pISSN 2005-8063
eISSN 2586-5854
2025. 3. 31.
Vol.17 No.1
pp. 51-61

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2025.17.1.051

Check for updates

Self-supervised learning-based Korean phoneme recognition for evaluating Korean pronunciation of non-native speakers*

Na Geng[1] · Heejung Na[1] · Jieun Park[1] · Jeong-Sik Park[2,**]

[1]*Department of English Linguistics, Hankuk University of Foreign Studies, Seoul, Korea*
[2]*Department of English Linguistics & Language Technology, Hankuk University of Foreign Studies, Seoul, Korea*

Abstract

To evaluate the Korean pronunciation of non-native speakers, it is essential to develop models capable of recognizing Korean phonemes and detecting pronunciation errors at the phoneme level. Self-supervised learning models, such as Wav2Vec2.0 and Whisper, which were trained on large-scale speech data, have demonstrated strong performance in Korean speech recognition. However, their phoneme recognition accuracy for non-native speakers may be limited because of the lack of labeled data reflecting the unique characteristics of non-native speech. In this study, we developed a Korean phoneme recognition model tailored for non-native speakers by fine-tuning the pretrained Whisper model with Korean language education data from AIHub. This dataset includes speech samples from non-native speakers of various nationalities. In particular, to address the issue of the low phoneme label accuracy in this corpus, we proposed a method to improve label quality by incorporating news data clearly articulated by native Korean news anchors with the AIHub data. The refined dataset was then used for further fine-tuning, resulting in improved phoneme recognition performance. Experiments on Korean phoneme recognition with non-native speakers showed a significant increase in accuracy compared to models trained without the refined data.

Keywords: self-supervised learning, non-native speakers, Whisper, Korean phoneme recognition, Korean pronunciation ev aluation

1. 서론

본 연구는 비원어민 한국어 학습자를 대상으로 음소 단위의 발음 오류를 자동으로 탐지하여 정밀한 피드백을 제공할 수 있는 음소 인식 모델을 개발하는 것을 목표로 한다. 한국어는 자음과 모음의 조합으로 음절이 형성되는 독특한 음운 체계를 가지고 있으며, 이는 비원어민 학습자에게 있어 발음 학습의 어려움을 가중시키는 주된 요인이 된다. 특히, 모국어와의 음운적 차이

* This work was supported by Hankuk University of Foreign Studies Research Fund and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2020R1A2C1013162).
** parkjs@hufs.ac.kr, Corresponding author
Received 24 January, 2025; Revised 20 February, 2025; Accepted 21 February, 2025

로 인해 대치, 생략, 첨가 등의 발음 오류가 자주 발생하며, 이는 발음 정확도와 의사소통 능력을 저하시킨다(Kim, 2006; Yang & Chung, 2014). 기존에 소개된 발음 평가 소프트웨어는 원어민인 한국인의 발음을 기준으로 학습자의 발음을 평가하였으며 비원어민 발음의 억양, 발화 스타일, 발음 특성을 충분히 반영하지 못하였다(Kim et al., 2022a). 이에 따라 비원어민 학습자의 발음을 효과적으로 평가하고 교정할 수 있는 기술적 도구의 필요성이 제기되고 있다.

음소 인식기는 발음 오류를 정밀히 분석하고 교정 피드백을 제공하는 데 핵심적인 역할을 한다. 영어의 경우에도 비원어민 영어 학습자의 발음 평가에 음소 인식 기술이 중요하게 활용되고 있다(Korzekwa et al., 2021; Leung et al., 2019). 음성 인식처럼 음소 인식 또한 대용량의 학습 데이터가 필요한데, 음소 단위의 정확한 레이블을 갖춘 데이터를 확보하는 것은 현실적으로 어렵다. 또한, 기존 음성 인식기와 달리 텍스트 문맥보다 발음 정보를 우선적으로 다뤄야 하며, 데이터 정제와 모델 설계 과정에서 상당한 기술적 어려움이 존재한다. 이러한 문제를 해결하기 위해 본 연구에서는 음성 인식 모델로 잘 알려진 Whisper 모델(Radford et al., 2023)의 사전 학습된 능력을 기반으로, 비원어민 한국어 화자의 발음을 반영하도록 모델을 미세 조정(fine-tuning)한다. 이를 위해 AIHub에서 제공하는 비원어민 발화 데이터와 한국어 데이터를 활용하여 데이터를 레이블링하고, 발음 정보를 정제하면서 반복 학습을 통해 비원어민의 한국어 음소 인식에 최적화된 모델을 구축한다.

Whisper 모델은 다국어 데이터를 기반으로 사전 학습된 음성 인식 모델로, 텍스트 변환 과정에서 문맥 의존적인 언어 모델을 사용하지 않고 발음 자체를 중심으로 작동한다. Kaldi(Ravanelli et al., 2019)나 ESPnet(Watanabe et al., 2018)과 같은 기존 오픈소스 모델 대비 우수한 성능을 보여, 비원어민 화자의 발음 변이를 안정적으로 처리함으로써 발음 오류를 효과적으로 탐지할 수 있을 것으로 기대된다.

본 연구는 Whisper 모델을 기반으로, 비원어민 학습자의 발음을 음소 단위로 인식하며, 발음 교정에 효과적인 피드백을 제공하는 시스템 개발을 목표로 한다. 이를 통해 한국어 학습자가 자신의 발음 오류를 명확히 이해하고 교정할 수 있는 실질적 지원 도구를 제공하고자 한다.

2. 관련 연구

2.1. 음소 인식 연구

음소 인식 기술은 학습자의 발화된 음소와 기준 음소 간의 차이를 분석하여 오류의 위치와 유형을 명확히 제시한다(Kim et al., 2022b; Leung et al., 2019; Peng et al., 2021; Zahran et al., 2023). 이를 통해 비원어민 학습자로 하여금 자신의 발음 오류를 이해하고 교정할 수 있는 효율적인 학습 효과를 제공한다.

음소 인식은 입력된 음성 데이터에서 발음을 구성하는 최소 단위인 음소(phoneme)를 식별하는 기술이다. 초기 연구는 HMM-GMM(hidden markov model-gaussian mixture model) 기반의 통계적 모델을 사용하여 음소 간의 관계를 모델링하였으나, 이 방법은 복잡한 음향 패턴을 처리하는 데 한계가 있었다(Golowich & Sun, 1998; Kannadaguli & Bhat, 2015). 이후 딥러닝 기반의 모델 학습 기법이 도입되면서 음소 인식 기술은 크게 발전하였다. CNN(convolutional neural network), RNN(recurrent neural network), transformer와 같은 신경망 모델은 음소 간의 복잡한 관계를 학습하여 기존 통계적 모델의 한계를 극복하였다. 특히, transformer 모델은 self-Attention 메커니즘을 통해 긴 문맥 정보를 효과적으로 학습함으로써 음소 인식의 정확도를 크게 향상시켰다(Vaswani et al., 2017). 하지만 이러한 딥러닝 기반 접근법은 대규모의 레이블링된 데이터에 크게 의존한다는 한계를 갖는다.

