JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 25, No. 2, pp. 547-555, Feb. 2024

아동 음성 인식 향상을 위한 인공지능 모델 개선

손 계 원[1] · 소 준 섭[1] · 고 주 은[1] · 이 진 우[1] · 이 정 록[2] · 신 원 선[3*]
[1]주식회사 비전21테크 연구원
[2]에이아이리더 기술이사
[3*]주식회사 비전21테크 대표이사

Enhanced AI Model to Improve Child Speech Recognition

Gyewon Son[1] · Junseop So[1] · Jooeun Ko[1] · Jin-Woo Lee[1] · JeongRok Lee[2] · Won-Sun Shin[3*]
[1]Researcher, Vision21Tech, Daejeon 34050, Korea
[2]Chief Technology Officer, AI Leader, Cheonan 30194, Korea
[3*]CEO, Vision21Tech, Daejeon 34050, Korea

[요 약]

아동의 음성 인식은 인간-컴퓨터 상호 작용, 교육적 기술에서 중요한 연구 주제로 부각 되고 있다. 아동의 발화는 성인의 발화와 다른 특징이 있어, 기존의 자동 음성 인식 (ASR) 모델은 아동의 음성을 정확하게 인식하는데 어려움을 겪는 경우가 많다. 이 연구에서는 Open AI의 Whisper 모델을 기반으로 4-7세 아동의 음성을 텍스트로 변환하였다. 특히, 아동과 성인의 발화 차이를 고려하여 모델의 성능을 개선하기 위해 데이터 정제와 데이터 셋을 구축하였다. 이러한 작업은 Whisper 모델의 성능을 아동 음성에 최적화하기 위한 학습 데이터 관점에서의 방법을 제시한다. 이 연구의 실험적 접근법은 Whisper 모델을 이용하여 아동의 음성 인식 성능을 향상시키는 방법을 탐구한다. 제시한 방법을 통해 아동 한국어 음성인식의 에러율을 84% 개선하였다.

[Abstract]

Child speech recognition has emerged as a significant research topic in the fields of human-computer interaction and educational technology. Children's utterances possess distinct characteristics from adults', often making it challenging for conventional automatic speech recognition (ASR) models to accurately recognize their speech. In this study, we utilized OpenAI's Whisper model to transcribe the voices of 4-7 year-old children into text. Specifically, considering the differences in speech between children and adults, we conducted data refinement and dataset construction to enhance the model's performance. These efforts present an approach to enhance the performance of the Whisper model for child speech recognition from the perspective of training data. Our method improved the error rate of Korean child voice recognition by 84%.

색인어 : 자동음성인식, 아동, 한국어, 음성 변조, 위스퍼

Keyword : Automatic Speech Recognition, Child, Korean, Voice Modulation, Whisper

http://dx.doi.org/10.9728/dcs.2024.25.2.547

Received 14 December 2023; **Revised** 24 January 2024
Accepted 31 January 2024

***Corresponding Author; Won-Sun Shin**
Tel: ※ 개인정보 표시제한
E-mail: vision21tech@naver.com

Ⅰ. 서 론

자동 음성 인식 (ASR; automatic speech recognition) 기술은 인간의 음성을 텍스트로 변환하는 기술이다. 음성 인식 기술은 음향 모델, 발음 사전 그리고 언어 모델의 3단계로 구분되어 독립된 구조로 모델링이 되어 왔다. 2010년대부터 딥러닝 기술이 발전하여 음성 인식의 정확도가 획기적으로 향상되었다. 2015년부터는 음향 모델, 발음 사전, 언어 모델을 하나의 딥러닝 모델로 표현하는 종단 간 (End-to-End)이 등장하여 주목받고 있다.

효과적인 아동 자동 음성 인식 시스템의 개발은 최근 몇 년 동안 중요해졌다. 예를 들어, 아동 자동 음성 인식의 발달은 소셜 로봇 (Social Robot)과 같은 대화형 시스템을 사용한 교육 환경에서 아동을 위한 교육 및 평가 도구 개발을 촉진시킬 수 있다[1]-[6].

현재 딥러닝 기반 자동 음성 인식 모델들은 성인들의 음성을 인식하는 것을 잘 수행하지만, 아동들의 음성을 인식하는 데는 큰 어려움이 있다. 아동 음성의 특징은 성인과 다르다. 유아 음성의 음의 높낮이, 성도, 포먼트와 같은 음성의 음향학적 상관성은 나이에 의존하는 체계적인 구조를 따른다. 유아의 음성 특징은 아이가 성장함에 따라 해부학적, 생리학적 변화로 인해 급속도로 발달하게 된다. 따라서 기존에 성인 대상으로 구축된 음성 데이터로 학습된 모델로 유아의 음성을 인식할 경우 인식률 저하가 일어날 수 있다[7]. 본 연구에서 Whisper-Base 모델을 4-7세 아동과 성인의 음성을 인식시켰을 때 CER이 각각 1.208과 0.177이었다.

