creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

문학석사 학위논문

교차 언어적 파인튜닝을 사용한 wav2vec 2.0 기반 발음 오류 검출

wav2vec 2.0-based Mispronunciation Detection Using Cross-lingual Fine-tuning

2022년 8월

서울대학교 대학원

언어학과 언어학전공

지 승 훈

교차 언어적 파인튜닝을 사용한 wav2vec 2.0 기반 발음 오류 검출

지도 교수 정 민 화

이 논문을 문학석사 학위논문으로 제출함
2022년 7월

서울대학교 대학원
언어학과 언어학전공
지 승 훈

지승훈의 문학석사 학위논문을 인준함
2022년 7월

위 원 장 _____ 이 호 영 _____ (인)

부위원장 _____ 정 민 화 _____ (인)

위 원 _____ 김 선 희 _____ (인)

초 록

컴퓨터 보조 발음 훈련(CAPT)의 세부 과제인 발음 오류 검출(MD)은 외국어 학습자의 발화를 입력받았을 때 각 음소에 대한 발음 오류 여부를 정확히 판단할 수 있는 계산 모델을 구축하는 것을 목표로 한다. 잘 훈련된 발음 오류 검출 모델이 출력하는 결과를 이용하여 외국어 교수자는 수동 또는 자동으로 학습자의 발음에 대한 교정 피드백을 제시할 수 있다. 현재 발음 오류 검출 분야의 주된 연구 경향은 심층 신경망 기반 음향 모델을 이용하여 음소 단위의 인식을 진행하는 것인데, 이때 음향 모델의 훈련 데이터는 학습 대상 언어로 된 문장 또는 단어를 읽는 비원어민 학습자 발화 데이터이다. 그러나 비원어민 학습자 발화 데이터는 많은 양을 확보하는 것이 매우 어렵다는 단점이 있어 심층 신경망 기반 모델의 학습에 충분한 양을 제공하지 못할 수 있다. 이와 같은 문제를 해결할 수 있는 가능성을 탐구하기 위해 본 연구에서는 비원어민 학습자 발화 데이터 대신 학습자 모어의 원어민 발화 데이터와 학습 대상 언어의 원어민 발화 데이터만을 사용하여 음향 모델을 훈련하는 방법론을 제안한다.

음향 모델은 최근 많은 음성 관련 과제에서 좋은 성능을 보이고 있으며 발음 오류 검출 분야에서도 많이 연구되고 있는 wav2vec 2.0을 이용하며, 학습 대상 언어는 한국어, 학습자 모어는 표준 중국어로 한다. wav2vec 2.0 모델은 사전 훈련과 파인튜닝 과정으로 나뉘어 학습을 진행하는데, 이미 학습되어 있는 사전 훈련 모델에 중국인 한국어 학습자 발화 데이터, 한국어 또는 중국어 원어민 발화 데이터, 한국어와 중국어 원어민 발화를 합친 데이터 등을 이용하여 각각 파인튜닝을 진행한 후 그 성능을 알아보고 인식 양상을 분석한다. 더불어 본 연구에서는 두 언어의 음소 단위 텍스트를 공통의 전사 단위로 표현할 수 있도록 IPA 체계를 전사 및 인식 단위로 한다. 이를 위해 한국어와 중국어의 음소 체계 및 음운 현상을 고려하여 각 언어 텍스트를 IPA 기호 연쇄로 변환하는 방법을 제안한다.

두 언어의 원어민 발화 데이터를 모두 이용하여 교차 언어적으로 파인튜닝한 모델은 음소 인식에서 정답률 최고 74.37%, 발음 오류 검출에서 F1-점수 최고 39.90%를 보였다. 이는 한국어 원어민 발화 데이터만을 이용하여 파인튜닝한 모델보다 각각 약 1.42%p, 1.44%p 높은 성능 수치이다. 하지만 기존의 방법론과 같이 중국인 한국어

학습자 발화 데이터로만 파인튜닝한 모델의 정답률 최고 80.11%, F1-점수 최고 50.00%에는 못 미치는 수치였다. 단 각 파인튜닝 모델을 정답 발음 오거부, 빈번하게 나타나는 중국인 한국어 학습자의 발음 오류 유형 등의 면에서 상세히 살펴보았을 때 각 파인튜닝 모델은 혼재된 인식 양상을 보였다.

주요어 : 발음 오류 검출, wav2vec 2.0, 교차 언어적 훈련, 음향 모델, 중국인 한국어 학습자, 비원어민 발화 데이터셋
학 번 : 2020-25664

목 차

1. 서론 1

2. 관련 연구 4
2.1. 발음 평가와 발음 오류 검출 4
2.2. 한국어와 중국어의 음소 체계 10
2.2.1. 한국어와 중국어의 자음 체계 10
2.2.2. 한국어와 중국어의 모음 체계 11
2.3. 중국인 한국어 학습자의 발음 오류 12
2.4. wav2vec 2.0 14

3. 실험 방법론 18
3.1. 데이터셋 18
3.1.1. KsponSpeech 18
3.1.2. AISHELL-1 18
3.1.3. L2KSC 19
3.2. 전사 텍스트 가공 20
3.2.1. KsponSpeech 20
3.2.2. AISHELL-1 21
3.2.3. L2KSC 22
3.3. 사전 훈련 모델 23
3.4. 파인튜닝 23
3.4.1. 베이스라인 환경 24
3.4.2. 단일 언어적 파인튜닝 환경 24
3.4.3. 교차 언어적 파인튜닝 환경 24
3.4.4. 파인튜닝 상세 24
3.5. 평가 척도 25
3.5.1. 음소 인식 26
3.5.2. 발음 오류 검출 26

4. 실험 29
4.1. 음소 인식 29
4.2. 발음 오류 검출 30
4.3. 발음 오류 유형에 따른 결과 분석 31
4.3.1. 어두 초성 파열음 및 치조 마찰음의 발음 오류 33
4.3.2. 단모음의 대치 오류 35
4.3.3. 종성의 탈락과 대치 오류 37

5. 논의 39
5.1. 결과 해석 39
5.2. 발음 오류 검출을 위한 비원어민 학습자 데이터셋 42

5.3. 의의 45
5.3.1. 원어민 발화만을 이용한 교차 언어적 발음 오류 검출 모델.....45
5.3.2. 모델 성능의 언어학적 분석 46
5.4. 한계와 후속 과제 47

6. 결론 49

참고 문헌.....50

부록 58

Abstract.....69

표 목차

표 2.1 한국어의 자음 체계 11
표 2.2 중국어의 자음 체계 11
표 2.3 한국어의 모음 체계 12
표 2.4 중국어의 모음 체계 12
표 3.1 실험에 사용한 데이터셋 분량 정리 20
표 4.1 파인튜닝 모델의 음소 인식 성능 29
표 4.2 파인튜닝 모델의 발음 오류 검출 성능 31
표 4.3 어두 초성 평 파열음의 정답 발음 인식 양상 33
표 4.4 어두 초성 평 파열음의 발음 오류 인식 양상 34
표 4.5 어두 초성 경 파열음의 정답 발음 인식 양상 34
표 4.6 어두 초성 경 파열음의 발음 오류 인식 양상 34
표 4.7 어두 초성 경 마찰음의 정답 발음 인식 양상 35
표 4.8 어두 초성 경 마찰음의 발음 오류 인식 양상 35
표 4.9 [ɯ] 모음의 정답 발음 인식 양상 35
표 4.10 [ɯ] 모음의 발음 오류 인식 양상 36
표 4.11 [o] 모음의 정답 발음 인식 양상 36
표 4.12 [o] 모음의 발음 오류 인식 양상 36
표 4.13 [ʌ] 모음의 정답 발음 인식 양상 37
표 4.14 [ʌ] 모음의 발음 오류 인식 양상 37
표 4.15 종성의 정답 발음 인식 양상 38
표 4.16 종성의 발음 오류 인식 양상 38

그림 목차

그림 2.1 wav2vec 2.0의 사전 훈련 14
그림 3.1 KsponSpeech 전사 텍스트 가공 과정 21
그림 3.2 AISHELL-1 전사 텍스트 가공 과정 22
그림 3.3 L2KSC 전사 텍스트 가공 과정 22
그림 5.1 인식 결과 정렬 예 41

수식 목차

수식 2.1 wav2vec 2.0의 대조 손실 함수 15
수식 2.2 wav2vec 2.0의 다양성 손실 함수 15
수식 2.3 wav2vec 2.0의 손실 함수 16
수식 3.1 정답률 및 정확도의 계산식 26
수식 3.2 FRR 및 FAR의 계산식 27
수식 3.3 정밀도, 재현율 및 F1-점수의 계산식 27
수식 3.4 DER, 검출 정확도 및 진단 정확도의 계산식 28

1. 서론

발음 오류 검출(mispronunciation detection; MD) 과제는 컴퓨터 보조 발음 훈련(computer-aided pronunciation training; CAPT) 분야의 한 세부 과제로, 훈련된 외국어 교수자가 인력으로 직접 진행하는 비원어민 외국어 학습자의 발음 오류 검출을 계산 모델이 대신하도록 하는 데 그 목적이 있다. 발음 오류 검출 모델은 외국어 학습자가 발화한 학습 대상 외국어 음성을 입력받아 각 음소 구간에 대한 발음 오류 여부를 이진 혹은 다중 분류를 통해 출력하며, 교수자는 이와 같은 출력 결과를 바탕으로 자동 혹은 수동으로 학습자에게 교정 피드백(corrective feedback)을 제공할 수 있다.

그러나 현재 발음 오류 검출 분야에는 해결이 필요한 문제들이 아직 많다. 가장 널리 연구되고 있는 발음 평가 방법론인 GOP(goodness of pronunciation; Witt & Young, 2000) 점수를 기반으로 발음 오류를 검출하는 방법론은 각 음소의 발음 오류 여부를 발음 평가 점수를 통해 믿을 만한 정도의 성능으로 구할 수 있지만 구체적인 발음 오류의 내용을 알 수 없어 교정 피드백을 제공하기는 힘들다(Yan et al., 2020). 한편 최근에는 다양한 음성 관련 분야에서 연구되고 있는 심층 신경망(deep neural network) 기반 음향 모델의 높은 성능에 힘입어, 이를 기반으로 발음 오류 검출 모델을 구축하고자 하는 연구가 많이 진행되고 있고 연구 경향 또한 GOP 점수 기반에서 음소 인식 기반의 모델 구축으로 많이 이동하였다. 이러한 인식 기반 발음 오류 검출 모델은 발음 오류 여부를 인식된 음소 정보를 통해 확인할 수 있다.

심층 신경망 기반 음향 모델 분야 연구에서는 wav2vec 2.0(Baevski et al., 2020)으로 대표되는 트랜스포머 구조(Vaswani et al., 2017) 기반의 wav2vec 자기 지도 학습 모델이 여러 가지 음성 관련 과제에서 좋은 성능을 발휘하고 있고, 발음 오류 검출 연구에서도 wav2vec 모델을 사용하는 연구들이 보고되고 있다(Peng et al., 2021; Lin & Wang, 2021; Wu et al., 2021; Xu et al., 2021; Peng et al., 2022). wav2vec 음향 모델은 먼저 구체적인 하위 과제(downstream task)에 관계없이 전사 텍스트가 없는 대량의 음성 데이터로 사전 훈련(pre-training)을 진행하며, 이후 하위 과제의 성격에 맞는 추가 계층을 정의하여 정답과 함께 훈련하는 파인튜닝(fine-tuning)을 진행하는 두

단계로 나누어 훈련한다.

단 이와 같은 심층 신경망 기반 모델은 훈련 데이터로서 대량의 비원어민 외국어 학습자 발화 데이터를 필요로 한다. 이와 같은 데이터는 원어민 발화 데이터에 비해 녹음에 참여할 화자를 모집하는 데 어려움이 있으며 발음 오류가 포함된 비원어민 발화를 전사하는 것 또한 음성학적으로 훈련된 전사자를 필요로 하므로 충분한 규모의 데이터 확보 및 구축에 큰 노력이 따른다. 따라서 현재 발음 오류 검출 분야에서는 영어를 학습 대상 외국어로 하는 데이터셋들(Zhao et al., 2018; Meng et al., 2007)이 주로 이용되고 있고, 이마저도 학습자의 모어 배경은 크게 제한되어 있다. 또한 일반적으로 외국어 발음 오류는 학습자의 모어 음소 체계에 영향을 받는 것이 큰 부분을 차지하기 때문에(Witt & Young, 2000) 특정 모어를 배경으로 하는 학습자 음성 데이터가 없다면 같은 언어를 모어 배경으로 하는 학습자의 발음 오류 검출에 더욱 문제가 될 여지가 많다.

본 연구는 위와 같은 연구 현황을 배경으로 하여 발음 오류 검출 모델을 구축하는 새로운 방법론을 제안하고, 그 적용 가능성에 대해 최초로 탐구하고자 한다. 종전의 점수 기반 방법론과 반대로 구체적인 발음 오류 양상 제공을 기대할 수 있는 음소 인식 기반의 심층 신경망 모델을 사용하되, 데이터셋 구축에 큰 어려움이 따르는 비원어민 외국어 학습자 발화 데이터셋이 아닌 학습자의 모어와 학습 대상 외국어의 원어민 발화 데이터셋을 함께 사용하여 모델을 훈련하는 것으로 심층 신경망 기반 음향 모델의 데이터 요구량을 충족하고자 하는 것이다.

다시 말해서, 본 연구에서는 많은 음성 언어 처리 과제에서 좋은 성능을 보이는 것으로 알려져 있는 wav2vec 2.0 음향 모델을 기반으로 하여 학습자 모어와 학습 대상 외국어의 원어민 발화 데이터로 훈련한 교차 언어적(cross-lingual) 환경의 음소 인식 기반 발음 오류 검출 모델이 어떤 성능을 보이는지 알아보고, 나아가 이와 같은 모델의 발음 오류 검출 과제 내 활용 가능성을 탐구하고자 한다.

본 논문에서 제기하는 연구 질문은 다음과 같다.

1. **발음 오류 검출 모델 훈련에서, 학습자 모어와 학습 대상 언어의 원어민 발화 데이터는 학습 대상 언어의 비원어민 학습자 음성 데이터를 대체할 수 있는가?**
2. **wav2vec 기반 음향 모델을 교차 언어적 방식으로 파인튜닝하였을 때 모델의 인식 양상은 어떠한가?**

위에서 서술한 방식대로 교차 언어적 환경에서 발음 오류 검출 모델을 훈련하려면 외국어 학습자가 학습 대상으로 삼은 외국어와 그 학습자의 모어를 설정하는 것이 필요하다. 본 연구에서는 발음 오류 검출 분야에서 영어에 비해 덜 활발히 연구되고 있는 한국어를 학습 대상 언어로 선택한다. 한편 통계에 의하면 2021년 한국 내 고등 교육 기관에서 공부하는 외국인 유학생 149,682명 중 중국 국적을 가진 학생이 43.9%(65,677명)의 비율을 차지하며(한국교육개발원, 2021), 2020년 한국어능력시험(TOPIK)에 응시한 외국 국적 응시자 52,483명 중 38.9%(20,438명)가 중국 국적자이다(교육부 국립국제교육원, 2020). 이를 근거로 한국어를 배우는 외국인 중 가장 많은 수를 차지하는 중국인의 모어인 표준 중국어(Mandarin Chinese, 이하 중국어)를 한국어 학습자의 모어로 설정하고자 한다. 한국어 및 중국어 원어민 발화, 중국인 한국어 학습자 발화의 전사 데이터는 각 언어의 음소 체계 및 음운 현상을 반영하여 같은 기호 체계를 공유하는 국제 음성 기호(IPA) 단위로 변환하고, 이것을 훈련 데이터로 사용한다.