2.2. 자기 지도 학습 기반 음소 인식

기존의 딥러닝 기반 음소 인식은 레이블링된 데이터에 대한 높은 의존성으로 인해 상대적으로 데이터가 부족한 비주류 언어나 비원어민 학습자의 발음 평가에 적용이 어려웠다. 이를 해결하기 위해 자기 지도 학습(self-supervised learning, SSL) 기술이 대안으로 제시되었다(Baevski et al., 2020; Kim & Kang, 2021). SSL은 레이블이 없는 대규모 데이터를 활용하여 음향적 특징을 학습하며, 특히 음소 간의 구조적 관계와 복잡한 음향 패턴을 처리하는 데 효과적이다. 이러한 특성으로 인해 SSL은 다국어 음성 인식에 널리 사용되면서 비주류 언어에서도 음소 인식 성능 개선을 위해 활용되기 시작하였다. Yang et al.(2022)는 Wav2Vec2.0 기반의 모멘텀 가상 레이블링(momentum pseudo-labeling)을 활용하여 데이터 레이블링 비용을 절감하고 발음 오류 탐지에서 기존 모델 대비 20% 이상 높은 성능을 기록하였다. Xu et al.(2021)은 다국어 음성 데이터를 활용하여 제로샷(zero-shot) 학습을 통해 레이블이 없는 데이터를 처리하고 조음 특징(articulatory features)을 학습하여 기존 방법 대비 약 30% 높은 음소 인식 정확도를 달성하였다. 이러한 연구는 SSL 기반 모델이 저자원 환경에서도 효과적인 음소 인식 성능을 달성할 수 있음을 보여준다.

2.3. 비원어민 한국어 음소 인식

음소 인식 연구는 영어와 같은 주류 언어를 중심으로 활발히 진행되어 왔다. 음소 단위의 레이블을 갖춘 풍부한 음성 데이터를 활용하여 HMM-GMM 기반의 초기 모델부터 딥러닝 기반의 최신 모델까지 다양한 연구 결과가 소개되었다. 반면, 한국어 음소 인식 연구는 상대적으로 제한적이며, 특히 비원어민을 대상으로 한 연구는 기초 수준에 머물러 있다.

Ryu et al.(2016)은 음향 모델을 활용하여 학습자의 발음을 강제 정렬(forced alignment)하고 이를 기반으로 음향적 특징을 분석하여 발음 평가 모델을 개발하였다. 이 연구는 다양한 언어적 배경을 가진 비원어민 학습자의 발음 오류를 정량적으로 평가하고 탐지하는 데 성공하며 음소 인식의 정확성을 높였다. Lin & Wang(2023)은 도메인 적대적 학습(domain adversarial training)을 활용하여 다양한 억양을 가진 학습자의 발음을 평가

하고 억양 간 일반화 성능을 향상시켰다. 이는 비원어민 학습자의 발음 데이터를 정밀히 분석하고 음소 인식 성능을 강화하는 데 기여하였다. Jang et al.(2023)은 음향 신호와 시각 데이터를 결합한 멀티모달 학습 모델을 제안하여 발음 오류 탐지의 정확성을 개선하였다.

비원어민 대상의 한국어 음소 인식 및 발음 평가와 관련된 연구가 제한적인 이유는 다음과 같다. 한국어는 자음과 모음의 결합으로 이루어진 독특한 음운 체계를 가지고 있으며, 받침 자음과 같은 추가적인 음운 요소가 포함되어 있어 음소 인식에서 영어와는 다른 접근이 필요하다. 이러한 한국어의 특성으로 비원어민 학습자의 발음 오류 패턴이 다양하게 나타나며, 이를 반영하여 발음 오류를 효과적으로 처리할 수 있는 연구가 필요하다.

비원어민 화자 대상의 한국어 음소 인식이 어려운 주된 이유는 음운 체계의 차이, 문맥 의존성, 데이터 부족에서 비롯된다. 첫째, 음운 체계의 차이는 한국어의 복잡한 음운 구조에서 기인한다. 자음과 모음의 결합, 받침 자음의 존재 등은 비원어민 학습자에게 발음 학습의 어려움을 가중시키며, 대치 오류(예: ‘ㄹ’을 ‘ㄴ’으로 대치), 생략 오류(받침 자음의 생략), 첨가 오류(불필요한 모음 추가)와 같은 발음 오류가 빈번히 발생하는 요인이 된다. 둘째, 문맥 의존성을 배제한 연구가 필요하다. 기존의 음소 인식 시스템은 인식 결과로 출력되는 텍스트 상의 문맥을 활용하여 오류를 탐지하고 교정하는 데 중점을 두지만, 비원어민의 발화에서는 발음 자체를 중점적으로 고려하여 오류를 탐지하는 기술이 필요하다. 셋째, 원어민 화자에 비해 비원어민 화자의 한국어 음성 데이터는 상당히 부족하며, 특히 음소 단위의 레이블을 갖춘 데이터는 찾기 어렵다. 이러한 데이터 부족 문제는 딥러닝 모델 학습에 어려움을 가중시키며 인식 성능의 한계로 이어진다.

본 연구에서는 학습 데이터의 부족으로 인식 성능 개선에 어려움을 겪는 다양한 패턴 인식 문제에서 최근 널리 사용되고 있는 자기 지도 학습 기법을 기반으로 비원어민 대상의 한국어 음소 인식 성능을 개선하는 방법을 제안한다. 이를 위해 먼저 본 연구에서 핵심적으로 사용하는 Whisper 모델에 대해 소개한다.

2.4. Whisper 모델의 개요와 특징

Whisper 모델은 OpenAI에서 개발된 범용의 음성 인식 모델로, 음성-텍스트 변환(STT), 다국어 번역, 음성 감지와 같은 다양한 작업을 수행할 수 있는 모델로 알려져 있다(Radford et al., 2023). Transformer 기반의 종단형(end-to-end) 구조를 채택하여 음성 데이터를 입력받아 텍스트로 직접 변환하며, 기존 파이프라인 모델보다 효율적이고 강인한 음성인식 모델이다. 약 680,000시간 이상의 대규모 다국어 데이터를 학습한 Whisper는 다양한 언어와 억양, 소음 환경에서도 우수한 성능을 보여 비원어민 화자의 발음 변이 또한 효과적으로 처리할 수 있는 가능성을 지닌다.

Whisper 모델은 다중 언어 데이터를 학습하면서 다양한 음소 패턴을 포괄적으로 다룰 수 있는 능력을 가지고 있다. 특히, 비원어민 학습자의 발음 변이를 반영하기 위해 추가적인 미세 조정을 통해 특정 언어에 최적화된 성능을 제공할 수 있다. 이는 복잡한 음운 체계를 가진 한국어 음소 인식에 Whisper 모델이 효과적으로 활용될 수 있음을 나타내는 특징이다. Whisper는 기존 음성 인식 모델이 음성-텍스트 전환에 초점을 맞춘 것과 달리, 음소 중심의 분석 및 인식에도 강점을 보인다.

한국어 음소 인식에서 Whisper 모델의 활용 가능성은 이미 여러 연구를 통해 입증되었다. 가령, Oh et al.(2023)은 AIHub에서 제공된 약 1,000시간 분량의 한국어 음성 데이터를 사용해 Whisper 모델을 fine-tuning하여 문자 오류율(character error rate, CER)을 크게 개선하였다.

대규모의 다국어 데이터를 통해 학습된 Whisper 모델은 한국어와 같은 비주류 언어에서도 뛰어난 성능을 보였으며, 특히 소음 환경, 억양 차이, 불규칙한 발화 패턴에서도 안정적인 성능을 나타냈다.

Whisper 모델은 대규모 데이터 학습과 다국어 지원 구조를 바탕으로, 비원어민 대상의 한국어 음소 인식 문제에서 다양한 발음 변이와 음소 단위의 분석을 처리하는데 핵심적인 역할을 할 것으로 예상된다. 본 연구는 Whisper 모델의 이러한 강점을 활용하여 비원어민 대상의 한국어 음소 인식에 최적화된 모델을 개발하는 것을 목표로 한다. 이를 위해 한국어 데이터를 사용하여 Whisper 모델을 추가적으로 미세 조정(fine-tuning)하고, 데이터 정제 및 개선 과정을 통해 비원어민의 발음을 처리하는 데 적합한 솔루션을 제안한다.

