언어 장애인을 위한 음성 보조 모델 연구

정윤아[1], 김영수[2], 김태희[3], 신예영[3], 정예린[3]
[1] 한양여자대학교 빅데이터학과 학부생
[2] 고려대학교 SW/AI 융합대학원 석사과정
[3] 한양여자대학교 스마트 IT 과 학부생
jya2470388@student.hywoman.ac.kr, kkyyss1278@gmail.com,
th040521@naver.com, yeriniove@naver.com, s97863395@gmail.com

Research on a Voice-Assisted Model for Language Disorders

Yoona Chung[1], Yeong-Soo Kim[2], Tae-Hee Kim[3], Ye-Young Sin[3], Ye-Rin Jung[3]
[1]Hanyang Women's University, Department of Big Data, Seoul, Korea
[2]Korea University, Graduate School of SW/AI Convergence, Seoul, Korea
[3]Hanyang Women's University, Department of Smart Information Technology, Seoul, Korea

요 약

언어 장애는 여러 장애인에게 흔히 동반되는 문제로, 이로 인해 의사소통에 어려움을 겪는다. 본 연구에서는 이러한 문제를 해결하기 위해 언어 장애인의 음성 데이터를 기반으로 한 음성 번역 모델을 구현하였다. 이 모델은 부정확한 음성을 정확한 텍스트와 음성으로 변환하여, 보다 원활한 의사소통을 가능하게 한다. 이를 통해 언어 장애를 가진 장애인들이 현대 사회에서 보다 독립적이고 효과적으로 소통할 수 있을 것으로 기대된다.

1. 서론

언어장애(Language disorder)는 언어의 이해와 표현 능력에 결함이 있는 상태로[1], 전체 장애인 중 0.9% 밖에 해당되지 않지만 다른 장애와 동반되는 경우가 많다[2]. 특히 뇌병변 장애인의 경우, 42.4%가 언어 장애를 동반함에도 불구하고[3], 이들 중 86.2%가 주요 의사소통 방식으로 '말'을 사용한다[4]. 언어장애를 가진 장애인들은 일상 생활과 의사소통에서 어려움을 겪으며, 병원 진료와 같은 생사 관련한 부분에서도 언어장애로 인해 비장애인에 비해 더 많은 진료 시간이 소요된다고 29.6%가 응답한 바 있다[5].

따라서 본 논문에서는 언어 장애인의 음성 데이터를 텍스트로 변환하고, 이를 소리 및 텍스트로 출력할 수 있는 학습 모델을 구현하였다. 이 모델을 통해 언어 장애인들의 의사소통, 업무, 일상생활 등 다양한 분야에 접목될 것으로 기대된다.

2. 구현

2.1. 데이터셋 선정

해당 기술을 구현하기 위해서는 먼저 음성 인식 모델을 학습시킬 수 있는 데이터셋이 필요하다. 본 연구에서는 한국지능정보사회진흥원에서 제공하는 AI-Hub 데이터 허브 사이트의 '구음장애 음성인식 데이터' 를 사용하였다. 이 데이터셋은 음성 데이터(.wav)와 라벨링된 텍스트 데이터(.json)로 구성되어 있으며, 라벨링 데이터 파일에는 텍스트뿐만 아니라 음성 파일에 대한 추가 정보도 포함되어 있다. 이를 바탕으로 지도학습을 통해 모델을 학습시키고자 하였다.