본 논문의 순서는 다음과 같다. 1장에서는 연구의 배경 및 논문의 내용을 개괄하고, 2장에서는 본 연구와 관련한 연구를 소개한다. 3장에서는 본 연구의 실험에서 사용되는 데이터셋, 모델 성능 평가를 위한 평가 척도와 실험 설계에 대해 서술하고, 4장에서는 실험 결과 및 성능을 보고한다. 5장에서는 실험 결과를 바탕으로 본 연구의 논의 사항과 의의, 한계를 선행 연구와 비교하여 서술하며, 6장에서 본 논문의 결론을 제시한다.

2. 관련 연구

2.1. 발음 평가와 발음 오류 검출

발음 오류 검출 과제를 해결하고자 하는 연구의 흐름은 크게 두 방법론으로 나눌 수 있다. 첫째는 발음 평가 척도인 goodness of pronunciation(GOP; Witt & Young, 2000) 점수로 대표되는 점수 기반 방법론이며, 둘째는 음성 인식기의 구체적인 출력 결과 또는 발음 오류 여부를 이용하는 인식 기반의 발음 오류 검출 방법론이다.

연구가 어떠한 방식의 음향 모델을 기반으로 진행되는지에 따라서는 세 방법론으로 나눌 수 있다. 첫째는 은닉 마르코프 모델(hidden Markov model; HMM)과 가우시안 혼합 모델(Gaussian mixture model; GMM) 기반의 전통적 HMM-GMM 방법론이고, 둘째는 HMM-GMM 모델의 GMM을 심층 신경망(deep neural network)으로 대체한 HMM-DNN 방법론이다. 한편 최근 컴퓨팅 성능이 발전하고 대규모 데이터셋을 구축할 수 있게 되면서 큰 규모의 심층 신경망 모델을 훈련할 수 있게 되었는데, 음성 인식 분야에서는 이 방법론이 기존 전통적 모델의 성능을 앞지르게 되었다. 오직 심층 신경망만을 사용하여 음향 모델을 구축하는 이 방법론을 종단 간(end-to-end) 방법론이라 하며 발음 오류 검출 모델에서도 주로 이용되는 방법론 중 하나가 되었다.

일반적으로 발음 오류 검출 모델은 크게 음향 모델과 강제 정렬(forced-alignment) 모듈로 구성되어 있으며, 외국어 학습자가 주어진 학습 대상 언어의 단어 또는 문장을 읽는 상황을 상정한다. 따라서 해당 학습자 발화가 어떤 단어 또는 문장을 읽은 것인지에 대한 정보가 미리 확보되어 있으므로 모델 입력은 발화 음성과 그 발화의 표준 음소(canonical phoneme) 또는 표준 음(canonical phone) 연쇄의 쌍으로 구성되어 있는 것이 보편적이다.

음향 모델이 발화 음성을 처리하기 전 입력 쌍은 강제 정렬 모듈에서 먼저 처리된다. 강제 정렬 모듈 또한 음향 모델 기반의 계산 모델로, 주어진 음성과 음소 연쇄 쌍을 입력받아 해당 음성 내 음소 구간 경계를 찾고 각 구간마다 음소 기호들을 차례대로 할당해 주는 작업을 수행한다.

GOP 점수를 예로 들어 점수 기반 방법론의 문제 해결 방법에 대해

설명하자면 다음과 같다. 점수 기반 방법론에서는 학습 대상 언어의 원어민 음성 발화 데이터셋으로 훈련된 음향 모델을 사용하여, 입력받은 비원어민 학습자 발화 내 각 음소 구간마다 학습 대상 언어의 모든 음소에 대한 사후 확률(posterior probability)을 계산하는 방식이라 할 수 있다. 달리 말하면 이와 같은 방법은 비원어민 외국어 학습자의 발화에 존재하는 각 발음이 학습 대상 언어 원어민의 그것과 비교하여 얼마나 차이가 있는지를 계산하는 것이다. GOP 점수는 발음 평가 척도로서 그 자체로 해당 음소의 발음 오류 여부를 결정할 수는 없다. 대신 모델에 임곗값(threshold)을 미리 정해 두어, 어떤 음소에 대한 GOP 점수가 임곗값을 넘는지 여부에 따라 해당 음소가 정확히 발음된 것인지, 발음 오류인지를 결정하도록 한다.

GOP 점수를 평가 척도로 하는 발음 오류 검출 모델 연구는 주로 서로 다른 음향 모델에 GOP 알고리즘을 적용하거나, GOP 알고리즘을 개선하는 추가적인 점수 계산 방식을 도입하는 방향으로 이루어진다. Hu, Qian & Wang(2015)은 전통적인 GMM-HMM 모델이 아닌 GMM-DNN 음향 모델을 사용하여 GOP 점수를 계산하는 영어 및 중국어 학습자의 발음 오류 검출 모델을 제안하였다. Zhang et al.(2018)은 중국어 학습자의 발음 오류 검출 과제를 풀기 위한 방법으로 GOP 점수 계산식에 스케일 팩터를 도입하고, 중국어 음절 구조의 특성을 이용한 성모 및 운모별 가중치를 사용해 최종적인 GOP 점수를 계산하는 방법을 제시했다. Huang et al.(2015)은 GOP 점수 기반 발음 오류 검출 모델을 훈련하는 방식으로 F1-점수를 최대화하는 목적 함수를 사용하였다. 가장 최근의 연구로는 Gong et al.(2022)를 들 수 있으며, 이 연구는 트랜스포머 구조(Vaswani et al., 2017)를 이용한 음향 모델을 통해 GOP 점수를 계산하는 모델을 제안하였다.

반면 인식 기반 방법론에서는 발음 오류가 포함되어 있는 비원어민 외국어 학습자의 음성 데이터셋으로 훈련된 음향 모델을 사용한다. 역시 강제 정렬된 학습자 발화의 각 음소 구간을 독립적으로 음향 모델에 입력한 후, 해당 구간의 음소 인식 결과를 직접 출력하는 다중 분류를 수행하거나 단순히 해당 음소 구간이 오류 발음으로 산출된 것인지의 여부를 출력하는 이진 분류를 수행하여 발음 오류를 검출한다. 특히 후자의 방법은 분류 기반 방법론이라고도 일컬어진다.

인식 기반 발음 오류 검출 모델은 종래 분류기를 이용한 이진 분류 과제, 즉 분류 기간 방법론으로 많은 연구가 진행되었다. Wei et al.(2009)은 서포트 벡터 머신(support vector machine)을 이진

분류기로 선택하여 발음 오류 여부를 검출하는 모델을 제시하였다. Hu, Qian, Soong & Wang(2014)은 심층 신경망으로 훈련된 음향 모델과 전이 학습을 이용한 로지스틱 회귀 분류기로 발음 오류 검출 모델을 구축하였다.

심층 신경망 기반 음향 모델이 각광받기 시작한 이후에는 이를 이용한 분류 기반 방법론과 함께 음소 단위의 인식 기반 방법론이 널리 연구되고 있다. 음소 단위 인식 기반 방법론에서는 일반적인 음성 인식기와 같이 모델이 개별적인 음소를 출력하는 것을 주된 작동 방식으로 하되, 인식을 위한 은닉 벡터를 산출하는 음향 모델에 다양한 구조의 심층 신경망을 적용하는 것이 주된 연구 경향으로 자리잡고 있다. 또한 이 방법론을 이용한 연구에서는 음향 모델 입장에서 외부 모듈인 강제 정렬 모듈을 사용하지 않음으로써 완전한 종단 간 모델을 실현하려는 시도가 많이 행해진다. 대표적으로 중국인 영어 학습자 발화를 데이터로 훈련한 CNN-RNN-CTC(Leung et al., 2019) 모델을 꼽을 수 있는데, 음성에서 추출한 스펙트로그램을 음향 특징으로 하여 합성곱 신경망(convolutional neural network), 순환 신경망(recurrent neural network)을 차례로 거치도록 하고, CTC 손실 함수(Graves et al., 2006)를 목적 함수로 이용하여 훈련한다. Lo et al.(2020)은 순환 신경망 구조에서 높은 성능 향상을 보이는 주의 집중 기제(attention mechanism)와 CTC 손실 함수를 이용하는 중국어 발음 오류 검출 모델을 제안하였는데, 이 모델은 hybrid CTC-attention 모델로 알려져 CNN-RNN-CTC 모델과 함께 많은 연구에서 베이스라인 모델로 인용되고 있다.

HMM 모델로 사후 확률을 계산하는, GOP 점수 등의 통계적 점수 기반 발음 평가 척도는 인간의 평가 점수와 높은 상관 관계를 나타내는 것으로 알려져 있다(Franco et al., 1997). 또한 음향 모델을 훈련하는데 원어민 발화 데이터를 사용하므로 음성 분야에서 더욱 보편적인 과제인 음성 인식 과제를 위해 이미 훈련된 음향 모델을 사용할 수도 있고, 새로운 모델을 훈련해야 하더라도 훈련 대상으로 삼을 수 있는 데이터셋이 풍부하다. 그러나 이 방법론은 별도의 강제 정렬 모듈에 의존해야 하기 때문에 앞뒤 음소에 대한 정보를 이용할 수 없고(Yan et al., 2020), 문제 정의에 따라 강제 정렬 오류로 인한 추가적인 모델 성능 저하가 나타날 수 있다(Mathad et al., 2021). 또한 인식 기반 방법론과는 달리 실험자가 미리 정한 임곗값에 따라 해당 개별 음소가 정확히 발음된 것이냐의 여부만을 출력할 수 있기 때문에, 일반적인

발음 평가에 사용되기에는 적절할 수 있겠으나 발음 오류 검출 등 발음 평가에서 더 나아가 교정 피드백을 제공할 수 있는 정도의 정보로는 부족하다.

반면 인식 기반 방법론은 훈련 데이터로 해당 언어를 배우는 비원어민 학습자가 발화한, 발음 오류가 포함되어 있는 발화 데이터셋을 직접 모델링하기 때문에 이상적인 계산 모델링 방법론이 될 수 있다. 그러나 훈련 데이터로서 비원어민 외국어 학습자의 발화 데이터셋을 구축하기 위해서는 학습 대상 언어를 배우고 있는 비원어민 화자의 실제 발화를 녹음해야 하는데, 일반적으로 어떤 언어를 외국어로서 학습하는 비원어민 학습자의 수는 그 언어를 모어로 사용하는 원어민 인구보다 훨씬 적으므로 원어민 발화 데이터셋에 비해 녹음자 섭외에 어려움이 따른다. 더불어 음향 모델의 안정적 훈련을 위해서는 같은 데이터셋의 발화 데이터들이 최대한 일관적인 녹음 환경에서 수집되어야 하는데, 비원어민 학습자 발화라는 데이터 특성상 물리적으로 같은 공간에 많은 비원어민 녹음자를 섭외하는 것에는 동일한 환경에서 원어민 녹음자를 섭외하는 것보다 더 큰 한계점이 존재한다.

발음 오류 검출과 같은 음성 언어 처리 분야의 지도 학습(supervised learning) 과제에서는 계산 모델이 입력받는 음성 데이터 외에도 모델이 출력한 결과가 정답인지의 여부를 알 수 있게 하는 정답 레이블이 필요하다. 음소 단위의 발음 오류 검출 과제에서 정답 레이블은 각 음소의 발음 오류 여부 또는 음성 데이터를 들리는 대로 받아 쓴 음소열(phoneme sequence)이다. 분류 기반 방법론을 사용하는 발음 오류 검출 모델이 아니라면 후자의 경우를 정답 레이블로 사용하여야 하는데, 발음 오류를 적절히 표시하기 위해서는 비원어민 학습자가 발화한, 부분적으로 오류가 포함되어 있을 수 있는 학습 대상 언어 음성을 들리는 대로 표시해야 한다. 이를 위해서는 음성학적으로 훈련되었으며 학습 대상 언어의 음소 체계 및 그 언어에 존재하지 않는 말소리, 혹은 학습자 모어의 음소 체계에 대한 지식을 함께 갖춘 전사자가 필요하고, 그 전사 과정도 일반적인 원어민 발화 데이터의 그것보다 난이도가 높다. 이들 전사자를 섭외하여 전사를 완료했다고 하더라도 전사자 전원이 일관적인 전사 텍스트를 생성하지 않았다면 같은 꼴의 발음 오류에 대해 서로 다른 레이블링이 진행될 수 있으므로 모델의 적절한 훈련에 제약이 따를 수 있다(Lee & Glass, 2015). 전통적인 기계 학습 모델과 비교해 훨씬 많은 양의 훈련 데이터를 필요로 하는 심층 신경망 기반 모델을 훈련하기 충분한 양의 데이터를

수집해야 한다면 그 어려움은 가중된다.

현재 발음 오류 검출 분야에서 가장 널리 이용되고 있는 데이터셋은 L2-ARCTIC(Zhao et al., 2018)이다. L2-ARCTIC 데이터셋은 영어 학습자의 비원어민 발화 데이터셋이며 최신 버전은 힌디어, 한국어, 표준 중국어, 스페인어, 아랍어, 베트남어의 모어 배경을 가진 총 24명의 화자의 발화를 수록하고 있다. 그중 각 화자마다 150개 발화가 단어와 음소 단위로 각각 전사 및 정렬되어 발음 오류 음소에 대해서는 표준 음소(canonical phoneme), 실제 음소, 발음 오류 종류가 함께 레이블링되어 있다.

하지만 24명의 화자가 150개 발화씩을 녹음한 L2-ARCTIC 데이터셋은 총 3시간가량의 분량으로, 이는 일반적으로 음향 모델을 훈련하기에 매우 적은 양의 데이터이다. 다른 영어 학습자 발화 데이터셋의 경우에도 데이터 규모의 차이는 L2-ARCTIC과 크지 않아서, 보통의 영어 발음 오류 검출 연구에서는 부족한 데이터량을 보충하기 위해 영어 원어민 발화 데이터셋이자 음소 단위로 전사되어 있는 TIMIT(Garofolo et al., 1993) 데이터셋을 합쳐 비원어민 영어 학습자 데이터셋과 전사 단위를 통일한 후 모델 학습에 사용한다(Leung et al., 2019; Feng et al. 2020; Lo et al., 2021; Wu et al., 2021; Lin & Wang, 2022; Peng et al., 2022). 이와 같은 방법을 사용하면 모델이 학습자 음성 데이터셋에 과적합될 우려는 줄지만, 합친 데이터셋 내에서 발음 오류 음소가 차지하는 비율이 줄어들기 때문에 모델이 표준 음소 패턴에 과적합되어 발음 오류 음소를 정답 발음 음소로 오인식할 가능성이 있다.