3. 자기 지도 학습 기반 비원어민 대상 한국어 음소 인식 모델 구축

본 장에서는 자기 지도 학습 모델을 사용하여 비원어민 대상의 한국어 음소 인식에 최적화된 모델을 구축하기 위해 수행한 연구 내용을 기술한다. 비원어민 및 원어민의 한국어 음성 데이터를 구축하고 구축된 데이터의 레이블링과 정제를 통해 모델을 구축하는 과정, 그리고 음소 단위 인식을 위해 G2P(grapheme to phoneme) 도구(Park, 2019)를 이용하여 글자 기반 표기를 발음 기반의 표기로 변환하는 방법을 소개한다.

3.1. 데이터셋 구축

비원어민 화자의 한국어 발화를 발음대로 인식할 수 있는 효과적이고 정확한 한국어 음소 인식 모델을 구축하기 위해서는 정확하게 레이블링된 다양한 데이터를 기반으로 학습이 이루어져야 한다. 그러나 비원어민 화자의 발화를 정확하게 레이블링한 데이터가 부족하기 때문에, 본 연구에서는 이를 보완하기 위해 두 가지 유형의 데이터를 수집하였다.

3.1.1. 비원어민 화자 발화 데이터

첫 번째 데이터셋은 다양한 언어권의 비원어민, 즉 L2 화자의 발화 데이터를 포함한다. 이 데이터는 AIHub에서 제공하는 비원어민 한국어 교육용 음성 데이터(이하 'L2 화자 데이터'로 칭함)로, 중국·일본을 제외한 아시아권(이하 ‘아시아권’), 중국·일

본권(이하 '중일권'), 유럽권, 그리고 영어를 모국어로 사용하는 영어권 데이터로 구성되어 있다(AIHub, 2022a, 2022b, 2022c, 2022d).

AIHub의 L2 화자 데이터는 발음 수준이 다양한 단어 및 문장 발화로 구성되어 있으며, 레이블은 음소와 단어 레이블을 제공하고 있다. 예를 들어, [나문닙]으로 발음되는 단어의 경우 단어 레이블은 '나뭇잎'으로, 음소 레이블은 'ㄴㅏㅁㅜㄴㄴㅣㅂ'으로 표기되어 있다.

3.1.2. 한국인 화자 발화 데이터

두 번째 데이터는 한국인 화자의 표준 발화 데이터이다. 이 데이터는 AIHub에서 제공하는 '뉴스 대본 및 앵커 음성 데이터'(이하 '뉴스 데이터'로 칭함; AIHub, 2022e)로 구성되어 있으며, 한국인 아나운서가 표준 발음에 맞추어 명료하게 발화한 뉴스 대본 리딩 데이터를 포함한다.

이 데이터를 사용하는 목적은 한국어의 정확한 발음 기준을 제시하여 모델이 한국어 음소를 더욱 정교하게 학습하도록 만들기 위함이다. 앞서 소개한 L2 화자 데이터의 경우, 다양한 발음 변이를 나타내는 비원어민 화자 음성의 특성 상 한국어 레이블이 부정확한 문제가 있는데, 뉴스 데이터는 L2 화자 데이터에서 발생하는 레이블 불일치 문제를 보완하고, 모델이 원어민 발음을 기준으로 학습할 수 있도록 균형을 맞추는 역할을 한다.

이처럼, 본 연구에서는 비원어민 화자와 한국인 화자의 발화를 모두 포함하는 데이터셋을 사용함으로써, 비원어민 화자의 다양한 발음 특성을 정확히 반영하는 한국어 음소 인식 모델을 효과적으로 구축할 수 있을 것으로 기대한다.

3.2. G2P(Grapheme to Phoneme)을 사용한 발음 기반 레이블 변환

Whisper와 같이 음성 인식을 목적으로 개발된 모델의 경우 인식 결과가 발음 표기가 아닌 어휘 사전에 정의된 단어 열로 출력되는 경향이 있다. 발음 평가를 위한 음소 인식을 대상으로 하는 본 연구에서는 발음 기반의 표기가 적합하기 때문에 음성 인식 모델이 출력한 글자(어휘 사전) 기반 표기를 실제 발음으로 변환하여 레이블을 수정하는 과정이 필요하다. 이를 위해 본 연구에서는 G2P 도구를 사용하여 발음 기반 레이블을 생성하였다. 이 과정에서 연음, 유음화, 비음화와 같은 한국어 발음 규칙을 적용하여, 가령 '좋겠다'는 '조켄따'로, '막내'는 '망내'로 변환된다.

G2P 변환의 목적은 음소 인식 오류율, 즉 PER(phoneme error rate) 계산의 정확도를 높이고, 문맥 의존성을 최소화하면서 발음 평가를 용이하게 하는 데 있다. 우선, G2P 변환은 PER 계산의 정확성을 높이는 데 기여한다. 발음이 사전적 표기와 일치하지 않는 경우에도 G2P를 적용함으로써 잘못된 오류율 계산을 방지할 수 있다. 또한, G2P는 문맥 의존성을 감소시키는 데에도 중요한 역할을 한다. 일반적으로 음성 인식 모델은 언어 모델과 함께 학습되기 때문에 문맥이나 언어적 규칙에 따라 발음이 왜곡되는 문제가 발생할 수 있다. 예를 들어, [망내]라는 발음의 단어를 문맥 규칙에 따라 '막내'로 출력하는 경우가 이에 해당한다. G2P 변환을 활용하여 생성된 발음 기반의 레이블을 통해 모델의 미세 조정(fine-tuning)을 수행한다면, 음성 인식을 위해 구축된 사전 학습 모델에 존재하는 문맥적 개입을 줄이고 발음 그대로의 음소 인식 성능을 기대할 수 있다. 끝으로, G2P를 통해 구축된 음소 기반 토큰은 다양한 억양, 방언, 그리고 발음 변이를 효과적으로 포괄할 수 있어 발음에 억양이 있는 비원어민의 개인화 모델에서도 높은 유연성을 제공할 수 있다.

본 연구에서 미세 조정에 사용되는 데이터 중 3.1.2절에서 소개한 뉴스 데이터의 경우 음소 단위의 레이블이 제공되지 않아 G2P 작업을 통해 발음 기반으로 레이블을 변환함으로써 모델이 한국어 발음 규칙을 보다 정확하게 학습할 수 있도록 한다. 학습 데이터에 G2P를 적용한 후, word delimiter '|', unknown word '[UNK]', 및 padding token '[PAD]'를 포함하여 총 1,245개의 토큰을 생성한다.

3.3. 자기 지도 학습 모델 기반 데이터 레이블 정제

2.3절에서 기술한 바와 같이 비원어민 화자의 한국어 음소 인식 연구가 제한적인 이유 중의 하나는 음소 단위의 레이블을 갖춘 비원어민 한국어 음성 데이터가 부족하기 때문이다. 본 연구에서 미세 조정에 활용하고자 하는 AIHub의 L2 화자 데이터는 음소 단위 레이블이 제공되는 자료이지만, 음소 레이블에 다량의 오류가 포함되어 원 자료를 그대로 사용할 때 인식 성능이 저하되는 문제가 확인되었다. 이와 관련된 실험 결과는 4장에서 소개한다. 따라서, 본 연구에서는 미세 조정에 활용되는 음성 데이터를 보다 효과적으로 모델 학습에 반영하기 위해, 레이블링 정제 과정을 수행한다.

3.3.1. L2 화자 데이터의 레이블 정제

음소 단위의 레이블을 수정하는 툴이 존재하지만 대부분 수동으로 청음하면서 개별 음소 단위로 작업하는 방식이기 때문에 본 연구에서는 대량의 L2 화자 음성 데이터의 음소 레이블을 자동으로 수정하는 방법을 제안한다. 음소 단위의 자동 레이블링을 위해서는 한국어 음소 인식을 전문적으로 수행하는 모델이 필요하나 음성 인식에 비해 음소 인식 모델은 찾기 어렵다. 따라서, 본 연구에서는 성능이 좋은 음성 인식 모델에서 1차 출력된 결과(단어열)로부터 G2P를 통해 발음을 표시하는 레이블을 생성한다.