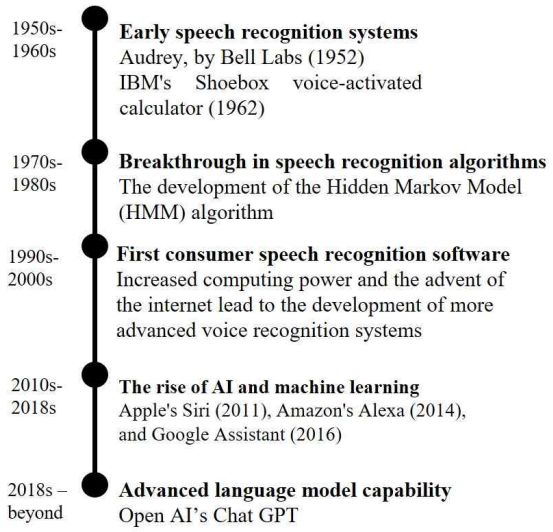

그림 1. 음성 인식 인공 지능의 역사
Fig. 1. The history of voice AI

이런 이유로 아동 음성 데이터로 학습을 한 음성 인식 모델이 필요하다. 하지만 현재 음성 인식 모델을 학습할 수 있는 성인용 음성 데이터는 많지만, 아동용 음성 데이터는 현저히 적다. 학습에 사용 가능한 레이블이 있는 아동의 음성 데이터를 얻는 것은 레이블이 있는 성인 음성 데이터를 얻는 것에 비해 매우 어렵다. 아동의 음성 데이터를 구축하는 것은 성인과 아동의 목소리 차이, 잡음의 정도, 단어를 이해하고 아동의 음성 데이터에 레이블을 달아야 하는 작업의 어려움으로 상당히 제한적이다[8].

많은 연구자들은 그동안 아동 음성 인식 성능을 개선하기 위해 다양한 시도를 해왔다. 그중 학습 데이터를 증가시키기 위한 데이터 증강 방법을 사용하였다. 예를 들어 TTS (Text-to-Speech) 기반 데이터 증강, GAN (Generative Adversarial Network) 기반 데이터 증강, 아동 음성의 기본 주파수 특징을 정규화하는 데이터 증강 그리고 속도 변화, 피치(pitch) 변화, 박자 변화, 소리 크기 변화, 소리의 울림 변화 등을 통한 데이터 증강 방법이 있다. 이러한 접근 방식은 초보적이지만 매우 효과적이다[9]-[21].

본 연구에서는 아동 음성 인식 성능을 높이기 위해 데이터의 변환을 통해 데이터의 다양성을 향상시키고 아동 음성 인식의 일반화 능력을 키우려 하였다. 또한, 사전 학습된 (Pre-trained) Whisper 모델을 미세 조정 (fine-tuning)을 함으로써 학습 시간과 리소스를 절약함과 동시에 더 높은 성능을 얻고자 하였다.

Ⅱ. 관련 연구

Wav2Vec2.0 모델의 경우, 자기주도학습 (SSL; Self-Supervised Learning)을 통해 대규모의 음성 데이터에서 음성 표현을 학습한다. 이 학습은 레이블이 지정되지 않은 데이터에서 음성 특징을 추출하는 데 사용되고 이를 통해 모델은 음성 신호에 대한 의미 있는 표현을 학습하게 된다. Wav2Vec2.0은 사전 학습으로 양질의 음성 표현을 학습하지만 자동 음성 인식 작업에 적용하기 위해서는 추가적인 미세 조정 학습이 필요하다. 미세 조정 학습은 레이블이 있는 음성 데이터를 사용하여 모델을 특정 작업에 맞게 조정하는 과정이다. 이 단계에서 모델의 디코더 부분이 미세 조정되고 입력 음성 데이터의 텍스트와의 일치를 개선하기 위해 학습한다. 레이블이 있는 데이터로 미세 조정 학습을 한 후에는 자동 음성 인식을 하기 위한 디코더가 추가되어 음성 신호를 텍스트로 디코딩하는 역할을 수행한다. 이 미세 조정 학습이 복잡한 과정이기에 모델의 유용성을 제한할 수 있다. 이러한 학습 방법은 학습 데이터 셋 내에서 성능을 향상시키는 패턴을 찾는 데 매우 능숙하지만 다른 데이터 셋에 일반화되지 않을 수 있다. 또한, 성인 데이터에 대한 인식 성능은 좋지만 그에 비교해 아동 데이터에 대한 인식 성능은 좋지 않다. 자동 음성 인식의 목표는 다양한 환경에서 신뢰성 있게 작동하는 것이어야 한다[22]-[23].