Witt & Young(2000)은 외국어 학습자가 발음 오류를 산출하는 원인을 크게 두 가지로 나누었는데, 첫째는 개별 학습자가 학습 대상 언어의 어떤 단어에 대한 발음 자체에 대해 익숙하지 않을 때, 둘째는 학습자 모어의 음소 체계에 영향을 받아 학습 대상 언어의 어떤 음소 발음을 자신의 모어 음소의 발음으로 대치하는 것이 그것이다. 이중 두 번째 원인으로 보아 L2-ARCTIC과 같이 학습자의 모어 배경이 다양한 데이터셋은 그것을 훈련 데이터로 한 발음 오류 검출 모델을 어떠한 환경에서 실제 적용하느냐에 따라 성능 개선 또는 저하 요인을 제공할 수 있다. 만약 훈련된 모델을 복수의 모어 배경을 가진 비원어민 학습자들의 발음 교육을 위해서 사용한다고 하면 다양한 언어의 음소 체계에 대한 지식을 가진 발음 오류 검출 모델이 이들 학습자의 발음 오류를 정확히 진단하는 일반화된(generalized) 성능을 가질 수 있을

것이라고 기대할 수 있다. 반면 모델을 적용하려는 비원어민 학습자 집단의 모어 배경이 하나의 언어뿐이라고 하였을 때, 해당 언어의 음소 체계에 존재하지 않는 음소를 학습한 것이 오히려 모델로 하여금 발음 오류 진단에 혼란을 초래할 수 있다.

발음 오류 검출 분야에서 본 연구와 관련도가 좀 더 높은 연구로는 세 가지를 소개할 수 있다. 먼저 중국인 한국어 학습자의 발음 오류 검출을 주제로 한 양승희(2016)는 중국인 한국어 학습자에게서 빈번히 나타나는 발음 오류 패턴을 분석하고 이를 HMM 기반 음향 모델 훈련에 이용하였다. 이 연구는 HMM 기반 음향 모델의 한 세부 모듈인 발음 사전(pronunciation dictionary)에 중국인 한국어 학습자의 발음 오류 패턴에 해당하는 음운 현상 지식을 추가해 주어 중국인 한국어 학습자의 단어 단위 발화 인식률을 높이고자 하였다. 더불어 한국어와 중국어 음소 체계 비교를 바탕으로 전사 데이터가 존재하지 않는 외국어로서의 한국어 음성 코퍼스(L2KSC; 이석재 외, 2006) 일부를 전사하여 실험 데이터로 이용하였다. 이 연구에서 제시한 방법론은 비원어민 외국어 학습자의 읽기나 자유 발화 음성을 음향 모델이 더 잘 인식할 수 있도록 하여 더 정확한 교정 피드백을 위한 정보를 제공할 수 있도록 하는 데에 그 활용 가치가 있다고 할 수 있다. 반면 본 연구는 발음 오류 자체를 정확히 인식하는 음향 모델을 구축하는 것을 목적으로 하므로 두 연구의 실험 성능을 직접 비교하는 것은 어렵다.

다음으로 Lee & Glass(2015)는 외국어 학습자 발화와 같은 비원어민 발화 데이터를 사용하지 않는 음소 기반의 발음 오류 검출 모델 방법론을 제안하였다. HMM 기반 음향 모델을 구성하는 기본 구조라 할 수 있는 유한 상태 변환기(finite state transducer)의 일종인 확장 인식 네트워크(extended recognition network; ERN)를 이용하여 음소 변이를 추가 정의하고 이를 영어 원어민 발화 데이터셋으로 훈련한 뒤, 중국인 영어 학습자의 발화를 입력받은 후 반복적으로 강제 정렬과 오류 후보 선택을 수행하면서 발음 오류 음소를 검출하는 방식이다.

교차 언어적 훈련을 연구한 것으로는 Joshi et al.(2015)을 들 수 있다. 학습 대상 언어를 인도 표준 영어, 학습자 모어를 구자라트어라 하고 학습 대상 언어의 음성 데이터 양이 제한되어 있다고 할 때, 미국 영어와 힌디어 등 다른 언어들의 원어민 발화 데이터를 이용하여 HMM-DNN 기반 발음 오류 검출 모델을 교차 언어적으로 훈련하는 방법론을 제시하였다. 단 이 연구는 모음의 발음 오류만을 검출 대상으로 했다는 점에서 모든 음소를 대상으로 하는 본 연구의 범위와는

차이가 있으며, 모델 훈련 방식에 있어서도 사전 훈련된 모델을 파인튜닝하는 데 있어 교차 언어적 방법론을 사용하고자 하는 본 연구와 다른 점이 있다.

2.2. 한국어와 중국어의 음소 체계

2.2.1. 한국어와 중국어의 자음 체계

한국어와 중국어의 자음 체계 및 그 변이음을 동시 조음에 의한 변동을 고려하지 않고 표로 정리하여 각각 보이면 표 2.1, 표 2.2와 같다(이호영, 1996; Duanmu, 2007). 변이음은 괄호 안에 표시하였다. 단 이호영(1996)에서는 한국어 평 파열음 /ㅂ/, /ㄷ/, /ㄱ/를 유성음 기호에 무성음 부호를 붙여 /b̥/, /d̥/, /g̊/와 같이 표기했으나, 본 연구에서는 별도의 부호를 사용하지 않고 무성음 기호인 /p/, /t/, /k/로 표기한다. 변이음을 포함하지 않는다면 한국어에는 반모음 3개를 제외하고 총 19개, 중국어에는 반모음 3개 및 성절 자음(syllabic consonant) /z/를 제외하고 총 19개의 자음 음소가 있다.

다른 언어들과 비교하여 한국어 자음 체계에서 가장 두드러지는 특징은 파열음 및 파찰음 음소가 기본적으로 모두 무성음이며, 기식성과 긴장성에 따라 평음, 격음과 경음의 세 가지 종류로 나뉜다는 것이다. 마찰음 또한 긴장성 여부에 따라 평음과 경음으로 나뉜다. 이와 같은 특징은 한국어의 평음, 격음, 경음을 중국어의 무기음과 유기음에 각각 어떻게 대응시켜야 하는지에 대한 문제의 원인이 된다(조민하, 2017; Zheng Yihan & 원미진, 2021).

파열음, 파찰음 및 마찰음은 어말에서 불파음으로 실행되고 조음 위치가 같은 평 파열음으로 중화된다. 비음도 마찬가지로 어말에서는 불파음으로 실현된다. 또 마찰음의 경우 전설 고모음 [i] 앞에서 구개음화가 일어나 경구개음으로서 나타난다. 마지막으로 유음은 기본적으로는 탄설음으로 나타나지만 어말에서 또는 자기 자신에 후행하는 초성 위치에서 설측음으로서 실현된다.

중국어의 자음 음소들 중 파열음, 파찰음 및 마찰음은 기식성 여부에 따라 무기음과 유기음으로 구별되며, 치조 파찰음과 유기 치조 마찰음의 구개음화된 변이음이 존재한다. 비음의 경우 치조음에 불파음으로서 변이음이 존재한다. 한국어와 다르게 중국어 자음 체계에는 권설음이 있기도 하다.

표 2.1 한국어의 자음 체계

		양순음		치조음		경구개음	연구개음		성문음
파열음	평음	p	(pʾ)	t	(tʾ)		k	(kʾ)	
	격음	pʰ		tʰ			kʰ		
	경음	p⁼		t⁼			k⁼		
파찰음	평음					tɕ			
	격음					tɕʰ			
	경음					tɕ⁼			
마찰음	평음			s		(ɕ)			h
	경음			s⁼		(ɕ⁼)			
유음	탄설음			ɾ					
	설측음			(l)					
비음		m	(mʾ)	n	(nʾ)		ŋ		
반모음	비원순음					j	ɥ		
	원순음						w		

표 2.2 중국어의 자음 체계

		양순음	순치음	치조음		권설음	경구개음	연구개음
파열음	무기음	p		t				k
	유기음	pʰ		tʰ				kʰ
파찰음	무기음			ts		tʂ	(tɕ)	
	유기음			tsʰ		tʂʰ	(tɕʰ)	
마찰음	무기음			z		ʐ		
	유기음		f	s		ʂ	(ɕ)	x
유음	설측음			l				
비음		m		n	(nʾ)			ŋ
반모음	비원순음						j	
	원순음						ɥ	w

한국어와 구별되는 중국어 자음 체계의 특징으로는 성절 자음이 있다. 성절 자음은 그 자체로 음절을 구성할 수 있는 성절성을 가지며, 중국어에는 총 두 개로 /z/ 및 /ʐ/가 존재한다. 한 음절에서 전자는 /s/, 후자는 /ʂ/가 초성일 때 나타난다.

2.2.2. 한국어와 중국어의 모음 체계

두 언어의 모음 체계에 대한 표를 제시하면 표 2.3 및 표 2.4와 같다. 표 2.1, 2.2와 같이 변이음은 괄호 안에 표시하였다.

표 2.3 한국어의 모음 체계

	전설모음		중설모음		후설모음	
	비원순	원순	비원순	원순	비원순	원순
고모음	i				ɯ	u
중고모음	e					o
저모음			a		ʌ	

표 2.4 중국어의 모음 체계

	전설모음		중설모음		후설모음	
	비원순	원순	비원순	원순	비원순	원순
고모음	y	i				u
중고모음					(o)	(ɤ)
중모음			ə			
저모음			a		(ɑ)	
권설 모음	ɚ					

두 언어 모두 모음 음소의 수가 몇 개인지에 대해서는 완전한 동의가 이루어지지 않았다. 한국어의 경우 7모음, 8모음, 9모음, 10모음 체계 등의 이론이 있는데(조문우, 2006), 본 연구에서는 이호영(1996)을 따르되 같은 연구에서 지적한 젊은 세대 서울 토박이들이 /ㅔ/와 /ㅐ/를 구별하지 못한다는 점, /ㅚ/와 /ㅟ/는 이중 모음으로 발음되기도 한다는 점을 받아들여 7모음 체계를 취한다. 따라서 본 연구에서 한국어 모음 음소는 7개이다.

중국어의 경우 Duanmu(2007)를 따라 5개의 기본 모음 음소 및 권설 모음 1개가 있다고 정의한다. 다만 이들 모음은 변이음을 가지고 있는 경우가 있다. 후설 원순 중고모음 [ɤ], 후설 비원순 중고모음 [o]는 중모음 /ə/의 변이음으로, /ə/가 개음절의 중성 위치에 있을 때 초성이 양순음 혹은 경구개음일 시 [o], 이외의 경우 [ɤ]로 나타난다. 후설 원순 저모음 [ɑ]는 중설 비원순 저모음 /a/의 변이음이다. 또한 중국어에는 5개의 모음 외에도 권설 모음 /ɚ/가 존재한다.

2.3. 중국인 한국어 학습자의 발음 오류

한국어 학습자의 발음 오류에 대한 연구 중에서는 중국인 학습자들을 대상으로 한 것이 대부분을 차지한다(원다송, 2016). 중국인 한국어 학습자의 발음 오류를 대상으로 하는 연구들 가운데서는 자음 중

파열음과 종성 비음, 그리고 단모음의 발음 오류에 대한 것들이 많은 수를 차지했다.

유효려(2008), 위원징 & 홍미주(2013), Zheng Yihan & 원미진(2021) 등은 중국인 한국어 학습자의 어두 파열음 발음 오류에 대해 연구하였다. 특히 평음 오류에 대해서 격음 또는 경음 중 어느 한 방향으로 오류가 빈번하게 발생하는 경향을 띠는 것이 아니라, 격음과 경음으로 잘못 발음하는 오류가 서로 빈번하게 산출되는 쌍방향적인 오류 양상이 관찰된다고 하였다. 이들 연구는 공통적으로 이와 같은 오류의 원인으로서 두 언어의 파열음 체계가 서로 다르기 때문임을 지적하고 있는데, 특히 Zheng Yihan & 원미진(2021)은 음운론적인 대조를 통한 연구들에서는 중국어 무기 파열음을 한국어 평 파열음에 대응하여야 한다고 주장하고, 음성학적 특성을 측정하는 방식으로 진행한 연구들에서는 경 파열음에 대응하여야 한다고 주장하는 등 이와 같은 문제에 대해 동의가 이루어지지 않았다고 하였다.

종성 오류에 대한 연구로는 유효려(2008), 위국봉(2014), 배영환 & 金成蘭(2019) 등을 꼽을 수 있다. 특히 유효려(2008)는 이전까지의 중국인 한국어 학습자의 발음 오류 연구에서 종성 비음에 대한 오류가 아주 많이 지적되고 있다고 하였다. 위국봉(2014)은 음소적 맥락을 고려하여 선행 모음에 따른 종성 비음의 발음 오류를 연구하였다. 배영환 & 金成蘭(2019)은 전체적인 종성 발음 오류를 어말이라는 환경에만 국한하지 않고 알아보았는데, 삭제 오류는 /ㄱ/와 /ㄹ/, 대치 오류는 /ㅇ/, 첨가는 /ㄴ/과 /ㅁ/에 대한 것이 두드러진다고 하였으며 이들 오류 양상은 또한 해당 음소를 포함하는 단어가 한자어인지의 여부에 따라서도 달라진다고 하였다.

마지막으로 모음 오류에 대한 연구로는 조문우(2006), 황미경 & 조남민(2021)을 소개할 수 있다. 두 연구는 한국어의 /ㅡ/, /ㅜ/, /ㅓ/, /ㅗ/가 중국인 학습자들이 가장 발음을 어려워하는 모음이라고 주장하였고, 특히 조문우(2006)는 음향음성학적인 방법으로 두 언어의 모음 특성을 분석하여야 중국인 한국어 학습자에 대한 효과적인 모음 발음 지도가 가능하다고 하였다.

이외에도 자음군에 해당하는 겹받침 오류에 대한 연구(박시균 & 최경복, 2014), 한국어 음절 구조를 원인으로 하는 발음 오류에 대한 연구(장향실, 2009; 장찌엔, 2017), 이중 모음의 발음 오류에 대한 연구(노채환, 2020) 등이 있었고, 음운 현상 규칙을 알맞게 적용하지 못하여 산출된 발음 오류에 대한 연구(정효주 & 김영주, 2011; 정효주

& 김영주, 2012; 양일, 2015; 조민하 & 여문리, 2017) 또한 다수 존재하였다.