다양한 사전 학습(pre-trained) 음성 인식 모델 중 대표적인 Whisper(Radford et al., 2023)와 Wav2vec2.0(Kim & Kang, 2021) 모델을 비교하여 가장 실제 발음에 가까운 출력을 제공하는 모델을 선택하였다. 실험 결과, Whisper 모델이 다른 모델에 비해 비원어민의 실제 발음을 잘 반영하는 경향을 보였다. Wav2Vec2.0과 같은 일반적인 음성 인식 모델은 언어 모델의 영향을 받아 어휘 사전에 존재하는 단어 열을 결과로 출력하는 경향이 있는 반면, Whisper 모델은 문법에 맞지 않거나 사전 외 단어 등 오류가 포함된 발화에 대해서도 실제 발음대로 결과를 출력하는 경향을 보인다. 또한, Wav2Vec2.0 계열의 모델은 지정

된 언어의 원어민 데이터만으로 모델 학습이 진행된 반면, Whisper 계열의 모델은 대규모의 다국어 데이터를 통해 학습이 진행되어 Whisper 계열의 모델이 비원어민 화자의 음성을 잘 처리하는 경향을 보인다. 예를 들어, “구하다”라는 단어를 [쿠하다]라고 발성한 비원어민 화자의 음성에 대해 Wav2vec2.0 모델은 문맥 정보를 반영하여 “구하다”로 출력하는 반면, Whisper 모델은 “쿠하다”로 인식한다. 이 같은 모델 특성을 반영하여 본 연구에서는 L2 화자의 발음을 그대로 레이블링하는데 유리한 모델로 Whisper 계열의 모델을 사용한다.

3.3.2. 사전 학습 모델 기반의 유효 데이터 선별 및 1차 레이블 정제

단어 단위의 음성과 문장 단위의 음성이 혼재된 L2 화자 데이터는 발화 길이의 편차가 크기 때문에, 전체 데이터를 사용하는 대신 2초에서 9초 사이의 데이터 중 약 200만 개(이하 ‘L2_2M’으로 칭함)를 선별하여 실험을 진행한다.

Whisper 모델을 통해 생성된 레이블에는 숫자, 한국어 외의 다른 언어, 인식 오류 등 여러 종류의 오류가 포함되어 있어, 이를 정제하기 위해 PER을 기준으로 필터링한다. 구체적으로, 음절 단위로 출력된 결과를 음소 단위로 변환하여 PER이 0% 초과 40% 이하인 데이터를 선별하여 해당 데이터와 레이블을 최종 데이터셋에 포함시킨다. PER이 0%인 데이터를 제외시킨 이유는, 언어 모델을 사용하여 훈련된 Whisper 모델은 문맥 의존성이 존재하므로 발음 오류가 있음에도 불구하고 정확한 레이블로 인식(즉, PER 0%)하는 경우가 빈번하게 발견되었기 때문이다. 또한, PER이 40% 이하인 데이터를 선택함으로써 일정 수준의 정확도를 보이는 충분한 양의 데이터를 확보할 수 있었다. 이러한 기준을 적용하여 약 68,000개의 L2 화자 데이터를 선별하고 1차 레이블 정제 결과를 획득하였다.

3.3.3. 미세 조정 모델 기반의 유효 데이터 선별 및 2차 레이블 정제

1차 레이블 정제에 사용한 사전 학습 모델은 비원어민 화자의 음성 특성이 반영되지 않은 모델로 정제된 레이블에 여전히 오류가 존재한다. 따라서, 1차 정제 과정에서 선별된 L2 화자 데이터와 원어민(한국인)의 정확한 발성 특성을 보이는 뉴스 데이터를 사용하여 사전 학습 모델을 미세 조정함으로써 화자(비원어민)와 언어(한국어)의 특징을 잘 표현하는 모델을 구축하고 이를 이용하여 2차 레이블 정제를 진행한다.

두 종류의 미세 조정 데이터, 즉 L2 화자 데이터와 뉴스 데이터의 조합을 통해 표 1과 같이 4개의 미세 조정 모델을 생성할 수 있다. ‘L2’ 모델과 ‘Kor’ 모델은 각각 비원어민 화자 데이터(L2 화자 데이터)와 한국어 원어민 발화 데이터(뉴스 데이터)만 사용하여 구축한 모델이다. ‘Kor-L2’와 ‘L2-Kor’ 모델은 두 종류의 데이터를 순차적으로 미세 조정에 사용하여 구축한 모델을 뜻한다. 이렇게 구축한 모델을 ‘모델1.0’이라 부르기로 한다.

표 1. 미세 조정 데이터 조합에 따른 모델 종류
Table 1. The model types categorized based on the composition of fine-tuning data

모델명	1차 미세 조정 데이터	2차 미세 조정 데이터
L2	L2 화자 데이터	-
Kor	뉴스 데이터	-
Kor-L2	뉴스 데이터	L2 화자 데이터
L2-Kor	L2 화자 데이터	뉴스 데이터

4개의 미세 조정 모델은 사용된 데이터의 특성에 따라 모델의 특성이 다르며 이로 인해 음소 인식 성능에 차이를 보인다. 본 연구진은 특성이 다른 모델임에도 불구하고 각 모델에서 동일한 출력 결과를 나타내는 데이터는 오류가 적고 신뢰성이 있는 자료라 판단한다. 이러한 가정을 기반으로 여러 모델에서 동일한 출력 결과를 나타내는 자료를 선별하고 선별된 자료를 미세 조정에 사용하여 구축된 모델을 통해 2차 레이블 정제를 진행한다.

4개 모델의 출력 결과를 모두 비교하여 데이터를 선별할 경우, 선별된 데이터의 양이 매우 부족하여 미세 조정 모델을 구축하는데 문제가 발생한다. 따라서 4가지 모델 중 성능이 가장 좋지 않은 ‘L2’ 모델을 제외한 나머지 3가지 모델(‘Kor’, ‘L2-Kor’, ‘Kor-L2’)을 대상으로 ‘L2-2M’의 각 데이터를 인식하고 그 결과를 비교하여 데이터를 선별하였다.

2차 레이블 정제를 위한 데이터 선별은 단어 데이터와 문장 데이터로 구분하여 진행한다. 단어 데이터의 경우 3가지 모델의 출력 결과가 모두 동일한 데이터를 선별한다. 이때, 단어 데이터의 균형을 유지하기 위해 중복된 레이블을 갖는 데이터의 수를 16개로 제한한다. 중복 레이블을 제거하는 이유는, 특정 단어의 데이터들에 대해 모델이 오류 발음을 일관되게 예측하는 경우 해당 단어의 잘못된 특성을 모델이 그대로 학습할 가능성이 높기 때문이다. 가령, ‘시계’라는 단어를 [시기]와 같은 오류 발음으로 예측하는 빈도가 비정상적으로 높은 경우, 모델은 ‘시계’를 [시기]로 학습했을 가능성이 크다고 볼 수 있다. 따라서, 미세 조정 데이터셋에서 단어의 빈도를 균형 있게 조정함으로써 모델이 특정 오류 예측으로 편향되는 문제를 줄일 수 있다.

문장 데이터의 경우 단어 데이터와 달리, 세 모델의 출력 결과가 정확히 일치하는 경우가 많지 않다. 따라서, 음절 단위의 인식률인 CER을 기준으로 데이터를 선별한다. 구체적으로, 각 모델의 출력 결과를 CER로 계산하여 모든 모델에서 15% 이하의 CER을 나타내는 데이터를 선별한다.

위의 과정을 통해 총 22,000개의 L2 화자 데이터(이하 ‘L2_22k’으로 칭함)를 선별하였다. 이후 선별된 데이터를 사용하여 다시 한번 미세 조정을 수행하여 새로운 모델을 구축하고(이를 ‘모델2.0’이라 칭함), 이 모델을 통해 2차 레이블 정제를 진행한다.