2.2. 데이터 전처리

우선, 각 음성 파일의 침묵-비침묵 구간을 나누는 작업을 진행하였다. 침묵 구간과 비침묵 구간에 대한 time 값을 return 받아 어떤 위치에서 음성이 발화되는지를 파악하였다. 이후 침묵 구간은 제외하고 비침묵 구간(발화 구간)만 .wav 파일로 각각 저장되도록 한다.

```
File ID-02-25-N-KSM-02-01-M-45-JL.wav - Segment 0: 0.00s to 6.05s -> Impediment_segment/output1_0.wav
File ID-02-25-N-KSM-02-01-M-45-JL.wav - Segment 1: 6.05s to 9.42s -> Impediment_segment/output1_1.wav
File ID-02-25-N-KSM-02-01-M-45-JL.wav - Segment 2: 9.42s to 14.85s -> Impediment_segment/output1_2.wav
File ID-02-25-N-KSM-02-01-M-45-JL.wav - Segment 3: 14.85s to 20.11s -> Impediment_segment/output1_3.wav
File ID-02-25-N-KSM-02-01-M-45-JL.wav - Segment 4: 20.11s to 27.37s -> Impediment_segment/output1_4.wav
File ID-02-25-N-KSM-02-01-M-45-JL.wav - Segment 5: 27.37s to 34.67s -> Impediment_segment/output1_5.wav
File ID-02-25-N-KSM-02-01-M-45-JL.wav - Segment 6: 34.67s to 39.60s -> Impediment_segment/output1_6.wav
File ID-02-25-N-KSM-02-01-M-45-JL.wav - Segment 7: 39.60s to 43.42s -> Impediment_segment/output1_7.wav
File ID-02-25-N-KSM-02-01-M-45-JL.wav - Segment 8: 43.42s to 49.40s -> Impediment_segment/output1_8.wav
File ID-02-25-N-KSM-02-01-M-45-JL.wav - Segment 9: 49.40s to 54.66s -> Impediment_segment/output1_9.wav
File ID-02-25-N-KSM-02-01-M-45-JL.wav - Segment 10: 54.66s to 57.85s -> Impediment_segment/output1_10.wav
```

(Figure 1) 음성 파일 전처리 과정 진행

이후 기존 텍스트 데이터를 불러와 문장 부호를 기준으로 문장을 분리한다. 분리된 문장은 전처리된 음성 파일과 DataFrame 으로 병합하고, 음성 데이터와 텍스트 데이터의 개수가 일치하는지 확인한다.

	start	end	text
0	0.447	6.052	나는 바지를 입고 단추를 채웁니다
1	0.300	3.366	책상 위에 가방이 있습니다
2	0.465	5.428	가방에 사탕과 연필을 넣을 거예요
3	0.471	5.259	아빠와 자동차를 타고 동물원에 갑니다
4	0.466	7.265	잘 다녀와 하면서 엄마가 뽀뽀를 해줍니다
..	...	...	...
186	0.481	7.849	차선을 넘나들며 치열하게 추격전을 벌였다
187	0.451	6.306	카메라와 쿠키를 들고 있는 남자가 가장 키가 크다
188	0.365	9.623	타조는 투명한 유리에서 칼날과 티끌을 발견했다
189	0.479	7.949	파란 눈과 하얀 피부 덕분에 첫인상이 푸근해보였다
190	0.470	6.792	하늘이는 후미진 골목 끝에서 히죽거렸다

[191 rows x 3 columns]

(Figure 2) 음성과 텍스트 데이터 DataFrame 병합

병합된 각 음성 파일과 텍스트 데이터의 파일명을 일치하게 변경하여 클라우드 내에 저장한다.

2.3. 모델 학습

전처리된 데이터셋을 활용해 음성 인식 모델에 지도학습을 진행한다. 학습에는 HuggingFace Transformers 와 OpenAI Whisper 모델을 사용하였다. 먼저, Whisper Feature Extractor 를 통해 음성 데이터를 padding 하고 truncating 해 길이를 맞추고, padding 된 오디오 배열을 log-Mel 스펙트럼으로 변환하는 전처리 작업을 수행하였다. Whisper model 은 5 가지 크기로 제공되는데, 학습 시 CER(문자 오류율)과 epoch 에 부담을 주지 않는 최소 크기인 base 모델을 사용하였다.

Size	Parameters	English-only model	Multilingual model	Required VRAM	Relative speed
tiny	39 M	tiny.en	tiny	~1 GB	~32x
base	74 M	base.en	base	~1 GB	~16x
small	244 M	small.en	small	~2 GB	~6x
medium	769 M	medium.en	medium	~5 GB	~2x
large	1550 M	N/A	large	~10 GB	1x

(Figure 3) Whisper model Size

이후 전처리된 데이터를 활용해 모델 학습을 진행한다. Whisper Processor 를 사용해 모델에 오디오 데이터를 입력하고, 검증 데이터셋의 평가 지표로 CER(Character Error Rate)를 설정하였다. 한국어는 자음과 모음이 결합된 교착어이기 때문에 CER 을 사용하는 것이 효과적이다. 데이터셋은 학습(train)과 테스트(test) 비율을 각각 0.8, 0.2 로 나누고, 검증(validation)은 테스트 데이터셋의 50%를 활용했다. 학습은 epoch 2000 으로 설정하여 진행하였다.