2.4. wav2vec 2.0

wav2vec 2.0(Baevski et al., 2020)은 트랜스포머(Transformer; Vaswani et al., 2017) 구조 기반의 심층 신경망 음향 모델이다. wav2vec 2.0의 전신 모델인 wav2vec(Schneider et al., 2019)은 역시 트랜스포머 구조 기반 자연어 처리 모델인 BERT(Devlin et al., 2018)의 MLM(masked language modeling) 및 대조 예측 부호화(contrastive predictive coding; van den Oord et al., 2018) 기법에서 착안하여 일반적으로 음향 모델에서 사용하는 MFCC(Mel-frequency cepstral coefficients), filter-bank, 스펙트로그램 등 별도의 특징 추출 과정을 대체하는 계층을 가진 심층 신경망 모델을 제안하였으며, wav2vec 2.0은 그 후속 연구에서 제안한 모델이다.

wav2vec 2.0의 학습은 사전 훈련(pre-training)과 파인튜닝(fine-tuning) 단계로 구분된다. 사전 훈련 단계는 정답 레이블 없이 음성 데이터만으로 자기 지도 학습(self-supervised learning)을 수행한다. 먼저 원시 오디오(raw audio) 데이터가 다층의 합성곱 특징 인코더(convolutional feature encoder) $f: \mathcal{X} \mapsto \mathcal{Z}$ 에 의해 처리되어 은닉 음성 표현(latent speech representation) $\mathbf{z}_1, \ldots, \mathbf{z}_T$ 로 변환되고, 이들은 다시 각각 트랜스포머 계층 $g: \mathcal{Z} \mapsto \mathcal{C}$ 을 지나 맥락화된 표현 $\mathbf{c}_1, \ldots, \mathbf{c}_\mathrm{T}$, 양자화 모듈 $\mathcal{Z} \mapsto \mathcal{Q}$ 을 지나 이산화된 표현 $\mathbf{q}_1, \ldots, \mathbf{q}_T$ 을 출력하여 자기 지도 학습 쌍을 구성한다.

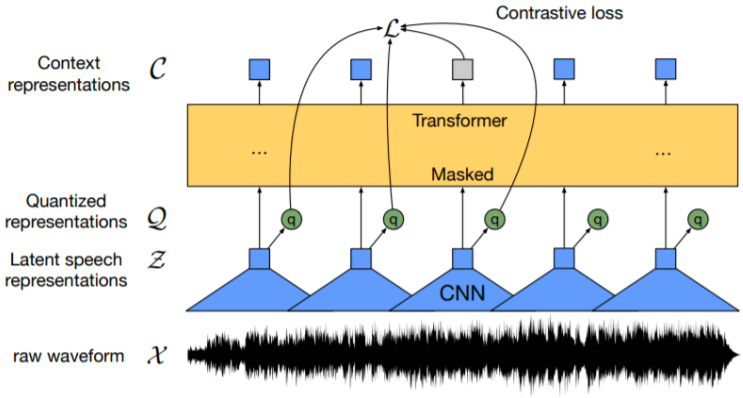

그림 2.1 wav2vec 2.0의 사전 훈련(Baevski et al., 2020)

여기서 일정 비율 p 만큼의 은닉 음성 표현 시점들은 트랜스포머 계층에 입력될 때 일정 길이 M 만큼 훈련된 마스크 벡터로 대체되며, 양자화 모듈에 입력될 때는 별도의 마스킹 과정을 거치지 않는다. 양자화 모듈은 입력받은 은닉 음성 표현과 검벨 소프트맥스(Gumbel softmax)를 이용하여 각각 V 개의 표제어(entry) $e \in \mathbb{R}^{V \times d/G}$ 를 가진 G 개의 코드북(codebook)에서 표제어 벡터 $e_1, \ldots, e_G$ 를 선택하고, 이를 모두 연결(concatenate)한 뒤 선형 변환 $\mathbb{R}^d \longmapsto \mathbb{R}^f$ 을 수행하여 최종적으로 이산화된 표현 $\mathbf{q} \in \mathbb{R}^f$를 얻어 내는 방식으로 작동한다.

이와 같은 자기 지도 학습 쌍 $(\mathbf{c}_1, \ldots, \mathbf{c}_\mathrm{T}, \mathbf{q}_1, \ldots, \mathbf{q}_T)$이 존재할 때, 사전 학습으로 하여금 풀도록 하는 문제는 트랜스포머 계층 입력에서 마스킹된 시점에 대한 양자화된 음성 표현을 찾는 대조 학습 과제(contrastive learning task)이고 이는 대조 손실 함수(contrastive loss) $\mathcal{L}_m$ 으로 나타낼 수 있다. t 시점에서 마스킹된 맥락화된 표현을 $\mathbf{c}_t$ 라고 했을 때, 대조 손실 함수 $\mathcal{L}_m$ 은 $\mathbf{c}_t$ 와, 정답 양자화된 음성 표현 $\mathbf{q}_t$ 와 양자화된 음성 표현 집합 $\mathbf{Q}_t$ 에서 샘플링된 K 개의 오답 선택지(distractor) $\tilde{\mathbf{q}}$ 로 이루어진 $K+1$ 개의 음성 표현 사이의 코사인 유사도를 계산하여 만들어지는 소프트맥스 함수로, 다음과 같이 정의된다.

$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)}$$

수식 2.1 wav2vec 2.0의 대조 손실 함수

더불어 wav2vec 2.0의 사전 훈련 손실 함수에서는 각 코드북 표제어의 균일화된 사용을 유도하도록 각 코드북 표제어에 대한 소프트맥스 분포의 엔트로피를 최대화하는 다양성 손실 함수(diversity loss) $\mathcal{L}_d$를 다음과 같이 추가 정의한다.

$$\mathcal{L} = \frac{1}{GV} \sum_{g=1}^{G} -H(\bar{p}_g) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{G} \bar{p}_{g,v} \log \bar{p}_{g,v}$$

수식 2.2 wav2vec 2.0의 다양성 손실 함수

따라서 최종적인 wav2vec 2.0의 사전 훈련 손실 함수 $\mathcal{L}$은 다음과 같이 정의된다.

$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_d$$

수식 2.3 wav2vec 2.0의 손실 함수

이때 α는 하이퍼파라미터이다.

파인튜닝 과정에서는 음성 인식, 음성 분류 등 다양하게 정의할 수 있는 하위 과제(downstream task)에 따라, 하위 과제를 푸는 데 적절한 형태로 추가 계층 및 전용 손실 함수를 정의한다. 이때 양자화 모듈은 사용되지 않는다. 예를 들어 단어·글자·음소 등 단위의 음성 인식을 하위 과제로 설정했다면 트랜스포머 계층 이후에 선형 계층이 추가로 정의되며, 이 계층은 입력받은 맥락화된 음성 표현을 파인튜닝 모델이 풀고자 하는 문제에서 정의한 레이블 집합의 크기와 같은 C차원의 벡터로 변환하여 각 레이블에 대한 모델의 확률분포 예측을 모델링하는 역할을 수행한다. 파인튜닝에서의 손실 함수는 하위 과제의 성격에 따라 달리할 수 있으나 wav2vec 2.0을 최초로 제시한 Baevski et al.(2020)에서는 글자 단위 음성 인식을 하위 과제로 하고, CTC 손실 함수(Graves et al., 2006)를 사용하였다. 사전 훈련과 파인튜닝 과정에서의 모델의 과제는 서로 다르므로, wav2vec 2.0 모델의 훈련은 전이 학습(transfer learning)에 속한다.

Baevski et al.(2020)은 wav2vec 2.0 모델이 효과적인 음성 특징 추출 방식을 심층 신경망으로 하여금 학습하게 하는 사전 훈련 방식을 기초로 하여 전사 텍스트가 적다고 하더라도 만족할 만한 성능을 낼 수 있다고 주장하였으며 발음 오류 검출 연구에서도 wav2vec 2.0 기반이 아닌 다른 모델들과 다르게 소량의 학습자 데이터셋만으로 파인튜닝을 진행하기도 한다(Xu et al., 2021; Yang et al., 2022). 사전 훈련의 경우 대량의 음성 데이터와 대규모의 컴퓨팅 자원을 필요로 하므로 많은 수의 사전 훈련 모델이 보고되고 있지는 않지만, 원 논문인 Baevski et al.(2020)에서 공개한 사전 훈련 모델 외에도 XLSR-53(Conneau et al., 2020) 등 다양한 언어의 음성 데이터를 이용하여 교차 언어적 사전 훈련을 진행한 모델 또한 공개적으로 이용할 수 있다.

wav2vec 모델을 음향 모델로 하는 발음 오류 검출 모델의 연구는 그 적용 가능성을 탐구하는 성격으로 이루어지고 있다. Peng et al.(2021) 및 Wu et al.(2021)은 그중 가장 기초적인 성격의 연구로, L2-ARCTIC과 TIMIT, 또는 중국인 영어 학습자 데이터셋인 CU-CHLOE(Meng et al., 2009)를 이용하여 파인튜닝을 진행하였다. 이들

연구에서는 강제 정렬 모듈을 추가하거나 별도의 모델 구조를 변경하지 않고 일반적인 wav2vec 2.0 기반의 음소 인식 실험을 수행했다. Xu et al.(2021)은 wav2vec 2.0 모델에 HMM-DNN 음향 모델 기반의 강제 정렬 모듈을 추가한 이진 분류 방식의 발음 오류 검출 모델을 제안하였다. Peng et al.(2022)은 학습자가 발화한 단어의 정보에 접근할 수 있다는 발음 오류 검출 상황의 특성을 이용해, wav2vec 2.0 모델을 오디오 인코더로, 트랜스포머 기반의 모델을 텍스트 인코더로 정의하여 두 정보를 혼합하여 사용하는 모델을 제시하였다. Yang et al.(2022)은 외부 모듈이 없는 wav2vec 2.0 모델만을 사용하되 파인튜닝 단계에 의사 레이블링(pseudo-labeling) 단계를 추가하는 방식으로 인식 기반 발음 오류 검출 모델의 성능을 제고하고자 하였다.

3. 실험 방법론

3.1. 데이터셋

실험에서는 한국어와 중국어 원어민 발화 데이터셋 및 이들의 성능을 비교·평가하기 위한 데이터셋으로 중국인 한국어 학습자의 한국어 비원어민 발화 데이터셋을 이용한다.

3.1.1. KsponSpeech

본 연구에서 학습 대상 언어로 설정한 한국어의 원어민 발화 데이터셋으로는 KsponSpeech(Korean Spontaneous Speech Corpus for Automatic Speech Recognition; Bang et al., 2020) 데이터셋을 사용한다. 이 데이터셋은 다양한 주제에 대해 화자 2,000명이 자유롭게 발화한 총 969시간의 음성으로 이루어져 있고, 각 발화를 전사한 한글 텍스트가 함께 제공되는 공개 데이터셋이다.

데이터셋은 훈련, 검증 및 시험 셋으로 이루어져 있으며 이중 시험 셋은 다시 상대적으로 깨끗한 음성 환경을 가진 것과 그렇지 못한 것으로 나뉘어 두 개 셋으로 존재한다. 단 본 연구에서는 검증과 시험 셋은 사용하지 않으며, 단일 언어적 파인튜닝 모델의 공평한 성능 비교와 교차 언어적 파인튜닝 모델을 학습할 시 이 데이터셋에 과적합되는 현상의 방지를 위해 중국어 원어민 데이터인 AISHELL-1 데이터셋과 발화 수 및 시간 분량 면에서 비슷한 규모를 유지하고자 하였다. 이에 따라 본 연구에서 이용하는 KsponSpeech 데이터셋은 전체 620,000개 발화로 이루어진 훈련 셋 중 117,790개를 무작위 선택한 것이다. 이들 발화의 총량을 시간 단위로 계산하면 약 183시간이다. 본래 개별 음성 발화 데이터는 pcm 확장자로 이루어진 파일이나 본 연구에서는 이것을 wav 확장자로 변환한 파일을 사용한다.

3.1.2. AISHELL-1

한국어 학습자의 모어로 설정한 중국어의 원어민 음성 데이터셋으로는 현재 중국어 음성 인식 연구에서 널리 사용되고 있는 공개 데이터셋인 AISHELL-1(Bu et al., 2017) 데이터셋을 사용한다.

데이터셋은 400명의 화자가 녹음한 총 165시간 분량의 발화 음성으로 이루어져 있으며, 경제, 과학·기술, 스포츠, 엔터테인먼트, 뉴스의 5개 분야를 주제로 한 텍스트를 각 화자가 읽어 녹음한 것이다. 각 발화 데이터에 대한 중국어 간체자 전사 텍스트가 함께 제공되고 전체 데이터셋에서 등장하는 단어에 대한 한어병음 및 성조 정보가 별도의 사전 형식으로 마련되어 있다.

이 데이터셋 또한 훈련, 검증과 시험 셋으로 이루어져 있으나 KsponSpeech 데이터셋과 같이 검증 및 시험 셋은 본 연구에서 이용하지 않는다. 이용하는 훈련 셋은 총 120,098개 발화, 약 151시간 분량이다.

3.1.3. L2KSC

L2KSC(이석재 외, 2006)는 외국어로서 한국어를 학습하는 비원어민 학습자 발화를 녹음한 음성 데이터셋이다. 총 29개의 모어를 배경으로 가진 357명의 화자가 녹음에 참여하였고 그중 51명의 화자는 한국어 모어 사용자로, 이들의 발화 음성 데이터를 대조군으로 이용할 수 있다. 이 데이터셋은 사전 정의된 학습자 모어 및 한국어 단어 리스트, 짧은 이야기, 주어진 상황에 대한 설명 등을 제재로 하는 총 6개의 하위 데이터셋으로 이루어져 있다.

본 연구에서는 전체 L2KSC 데이터셋에서 중국어를 모어로 사용하는 79명의 화자 데이터만을 사용하고, 그들이 300개의 사전 정의된 한국어 1~2어절 리스트를 읽어 녹음한 데이터로 구성된 하위 데이터셋을 사용한다. 앞으로 이와 같은 L2KSC의 일부분만을 L2KSC 데이터로 칭한다.

L2KSC 데이터셋은 화자의 성별 또는 한국어 능력에 따라 파일들이 구분되어 있기는 하지만 별도의 훈련, 검증 및 시험 셋이 정의되어 있지 않으므로 본 연구에서 임의로 지정하였다. 훈련, 검증, 시험 셋은 각각 화자의 구성은 겹치지만 단어의 구성은 겹치지 않도록 하였다. 이중 훈련 셋은 베이스라인 모델의 훈련 데이터로, 검증 및 시험 셋은 베이스라인 모델을 포함한 모든 파인튜닝 모델에 대한 검증 및 시험 데이터로 사용하였다.

위에서 서술한 세 데이터셋 중 본 연구의 실험에서 사용한 부분에 대한 정보를 표로 설명하면 표 3.1과 같다.