3.3.4. 추가 레이블 정제

미세 조정 모델을 사용하여 유효한 데이터를 선별하고 이를 통해 다시 구축한 미세 조정 모델을 사용하여 정제된 음소 레이

블의 정확도는 상당히 개선되었다. 이와 관련된 상세한 실험 결과는 4장에서 소개한다. 본 연구진은 이러한 과정을 다시 반복하여 수행함으로써 레이블의 정확도가 더욱 개선되고 음소 인식 모델 성능 또한 개선될 것이라 판단하여, 동일한 과정으로 두 차례 추가적인 레이블 정제 작업을 진행한다. 단계를 진행하면서 정제된 레이블에 의해 미세 조정 모델 또한 갱신되며, 갱신 모델을 구분하기 위해 '모델 3.0', '모델 4.0'과 같이 명명한다. 표 2는 총 5 차례의 레이블 정제 단계를 진행하면서 각 단계에서 사용된 모델, 선별된 데이터셋의 크기 및 선별된 데이터를 사용하여 생성된 미세 조정 모델을 정리한 것이다.

표 2. 단계별 미세 조정 데이터와 레이블 정제에 따른 모델 갱신
Table 2. Fine-tuning model update according to sequential data selection and label refinement

레이블 정제 단계	정제에 사용한 모델	선별된 L2 화자 데이터셋 크기 (k)	생성된 모델
1차	Whisper-large	68	모델1.0
2차	모델1.0	22	모델2.0
3차	모델2.0	17	모델3.0
4차	모델3.0	15	모델4.0
5차	모델4.0	16	모델5.0

각 단계 별로 선별된 데이터셋은 미세 조정 모델을 갱신하는데 사용되는데 데이터셋의 규모가 적을 경우 모델의 정확도에 문제가 발생하므로 3차 이후의 정제 단계에서는 데이터 선별에 사용되는 기준(가령, CER 성능 기준 등)을 변경하면서 데이터셋의 크기를 균등하게 유지한다. 5차 이후에는 선별된 데이터셋의 크기가 거의 변하지 않아 더 이상의 정제는 의미가 없는 것으로 판단하여 레이블 정제는 5차까지 수행한다. 단계 별로 정제된 레이블과 선별된 데이터를 통해 갱신된 모델의 성능은 다음 장에서 소개한다.

4. 실험 결과

이 장에서는 3장에서 기술한 레이블 정제 단계에 따라 생성된 모델의 유효성을 검증하기 위해 음소 인식 실험을 수행한 결과를 소개한다. 각 실험에 사용한 평가용 데이터는 모델 학습에 사용되지 않은 800개의 L2 화자 데이터와 400개의 한국어 원어민 화자 데이터로 고정하여 성능 평가를 진행하였다.

4.1. 미세 조정 데이터 유형 별 모델 성능 분석

3.1.1절에서 기술한 바와 같이 본 연구에 사용한 미세 조정 데이터는 두 종류(L2 화자 데이터, 뉴스 데이터)이며, 데이터의 조합에 따라 구축된 미세 조정 모델의 성능을 먼저 분석하였다. 표 1에 기술한 네 종류의 모델('L2', 'Kor', 'Kor-L2', 'L2-Kor') 외에 두 데이터셋을 동시에 사용하여 구축한 미세 조정 모델('Combined')을 추가하여 총 다섯 개의 모델의 성능을 비교하였다. 각 모델은 레이블 정제 단계 별로 선별된 데이터셋에 따라 성능 차이를 보였으며, 가장 좋은 성능을 나타내는 데이터셋과 이를 사용하여 구축된 모델을 선정하여 비교하였다. 표 3은 두 종류의 평가 데이터(800개의 L2 화자 데이터와 400개의 한국인 화자 뉴스 데이터) 각각에 대한 5가지 모델의 성능을 정리한 것이다.

'baseline'은 Whisper 모델을 사용하여 1차로 선별한 68k 규모의 L2 화자 데이터(L2_68k)로 구축한 미세 조정 모델이며, L2 화자 데이터에 대해 12.5%의 PER을 나타내며 다른 모델에 비해 가장 낮은 성능을 보였다. 이에 비해, 3차 레이블 정제 단계에서 선별된 17k 규모의 L2 화자 데이터(L2_17k)로 구축된 미세 조정 모델은 5.18%로 눈에 띄게 성능이 개선되었는데, 이는 단계별 정제 과정을 통해 수정된 L2 화자 데이터의 레이블의 정확도가 개선됨에 따라 모델 또한 크게 개선된 결과로 해석할 수 있다. 반면, 한국인 화자의 뉴스 데이터를 사용하여 구축된 미세 조정 모델은 6.07%의 PER을 나타내 오히려 성능이 저하됨을 보였다.

표 3. L2 화자 데이터 및 뉴스 데이터 조합에 따른 미세 조정 모델 성능(phoneme error rate, PER) 비교
Table 3. Performance comparison of fine-tuning models based on combinations of L2 speakers and news data

모델 종류	미세 조정 데이터셋	L2 화자 데이터 성능 (%)	한국인 화자 뉴스 데이터 성능 (%)
baseline	L2_68k	12.50	20.71
L2	L2_17k	5.18	14.10
Kor	Kor_1k	6.07	**0.46**
Kor-L2	Kor_12k, L2_16k	4.68	8.86
L2-Kor	L2_17k, Kor_12k	3.87	1.33
Combined	L2_16k, Kor_12k	**3.22**	1.22

다음으로 L2와 Kor 데이터를 교차 사용한 Kor-L2와 L2-Kor 모델은 L2 화자 데이터에 대해 4.68%와 3.87%의 PER을 보여 L2 데이터만 사용한 모델에 비해 성능이 크게 개선되었다. 이는 한국인 화자가 정확하게 발성한 뉴스 데이터가 L2 화자 데이터에 부족한 한국어 음소 특성을 보완함으로써 L2 화자의 한국어 발음 특성이 보다 효과적으로 모델에 반영된 것으로 분석된다.

L2 화자 데이터와 뉴스 데이터를 동시에 사용하여 구축된 Combined 모델은 3.22%의 PER을 달성하여 모든 모델 중 L2 화자 데이터에 대해 가장 좋은 성능을 보였다. 이는 L2 화자 데이터와 한국어 원어민의 뉴스 데이터가 미세 조정에 함께 사용됨으로써, 모델이 L2 화자 발음의 다양한 변이뿐만 아니라 정확한 한국어 발음 특성을 효과적으로 학습에 반영하는 것으로 분석된다.

한국인 화자의 뉴스 데이터로 평가한 결과에서는 한국인 뉴스 데이터를 사용하여 구축된 미세 조정 모델이 0.46%의 PER을 나타내며 거의 100%의 인식률을 보였다. 이는 발음 변이가 거의 없는 원어민의 낭독체 뉴스 데이터의 특성이 학습에 반영되었기 때문이다. 그러나 'L2-Kor'와 'Combined' 모델 또한 각각 1.33%와 1.22%의 높은 성능을 보였다.

종합적으로 분석한 결과, L2 화자 데이터와 뉴스 데이터에서

모두 우수한 성능을 보이는 모델은 ‘Combined’ 모델로 확인되었다. 이는 두 데이터 유형 간의 차이를 효과적으로 학습하고, 다양한 음성 환경에서도 높은 일반화 능력을 발휘하기 때문으로 판단된다. 이 같은 결과는 L2 화자와 한국어 원어민 화자의 음소 인식에 모두 최적화된 모델을 구축하기 위해서는 데이터셋의 선택이 중요한 요소로 작용하며, 특히 다양한 발음 데이터를 균형 있게 결합하는 미세 조정 모델 구축 전략이 필요함을 입증하는 결과이다. 이후의 실험은 가장 최적의 성능을 보이는 ‘Combined’ 모델의 결과를 통해 분석한다.

