차후에 입력되는 음성을 0~9 중의 하나의 숫자로 인식하도록 한다. 이 연구에서는 신경망을 통해서 음성인식을 진행하고 최적의 synapse 의 가중치를 구함으로 음성 인식 기능의 향상을 추구하였다. [4]

2. 제안하는 어플리케이션의 구성

2.1 어플리케이션 기본 구성

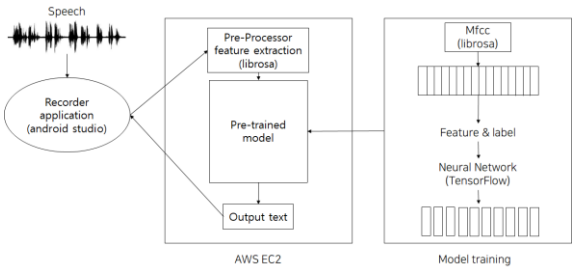

그림 1. 어플리케이션의 전체 구조

전체 구성은 프론트엔드인 어플리케이션과 백엔드인 서버로 구성된다. 프론트엔드는 서버에서 정답으로 선정된 결과값을 어플리케이션에서 받아 전자 음성으로 출력한다. 이때 입력되는 음성은 선정된 명령어들 중 하나로 가정한다. 이는 해당 프로그램이 고립 단어 인식 방법을 통해서 미리 지정된 명령어들 중의 하나로 답을 출력하기 때문이다.

백엔드인 서버는 pre-trained model 을 포함한 파이썬 코드들을 포함한다. 서버는 통신을 통해 전송받은 음성을 저장한다. 서버의 프로그램은 MFCC 를 통해 전송받은 음성의 특징을 추출한 후, 이를 학습된 신경망에 입력하여서 정해진 class 중 하나로 정답을 출력한다. 정답으로 선택됐을 때 확률이 0.3 를 넘을 경우에만 결과값으로 출력하고 0.3 이하인 경우는 재입력을 요구한다.

이때 신경망은 고립 단어 인식 방법을 통해서 학습되어진다. 학습하는 class 의 경우 AI 기기에서 많이 사용되는 명령어들을 기준으로 선정하였다. 선정된 명령어들은 다음과 같다.

'뉴스, 리모컨, 소리 작게, 소리 크게, 시간, 오늘 날씨, 오늘 일정, 지니야, 클로바'

이상의 9 개의 명령어를 학습하고 이중에서 입력값을 통해 신경망이 결정한 class 를 정답으로 출력하도록 한다.

3. 학습 신경망의 구조

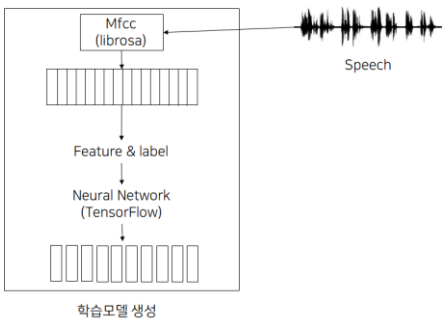

그림 2. 모델 학습 구조

한 명의 조음 장애인으로부터 9 개의 명령어에 대해 100 개의 음성 데이터를 모았다. 음성 데이터들 중 70 개의 음성 데이터는 신경망 학습에 사용하였고, 30 개의 음성 데이터는 학습 모델 테스트에 사용했다. 각 음성 데이터는 librosa 를 사용하여 MFCC 방식으로 특징추출을 한다. 추출한 명령어의 특징마다 각각의 해당하는 명령어를 라벨링을 해준다. 라벨링된 데이터들은 TensorFlow 를 통하여 신경망을 학습하여서 pre-trained model 을 생성한다.

3.1 음성의 특징 추출

신경망의 학습을 위해서 음성 데이터의 특징을 추출하게 된다. 신경망에 입력되기전에 음성은 MFCC 방식을 통해서 특징이 추출되어서 사용된다. 이때 음성의 추출을 위해서 Python 에서 제공해주는 음성 처리 library 인 librosa 를 사용하였다.