본 연구에서는 2022년 9월에 Open AI가 공개한 Whisper 음성 인식 모델을 사용하였다. 이 모델은 약한 지도 학습으로

학습된 음성 인식을 확장하여 레이블이 지정된 음성 데이터를 학습하였다. 그림 2와 같이 Whisper는 안정적으로 확장 가능한 인코더-디코더 Transformer를 사용하였고 레이블이 있는 오디오 데이터를 680,000시간 사용하여 학습을 시킨 모델이다. 이 모델은 데이터 집합 별 미세 조정 학습 없이 양질의 결과를 얻을 수 있었다. 이것은 영어뿐만 아니라 다국어 및 여러 작업에 적용된다. 680,000시간의 오디오 데이터 중에 117,000시간이 96개의 다양한 언어를 다루며, 125,000시간은 번역 데이터로 구성되었다. Whisper는 여러 음성 인식 모델과 비교하였을 때 좋은 성능을 보이며, 한국어의 학습 데이터의 양이 10,000시간으로 학습 데이터 양이 가장 많은 사전 학습 모델로 Zero-Shot Performance로 한국어 인식 에러율 (WER)이 0.15정도이다. 또한, 잡음이 있는 음성 데이터를 학습 데이터로 사용하였기에 잡음에 대해서 강건하다. 이러한 점은 4-7세 아동의 음성 인식을 연구하는 본 연구의 음성 환경에 적합하다[22].

본 연구에서는 4-7세 아동의 음성 인식 성능을 높이고 학습 시간을 절약하기 위해 전이 학습을 이용하였다[24]-[25].

Ⅲ. 연구 과정

본 연구에서는 Open AI에서 공개한 Whisper 음성 인식 모델을 사용하였다[26]. 이 모델에 4-7세 아동의 한국어 음성 데이터를 이용한 미세 조정 학습으로 아동의 한국어 음성 인식 성능을 개선하고자 하였다.

연구 윤리와 아동의 특성 등의 사유로 아동 음성 데이터 셋을 구축하는 일은 어려움이 따른다. 국내에서 연구 목적으로 공개한 음성 데이터의 대부분이 성인의 음성 데이터이며, 음성 인식 서비스를 제공하는 기업에서도 데이터 수집의 어려움과 비용 등의 문제로 성인 음성을 중심으로 음성 인식 엔진을 개발하고 있는 실정이다[27]. 수집이 어려운 아동 음성 데이터를 최소한으로 이용하여 음성 인식률을 개선한다. 이를 위해 데이터 변환과 성인 데이터 합성 기법 등을 활용하여 데이터 셋을 구축한다. 정제된 데이터를 학습하여 성능 개선을 확인한다.

3-1 데이터 수집 및 전처리

아동 음성 데이터는 AIHub에서 제공하는 공공데이터 중 '한국어 아동 음성 데이터'[28]를 학습 데이터로 선정하였다. 성인 음성 데이터는 AIHub의 공공데이터 중 '자유 대화 음성(일반남여)'[29]를 이용하였다.

아동 데이터는 4~12세 나이로 단계적으로, 그 외의 연령은 U4(4세 미만), O12(12세 초과)로 구분되어 있다. 음성의 음향학적 상관성은 나이에 의존하는 해부학적, 생리학적 구조를 따른다[7]. 또한, 7세 이하 미취학 아동은 글자와 소리 간의 관계, 문법 및 억양에 대한 학습이 부족하여 취학 아동 및 성인에 비해 발음이 부정확할 수 있다. 다른 연령대에 비해 다른 특성을 지닌 유아 음성 인식의 강건성을 위해 4~7세 아동의 데이터만 선별하였다. 발화수 기준으로 학습에는 50,000개, 테스트에는 1,000개의 테이터를 사용하였다. 학습 데이터와 검증 데이터는 9대 1 비율로 설정하였다.

데이터는 공공데이터 중 아동의 발화 비율이 높고 메타 데이터를 통해 발화 시간을 특정할 수 있는 것을 대상으로 하였다. 성인 데이터는 해당 공공데이터에서 스튜디오로 분류한 카테고리의 음성을 사용하였다. 아동과 동일하게 학습과 검증 데이터는 9대 1로 나누어 사용하였다. 사용된 연령 별 데이터는 표 1과 같다.

표 1. 연령별 데이터의 분포
Table 1. Number of sentences distributed by age of the data

Age	Train Data Set (ea)	Test Data Set (ea)
4	3,698	107
5	15,166	301
6	11,447	210
7	19,689	382
Total	50,000	1,000

해당 데이터들은 발화 별로 음성이 구분되어 있어 하나의 파일에서 레이블과 맞춰 자르는 작업 중에 생길 수 있는 문제가 줄어 데이터의 강건성을 확보하였다.

공개 데이터 오디오 파일을 이름을 기준으로 짝이 없는 음성 또는 레이블 데이터에 대한 제거를 실시하였다. 분리된 파일의 무결성 검사 실시 후 음성 데이터의 문장 레이블 확인을 위해 샘플링 검사를 실시하였다.

Python의 soundfile, librosa, pydub 라이브러리를 이용하여 wav 파일에 대한 무결성을 교차 검증하여 결손 파일을 제거하였다. 무결성이 확인된 데이터에 대해 데이터 셋 별로 200개 샘플링으로 음성과 레이블의 일치를 확인하였다.

음성의 시작 전 공백이 0초에서 5.78초까지 편차가 커서 음성 데이터를 기반으로 음성의 공백을 제거하였다. 음성의 앞뒤 공백을 잘라낸 후 0.5초씩 묵음을 삽입하여 데이터를 규격화시켰다. 이 데이터를 베이스라인으로 데이터 변환을 실시하였다. 모든 변조는 librosa 라이브러리를 이용하여 진행하였다.