4.2. 레이블 정제 단계별 모델 성능 비교

본 연구에서는 단계별로 레이블을 정제하고 그에 따라 미세 조정 모델을 갱신하는 과정을 반복적으로 수행하면서 비원어민에 최적화된 한국어 음소 인식 모델을 구축하는 방법을 제안하였다. 본 장에서는 레이블 정제 단계 별로 구축된 모델의 성능을 비교한다. 표 2에서 정리한 바와 같이, 다섯 단계의 레이블 정제 과정을 통해 다섯 개의 모델(‘모델1.0’ ~ ‘모델5.0’)이 생성

표 4. 미세 조정 모델에 적용된 데이터셋과 모델 별 성능(phoneme error rate, PER) 비교
Table 4. Performance comparison of models fine-tuned by different data sets

모델 번호	적용된 미세조정 데이터셋	L2 화자 데이터 성능 (%)	한국인 화자 뉴스 데이터 성능 (%)
모델1.0	Kor_12k, L2_68k	5.32	1.55
모델2.0	Kor_12k, L2_22k	4.47	1.39
모델3.0	Kor_12k, L2_17k	4.27	1.26
모델4.0	Kor_12k, L2_15k	4.08	1.31
모델5.0	Kor_12k, L2_16k	3.22	1.22

되었으며, 4.1장에서 성능이 가장 좋은 모델로 분석된 ‘Combined’ 모델을 사용하여 성능 평가를 진행하였다. 표 4는 각 미세 조정 모델 별로 적용된 데이터셋과 해당 모델의 성능을 정리한 것이다.

표 4에서 확인할 수 있듯이, 본 연구에서 제안한 여러 단계의 레이블 정제를 통해 L2 화자 및 뉴스 데이터에 대한 음소 인식 성능이 점진적으로 개선되었다. 첫 번째 미세 조정 모델(‘모델1.0’)의 성능은 L2 화자와 뉴스 데이터에서 각각 5.32%, 1.55%의 성능을 보였으나, 마지막 모델인 ‘모델5.0’에서는 3.22%, 1.22%로 크게 개선되었다. 이 같은 결과는 각 단계별로 생성되는 미세 조정 모델이 학습 데이터 레이블의 정확도를 높이고 개선된 레이블을 적용하여 다음 단계에서 구축되는 미세 조정 모델의 성능이 지속적으로 개선됨을 의미한다.

4.3. Dropout 조정을 통한 문맥 의존성 감소

본 연구에서는 Whisper 계열의 음성 인식 모델을 사전 학습 모델로 사용하여 이를 기반으로 미세 조정 모델을 구축하였다. 대부분의 사전 학습 모델이 음성 인식을 목적으로 구축된 모델이기 때문에 문맥 의존성을 보임으로써 음소 인식의 정확도가 저하되는 경향을 보인다. Whisper 모델은 텍스트 변환 과정에서 문맥 의존적인 언어 모델을 사용하지 않고 발음 자체를 중심으로 작동하는 모델로 알려져 있지만 음성 인식 모델의 특성으로 문맥 의존성이 발견되는 경우가 있다. 본 연구에서는 3.2장에서 소개한 G2P를 통해 문맥 의존성을 줄이는 방법을 제안하였으며, 추가적으로 모델 학습 과정에서 dropout 조정을 통해 문맥 의존성의 감소 여부를 살펴 보았다.

문맥 의존성은 모델이 주변 발음이나 단어의 언어적 맥락에 지나치게 의존하여 실제 발음을 왜곡하거나 부정확한 음소 인식 결과를 생성하는 요인이 된다. 신경망 모델 학습의 파라미터로 사용되는 dropout은 과적합(overfitting)을 방지하기 위한 정규화 기법으로 모델이 학습 데이터에 지나치게 특화되지 않도록 하기 위해 학습 과정에서 뉴런(혹은 노드)을 랜덤하게 비활성화하는 방법이다. 본 연구진은 데이터의 의존성을 줄이는 파라미터인 dropout 값을 조정함으로써 문맥 의존성을 줄일 수 있을 것으로 판단하였으며 이와 관련된 실험을 진행하였다. 표 4에서 살펴본 다섯 개의 모델 중 성능 개선의 여지가 있는 ‘모델2.0’을 대상으로 dropout 값 변화에 따른 성능 변화를 확인하였다.

표 5. Dropout 값의 변화에 따른 성능(phoneme error rate, PER) 비교
Table 5. Performance comparison according to dropout values

Dropout 값	L2 화자 데이터 성능 (%)	한국인 화자 뉴스 데이터 성능 (%)
0.1	4.52	1.16
0.2	4.91	1.30
0.3	4.54	1.26
0.4	4.59	1.33
0.5	4.46	1.10
0.6	4.65	1.39

표 5에 제시된 실험 결과에서, dropout 값에 따라 인식 성능이 변화함을 확인할 수 있으며 이는 dropout이 모델의 문맥 의존성에 일부 영향을 미치고 있음을 나타낸다. Dropout 값이 0.5일 때 두 종류의 평가 데이터에서 모두 가장 높은 성능을 보였으나, 0.2일 때 성능이 저하되는 특징을 보였다. Dropout 값과 성능 사이에 비례 또는 반비례의 특성은 보이지 않는데, 이는 dropout 비율이 지나치게 낮거나 높은 경우 문맥 의존성보다는 과적합이나 모델 학습 저하의 영향을 더욱 크게 받기 때문으로 판단된다.

4.4. 인식 결과 분석 및 개선 방안

지금까지는 PER 기준에 따라 전체 데이터의 성능을 일괄적으로 비교하였으나, 실제 인식 결과의 차이를 구체적으로 살펴보기 위하여 평가에 사용된 몇 가지 샘플 데이터를 정하여 레이블 정제 단계에 따라 생성된 두 종류 모델의 성능을 비교하였다. 즉, 다섯 개의 모델 중 가운데 위치한 ‘모델3.0’의 출력 결과와 가장 좋은 성능을 나타내는 ‘모델5.0’의 출력 결과를 비교하였다. 결과는 표 6과 같다.

표 6. ‘모델3.0’과 ‘모델5.0’의 출력 결과 비교
Table 6. Comparison of output results of ‘Model3.0’ and ‘Model5.0’

샘플 데이터 레이블	‘모델3.0’의 출력 결과	‘모델5.0’의 출력 결과
내일늗쯔시면안 됨니다	내일내주시면안 됨니다	내일눋쯔시면안됨 니다
칼등해겨를위에 어떤노려글하실 예정이심니까	갈등해겨를위에 어떤노려글하실 예정이십니다	갈등해겨를위의어 떤노려글하실예정 이심니까
괄력꽈공괄려긔 차이저미뭔지아 세요	궐력꽈공꿘녀긔 차이저미뭔지아 세요	괄력꽈공괄려긔차 이저미뭔지아세요
불마니이쓰면직 접가서말해봐	불마니이쓰면직 쩝와서말해봐	불마니이쓰면직첩 가서말해봐
옌날	연날	옌나에
왜이	회의	왜이

표 6에 제시된 6개의 샘플 데이터를 대상으로 두 모델의 출력 결과를 분석한 결과, ‘모델5.0’이 ‘모델3.0’보다 전반적으로 음소 열을 정확하게 출력하는 것으로 나타났다. 첫 번째 예에서 ‘모델3.0’은 [늗쯔]를 “내주”로 대체하며 문법적으로 자연스러운 결과를 출력함을 보였으나, 이는 실제 발음과 달리 음소를 잘못 인식한 결과이다. 반면, ‘모델5.0’은 동일한 발음을 “눋쯔”로 인식하여 실제 발음과 더 가까운 결과를 보였다.