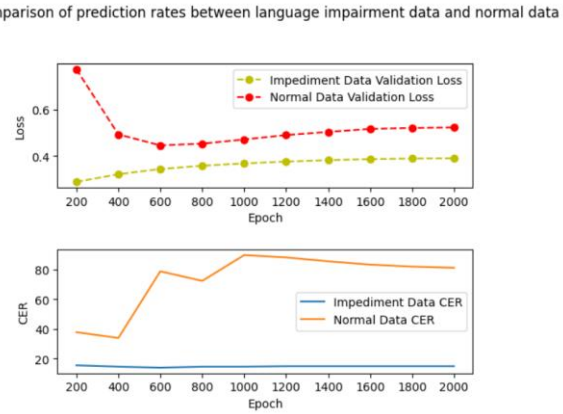
Step	Training Loss	Validation Loss	Cer
200	0.042600	0.289628	15.506329
400	0.001600	0.322463	14.556962
600	0.000600	0.344684	13.924051
800	0.000300	0.358797	14.556962
1000	0.000200	0.368638	14.556962
1200	0.000200	0.376486	14.873418
1400	0.000200	0.382673	14.873418
1600	0.000100	0.386912	14.873418
1800	0.000100	0.389570	14.873418
2000	0.000100	0.390634	14.873418

(Figure 4) Training 진행

3. 실험

모델이 제대로 동작하는지 확인하기 위해 반응형 웹앱을 구현하였다. 클라이언트에서 받은 음성 데이터를 모델에 입력하여 텍스트로 변환하고 예측을 진행한다.

실험 결과, 구음 장애 데이터를 학습한 모델과 일반인의 음성 데이터를 학습한 모델 간의 예측 오차(loss)는 각각 0.39 와 0.52, CER 값은 14.9 와 81.2 로 나타났다. 동일한 데이터셋 비율로 학습을 진행하고 예측을 수행했음에도, loss 와 CER 에서 큰 차이가 발생했다. 이는 구음 장애 데이터셋으로 fine-tuning 을 진행한 후, validation loss 와 CER 값에서 큰 차이가 나타났기 때문에 예측 정확도가 크게 향상되었음을 의미한다.



(Figure 5) 데이터셋 간 예측 오차 및 CER 비교

4. 결론

본 연구에서는 언어 장애인의 음성 데이터를 활용해 학습 모델을 생성하고, 이를 기반으로 음성 인식 모델을 구현하였다. 의사소통 능력은 현대 사회에서 필수적인 요소이며, 언어 장애인들은 이 음성 인식 모델을 활용해 자신의 의사를 보다 명확하게 표현함으로써 독립적인 삶을 영위할 수 있을 것이다. 향후, 이 논문을 바탕으로 다양한 장애 유형의 발화 데이터를 추가로 학습시키면 모델의 성능을 더욱 개선할 수 있을 것으로 기대된다.

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

참고문헌

[1] 이진경 and 허명진. "키워드 접근법을 활용한 언어학습장애아동의 지역중심 사회교과어휘 중재 단일연구." 한국청각언어장애교육연구, vol. 13, no. 2, pp. 85-96, 2022.

[2] 보건복지부, “2023 년 등록 장애인 현황 통계”, 2023.

[3] 서원선. "뇌병변장애인의 의사소통경험에 대한 현상학적 연구: Giorgi 의 현상학 분석을 중심으로." 장애인복지연구, vol. 12, no. 2, pp. 1-17, 2021.

[4] 한국보건사회연구원, “2017 장애인 실태조사”, 2017.

[5] 국민건강보험 일산병원, “장애인, 비장애인 환자에 대한 외래 진료시간, 형태 비교 및 장애인의 외래 진료 영향 인자연구” , 2020.