그림 2의 좌측 하단에 있는 Log-Mel Spectrogram 형태로 데이터를 만들기 위해 허깅페이스(Hugging Face)의 데이터 셋 라이브러리를 이용하였다. 25밀리 초 단위로 음성을 분할하여 80개 빈으로 구성하였다. 음성 파일과 해당 레이블을 매칭시켜 데이터 추가 전처리 작업을 진행하였다. 이후 그림 2의 구조를 따라 전이 학습이 진행된다.

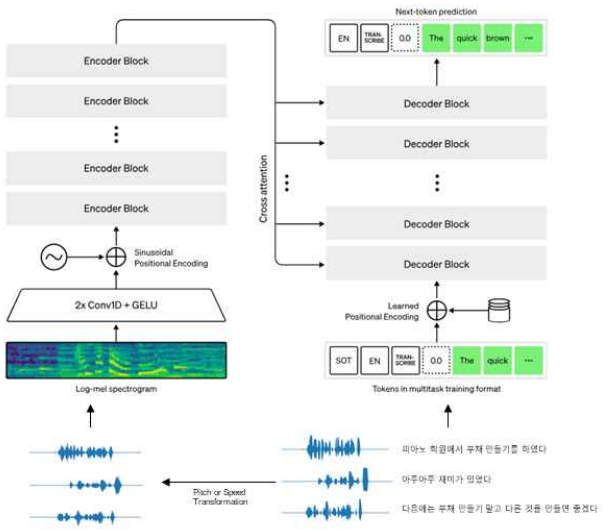

그림 2. Whisper 모델의 구조 [22]
Fig. 2. Whisper model architecture [22]

3-2 학습 모델과 학습 조건

Whisper 모델은 표 2와 같이 파라미터 수에 따라 여러 가지 크기의 모델이 있다[30]. Whisper 모델 중에 접근성이 좋은 Base 모델을 사전 학습 모델(Pre-trained Model)로 선택하였다. Whisper의 Base 모델은 한국어가 포함된 다국어가 학습된 모델이다.

표 2. Whisper 모델의 파라미터와 학습된 언어 [30]
Table 2. Parameters and trained languages of Whisper model [30]

Size	Parameters (M)	English-Only	Multilingual model
tiny	39	√	√
base	74	√	√
small	244	√	√
medium	769	√	√
large	1550		√

허깅페이스(Hugging Face)에서 제공하는 API와 라이브러리를 통해 Whisper 모델을 미세 조정 학습하였다. 이를 이용하여 Whisper 모델에 한국어 아동 데이터 셋으로 미세 조정하여 아동 음성 인식 성능 개선을 시도하였다[31].

실험의 조건은 표 3과 같이 설정하였고 본 연구에서 설정한 여러 데이터 셋으로 미세 조정하여 각각을 비교하였다. 허깅페이스 포럼을 참고하여 학습률은 Whisper-base 모델에서 권장되는 학습률인 2.5e-5로 설정하였다. 학습 배치 크기는 GPU의 구조상 8의 배수가 이점을 갖기 때문에 RTX 4090 메모리 24기가의 허용치 내에서 가장 큰 값으로 설정하였다.

표 3. 실험 세팅
Table 3. The setting of experiments

Parameter	Value
Learning Rate	2.5e-5
Epoch	8
Train Batch	48
Evaluation Batch	16
GPU	RTX4090 1ea
OS	Ubuntu 22.04.3 LTS

3-3 평가 지표

WER (Word Error Rate)와 CER (Character Error Rate)의 계산 방법은 거의 동일하다. 띄어쓰기로 구분되는 토큰들의 총 개수에 대비되는 삽입(I), 삭제(D), 교체(S)의 수가 얼마나 많은지를 계산하는 것이다. 차이가 있다면 WER은 단어가 토큰이 되며, CER은 문자가 토큰이 되는 것이다.

$$CER = \frac{S + D + I}{N} \tag{1}$$

S : minimum number of subtitutions
D : minimum number of deletions
I : minimum number of insertions
N : total number of characters

음성 데이터에 대한 유사도를 비교할 때 음성을 통해 텍스트로 변환된 데이터는 여러 변수가 존재한다. 영어를 텍스트로 변환한 경우에는 텍스트가 알파벳으로 이루어져 있고 띄어쓰기도 명확하여 유사도를 측정하기가 상대적으로 어렵지 않다. 한국어의 경우 초성, 중성, 종성이 하나의 글자를 이루고 있으며, 정확한 띄어쓰기가 어렵다. 또한, 두음법칙 연음법칙 등으로 의미 전달은 되지만 정확한 한글 표기와 발음상의 다른 점이 많은 경우가 많다. 따라서 CER이 WER보다 한국어 음성 인식의 평가 방법으로 적합하다[32]-[33].

Ⅳ. 연구 결과

4-1 결과 분석

표 4와 같이 미세 조정에 사용하는 학습 데이터의 종류를 다르게 하여 CER을 비교하였다. 표 4와 표 5에서 볼 수 있듯 미세 조정을 하지 않은 1번은 음성 인식 테스트만 하였고, 나머지는 8가지 데이터 셋으로 미세 조정을 진행하였다.