두 번째 예에서, ‘모델3.0’은 [예정이심니까]를 “예정이십니다”로 인식하였는데, 문맥적으로는 적합하지만 실제 발음을 정확히 반영하지 못한 결과이다. 반면, ‘모델5.0’은 [위에]를 “위의”로 잘못 출력하며 발음 유사성으로 인해 오류가 발생했지만, [예정이심니까]를 정확히 출력하며 L2 화자의 실제 발화를 그대로 출력하였다.

세 번째 예에서 ‘모델3.0’이 [괄력]을 “궐력”, [공괄려]를 “공꿘녀”로 잘못 인식하였다. 인식 결과인 “궐력”은 문맥적으로 단어 ‘권력’의 올바른 발음이지만, 실제 발음 오류를 그대로 반영하지 않고, 문맥적 자연스러움이나 언어적 규칙에 따라 보정하려는 경향을 나타낸다. 즉, 모델이 실제 발음을 왜곡하고, 문맥적으로 더 자연스럽거나 표준적인 발음을 선택한 것으로 해석된다. 이러한 문제는 미세 조정 데이터에 포함된 레이블 오류 또는 모델의 문맥 의존성에서 기인했을 가능성이 크다. 반면, ‘모델5.0’은 L2 화자의 실제 발음을 정확히 인식하며 정답 레이블과 일치하는 결과를 보였다. 이는 모델 갱신에 따른 레이블 정제가 효과적으로 이루어졌음을 의미한다.

다음 예에서, ‘모델3.0’은 [직접가서]를 “직쩝와서”로 변형하여 출력하였다. [가]를 [와]로 대체한 것은 문맥 의존성을 과도하게 활용했음을 뜻한다. 반면, ‘모델5.0’은 [직접가]를 “직첩가”로 인식하였는데 이는 개선된 모델이 발음 경계 처리를 보다 효과적으로 다루고 있음을 보여주는 결과이다.

마지막 두 개의 예는 단어 사례이다. 다섯 번째 예에서 ‘모델3.0’은 [옌날]을 “연날”로 단순화하여 발음을 처리하였다. 이는 모델이 실제 발음의 음소적 복잡성을 반영하는 대신, 단순화된 발음으로 대체한 결과로 해석된다. 반면, “옌나에”로 출력한 ‘모델5.0’은 “옌”을 정확히 인식하였지만 [날]이 “나에”로 대체된 것은 발음 경계 처리의 한계를 보여주고 있다. 이러한 결과는 개선 후의 모델이 일부 복잡한 음소 패턴에 대해서는 향상된 인식 성능을 보이지만, 여전히 예측에서 벗어나는 변이에 대해서는 정확하게 대응하는 능력이 부족함을 뜻한다.

여섯 번째 예에서, ‘모델3.0’은 [왜이]를 “회의”로 잘못 인식하였다. 이는 문맥적으로 더 자연스러운 단어를 선택한 결과로 해석되며, 과도한 문맥 의존성이 작용한 사례로 볼 수 있다. 이러한 결과는 모델이 독립적인 발음 인식보다는 언어적 규칙과 문맥에 지나치게 의존함을 뜻한다. 반면, ‘모델5.0’은 “왜이”를 정확히 출력하며 정답 레이블과 일치하는 결과를 나타냈다. 이는 개선 후 모델이 문맥 의존성을 줄이고, 발음의 독립적 처리를 강화한 결과로 해석된다.

지금까지 살펴본 바와 같이, ‘모델3.0’은 ‘모델5.0’에 비해 문맥 의존성을 더 많이 반영하여 문법적으로 자연스러운 결과를 인식하려는 경향을 보였으며, 이로 인해 실제 발음과 불일치하는 오류가 빈번히 발생하였다. 반면, ‘모델5.0’은 상대적으로 문맥에서 벗어나 발음을 독립적으로 처리하여 실제 발음과 더 가까운 결과를 도출하였다. 이 두 모델의 출력 결과 비교를 통해 데이터 레이블의 정제에 따른 음소 단위 레이블의 개선으로 모델이 문맥 의존성을 줄이고 발음 기반의 학습을 더 충실히 수행함을 확인할 수 있다.

5. 결론 및 향후 계획

본 연구에서는 단계별 레이블 정제를 통해 비원어민 대상 한국어 발음 평가를 위한 음소 인식 성능을 개선하는 모델 구축 과정을 제안하였다. 자기 지도 학습 기법을 활용하여, 사전 학습 모델인 Whisper 기반의 음성 인식 모델을 한국어 데이터로 미세 조정하는 모델 구축 방식을 사용하였다. 미세 조정 과정에서 비원어민이 발화한 한국어 데이터만 단독으로 사용하기보다는 비원어민 발화 데이터와 한국인(원어민)이 발화한 뉴스 데이터 등 다양한 발음 데이터를 균형 있게 결합하여 광범위한 화자의 발음 특성을 반영하는 통합적 학습이 비원어민 화자 데이터에서 인식 성능 개선에 중요한 역할을 한다는 점을 확인하였다.

또한, G2P 방식을 활용하여 한국어의 문법적 규칙보다 발음 규칙을 모델에 적용함으로써, 문맥에 대한 모델의 의존성을 효과적으로 줄일 수 있었다. 끝으로, 본 연구에서 가장 핵심적으로 진행한 단계별 데이터 레이블 정제와 유용한 데이터 선별을 통해 미세 조정 모델을 지속적으로 갱신함으로써 비원어민 대상의 한국어 음소 인식 성능을 크게 향상시킬 수 있음을 확인하였다.

향후 연구로, 여전히 존재하는 레이블 오류를 처리할 수 있는 추가적인 방법을 통해 모델을 더욱 개선할 수 있을 것으로 예상된다. 또한, 모델의 일반화 성능을 더 정확하게 확인하기 위해서는 새로운 음성 코퍼스를 활용하여 추가적인 평가를 진행할 필요가 있다.

References

AIHub. (2022a). Korean voice data for educational Asian language users. Retrieved from https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71479

AIHub. (2022b). Korean voice data from native Chinese and Japanese speakers for educational purposes. Retrieved from https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71490

AIHub. (2022c). Korean speech data from native European speakers for educational purposes. Retrieved from https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71489

AIHub. (2022d). Korean speech data from native English speakers for educational purposes. Retrieved from https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71469

AIHub. (2022e). News script and anchor voice data. Retrieved from https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71557

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449-12460.

Golowich, S. E., & Sun, D. X. (1998, October). A support vector/hidden Markov model approach to phoneme recognition. *Proceedings of the ASA Statistical Computing Section* (pp. 125-130). Dallas, TX.

Jang, J. S., Lim, B. Y., & Kwon, H. Y. (2023). Multimodal learning model for detecting pronunciation error segments of childrens and foreigners speech data. *Korean Institute of Information Scientists and Engineers, 29*(8), 396-401.

Kannadaguli, P., & Bhat, V. (2015, March). A comparison of Gaussian mixture modeling (GMM) and hidden Markov modeling (HMM) based approaches for automatic phoneme recognition in Kannada. *Proceedings of 2015 International Conference on Signal Processing and Communication (ICSC)* (pp. 425-430). Noida, India.

Kim, E. (2006). A study on the diagnosis & evaluation for pronunciation errors of Korean language learners. *Korean Language Education*, *17*(1), 71-99.

Kim, E., Jeon, J. J., Seo, H., & Kim, H. (2022a). Automatic pronunciation assessment using self-supervised speech representation learning. arXiv*, https://doi.org/10.48550/*arXiv. 2204.03863

Kim, J., & Kang, P. (2021). K-wav2vec 2.0: Automatic speech recognition based on joint decoding of graphemes and syllables. arXiv, https://doi.org/10.48550/arXiv.2110.05172

Kim, S. Y., Min, H., & Choi, H. W. (2022b). A strategic design and construction of a non-native voice data set of Korean speech for AI model training. Journal of Languistics Science, *100*, 63-88.