표 3.1 실험에 사용한 데이터셋 분량 정리

데이터셋	훈련 셋	검증 셋	시험 셋
KsponSpeech	117,790 발화 (183시간 11분 44초)	X	X
AISHELL-1	120,098 발화 (150시간 51분 7초)	X	X
L2KSC	13,141 발화 (5시간 55분 42초)	469 발화 (12분 22초)	469 발화 (12분 42초)

3.2. 전사 텍스트 가공

한국어와 중국어 두 언어의 발화 데이터셋을 모두 이용하여 교차 언어적 방식으로 음향 모델을 훈련하기 위해서는 두 데이터셋의 전사 텍스트가 같은 기호 체계를 가져야 한다. 본 연구에서는 국제 음성 기호(international phonetic alphabet; IPA) 체계가 두 언어의 음성 발화를 가장 중립적이고 일관적으로 나타낼 수 있는 표기 체계라고 판단하여, IPA 기호 체계를 모든 데이터셋의 전사 단위로 하여 실험을 진행하고자 한다. 단 본 연구에서 사용하는 데이터셋 중 어떤 것도 IPA를 전사 단위로 취하고 있지 않으므로 이들 전사 텍스트를 IPA 텍스트로 변환하는 과정이 필요하다. 각 데이터셋에 따라 본래의 전사 텍스트를 IPA 단위로 변환하는 과정을 제시하면 다음과 같다. 구체적인 각 데이터셋별 원 전사 단위와 IPA 기호의 대응 표는 부록 A에 수록한다.

3.2.1. KsponSpeech

KsponSpeech 데이터셋에서 제공하는 전사 데이터는 기본적으로는 한글로 이루어져 있으나, 숫자 및 로마자와 같은 타 문자 체계에 속한 텍스트와 잡음, 고쳐 말하기 등을 표시하기 위한 특수 기호 또한 존재한다. 이 전사 데이터는 먼저 다음과 같은 과정을 거쳐 한글로만 구성되어 있도록 1차 전처리를 수행했다.

1. 비언어적 또는 준언어적 표현임을 표시하는 특수 기호는 삭제한다.
2. 숫자 및 별도 기호로 표기하는 단위 등은 병기되어 있는 한글 전사를 선택한다.

원 텍스트	한 b/ (20분/이십 분) 봤지 TV를.
1차 전처리	한 이십 분 봤지 티비를
음운 현상 적용	하 니십 뿐 봗찌 티비를
자모 단위 분리	ㅎㅏㄴㅣㅅㅣㅂㅃㅜㄴㅂㅘㄷㅉㅣㅌㅣㅂㅣㄹㅡㄹ
IPA 단위 변환	hanicip'p=un'pwat'tɕ=itʰipiɾɯl

그림 3.1 KsponSpeech 전사 텍스트 가공 과정

3. 로마자로 표기되어 있는 영어 등 외국어는 해당하는 단어의 한글 표기로 변환한다.

1차로 전처리된 한글 전사 텍스트들은 다시 다음과 같은 과정을 거쳐 최종적으로 IPA 텍스트로 변환한다.

4. Python 패키지인 g2pK[①]를 이용하여 한글 표준 발음법에서 정하는 음운 현상을 적용한 한글 텍스트로 바꾼다. g2pK를 이용하여 음운 현상을 적용하는 과정에서 오류를 동반하는 텍스트의 경우 오류 부분에 수동으로 음운 현상을 적용하였다.
5. Python 패키지인 jamo[②]를 이용하여 음절 단위의 한글 완성형 텍스트를 초·중·종성 자모 단위의 한글 조합형 텍스트로 바꾸고 띄어쓰기 단위를 무시한다.
6. 분리된 한글 자모를 각각 대응하는 IPA 기호로 바꾼다. 한글 자모의 IPA 대응은 이호영(1996) 및 이호영(1996)을 참고하여 L2KSC 데이터셋을 전사한 양승희(2016)를 참고하였으며, 그 내용은 2.2.에서 서술하는 바와 같다.

3.2.2. AISHELL-1

AISHELL-1 데이터셋의 전사 데이터는 발음 사전과 단어 단위로 띄어쓰기된 중국어 간체자 전사 텍스트로 이루어져 있다. 이중 발음 사전은 사용하지 않고, 아래에서 서술하는 과정을 통해 중국어 간체자 텍스트를 IPA로 변환하였다. 단 한자를 바로 IPA로 변환하지 않고, 중간 단계로 한어병음 단계를 거쳤다. 변환된 한어병음 단위의 텍스트는 추가적인 처리 없이 IPA 기호로 일대일 대응이 가능하다.

---

① https://github.com/Kyubyong/g2pK
② https://pypi.org/project/jamo/

원 텍스트 我们 可以 暂时 假设 一 下
한어병음 변환 wo men ke yi zan shi jia she yi xia
IPA 단위 변환 w o m ə n˥˩ kʰ ɤ i ts a n˥˩ ʂ ʐ̩ tɕ j ɑ ʂ ɤ i ɕ j ɑ

그림 3.2 AISHELL-1 전사 텍스트 가공 과정

1. Python 패키지인 pypinyin [③] 을 이용하여 중국어 간체자 텍스트를 성모·운모 단위 한어병음 텍스트로 바꾼다. 현재 공개되어 있는 한어병음 변환 패키지 중에서 pypinyin이 가장 정확한 변환 결과를 보였다. 표준 한국어에는 성조가 존재하지 않으므로, 병음 텍스트 내 성조 정보는 버린다. 역시 원본 데이터에 있는 공백 단위는 무시한다.
2. 분리된 한어병음을 IPA 기호로 바꾼다. IPA 기호로의 변환에 있어서는 Duanmu(2007)를 참고하였으며, 그 내용은 2.2.에서 서술하는 내용과 같다. 텍스트 중 로마자가 존재하는 경우 수동으로 IPA 기호로 바꾸었다.

3.2.3. L2KSC

원본 L2KSC 데이터셋에는 전사 텍스트가 제공되지 않는다. 단 본 연구에서 이용하는 부분에 한해 양승희(2016)에서 전사 작업을 진행한 바 있다. 이 연구에서 정의한 전사 단위는 한국어 PLU(phone-likely unit)이며 전체 PLU 셋의 구체적 내용 정의는 이호영(1996)을 참고하였으며 중국인 학습자의 모어에 영향 받은 발음 오류를 전사하기 위해 6개의 추가적인 PLU를 정의하였다. 각 학습자 음성은 소리 나는 대로 전사하였다. 학습자가 읽은 텍스트에 대한 표준 발음 또한 같은 셋의 PLU 연쇄로 제공되고 있는데, 이중 잘못되어 있는 것은 본 연구에서 직접 수정하였다.

이 전사 데이터는 음운 현상이 이미 적용되었거나 발음 오류가 존재하는 발화를 소리 나는 대로 받아 적은 것이기 때문에, 양승희(2016)

원 텍스트(표준 발음) sil K WW R AX CH IY M AA NN sil
IPA 단위 변환(표준 발음) k ɯ ɾ ʌ tɕʰ i m a n˥˩
원 텍스트(학습자 발화) sil K WW L AX_NX CH IY M AA NN sil
IPA 단위 변환(학습자 발화) k ɯ l ʌ ŋ˥˩ tɕʰ i m a n˥˩

그림 3.3 L2KSC 전사 텍스트 가공 과정

---

③ https://pypi.org/project/pypinyin/

에서 진행한 HMM 기반 음성인식기의 훈련을 위해 텍스트 앞뒤로 추가되어 있는 발화 공백(silence) 토큰을 제거하는 등 간단한 전처리 과정을 제외하면 별도의 중간 단계는 필요치 않다. PLU를 IPA로 변환하는 한국어 PLU의 경우 3.2.1.의 6, 추가 PLU의 경우 3.2.2.의 2와 같은 방식을 취하였다.

3.3. 사전 훈련 모델

실험은 사전 훈련된 wav2vec 2.0 기반 모델을 위에서 설명한 세 데이터셋을 이용해 파인튜닝하는 방식으로 진행한다.

사전 훈련된 모델은 wav2vec 2.0 Large(Baevski et al., 2020)와 XLSR-53(Conneau et al., 2020)을 이용하였다. wav2vec 2.0 Large는 960시간 분량의 영어 원어민 발화 데이터셋인 LibriSpeech(Panayatov et al., 2015)로 사전 훈련되었으며, 이 모델을 역시 영어 원어민 발화 데이터셋인 TIMIT(Garofolo, 1994)에 대해 음소 인식 파인튜닝을 진행하였을 때 언어 모델 없이 8.3%의 시험 데이터 음소 오류율(phoneme error rate)을 보인다. XLSR-53 모델은 세 개의 데이터셋을 합친 53개 언어, 56,000시간의 다언어 원어민 발화 데이터셋으로 사전 훈련되었고, 역시 다언어 원어민 발화 데이터셋인 CommonVoice(Ardila et al., 2019)에 대한 파인튜닝 성능 평균 7.6%의 음소 오류율, BABEL(Gales et al., 2014)에 대한 파인튜닝 성능 평균 22.4%의 글자 오류율(character error rate)을 보인다.

두 사전 훈련 모델은 그 구조와 규모가 같으며, 사전 훈련 방식 또한 같다. 구체적인 모델 설정은 Baevski et al.(2020)과 Conneau et al.(2020)을 따른다.

3.4. 파인튜닝

본 연구에서 제기하는 연구 질문에 따라 파인튜닝 실험을 설계한다. 파인튜닝 설정은 크게 베이스라인 환경, 단일 언어적 환경 및 교차 언어적 환경으로 나뉜다. 각 설정은 모델의 파인튜닝 대상 훈련 데이터셋만을 달리하며, 이외 환경은 같도록 한다.

파인튜닝 모델은 사전 훈련 모델의 마지막 트랜스포머 블록 이후 한 개의 선형 계층을 추가하여, 주어진 발화 음성 파형을 입력으로 받아 IPA 기호 연쇄를 출력하도록 훈련하며, 목적 함수로는 CTC 손실

함수(Graves et al., 2006)를 사용한다. 각 파인튜닝 모델이 학습하는 IPA 기호 체계는 모두 동일하다.

3.4.1. 베이스라인 환경

먼저 일반적인 발음 오류 검출 모델의 학습 방식과 동일하게, 중국인 한국어 학습자 발화 데이터셋으로 파인튜닝하는 베이스라인 환경을 설정한다. 베이스라인 환경에서 파인튜닝된 모델은 단일 언어적 또는 교차 언어적 환경에서 파인튜닝한 모델들의 성능을 비교하는 기준으로 사용한다. 베이스라인 환경은 L2KSC 데이터셋의 훈련 셋만으로 파인튜닝을 진행한 것이며, 이를 L2KSC 모델이라 한다.

3.4.2. 단일 언어적 파인튜닝 환경

사전 훈련된 wav2vec 2.0 기반 모델들을 중국인 한국어 학습자의 모어인 중국어와, 학습 대상 언어인 한국어 원어민 음성 데이터만을 각각 사용하여 파인튜닝했을 때의 성능을 알아보기 위해 총 두 개의 단일 언어적 파인튜닝 환경을 설정한다. 한국어 원어민 발화 데이터셋인 KsponSpeech 데이터셋만을 사용하여 파인튜닝하는 것을 KsponSpeech 모델, 중국어 원어민 발화 데이터셋인 AISHELL-1 데이터셋만을 사용하여 파인튜닝하는 것을 AISHELL-1 모델이라 한다.

3.4.3. 교차 언어적 파인튜닝 환경

다음으로는 중국인 한국어 학습자의 모어인 중국어 및 학습 대상 언어인 한국어 원어민 발화 데이터셋을 모두 사용하는 교차 언어적 파인튜닝 환경을 설정한다. 교차 언어적 파인튜닝 환경의 모델들은 KsponSpeech와 AISHELL-1 데이터셋을 함께 사용하여 파인튜닝한다. 이것을 Cross-lingual 모델이라 한다.

3.4.4. 파인튜닝 상세

각 파인튜닝 환경의 모델은 1개의 GPU로 훈련하였다. 단 4개의 GPU 환경에서 훈련하는 것을 시뮬레이션하기 위해 손실 함숫값 계산을 4번 수행할 때마다 가중치를 업데이트하였으며, 이러한 업데이트를

100,000번 진행하였다. 최적화 알고리즘으로는 Adam 알고리즘(Kingma & Ba, 2014)을 사용하였으며, $\gamma = 3 \times 10^{-5}, \beta_1 = 0.9, \beta_2 = 0.98$이다. 단 학습률인 γ값은 학습률 스케줄러에 의해 조정되어 첫 10,000회 업데이트 동안에는 $\gamma \times 0.01$ 에서 γ 까지 상승하고, 이후 40,000회 업데이트 동안에는 유지되다가 마지막 50,000회 업데이트 동안에 $\gamma \times 0.05$까지 하강하도록 하였다.

Baevski et al.(2020)에서 제시하는 파인튜닝 시 마스킹 방법 또한 적용하였다. 각 시점은 0.65의 확률로 이후 10개 시점 동안 마스킹되며 각 채널 또한 0.5의 확률로 이후 64개 채널이 마스킹된다. 0.1 확률의 LayerDrop(Fan et al., 2019), 0.1 확률의 활성화 함수 드롭아웃을 적용하였다. 단 별도의 특징 인코더 가중치 동결은 적용하지 않았다.

일반적인 음성 인식 모델에서는 음향 모델에서 처리된 출력 텍스트 연쇄를 언어 모델을 통하여 보정하는 과정을 거치기도 한다. 언어 모델은 음성 데이터를 전사한 텍스트 및 글말 형태로 존재하는 텍스트 등을 학습 대상 데이터로 하여 단어 혹은 글자 단위의 연쇄 패턴을 모델링하는 방식으로 구축한다. 그러나 본 연구에서 다루는 데이터는 발음 오류를 포함하고 있으므로, 오류가 없거나 자연스러운 텍스트 연쇄 패턴을 모델링하는 것을 학습 목표로 하는 언어 모델의 존재가 오히려 성능 저하에 영향을 줄 수 있다. 따라서 본 연구의 실험에서는 별도의 언어 모델은 사용하지 않았다.

파인튜닝 환경에 관계없이 검증 데이터는 L2KSC 검증 셋으로 설정하였으며, 2,500번의 가중치 업데이트 시마다 검증 데이터에 대해 음소 오류율(phoneme error rate)을 계산하였다. 음소 오류율은 음소 단위의 삽입 오류, 삭제 오류, 대치 오류의 개수를 합한 것을 전체 정답 음소의 개수로 나눈 값이다. 100,000회의 가중치 업데이트가 완료되었을 때를 기준으로 검증 데이터에 대한 음소 오류율이 가장 낮을 때의 가중치 값을 최종 모델로 선택하였다. 모든 파인튜닝 실험은 fairseq(Ott et al., 2019)를 통해 진행하였다.

3.5. 평가 척도

각 모델의 성능 평가는 음소 인식과 발음 오류 검출의 두 단계로 나누어 진행하고, 각 단계에서의 평가 척도는 아래와 같이 정의한다. 음소 인식 척도는 각 파인튜닝 모델이 L2KSC 시험 셋에서 나타나는 중국인 한국어 학습자 발화의 각 음소를 발음 오류 여부와 상관없이

제대로 예측했는지를 알아보는 척도이며, 발음 오류 검출 척도는 각 음소가 발음 오류인지 여부에 따라 경우를 나누어 모델의 예측 성능을 알아보는 척도이다.

성능 평가를 위한 시험 데이터는 L2KSC의 시험 셋이다. 성능 평가를 위해 정답 IPA 연쇄와 모델 출력 IPA 연쇄는 Needleman-Wunsch 알고리즘(Likic, 2008)을 사용하여 정렬한 후 각 척도 계산을 수행하였고, 계산한 수치는 모두 소수점 셋째 자리에서 반올림하였다.