아동의 음성 데이터 발화수 기준 50,000개를 미세 조정한 결과 CER 0.196으로 미세 조정하지 않은 것과 비교하여 큰

표 4. 음성 데이터 변환에 따른 음성 인식 결과
Table 4. Speech recognition results for voice data transformation

Model No.	Finetuning Dataset	Data Size(ea)	CER	REF	HYP
1	N/A	–	1.208	하염없이 흔들어요	하얀 옵셀데라요
			0.177 (test : Adult)		–
2	Baseline, Original	50,000	0.196		하얀없이 흔들어요
3	Random Speed (*0.9, *1, *1.1)	50,000	0.239		하얀 업시원 흔들어요
4	Random Speed (*0.8,*0.9, *1, *1.1, *1.2)	50,000	0.206		하얀 업시을 흔들어요
5	Random Pitch (−0.3, 0, 0.3)	50,000	0.285		하얀 업시원을 들어요
6	Random Pitch (−0.3, −0.2, −0.1, 0, 0.1, 0.2, 0.3)	50,000	0.207		하얀 업시 흔들어요

*The purpose of our study is to recognize Korean children's voices. So, the reference data and predicted values(hypothesis) used were written in Korean. When written in English, it was written in Korean because it was difficult to see the difference in predicted values between models.

표 5. 아동 및 성인 음성 데이터의 통합적 학습에 따른 음성 인식 결과
Table 5. Speech recognition results for integrated training child and adult speech data

Model No.	Finetuning Dataset	Data Size(ea)	CER	REF	HYP
7	Child 50K + Adult 10K	60,000	0.191	하염없이 흔들어요	하얀없이 흔들어요
8	Child 50K + Adult 20K	70,000	0.194		하얀없이 흔들어요
9	Random Speed 50K (*0.8,*0.9, *1, *1.1, *1.2) + Adult 10K	60,000	0.228		하얀 업시금을 봐요
10	Random Pitch 50K (−0.3, −0.2, −0.1, 0, 0.1, 0.2, 0.3) + Adult 10K	60,000	0.200		하얀없이 흔들어요

개선이 있었다.

데이터를 증강하는 것과 달리, 데이터의 다양성을 기대하며 속도와 피치 변환을 거친 데이터를 학습에 이용하였다. 변환의 조건은 Ko 등[34]의 실험을 참조하였다. 표 4의 모델 3은 모델 2의 Baseline의 데이터를 같은 비율로 0.9배속, 1배속, 1.1배속으로, 모델 4는 0.8에서 1.2까지 5가지 속도로 무작위로 변환하였다. 모델 5와 6도 피치에 대해서 각각 3가지, 7가지로 균등한 비율로 변환을 적용하였다. 데이터의 수를 증강하지 않고 변환하여 학습 데이터로 이용한 경우 Baseline과 비교하였을 때 성능이 개선되지 않았지만, 변환의 범위에 따라 에러율의 차이를 보였다. 속도 변환의 경우, 3가지보다 더 넓은 범위의 5가지 속도로 변환한 모델이 성능이 14% 개선된 것을 확인하였다. 피치 변환은 3가지보다 7가지의 변환을 적용했을 때 에러율이 27% 낮아졌다. 데이터를 증강하지 않고 변환만으로 인공지능 모델의 성능을 개선할 수 있음을 확인하였다. 한정된 자원을 사용할 때 모델의 성능을 개선할 수 있는 방법을 얻었다고 볼 수 있다. 이것을 얻기 위해 최적의 변환 범위 등 변환 조건을 찾는 것이 중요하다.

표 5의 모델 7은 아동 5만 개와 성인 1만 개의 데이터를 학습에 이용하였다. 성인 데이터와 아동 데이터를 같이 학습한 경우, 아동 데이터만 학습했을 때보다 아동 음성 인식 에러율이 Baselilne과 비교하였을 때 2.6% 개선되었다. 성인 데이터는 높은 다양성으로 구성되어 있어 모델이 다양한 음성 특징 및 언어적 패턴을 학습하여 일반화하는 데 도움이 되었을 것으로 보인다. 하지만 성인 데이터를 20,000개로 증가시킨 모델 8이 모델 7보다 성능이 감소하였다. 이것은 아동과 성인의 언어 및 발음은 다를 수 있어 성인 데이터를 과도하게 사용하면 아동의 특수한 언어 특징을 제대로 학습하지 못할 수 있음을 보여준다. 성인 데이터가 아동의 음성 인식 학습에 도움을 주는 적정한 양이 있다는 것을 추론할 수 있다.

모델 9와 10은 아동 음성 데이터 5만 개를 각각 속도와 피치 기준으로 변환한 후 성인 음성 데이터 1만 개를 같이 학습하였는데 아동 음성 인식 CER이 각각 11%, -3%의 변화가 있었다. 데이터 변환 조건에 따라 성인 음성이 아동 음성 인식에 기여하는 바가 다른 것을 확인하였다.

4-2 향후 계획

본 연구에서는 제한된 시스템의 자원 때문에 상대적으로 작은 모델에 대해서만 실험을 실시하였다. 그럼에도 불구하고 기존 베이스라인 모델과 미세 조정한 모델의 아동 음성 인식률이 개선된 점을 확인하였다.