Korzekwa, D., Lorenzo-Trueba, J., Zaporowski, S., Calamaro, S., Drugman, T., & Kostek, B. (2021, June). Mispronunciation detection in non-native (L2) English with uncertainty modeling. *Proceedings of ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8135-8139). Toronto, Canada.

Leung, W. K., Liu, X., & Meng, H. (2019, May). CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis. *Proceedings of ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8132-8136). Brighton, UK.

Lin, B., & Wang, L. (2023, October-November). Multi-accent pronunciation assessment based on domain adversarial training. *Proceedings of 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 2424-2428). Taipei, Taiwan.

Oh, C., Kim, C., & Park, K. (2023). Building robust Korean speech recognition model by fine-tuning large pretrained model. *Phonetics and Speech Sciences, 15*(3), 75-82.

Park, K. (2019). g2pK: g2p module for Korean [Computer program]. Retrieved from https://github.com/Kyubyong/g2pk

Peng, L., Fu, K., Lin, B., Ke, D., & Zhang, J. (2021, August-September). A study on fine-tuning wav2vec2.0 model for the task of mispronunciation detection and diagnosis. *Interspeech* (pp. 4448-4452). Brno, Czechia.

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023, Jul). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning (ICML)* (pp. 28492-28518). Honolulu, HI.

Ravanelli, M., Parcollet, T., & Bengio, Y. (2019, May). The pytorch-kaldi speech recognition toolkit. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6465-6469). Brighton, UK.

Ryu, H., Hong, H., Kim, S., & Chung, M. (2016, December). Automatic pronunciation assessment of Korean spoken by L2 learners using best feature set selection. *Proceedings of 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. Jeju, Korea.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need. *Proceedings of 31st Conference on Neural Information Processing Systems (NIPS)*. CA, USA.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno,

Y., Soplin, N. E. Y., ...Ochiai, T. (2018). ESPnet: End-to-end speech processing toolkit. arXiv. *https://doi.org/10.48550/*arXiv.1804.00015.

Xu, Q., Baevski, A., & Auli, M. (2021). Simple and effective zero-shot cross-lingual phoneme recognition. arXiv. *https://doi*.org/10.48550/arXiv.2109.11680.

Yang, S. H., & Chung, M. (2014). Prediction of Chinese learners' Korean pronunciation variations based on contrastive analysis. *Annual Conference on Human and Language Technology* (pp. 206-210).

Yang, M., Hirschi, K., Looney, S. D., Kang, O., & Hansen, J. H. L. (2022). Improving mispronunciation detection with wav2vec2-based momentum pseudo-labeling for accentedness and intelligibility assessment. arXiv. *https://doi.org/10.48550/*arXiv.2203.15937.

Zahran, A. I., Fahmy, A. A., Wassif, K. T., & Bayomi, H. (2023). Fine-tuning self-supervised learning models for end-to-end pronunciation scoring. *IEEE Access, 11*, 112650-112663.

• **경나 (Na Geng)**
한국외국어대학교 박사과정
서울특별시 동대문구 이문로 107
Tel: 010-2436-6603
Email: gengna0324@gmail.com
관심분야: 음성공학

• **나희정 (Heejung Na)**
한국외국어대학교 석사
서울특별시 동대문구 이문로 107
Email: 92031l3@naver.com
관심분야: 음성공학

• **박지은 (Jieun Park)**
한국외국어대학교 석사
서울특별시 동대문구 이문로 107
Email: ppae0216@naver.com
관심분야: 음성공학

• **박정식 (Jeong-Sik Park)** 교신저자
한국외국어대학교 교수
서울특별시 동대문구 이문로 107
Tel: 02-2172-8814
Email: parkjs@hufs.ac.kr
관심분야: 음성 처리 기술, 머신 러닝, 인공지능

비원어민 한국어 발음 평가를 위한 자기 지도 학습 기반 한국어 음소 인식*

경 나[1] · 나 희 정[1] · 박 지 은[1] · 박 정 식[2]

[1]한국외국어대학교 영어학과, [2]한국외국어대학교 ELLT학과

국문초록

비원어민의 한국어 발음 평가를 위해서는 음소 인식뿐만 아니라 발음 오류를 정확하게 탐지할 수 있는 모델이 필요하다. 자기 지도 학습(self-supervised learning) 기반 음성 인식에서 대량의 음성 자료를 통해 구축된 사전 학습(pre-trained) 모델은 정밀한 음성 인식을 가능하게 하는 것으로 알려져 있다. 특히 Wav2Vec2.0 및 Whisper와 같은 모델들은 여러 연구에서 우수한 한국어 음성 인식 성능을 보였으며, Whisper 모델은 특히 뛰어난 성능을 나타냈다. SUPERB 벤치마크를 통해 다양한 사전 학습 모델을 비교한 결과, 음소 인식 분야에서의 성과도 입증되었다. 그러나 비원어민 화자의 발화 특성을 반영하여 실제 발화에 정교하게 맞춘 레이블 데이터의 부족으로 비원어민의 한국어 음소 인식 성능을 높이는 데 한계가 존재한다. 따라서, 비원어민의 한국어 발음 평가 모델을 구축하기 위해서는 정확한 레이블을 갖춘 데이터의 확보가 중요하다. 본 연구에서는 사전 학습된 Whisper 모델을 활용하여 비원어민의 한국어 발음 평가를 위한 한국어 음소 인식 모델을 개발한다. AIHub에서는 아시아, 중국, 일본, 유럽, 영어권 비원어민의 한국어 교육용 음성 데이터를 대량으로 제공하고 있어 이를 모델의 미세 조정(fine-tuning)을 위한 데이터로 활용한다. 그러나, 제공된 음소 레이블의 정확도가 매우 떨어지는 문제가 있어, 한국인이 명료하게 발음한 “뉴스 대본 및 앵커 음성 데이터”를 추가로 활용하여 정확한 한국어 음소 발음을 학습시킨다. 이 두 데이터로부터 구축한 모델을 통해 비원어민 한국어 음성 데이터의 음소 레이블을 실제 발음에 맞게 수정하고, 이를 다시 미세 조정에 적용하여 비원어민 한국어 음소 인식 모델을 구축한다. 이 같은 과정을 몇 차례 단계별로 수행하여 미세 조정 모델을 지속적으로 갱신한다. 최종적으로 구축한 모델의 유효성을 평가하기 위해 비원어민의 한국어 발화 음성과 원어민의 한국어 음성을 대상으로 음소 인식 실험을 진행한 결과, 기본 모델에 비해 음소 인식 성능이 유의미하게 향상되었음을 확인하였다.

핵심어: 자기 지도 학습, Whisper 모델, 한국어 음소 인식, 한국어 발음 평가

참고문헌

김수연, 민효민, 최혜원(2022). 인공지능 학습을 위한 외국인의 한국어 발화 음성 데이터 구축 방안. *언어과학연구*, 100, 63-88.

김은애(2006). 한국어 학습자의발음 오류 진단 및 평가에 관한 연구. *한국어교육*, *17*(1), 71-99.

오창한, 김청빈, 박기영(2023). 대형 사전훈련 모델의 파인튜닝을 통한 강건한 한국어 음성인식 모델 구축. *말소리와 음성과학*, *15*(3), 75-82.

양승희, 정민화(2014). 대조 분석 기반의 중국인 학습자의 한국어 발음 변이 양상 예측. *제26회 한글 및 한국어 정보처리 학술대회 논문집* (pp. 206-210).

장재석, 임보영, 권혁윤(2023). 아동 및 외국인 음성 데이터의 발음 오류 구간 검출을 위한 멀티모달 학습 모델. *정보과학회 컴퓨팅의 실제 논문지*, *29*(8), 396-401.

* 이 논문은 한국외국어대학교 교원연구지원사업, 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1A2C1013162).