3.5.1. 음소 인식

음소 인식 단계에서는 정답률(correct rate)과 정확도(accuracy) 척도로 각 파인튜닝 모델의 성능을 평가한다. 각 음성 데이터의 전사 텍스트에 존재하는 정답 IPA 기호의 개수를 N, 모델의 예측 IPA 연쇄 중 대치 오류(substitution error)의 개수를 S, 삭제 오류(deletion error)의 개수를 D, 삽입 오류(insertion error)의 개수를 I라고 했을 때, 정답률 및 정확도를 구하는 식은 아래와 같다.

$$\text{정답률} = \frac{N - S - D}{N}$$

$$\text{정확도} = \frac{N - S - D - I}{N}$$

수식 3.1 정답률 및 정확도의 계산식

3.5.2. 발음 오류 검출

발음 오류 검출 단계에서는 중국인 한국어 학습자 발화에서 나타나는 음소의 발음 오류 여부를 고려한 성능 평가를 진행한다. 먼저 정답 IPA 연쇄와 파인튜닝 모델의 출력 IPA 연쇄를 정렬한 후 모델이 각 음소를 발음 오류라고 예측하였는지의 여부에 따른 평가 척도를 아래와 같이 정의한다.

어떤 한국어 단어 W를 읽은 학습자 발화 X에 대한 전사 텍스트인 IPA 연쇄를 $Y = \{y_1, y_2, y_3, \dots, y_T\}$, W에 대한 표준 음소(canonical phones)를 $Y' = \{y'_1, y'_2, y'_3, \dots, y'_{T'}\}$라고 하고, 같은 발화 X에 대한 파인튜닝 모델의 출력값인 IPA 연쇄를 $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \hat{y}_3, \dots, \hat{y}_{\hat{T}}\}$라 하여 Y와 Y', $\hat{Y}$를 정렬하였을 때 어떠한 시점 t에 대해서,

1. y_t 가 정답 발음이고($y_t = y'_t$) $\hat{y}_t = y_t$ 이면 정인식(true acceptance; TA)이다.
2. y_t 가 발음 오류이고($y_t \neq y'_t$) $\hat{y}_t = y'_t$ 이면 오인식(false acceptance; FA)이다.
3. y_t가 발음 오류이고($y_t \neq y'_t$) $\hat{y}_t \neq y'_t$이면 정거부(true rejection; TR)이다.
4. y_t가 정답 발음이고($y_t = y'_t$) $\hat{y}_t \neq y_t$이면 오거부(false rejection; FR)이다.

위의 정의로부터 오거부율(false rejection rate; FRR) 및 오인식률(false acceptance rate; FAR)을 아래 수식과 같이 정의할 수 있다.

$$FRR = \frac{FR}{TA + FR}$$
$$FAR = \frac{FA}{FA + TR}$$

수식 3.2 FRR 및 FAR의 계산식

또한 모델이 발음 오류로 예측한 것들 중 실제로 발음 오류인 것의 비율을 정밀도(precision)로, 실제로 발음 오류인 것들 중 모델이 발음 오류로 예측한 것의 비율을 재현율(recall)로 정의할 수 있으며, 정밀도와 재현율의 조화 평균인 F1-점수(F1-score)를 아래 수식과 같이 정의할 수 있다.

$$\text{정밀도} = \frac{TR}{TR + FR}$$
$$\text{재현율} = \frac{TR}{TR + FA} = 1 - FAR$$
$$F1\text{-점수} = 2 \times \frac{\text{정밀도} \times \text{재현율}}{\text{정밀도} + \text{재현율}}$$

수식 3.3 정밀도, 재현율 및 F1-점수의 계산식

한편 정거부를 세분화하여 정진단(correct diagnosis; CD)과 진단 오류(diagnosis error; DE)으로 나눌 수 있다. $y_t \neq y'_t$일 때 $\hat{y}_t = y_t$인 경우, 즉 모델이 해당 음소를 정답 발음이 아닌 것으로 제대로

예측하였고 동시에 그 발음 오류 음소의 내용까지 정확히 맞추었다면 정진단이고, 그렇지 않은 경우 진단 오류이다. 진단 오류율(detection error rate; DER), 검출 정확도(detection accuracy), 진단 정확도(diagnosis accuracy)는 아래 수식과 같이 정의한다.

$$DER = \frac{DE}{CD + DE}$$

$$\textit{검출 정확도} = \frac{TA + TR}{TA + FR + FA + TR}$$

$$\textit{진단 정확도} = \frac{CD}{CD + DE} = 1 - DER$$

수식 3.4 DER, 검출 정확도 및 진단 정확도의 계산식

4. 실험

4.1. 음소 인식

표 4.1 파인튜닝 모델의 음소 인식 성능

파인튜닝 모델	정답률	정확도
사전 훈련 모델: wav2vec 2.0 Large		
KsponSpeech	72.74	65.14
AISHELL-1	51.57	38.40
Cross-lingual	*73.15*	*66.43*
L2KSC	**74.92**	**70.87**
사전 훈련 모델: XLSR-53		
KsponSpeech	72.95	64.96
AISHELL-1	53.70	41.46
Cross-lingual	*74.37*	*67.04*
L2KSC	**80.11**	**76.70**

각 파인튜닝 모델의 음소 인식 성능 평가 결과는 표 4.1과 같다. 사전 훈련 모델에 관계없이 교차 언어적 환경의 Cross-lingual 모델 성능이 정답률과 정확도 두 척도 모두에서 단일 언어적 환경의 모델보다 높았다. 그러나 이 모델의 성능이 베이스라인 모델의 그것에는 미치지 못했다. 사전 훈련 모델이 wav2vec 2.0 Large일 때 베이스라인 모델의 성능은 정답률 74.92%, 정확도 70.87%이었고, XLSR-53일 때 정답률 80.11%, 정확도 76.70%이었다.

사전 훈련 모델을 기준으로 각 파인튜닝 모델의 성능을 비교해 보면, wav2vec 2.0 Large 모델로 파인튜닝한 것보다 XLSR-53 모델로 파인튜닝한 것이 약간 더 높은 성능을 보였다. 이것은 영어 데이터로만 사전 훈련하였을 때보다 다언어 데이터를 통해 사전 훈련을 진행하였을 때 음소 인식에서 더 좋은 성능을 보인다는 Conneau et al.(2020)의 실험 결과와 부합하는데, 이 연구에서는 그 이유를 교차 언어적 전이(multi-lingual transfer)로 인하여 모델이 각 음소에 대한 범주화를 더 잘 수행하기 때문이라고 주장하였다. 단 본 연구에서 이용한 단일 언어 사전 훈련 모델은 영어 데이터로 학습한 것으로 최종 시험 데이터를 구성하는 언어인 한국어와는 다르기 때문에 본 실험의 결과 또한 Conneau et al.(2020)에서 주장하는 것과 같은 이유로

도출된 것인지에 대해서는 더욱 신중한 분석이 필요하다. 더불어 단일 언어적 파인튜닝 모델과 교차 언어적 파인튜닝 모델의 정답률과 정확도 사이에서는 최대 3%p가량의 성능 차이만 보이는 반면 베이스라인 모델과는 최대 6%p에 가까운 차이를 보인다는 점 또한 눈여겨볼 만하다.

또한 한국어 단일 언어적 파인튜닝 모델이 중국어 모델과는 성능 차이가 큰데, 이 모델과 교차 언어적 파인튜닝 모델의 성능 차는 크게 나지 않는지에 대해서도 분석할 여지가 있다. 가장 큰 이유로는 발음 오류가 포함되어 있는 한국어 단어 음소열의 패턴이 한국어 원어민 자유 발화 음소열의 패턴과 크게 다르지 않기 때문이라는 점을 들 수 있다.

시험 데이터에 대한 각 파인튜닝 모델의 구체적인 음소 인식 결과를 혼동 행렬로 나타낸 결과는 부록 B와 C에서 제시한다. 혼동 행렬 구성 역시 Needleman-Wunsch 알고리즘을 통해 음소 단위로 정렬한 정답과 모델 출력 텍스트 쌍을 이용하였다. 혼동 행렬에서 나타나고 있는 각 파인튜닝 모델의 인식 양상을 정리하면, 단일 언어적 파인튜닝 모델의 경우 각 모델 스스로가 훈련 대상으로 삼았던 데이터셋에 나타나는 IPA 기호만을 출력하는 경향을 보였다. 교차 언어적 모델과 베이스라인 모델 또한 대체로 비슷한 양상을 보였으나, 베이스라인 모델이 교차 언어적 모델과 비교하였을 때 중국어에 영향을 받은 학습자의 발음 오류를 뚜렷하게 잘 인식하는 모습을 보이지는 않았다.

4.2. 발음 오류 검출

각 파인튜닝 모델의 발음 오류 검출 성능 평가 결과는 표 4.2와 같다. 베이스라인 모델은 사전 훈련 모델에 상관없이 FAR, 재현율을 제외한 모든 척도에서 가장 좋은 성능을 보였다. 교차 언어적 파인튜닝 모델과 단일 언어적 파인튜닝 모델을 비교하면 교차 언어적 모델이 단일 언어적 모델에 비해 FRR, 정밀도, F1-점수, 검출 정확도, 진단 정확도 면에서 좋은 성능을 보였으며, 중국어 데이터셋으로만 파인튜닝된 모델은 교차 언어적 모델보다 더 나은 FAR과 재현율 성능을 보였다. 또 흥미로운 점은 베이스라인 모델이 시험 데이터와 같은 데이터셋에 속하는 데이터로 파인튜닝되었음에도 불구하고 모든 파인튜닝 모델 중 FAR과 재현율 성능이 가장 떨어진다는 것이다. 외국인 학습자 발화 데이터셋으로 음향 모델을 훈련하였더라도 일부 측면에서는 비교적 떨어지는 성능을 보일 수 있다는 점을 확인할 수 있다.

표 4.2 파인튜닝 모델의 발음 오류 검출 성능

파인튜닝 모델	FRR	FAR	DER	정밀도	재현율	F1-점수	검출 정확도	진단 정확도
사전 훈련 모델: wav2vec 2.0 Large								
KsponSpeech	32.39	14.82	37.97	24.82	85.18	38.44	69.57	62.03
AISHELL-1	60.81	**10.03**	58.20	14.41	**89.97**	24.85	44.38	41.80
Cross-lingual	*31.13*	*13.19*	*36.39*	*25.80*	*86.81*	*39.77*	*70.86*	*63.61*
L2KSC	**26.56**	17.03	**28.81**	**28.63**	82.97	**42.57**	**74.52**	**71.19**
사전 훈련 모델: XLSR-53								
KsponSpeech	32.98	12.57	35.00	24.65	87.54	38.46	69.26	65.00
AISHELL-1	57.91	**10.51**	51.11	14.89	**89.49**	25.53	46.91	48.89
Cross-lingual	*30.67*	*14.32*	*33.44*	*26.00*	*85.68*	*39.90*	*71.16*	*66.56*
L2KSC	**20.48**	15.61	**28.53**	**35.52**	84.39	**50.00**	**80.09**	**71.47**

4.3. 발음 오류 유형에 따른 결과 분석

각 파인튜닝 모델의 전반적인 음소 인식과 발음 오류 검출 성능을 앞 절에서 알아보았다. 이 절에서는 중국인 한국어 학습자의 발음 오류를 주제로 한 선행 연구들에서 공통적으로 제시하는 빈번한 발음 오류 유형 중 본 연구의 시험 셋으로 알아볼 수 있는 것들에 대해 각 모델의 구체적인 인식 양상을 살펴본다. 이때 파인튜닝 모델의 사전 훈련 모델은 Baevski et al.(2020)에서 제안하는, wav2vec 2.0 기반 음향 모델의 가장 기본적인 사전 훈련 모델이라 할 수 있는 wav2vec 2.0 Large로 한다.

발음 오류 유형에 따른 결과 분석 과정에서는 평가 척도 산출 및 혼동 행렬 구성에서 사용하였던 정렬 알고리즘은 사용하지 않았다. 정렬 알고리즘을 사용한 일대일 대응 방식의 텍스트 비교는 규모가 큰 인식 결과 데이터를 수치화하여 그 성능을 비교하기에는 편리하지만 직관에 맞는 분석 결과를 얻기에는 힘든 한편, 본 연구의 시험 데이터는 직접 모든 결과를 분석하기에 충분한 규모를 가지고 있으며 강제 정렬 모듈이 없는 각 파인튜닝 모델의 특성상 구체적인 인식 양상을 알아보기 위해서는 일대다 대응 방식의 결과 분석이 필요하기 때문이다. 파인튜닝 모델의 인식 양상을 알아보기 위해, 구체적인 분석의 의미가 다른 모델에 비해 떨어진다고 판단한 AISHELL-1 모델을 제외한 KsponSpeech, Cross-lingual, L2KSC 파인튜닝 모델에서 출력한 모든

음소 인식 결과 IPA 텍스트를 정답 IPA 텍스트와 수동으로 비교하였다.

먼저, L2KSC 데이터셋의 시험 셋이 포함하고 있는 한국어 발화와 그 표준 발음에 대해 본 연구에서 생성한 IPA 기호 연쇄는 다음과 같다.

1. 효도[hjoto]
2. 가득하다[katɯkʰata]
3. 극장[kɯkˀtɕ̝aŋ]
4. 컴퓨터[kʰʌmˀpʰjutʰʌ]
5. 꽃[k̝otˀ]
6. 꽤 많다[k̝wemanˀtʰa]
7. 무쇠[muɕwe]
8. 밝아요[palkajo]
9. 편리합니다[pʰjʌllihamˀnita]
10. 씨앗[ɕ̝iatˀ]

또 2장에서 서술한 바와 같이 선행 연구들에서 주로 지적한 발음 오류들을 참고하면(조문우, 2006; 유효려, 2008; 배영환 & 金成蘭, 2019; 위원징 & 홍미주, 2013; 위국봉, 2014; 양승희, 2016; 황미경 & 조남민, 2021; Zheng Yihan & 원미진, 2021) 위 어절들을 발음한 중국인 한국어 학습자 발화에서 확인할 수 있는 발음 오류 및 그 발음 오류를 포함할 수 있는 시험 셋의 발화들을 보이면 다음과 같다.

1. 어두 초성 평 파열음의 발음 오류: 가득하다, 극장, 밝아요
2. 어두 초성 경 파열음의 발음 오류: 꽃, 꽤 많다
3. 어두 초성 치조 마찰음의 발음 오류: 씨앗
4. /ɯ/ 모음의 대치 오류: 가득하다, 극장
5. /o/ 모음의 대치 오류: 꽃, 효도
6. /ʌ/ 모음의 대치 오류: 컴퓨터
7. 어말 종성 자음의 탈락과 대치 오류: 극장, 꽃, 씨앗

이 가운데 어두 초성에 나타나는 파열음 및 마찰음의 발음 오류, 모음의 대치 오류를 한데 묶고, 어말 종성 자음의 탈락과 대치 오류에 대한 분석을 마지막으로 제시하고자 한다.