현재 연구 결과를 바탕으로 미세 조정한 모델의 성능 향상을 위해서 실제 4-7세 아동 음성을 녹음해 활용할 예정이다. 연령대 별로 발화 특징이 다르므로 이 부분을 고려해 데이터 셋을 구성할 예정이며, 이를 통해 더 강건한 아동 음성 인식 모델을 기대한다.

본 연구에서는 속도 또는 피치 변환을 이용해 데이터 변환을 진행하였지만, 더 다양한 데이터를 수집하기 위해 RVC (Realtime Voice Changer), Voice Transfer 등 딥러닝 모델을 기반으로 기존에 수집된 성인 데이터를 아동 음성 특징을 가진 데이터로 변환하는 등 다른 변환 방법을 이용할 예정이다. 특히, 본 연구에서 성인 데이터가 아동 음성 인식에 미치는 영향을 주는 부분을 확인했다. 한국어를 유창하게 발음하는 성인 데이터와 발음이 상대적으로 좋지 않은 유아의 음성 데이터를 같이 학습하였을 때, 모델 학습에 긍정적인 효과가 기대되어 가설 확인을 위한 추가 연구도 진행할 예정이다.

이후 추가 시스템 자원이 준비된다면, Whisper-Large 모델을 기반으로 실험을 진행해 현재보다 더 높은 정확도를 구현할 수 있을 것으로 예상된다. 이를 이용해 어린이를 위한 교육, 게임, 음성 인식 기반 응용 프로그램 등에 적용할 수 있는 방법들을 연구할 예정이다.

V. 결 론

본 연구에서는 아동의 음성 인식 정확도를 향상시키기 위해 여러 조건의 아동 음성 데이터로 Whisper-Base 모델을 미세 조정하였다. 다양한 성인 음성으로 훈련된 Whisper 사전 훈련 모델을 4-7세 아동의 음성 데이터 셋을 활용하여 미세 조정하였으며, 이를 통해 성인과 구별되는 아동의 독특한 발음을 더욱 효과적으로 처리할 수 있게 되었다. 미세 조정된 모델은 아동 음성에 대한 문자 오류율(CER)을 감소시켜, 정확도를 크게 개선하였다.

4-7세의 아동 데이터 50,000개와 성인 데이터 10,000개로 이루어진 학습 데이터로 얻은 모델의 성능은 본 연구에서 가장 개선된 결과로 아동 음성 인식에 성인 음성 데이터가 도움이 되는 것을 확인하였다.

또한, 데이터 증강 없이 변환을 통한 데이터의 다양성을 확보한 데이터 셋으로 확인한 모델의 성능은 아동 음성 인식 성능의 개선 가능성을 확인하였다. 이러한 방법은 음성 인식 모델이 다양하고 복잡한 아동 음성을 인식하는데 도움을 줄 수 있을 것으로 기대한다.

결론적으로, 본 연구에서는 Whisper 모델의 미세 조정을 위한 데이터 셋 구성 방법과 데이터 변환 기법을 중점적으로 다루었다. 이를 통해 한국어를 사용하는 아동의 음성 인식이 84% 개선된 실험 결과를 얻었다. 이러한 결과는 음성 인식 기술의 한국어에 대한 정확성 및 효율성을 향상시킬 수 있음을 시사한다. 특히, 교육, 게임, 음성 인식 기반 응용 프로그램과 같은 다양한 분야에서 어린이와 AI 시스템 간의 상호작용이 강화될 것으로 예상되는 현 시점에서, 이는 아동을 대상으로 하는 기술의 접근성과 유용성을 향상시키는 데 중요한 역할을 할 것으로 기대한다. 또한, 어린이 음성 인식 기술의 발전에 기여함으로써, 향후 다양한 응용 분야에서의 실질적인 변화를 가져올 것으로 기대한다.

감사의 글

이 연구는 과학기술정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 구축된 “한국어 아동 음성 데이터”, “자유대화 음성(일반남여)”을 활용하여 수행된 연구입니다. 본 연구에 활용된 데이터는 AI 허브(aihub.or.kr)에서 다운로드 받으실 수 있습니다.

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2023-00251378).

참고문헌

[1] J. Tepperman, J. Silva, A. Kazemzadeh, H. You, S. Lee, A. Alwan, and S. Narayanan, “Pronunciation Verification of Children’s Speech for Automatic Literacy Assessment,” in *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh: PA, pp. 845-848, September 2006. https://doi.org/10.21437/Interspeech.2006-286

[2] H. T. Bunnell, D. M. Yarrington, and J. B. Polikoff, “STAR: Articulation Training for Young Children,” in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, Beijing, China, pp. 85-88, October 2000. https://doi.org/10.21437/ICSLP.2000-757

[3] G. Yeung, A. Afshan, K. E. Ozgun, K. Kaewtip, S. M. Lulich, and A. Alwan, “Predicting Clinical Evaluations of Children’s Speech with Limited Data Using Exemplar Word Template References,” in *Proceedings of the 7th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2017)*, Stockholm, Sweden, pp. 161-166, August 2017. https://doi.org/10.21437/SLaTE.2017-28

[4] S. Spaulding, H. Chen, S. Ali, M. Kulinski, and C. Breazeal,“A Social Robot System for Modeling Children’s Word Pronunciation,” in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS ’18)*, Stockholm, Sweden, pp. 1658-1666, July 2018.