3.2 신경망 학습

신경망은 미리 지정된 9 개의 class 를 classification 하도록 학습을 진행하였고, MLP 신경망을 사용하였다. output layer 로 Softmax 함수를 사용하여서 각 class 에 대한 결과값을 0~1 사이의 값으로 출력한다. 이를 통해서 출력된 output 중 가장 큰 값을 가진 class 를 정답 class 로 인식하고 출력하도록 한다. 이때 최종 결과값이 가지는 일치율이 30% 이하인 경우에 음성을 인식하지 못했다는 결과를 출력하도록 한다.

3.3 유사한 단어의 학습

[소리 작게, 소리 크게], [오늘날씨, 오늘일정]과 같이 유사한 명령어의 경우 인식률이 낮았다. 이를 해결하기 위해서 해당 명령어들을 pair 로 two class classification 학습을 진행하여 각 명령어에 대한 인식을 진행하였다. 유사한 형태의 두 명령어에 대해서 two class classification 모델로 인식을 진행한 경우 기존의 9 class classification 보다 좋은 인식률을 보였다. 이를 통해 학습된 모델을 기존의 모델에 추가하여 인식을 진행하였다. 입력 값이 기존의 모델을 통해서 1 차적인 결과값을 출력하게 된다. 이때, 1 차 결과값이 유사한 명령어 class 인 경우, 추가로 two class classification 모델을 거쳐서 다시 한번 결과값을 출력하도록 한다. 이때 two class classification 을 거치게 되는 class 는 [소리 작게, 소리 크게], [오늘날씨, 오늘일정] 이다.

4. 결론 및 향후 연구

본 논문에서는 고립 단어 인식 방법을 사용하여 조음 장애인의 음성 인식을 위한 어플리케이션을 제안하였다. 기존의 음소 단위 음성 인식을 하는 음성 인식기와는 달리 조음 장애인의 음성을 단어 단위로 학습하여 인식하고 그에 대한 인식 결과를 표현할 수 있는 어플리케이션을

제작하였다. 이를 통하여 각각의 특성이 다른 조음 장애인의 경우에도 어플리케이션을 통해서 음성인식 인터페이스를 사용할 수 있는 방법을 제안하였다. 아래는 기존의 AI 스피커 '클로바'와 인공지능 비서 '빅스비', '시리'와 본 논문에서 제안한 방법을 통해 개발한 어플리케이션의 인식률을 비교한 내용이다. 실험은 각 class 별로 30 번씩을 들려주어서 일치하였는 지를 확인하였다. 아래의 일치율은 전체 test case 에 대한 일치율이다.

	일치율
클로바	8.52%
빅스비	35.56%
시리	*
고립 단어 인식을 통한 음성인식	75.56%

(* 시리는 녹음된 음성을 인식하지 못함)

표 1. 각 기기에 따른 test 일치율

향후 조음 장애인이 직접 본인의 목소리로 사용하고자 하는 명령어의 학습하는 기능을 추가하고, 이를 통해서 인공지능 비서나 AI 스피커 등의 음성 인식을 사용한다면 이전보다 일상생활에서의 편리함을 느낄 것이다.

5. 참고문헌

[1] 이기희, 임인칠. 화자적응 신경망을 이용한 고립단어 인식 (Isolated Word Recognition Using a Speaker-Adaptive
Neural 전자공학회논문지-B, 32(5), 765-776, (1995). 8.52% 35.56% * 90.00%

[2] 이성주, 강병옥, 정훈, 정호영, 박전규. 심층신경망 구조 기반 음성인식 시스템을 위한 변형된 MFCC 특징추출방법. 한국정보과학회 학술발표논문집, 710712. (2015).

[3] 정성훈. 도어락에 적용한 화자 종속형 음성 인식 시스템 구현에 관한 연구 (A study on the implementation of speech recognition system for speaker dependent applied to doorlock). 한국해양대학교 대학원. (2005).

[4] 박정운, 정호선. IDMLP 신경회로망을 이용한 화자종속 고립단어 인식 시스템 (Speaker-Dependent Isolated Word Recognition System Using IDMLP Neural Network). 대한전자공학회 학술대회, 293-303. (1991). Network).