4.3.1. 어두 초성 파열음 및 치조 마찰음의 발음 오류

2장에서 서술했듯 평음, 격음, 경음의 삼중 대립을 가진 한국어의 파열음 체계를 유기음과 무기음의 이중 대립을 보이는 중국어 파열음 체계에 어떻게 대응시킬 것인지에 대해서는 연구마다 의견이 나뉜다. 치조 마찰음의 경우에도 마찬가지로, 한국어의 평음 및 경음과 같은 대립이 중국어의 치조 마찰음에는 존재하지 않기 때문이다. 이는 한국어를 공부하는 중국인 학습자 입장에서도 각 음을 어떻게 지각하고 발음해야 하는지에 있어서 어려움을 초래하게 된다.

L2KSC 시험 셋에서도 이로 인한 발음 오류를 확인할 수 있다. '가득하다', '극장', '밝아요', '꽃', '꽤 많다', '씨앗'을 발음한 총 271개 발화 중 31개의 발화에서 발음 오류를 관찰할 수 있으며, 이들 발음 오류는 평음을 격음으로(4개), 평음을 경음으로(2개), 경음을 평음으로(20개), 경음을 격음으로(2개) 잘못 발음한 것들로 그 세부 유형을 나눌 수 있다. 이들 발음 오류에 대한 각 KsponSpeech, Cross-lingual, 그리고 L2KSC 모델의 자세한 인식 양상을 학습자의 정답 발음과 발음 오류의 경우로 나누어 정리한 것을 표로 제시하면 아래와 같다.

어두 초성 평 파열음의 경우, 학습자가 오류 없이 발음한 것의 정인식률은 한국어 원어민 발화 데이터만을 이용해 파인튜닝한 모델의 그것이 가장 높았다. 오히려 교차 언어적 파인튜닝 모델과 베이스라인 모델이 각 파열음을 같은 조음 위치의 격음 또는 경음으로 잘못 판단하는 경우가 많았다. 반면 KsponSpeech 모델은 발음 오류인 음소를 보았을 때 오류 음소로 제대로 판단하는 경향이 떨어져, 그것들을 제대로 발음한 것이라고 잘못 판단하는 빈도가 높았다.

표 4.3 어두 초성 평 파열음의 정답 발음 인식 양상

파인튜닝 모델	정인식	발성 유형 오거부		조음 방법 오거부	조음 위치 오거부	삭제	선행 모음 삽입	식별 불가
		격음	경음					
KsponSpeech	115	0	7	4	3	0	1	2
Cross-lingual	107	10	8	2	2	0	1	2
L2KSC	93	13	15	0	7	2	0	2

표 4.4 어두 초성 평 파열음의 발음 오류 인식 양상

파인튜닝 모델	정진단	오인식	진단 오류
KsponSpeech	2	6	0
Cross-lingual	3	5	0
L2KSC	4	4	0

표 4.5 어두 초성 경 파열음의 정답 발음 인식 양상

파인튜닝 모델	정인식	발성 유형 오거부		조음 방법 및 위치 오거부	삭제
		평음	격음		
KsponSpeech	24	53	0	0	2
Cross-lingual	17	62	0	0	0
L2KSC	50	26	2	2	1

표 4.6 어두 초성 경 파열음의 발음 오류 인식 양상

파인튜닝 모델	정진단	오인식	진단 오류
KsponSpeech	11	1	3
Cross-lingual	12	2	1
L2KSC	11	2	2

경 파열음 [k˭]는 단일 언어적 모델과 교차 언어적 모델 모두에서 평 파열음 [k]로 잘못 인식되는 빈도가 매우 높았고, 대신 평 파열음으로 잘못 발음한 것을 제대로 판단하는 빈도도 높았다. 그러나 베이스라인 모델은 나머지 파인튜닝 모델의 경향과 달리 정답 발음의 경 파열음도 비교적 정확히 인식하고, 발음 오류인 평 파열음도 잘 인식하는 양상을 보였다.

경 마찰음 [ɕ˭]의 인식 양상에서는 세 모델이 음의 긴장성 여부를 구분하는 것을 어려워했다. 원어민 데이터로만 파인튜닝한 모델들은 이들을 거의 모든 경우에 평 마찰음 [ɕ]로 잘못 인식하였고 베이스라인 모델 또한 50% 정도의 정확도를 보였다. 중국인 한국어 학습자들이 생성한 이것의 발음 오류는 모두 긴장성을 누락하고 평 마찰음으로 발음한 것이었는데, 이들을 인식하는 데 있어서도 교차 언어적 모델을 제외한 두 모델은 어려움을 겪었다.

위에서 살펴본 바에 따르면 각 모델은 발성 유형에 따라 구분되는 한국어의 어두 파열음 및 마찰음에서, 기식성과 긴장성 여부를 구분하는 것을 수행하는 능력이 떨어진다는 분석 결과를 도출할 수 있다. 정답 발

표 4.7 어두 초성 경 마찰음의 정답 발음 인식 양상

파인튜닝 모델	정인식	발성 유형 오거부(평음)	조음 방법 및 위치 오거부	조음 방법 및 발성 유형 오거부
KsponSpeech	7	32	0	0
Cross-lingual	0	39	0	0
L2KSC	19	18	1	1

표 4.8 어두 초성 경 마찰음의 발음 오류 인식 양상

파인튜닝 모델	정진단	오인식	진단 오류
KsponSpeech	5	3	0
Cross-lingual	8	0	0
L2KSC	4	3	1

음의 경우에는 한국어 원어민 발화 데이터셋로만 파인튜닝한 모델이 교차 언어적 모델에 비해 비교적 좋은 성능을 보이지만, 이것은 발음 오류 여부에 관계없이 다소 무차별적으로 출력 결과를 적용하기 때문이라고 볼 수 있기 때문에 실질적으로 단일 언어적 파인튜닝 모델이 더 성능이 낫다고 이야기하기는 어렵다.

4.3.2. 단모음의 대치 오류

양승희(2016)는 선행 연구 정리를 통해 중국어 학습자들은 한국어 단모음을 발음할 때 /ɯ/는 /ʌ/, /u/ 또는 이중 모음, /o/는 /ʌ/ 또는 /u/, /ʌ/는 /o/로 각각 잘못 발음하는 발음 오류 패턴을 보인다고 하였으며, 조문우(2006)와 황미경 & 조남민(2021)은 중국인 한국어 학습자들이 모음 /ㅡ/, /ㅜ/, /ㅓ/, /ㅗ/에 대해 가장 오류를 많이 산출한다고 하였다. 발음 오류 검출 모델에 있어서는 이들 발음 오류를 정확히 검출하는 것이 전체적인 성능에 큰 영향을 미칠 것이다. 세 모음의 인식 양상에 대한 각 파인튜닝 모델의 결과를 정리하면 표 4.9부터 4.14까지와 같다.

표 4.9 [ɯ] 모음의 정답 발음 인식 양상

파인튜닝 모델	정인식	조음 위치 오거부	모음 연쇄 오거부	식별 불가
KsponSpeech	75	11	2	1
Cross-lingual	74	15	0	1
L2KSC	79	10	1	0

표 4.10 [ɯ] 모음의 발음 오류 인식 양상

파인튜닝 모델	정진단	오인식	진단 오류
KsponSpeech	0	1	2
Cross-lingual	1	0	2
L2KSC	2	0	1

표 4.11 [o] 모음의 정답 발음 인식 양상

파인튜닝 모델	정인식	조음 위치 오거부	모음 연쇄 오거부	식별 불가	삭제
KsponSpeech	39	35	7	9	1
Cross-lingual	53	38	1	0	0
L2KSC	59	27	4	1	0

표 4.12 [o] 모음의 발음 오류 인식 양상

파인튜닝 모델	정진단	오인식	진단 오류
KsponSpeech	1	1	0
Cross-lingual	0	2	0
L2KSC	1	1	0

‘가득하다’, ‘극장’의 발음으로 본 중국인 한국어 학습자의 모음 [ɯ]에 대한 각 파인튜닝 모델의 인식 양상이다. 대체로 인식률이 높았으며, 조음 위치를 잘못 인식한 경우 다양한 단모음을 출력했지만 모델에 관계없이 [o] 또는 [u] 모음으로 잘못 판단한 경우가 대부분을 차지했다. 한편 단모음의 연쇄 또는 단모음과 반모음의 연쇄와 같은 모음 연쇄로 구간이 잘못 인식되는 경우도 있었는데, [ɯu], [ɯʌ], [ɯjʌ]와 같은 모음 연쇄를 하나씩 찾아볼 수 있었다.

[o] 모음 인식의 경우 양승희(2016)의 연구와 연관지을 수 있는 점이 존재했다. [o] 정답 발음을 잘못 인식하는 경우에 각 모델에 걸쳐 주로 출력하는 오류 음소가 존재했는데, KsponSpeech와 Cross-lingual 모델은 [ʌ](각각 조음 위치 오인식 35개 중 24개, 38개 중 31개), 베이스라인 모델은 [u](27개 중 19개)로 잘못 인식하는 경향이 컸다. 오류 발음에 있어서도 [o]를 [u]로 잘못 발음한 2개의 경우에 대해 오인식 결과는 모두 [o]로 잘못 인식한 것이었다.

[ʌ] 모음 인식에 있어서도 파인튜닝 모델 간에 약간의 차이를 보였다. 제대로 발음한 것의 경우 KsponSpeech와 Cross-lingual 모델

표 4.13 [ʌ] 모음의 정답 발음 인식 양상

파인튜닝 모델	정인식	조음 위치 오거부	모음 연쇄 오거부	식별 불가	삭제
KsponSpeech	41	41	6	3	0
Cross-lingual	37	46	4	4	0
L2KSC	25	62	1	1	2

표 4.14 [ʌ] 모음의 발음 오류 인식 양상

파인튜닝 모델	정진단	오인식	진단 오류	식별 불가
KsponSpeech	1	0	1	1
Cross-lingual	1	1	1	0
L2KSC	1	1	1	0

은 [a](각각 조음 위치 오인식 41개 중 29개, 46개 중 33개)로, 베이스라인 모델은 [o](62개 중 42개)로 혼동하는 양상을 보였다. 특히 베이스라인 모델은 이 오류가 전체 인식 결과의 대부분을 차지해 전체적인 인식률이 떨어지는 경향을 보였다.

이상 단모음에 대한 각 모델의 인식 결과를 정리해 보았다. 모델의 성능과 관계없이, 중국인 한국어 학습자가 [o]와 [ʌ]를 혼동하여 발음하는 오류처럼 파인튜닝 모델들도 이들 학습자의 두 모음을 헷갈려하는 양상을 보였다. 단 이러한 현상에 대해 하나의 의견을 제시할 수 있는데, 같은 후설 모음이면서 혀의 높이에서만 약간의 차이를 보이는 두 모음을 중국인 한국어 학습자들이 각각 발음한다면 인간뿐 아니라 음향 모델 또한 이것을 지각하는 데 어려움을 겪을 수도 있을 것이다.

4.3.3. 종성의 탈락과 대치 오류

한국어의 자음 중 종성으로 올 수 있는 음은 [k̚], [n̚], [t̚], [l], [m̚], [p̚], [ŋ]의 일곱 개인 반면 중국어는 [n̚], [ŋ]의 두 개 음만이 종성으로 올 수 있다(한성우, 2008). 중국인 한국어 학습자들은 이 두 개 음이 아닌 다른 음들이 종성으로 왔을 때 이것을 발음하는 데 어려움을 겪을 수 있고, 이들 종성의 발음 오류로는 다양한 것들이 보고된다(양승희, 2016). 본 연구에서는 ‘극장’과 ‘씨앗’ 두 단어의 어말 종성을 대상으로

표 4.15 종성의 정답 발음 인식 양상

파인튜닝 모델	정인식	조음 위치 오거부	후행 모음 삽입	삭제	식별 불가
KsponSpeech	41	12	1	20	9
Cross-lingual	38	4	0	37	1
L2KSC	62	13	0	5	0

표 4.16 종성의 발음 오류 인식 양상

파인튜닝 모델	정진단	오인식	진단 오류
KsponSpeech	9	0	3
Cross-lingual	8	0	4
L2KSC	9	0	3

중국인 한국어 학습자의 정답과 오류 발음을 각 파인튜닝 모델이 어떻게 인식하는지 알아보고자 한다.

'극장'의 경우 종성 [ŋ]에 대한 발음 오류는 그것을 [n']으로 잘못 발음한 것으로 한 개가 있었으며, 나머지 11개의 발음 오류는 모두 '씨앗'에 대한 발음 오류이다. 상세히는 [t']을 [k']으로 발음한 조음 위치 오류가 7개, 발음하지 않은 삭제 오류가 4개 있었다.

각 모델은 대체로 종성 [ŋ]은 잘 인식하였으나, [t']는 그 경향성이 나뉜다. 베이스라인 모델은 이것을 [k']로 오인식하는 경우가 전체 35개 중 11개 있었다. 반면 KsponSpeech 모델은 16회의 삭제 오류를 보였으며 [k'], [p'] 등 다른 조음 위치의 종성으로 잘못 인식하거나 후행 모음 [i]가 삽입되는 등 다양한 오류 패턴을 보였다. Cross-lingual 모델은 이것을 하나도 맞추지 못했는데, 35개 중 33개의 경우에 삭제 오류를 보였다.

5. 논의

5.1. 결과 해석

본 연구에서 제안한 것처럼 학습 대상 언어와 학습자 모어 원어민 발화 데이터셋을 모두 이용해 교차 언어적 방식으로 파인튜닝한 wav2vec 2.0 기반 음소 단위 발음 오류 검출 모델은 학습 대상 언어 또는 학습자 모어 원어민 발화 데이터셋 중 하나만을 이용한 단일 언어적 파인튜닝 모델보다는 더 좋은 성능을 보였지만, 일반적으로 발음 오류 검출 모델에서 사용하는 방식인 비원어민 학습자 발화 데이터셋을 이용하는 것으로 파인튜닝한 베이스라인 모델보다는 FAR 및 재현율을 제외하면 전체적으로 낮은 성능을 보였다. 중국인 한국어 학습자가 빈번하게 산출하는 발음 오류 유형에 따라 구체적인 인식 양상을 보았을 때는 파열음 및 마찰음의 발성 유형, 모음의 조음 위치, 종성의 존재 등을 정확히 판단하는 데 어려움을 겪었고, 특히 어두 초성 경 마찰음을 모두 평음으로 잘못 예측하거나 종성의 경우 전체 정답 발음의 절반 가까이를 인식하지 못하였다. 그러나 어두 평 파열음의 정답 발음 인식, 어두 초성 경 마찰음의 발음 오류 인식, [ʌ] 모음의 인식에 있어서는 베이스라인보다 더 나은 성능을 보였으며, 어두 초성 경 파열음의 발음 오류 인식, [ɯ]와 [o] 모음의 인식, 종성의 발음 오류 인식 등의 경우 베이스라인 모델과 크게 차이를 보이지 않았다. 즉 두 언어의 원어민 발화 데이터셋만을 가지고 훈련한 모델이 비원어민 학습자 발화 데이터셋으로 직접 훈련한 것과 비교하여 일관적으로 뒤지는 성능을 보이지는 않았다는 것이다.