[5] G. J. Yeung, A. L. Bailey, A. Afshan, M. Q. Pérez, A.

Martin, S. Spaulding, ... and C. L. Breazeal, "Toward the Development of Personalized Learning Companion Robots for Early Speech and Language Assessment," in *Proceedings of 2019 Annual Meeting of the American Educational Research Association (AERA)*, Toronto, Canada, April 2019. https://doi.org/10.3102/1431402

[6] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, ... and T. Belpaeme, "Child Speech Recognition in Human-Robot Interaction: Evaluationsand Recommendations," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)*, Vienna, Austria, pp. 82-90, March 2017. https://doi.org/10.1145/2909824.3020229

[7] J.-K. Yoo and K.-M. Lee, "Comparison of Adult and Child's Speech Recognition of Korean," *Journal of the Korea Contents Association*, Vol. 11, No. 5, pp. 138-147, May 2011. https://doi.org/10.5392/JKCA.2011.11.5.138

[8] F. Claus, H. G. Rosales, R. Petrick, H.-U. Hain, and R. Hoffmann, "A Survey about Databases of Children's Speech," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, Lyon, France, pp. 2410-2414, August 2013. https://doi.org/10.21437/INTERSPEECH.2013-561

[9] S. Shahnawazuddin, N. Adiga, H. K. Kathania, and B. Tarun Sai, "Creating Speaker Independent ASR System through Prosody Modification Based Data Augmentation," *Pattern Recognition Letters*, Vol. 131, pp. 213-218, March 2020. https://doi.org/10.1016/j.patrec.2019.12.019

[10] W. Wang, Z. Zhou, Y. Lu, H. Wang, C. Du, and Y. Qian, "Towards Data Selection on TTS Data for Children's Speech Recognition," in *Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*, Toronto, Canada, pp. 6888-6892, June 2021. https://doi.org/10.1109/ICASSP39728.2021.9413930

[11] V. Kadyan, H. Kathania, P. Govil, and M. Kurimo, "Synthesis Speech Based Data Augmentation for Low Resource Children ASR," in *Proceedings of the 23rd International Conference on Speech and Computer (SPECOM 2021)*, St. Petersburg, Russia, pp. 317-326, September 2021. https://doi.org/10.1007/978-3-030-87802-3_29

[12] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad, "Voice Conversion Based Data Augmentation to Improve Children's Speech Recognition in Limited Data Scenario," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association (Interspeech 2020)*, Shanghai, China, pp. 4382-4386, October 2020. https://doi.org/10.21437/Interspeech.2020-1112

[13] D. K. Singh, P. P. Amin, H. B. Sailor, and H. A. Patil, "Data Augmentation Using CycleGAN for End-to-End Children ASR," in *Proceedings of the 29th European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, pp. 511-515, August 2021. https://doi.org/10.23919/EUSIPCO54536.2021.9616228

[14] N. Jia, C. Zheng, and W. Sun, "Speech Synthesis of Children's Reading Based on CycleGAN Model," *Journal of Physics: Conference Series*, Vol.1607, 012046, 2020. https://doi.org/10.1088/1742-6596/1607/1/012046

[15] R. Serizel and D. Giuliani, "Vocal Tract Length Normalisation Approaches to Dnn-Based Children's and Adults' Speech Recognition," in *Proceedings of 2014 IEEE Spoken Language Technology Workshop (SLT)*, South Lake Tahoe: NV, pp. 135-140, December 2014. https://doi.org/10.1109/SLT.2014.7078563

[16] G. Yeung, R. Fan, and A. Alwan, "Fundamental Frequency Feature Normalization and Data Augmentation for Child Speech Recognition," in *Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*, Toronto, Canada, pp. 6993-6997, June 2021. https://doi.org/10.1109/ICASSP39728.2021.9413801

[17] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving Children's Speech Recognition through out-of-Domain Data Augmentation," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech 2016)*, San Francisco: CA, pp. 1598-1602, September 2016. https://doi.org/10.21437/INTERSPEECH.2016-1348

[18] G. Chen, X. Na, Y. Wang, Z. Yan, J. Zhang, S. Ma, and Y. Wang, "Data Augmentation for Children's Speech Recognition: The "Ethiopian" System for the SLT 2021 Children Speech Recognition Challenge," arXiv: 2011.04547v1, November 2020. https://doi.org/10.48550/arxiv.2011.04547

[19] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech 2019)*, Graz, Austria, pp. 2613-2617, September 2019. https://doi.org/10.21437/Interspeech.2019-2680

[20] V. P. Singh, H. Sailor, S. Bhattacharya, and A. Pandey, "Spectral Modification Based Data Augmentation for

Improving End-to-End ASR for Children's Speech," arXiv:2203.06600v1, March 2022. https://doi.org/10.48550/arXiv.2203.06600

[21] T. Rolland, A. Abad, C. Cucchiarini, and H. Strik, "Multilingual Transfer Learning for Children Automatic Speech Recognition," in *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France, pp. 7314-7320, June 2022.

[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning (ICML '23)*, Honolulu: HI, pp. 28492-28518, July 2023. https://doi.org/10.48550/arXiv.2212.04356

[23] R. Jain, A. Barcovschi, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A Wav2Vec2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition," *IEEE Access*, Vol. 11, pp. 46938-46948, May 2023. https://doi.org/10.1109/ACCESS.2023.3275106

[24] J. Kunze, L. Kirsch, I. Kurenkov, A. Krug, J. Johannsmeier, and S. Stober, "Transfer Learning for Speech Recognition on a Budget," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*, Vancouver, Canada, pp. 168-177, August 2017. https://doi.org/10.18653/v1/W17-2620

[25] P. G. Shivakumar and P. Georgiou, "Transfer Learning from Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations," *Computer Speech & Language*, Vol. 63, 101077, September 2020. https://doi.org/10.1016/J.CSL.2020.101077

[26] OpenAI. Introducing Whisper [Internet]. Available: https://openai.com/research/whisper.

[27] J.-W. Kim and H.-Y. Jung, "End-to-End Speech Recognition Models Using Limited Training Data," *Phonetics and Speech Science*, Vol. 12, No. 4, pp. 63-71, December 2020. https://doi.org/10.13064/KSSS.2020.12.4.063

[28] AI-Hub. The Voice Data of Korean Children [Internet]. Available: https://aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=data&dataSetSn=540.

[29] AI-Hub. Conversational Voice (adults) [Internet]. Available: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=109.

[30] GitHub. Whisper/Model-Card.md at Main: Openai/Whispter [Internet]. Available: https://github.com/openai/whisper/blob/main/model-card.md.

[31] Hugging Face. Fine-Tune Whisper for Multilingual ASR with Hugging Face Transformers [Internet]. Available: https://huggingface.co/blog/fine-tune-whisper#fine-tuning-whisper-in-a-google-colab.

[32] S.-Y. Min, K.-H. Lee, D.-S. Lee, and D.-Y. Ryu, "A Study on Quantitative Evaluation Method for STT Engine Accuracy based on Korean Characteristics," *Journal of the Korea Academia-Industrial Cooperation Society*, Vol. 21, No. 7, pp. 699-707, July 2020. https://doi.org/10.5762/KAIS.2020.21.7.699

[33] C. Oh, C. Kim, and K. Park, "Building Robust Korean Speech Recognition Model by Fine-Tuning Large Pretrained Model," *Phonetics and Speech Science*, Vol. 15, No. 3, pp. 75-82, September 2023. https://doi.org/10.13064/KSSS.2023.15.3.075

[34] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio Augmentation for Speech Recognition," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech 2015)*, Dresden, Germany, pp. 3586-3589, September 2015. https://doi.org/10.21437/Interspeech.2015-711

손계원(Gyewon Son)

2020년 : 교통대학교(공학사 소프트웨어)

2020년~2023년: 국방부 육군 통신장교
2023년~현 재: 주식회사 비전21테크
※관심분야 : 인공지능, 신호처리

소준섭(Junseop So)

2017년 : 한양대학교 대학원
(컴퓨터공학 석사 중퇴)

2018년~2023년: 팡퍼니
2023년~현 재: 주식회사 비전21테크
※관심분야 : 인공지능, 생성모델

고주은(Jooeun Ko)

1999년 : 서강대학교(이학사 화학과)
2001년 : 서강대학교 대학원
(이학석사 화학과)

2001년~2008년: LG화학 기술연구원
2020년~현 재: 주식회사 비전21테크
※관심분야 : 자연어 처리

이진우(Jin-Woo Lee)

2021년 : 우송대학교
(공학사 컴퓨터 정보 보안)

2022년~현 재: 주식회사 비전21테크
※관심분야 : 사운드 엔지니어링, 3D 애니메이팅

이정록(JeongRok Lee)

2004년 : 한국항공대학교
(공학사 항공전자공학)
2008년 : 한국항공대학교 대학원
(석사 항공전자공학)

2004년~2012년: 한터 R&F
2012년~2014년: 스마트시스템
2014년~2016년: 자우텍
2016년~2017년: 우주씨앤티
2017년~2018년: 주식회사 다이노
2018년~2022년: 주식회사 엠젠솔루션
2022년~현 재: 주식회사 에이아이리더
※관심분야 : IoT(Internet of Thing), 인공지능(Artificial Intelligence), 스마트시티 안전(Smartcity Security)

신원선(Won-Sun Shin)

1997년 : 숙명여자대학교 대학원
(석사 전산학)

1996년~2008년: 숙명여자대학교 아태여성정보통신원
2015년~2020년: 공룡컴 평생교육원
2020년~현 재: 주식회사 비전21테크
2022년~현 재: 충북대학교 전파통신공학 박사과정
※관심분야 : 자동음성인식(ASR), 감정인식(Emotion Recognition), 대화시스템(Dialogue Systems) 등