실험에서 보고한 여러 척도 중 FAR과 재현율에 관심을 가질 필요가 있다. 기계 학습 모델은 훈련 셋으로 사용한 데이터의 분포를 모델링하기 때문에, 본 연구의 파인튜닝 모델 또한 훈련 셋으로 사용하는 데이터에 따라 모델의 시험 셋에 대한 최종 성능이 달라질 것이다. 본 연구에서 진행한 실험에서도 L2KSC 데이터셋을 시험 셋으로 하였을 때 대부분의 척도에서 베이스라인 모델의 성능이 가장 좋았으며, 다음으로 교차 언어적 모델과 단일 언어적 모델이 성능이 좋았다. 특히 단일 언어적 모델 사이에서도 KsponSpeech 모델의 성능은 교차 언어적 모델과는 큰 차이가 나지 않았으나 AISHELL-1

모델은 나머지 모델과 큰 폭의 성능 차이를 보였다. 이는 각 파인튜닝 모델의 시험 셋인 L2KSC 데이터셋이 일부 발음 오류를 제외하면 한국어 발화의 음성적 특징, 음소 기호와 음소 연쇄 패턴으로 이루어져 있고, KsponSpeech 데이터셋은 같은 언어의 원어민 발화 데이터라는 특성상 L2KSC 데이터셋과 그 분포가 비교적 비슷한 반면 AISHELL-1 데이터셋은 크게 다른 분포를 갖고 있기 때문이다. 그러나 FAR과 재현율 척도에서는 반대로 AISHELL-1 모델이 가장 높은 성능을 보이고 베이스라인 모델의 성능이 가장 낮은 것을 관찰할 수 있다. 즉 실제로 발음 오류인 음소를 오류가 아니라고 잘못 판단할 가능성이 AISHELL-1 모델에서 가장 낮고 베이스라인 모델에서 가장 높다는 것이다.

앞에서 설명한 내용과 연관 짓는다면 AISHELL-1 모델의 FAR 및 재현율 성능이 높은 것은 쉽게 예측할 수 있다. 이 모델은 중국어 원어민 발화 데이터에만 노출되었기 때문에, 모델 시험 시 L2KSC 데이터셋의 발화 음성을 입력받더라도 각 음소의 음성적 특성과 흡사하지만 AISHELL-1 데이터셋의 음소 기호 및 연쇄 패턴을 띠는 결과를 출력할 가능성이 높다. AISHELL-1 모델은 중국인 한국어 학습자의, 중국어에 영향을 받아 발생하는 발음 오류를 제대로 예측할 가능성이 높지만, 동시에 정확히 발음한 음소에도 중국어 음소와 관련한 예측을 내놓을 가능성 역시 높다는 것이다. 이는 본 연구의 실험 결과 중 FRR과 정밀도 측면에서는 AISHELL-1 모델의 성능이 다른 모델에 비해 현저히 나쁘다는 점과 통한다.

반면 교차 언어적 모델과 베이스라인 모델의 FAR 및 재현율을 비교했을 때 베이스라인 모델의 수치가 더 떨어지는 점은 주목할 만한데, 교정 피드백이 필요한 발음 오류 음소를 베이스라인 모델이 오히려 지나칠 가능성이 높다는 것이다. 베이스라인 모델은 시험 셋과 같은 L2KSC 데이터셋을 이용하여 직접 오류 발화를 모델링했다는 점을 고려하면 이는 쉽게 예상할 수 없는 결과이다.

이러한 현상의 원인으로 L2KSC 데이터셋의 특성을 지적할 필요가 있다. 본 연구에서 사용한 L2KSC 데이터셋은 최대 2어절의 짧은 발화만으로 구성되어 있어 개별 발화 음성의 길이가 KsponSpeech 데이터셋의 그것에 비해 매우 짧으며, 전체 데이터셋의 분량 또한 KsponSpeech가 30배가량 많다. 또한 각 화자마다 동일한 발화 리스트를 읽어 녹음하였기 때문에 훈련 셋 안에 같은 말을 읽은 발화가 거듭 실려 있으며, 짧은 발화를 녹음하였다는 특성상 개별 발화에

포함되어 있는 음소 분포 또한 다양하지 않다. 한편 훈련에 사용된 CTC 손실 함수는 각 출력 단위 간에 조건부 독립(conditional independence)을 가정하고 있지만(Graves et al., 2006), 트랜스포머 구조의 자기 주의 집중 기제(self-attention mechanism)가 각 시점 간 문맥 관계를 효과적으로 모델링할 수 있다(Vaswani et al., 2017). 따라서 L2KSC 데이터셋 내에서 더 많은 비율을 차지하는, 정답 발음으로만 구성된 발화의 음소 패턴에 대한 문맥 정보를 모델이 우세하게 학습하게 된다면 발음 오류가 일부 포함된 발화에도 정답 발음 발화 또는 표준 발음과 같은 문맥 정보 계산을 적용하게 될 것이다. 이것은 일종의 과적합 문제로 볼 수 있다. 음소 연쇄의 정확한 인식을 위해서는 음성적 맥락을 고려하는 것이 큰 도움이 되지만, 발음 오류가 포함된 비원어민 발화 음성을 인식하는 문제에 있어서는 여기에 더해 종종 나타나는 발음 오류 음소까지 강건하게 검출하여야 한다. 이때는 오히려 너무 넓은 문맥 정보를 고려할 시 오인식이 일어날 수 있다.

부록 B와 C에서는 각 파인튜닝 모델의 전체 인식 양상을 혼동 행렬로 구성하였다. 혼동 행렬은 인식 기반 발음 오류 검출 모델 연구에서 실험 결과의 일부로서 일반적으로 제시된다. 본 연구에서 보고하는 혼동 행렬에서는 어떠한 환경에서 파인튜닝을 진행하였는가에 따라 그 인식 양상의 차이를 한눈에 확인할 수는 있으나, 각 실제 음소를 어떤 음소로 인식하였는지의 여부를 정확히 비교한 결과로서 받아들이기에는 한계가 따른다. 본 연구에서는 각 모델의 인식 및 출력 단위를 자음, 모음, 반모음 등에 상관없이 모두 개별적인 하나의 음소 단위로 정하였고, 이러한 인식 단위에 따라 일반적인 음성 인식 및 발음 오류 검출 분야의 성능 분석 방법론과 같이 일대일 대응 방식의 정렬 알고리즘을 사용하였기 때문에, 그림 5.1의 예와 같이 실제로 단모음인 음소를 이중모음으로 인식하였다면 인식 결과는 두 개의 IPA 단위로 나누어져, 정답 IPA 연쇄와 정렬하여 비교할 때 하나의 대치 오류와 하나의 삽입 오류로 계산되기 때문이다. 이처럼 전사 및 인식 단위에 따라 정렬 결과가 달라지고 최종적인 성능 계산에도 영향을 미칠 수 있다는 것을 지적해야 한다.

정답 IPA	h	j	**o**		t	**o**	
인식 결과 IPA	ɕ	j	**ə**	**u**	t	**ə**	**u**
오류 여부	S	C	**S**	**I**	C	**S**	**I**

* C: 정답, S: 대치 오류, I: 삽입 오류

그림 5.1 인식 결과 정렬 예

5.2. 발음 오류 검출을 위한 비원어민 학습자 데이터셋

본 절에서는 발음 오류 검출 모델을 효과적으로 훈련하기 위해서는 어떠한 성격을 가진 비원어민 학습자 데이터셋을 사용해야 하는가에 대한 논의를 진행하고자 한다. 2장 및 앞 절에서 지적하였듯 비원어민 한국어 학습자 발화 데이터셋은 존재하는 것이 많지 않으며, 존재한다 하더라도 별도의 전사 텍스트가 없어 발음 오류 검출 모델 훈련에 그대로 사용하기에 부적합한 것들이 있기도 하다. 더불어 일반적인 한국어 원어민 발화 데이터셋과 비교하여 데이터셋 규모가 아주 작은데, 이는 심층 신경망 기반 음향 모델을 훈련시키기에 적합한 양이 아니다. 이미향 & 안미애(2021)는 L2KSC 데이터셋을 대표적인 비원어민 한국어 학습자 데이터셋으로 꼽으나 그 구축 목적이 한국어 학습자 발음 오류에 대한 이론적 연구에 있고, 발음 오류 검출 모델 훈련과 같은 실제적 이용을 목적으로 설계된 것이 아니라고 말하였다. 또한 한국어 발음 교육 분야의 연구 측면에 있어서도 인문학, 교육학 분야에서의 발음 오류에 대한 이론적 연구가 대부분을 차지하고 공학 분야에서의 연구나 학제적인 접근이 있는 경우는 이론적 연구에 비해 턱없이 부족하다고 지적하였으며, 인식 기반 발음 오류 검출 모델에 적합한 비원어민 한국어 학습자 발화 데이터셋을 구축할 필요성을 제기하였다.

음성 인식 과제에서 강건한 성능을 보이는 심층 신경망 기반 음향 모델을 사용하는 현재의 발음 오류 검출 연구 추세로 볼 때, 그 모델이 학습 및 성능 검증에 사용할 데이터셋은 다음과 같은 특성을 가져야 한다. 첫째, 전체 데이터셋의 규모가 방대해야 하며, 개별 발화는 충분한 길이를 가져야 한다. 심층 신경망 기반의 모델은 일반적인 기계 학습 기반 모델에 비해 더 많은 양의 데이터로 훈련하여야 안정적인 성능을 낼 수 있다는 것은 잘 알려져 있다. 소량의 파인튜닝 데이터만으로도 높은 성능을 기대할 수 있는 wav2vec 기반의 음향 모델 역시 파인튜닝 데이터의 분량이 많아질수록 성능이 점진적으로 상승하는 것은 다른 모델과 크게 다르지 않다(Baevski et al., 2020). L2-ARCTIC 데이터셋은 전체 3시간 분량으로 원어민 발화 데이터셋과 비교했을 때 아주 적은 양이지만 약 5.4시간 분량의 TIMIT 데이터셋과 함께 훈련하는 방식으로 분량 문제를 해결한다. 그러나 발음 오류 검출 모델을 위한 가장 이상적인 비원어민 학습자 발화 데이터셋은 그 전체가 비원어민 학습자 발화로 이루어져 있어야 할 것이다. 한편 본 연구에서 사용한 L2KSC 데이터셋은 각 발화 데이터가 최대 2어절의 짧은

음성으로 구성되어 있는 반면 L2-ARCTIC 데이터셋의 각 발화 데이터는 평균 3.7초 길이의 문장 발화로 구성되어 있고, 한국어 원어민 데이터셋인 KsponSpeech의 개별 발화당 평균 길이는 약 5.6초이다. L2KSC 데이터셋을 이용한 본 연구의 실험 결과 등을 이와 함께 고려했을 때, 한국어 발음 오류 검출 모델을 성공적으로 훈련할 수 있는 데이터셋은 최소 10시간 분량의 비원어민 한국어 학습자 발화만을 담고 있음과 동시에 개별 발화는 3~5초 길이의 문장 단위 발화 녹음일 것을 제안한다. 더불어 본 연구에서는 음소 분포 여부를 고려하지 않고 L2KSC 데이터를 발화 내용 단위로 무작위로 나누어 훈련, 검증 및 시험 셋을 구성하였는데, 그 결과 시험 셋이 음성적으로 균형 잡혀 있지(phonetically balanced) 않아 보다 넓은 범위의 발음 오류에 대한 결과 분석을 진행할 수 없었다. 따라서 비원어민 학습자 발화 데이터셋을 구성할 시 음성적 균형 및 학습 대상 언어와 학습자 모어를 고려한 발음 오류 유형을 고려하여 훈련, 검증, 시험 셋을 구성함으로써 더욱 신뢰할 만한 모델 성능 평가가 가능하도록 하여야 한다.

둘째, 발음 오류 검출 모델의 성능 제고와 결과 분석을 위한 최적의 전사 단위를 탐구하여야 한다. 본 연구에서는 한국어와 중국어 음소를 모두 표기할 수 있는 IPA 기호를 전사 단위로 이용하였는데, 두 언어에 대한 음운론 및 음성학 분야의 선행 연구들을 참고하여 음소 셋을 구성하였으며 자음, 모음 및 반모음 단위로 그 인식 단위를 최대한 나누었다. 그 결과 이중 모음을 하나의 인식 단위로 하는 양승희(2016) 등과 비교했을 때 더 세밀하고 구체적인 인식 결과를 얻어 교정 피드백 제공에 더 효과적일 수 있었다. 그러나 앞 절에서 언급하였듯 정렬 문제 등으로 인하여 더 낮은 성능 점수가 산출되거나 혼동 행렬 구성이 직관과 어긋나는 등 결과 분석이 어려워질 우려가 있다. 이를 위해서는 Kondrak(2000)에서 제안하는 바와 같이 음성적 자질을 고려한 정렬 방식 등 모델의 인식 단위 일관성을 해치지 않으면서 결과 분석을 정확하고 용이하게 할 수 있는 방안을 고려해야 한다.

또한 계산 모델의 궁극적인 목표는 해결하고자 하는 문제에 대한 정확도를 높이는 것이겠지만, 발음 오류 검출 문제의 특성상 발음 오류를 검출하고 그 내용을 제대로 진단하는 것에만 초점을 두어서는 안 된다. 현대 외국어 발음 교육의 목표는 학습자로 하여금 원어민과 똑같은 발음을 하도록 교육하는 것보다는 원어민과 의사소통이 가능한 정도의 발음을 할 수 있도록 훈련시키는 것이며, 학습자가 원어민과 같은 발음을 습득하도록 유도하는 것은 학습자에게 부담을 주어 그의

외국어 학습 의욕을 떨어지게 할 수 있다(이다연, 2017). 따라서 발음 오류를 정답 발음이라고 잘못 판단하는 경우가 있더라도 발음 오류를 지나치게 엄격히 검출하는 것은 지양하도록 모델을 훈련시킬 필요 또한 있다. 이는 전사 단위를 얼마나 효과적으로 정의하느냐의 문제와 밀접하게 관련되어 있다. 예를 들어 본 연구에서는 한국어 모음 음소의 갯수를 7개로 줄이거나, 파열음이 유성음 사이에서 유성음화되는 현상(이호영, 1996) 등을 전사 텍스트에 반영하지 않는 등 전사 단위를 간략화하기도 하는 한편, 양승희(2016)와 달리 한국어 자음 /ㅅ/, /ㅆ/가 전설 모음 또는 반모음 /j/ 앞에 올 때 각각 변이음 [ɕ], [ɕ͈]으로 실현되는 현상(이호영, 1996)을 반영하는 등 전사 단위를 반대로 정밀화하기도 하였다. 어떤 음소들의 전사 단위를 간략화하면 음향 모델은 해당 음소들에 대한 변이에 관대해질 것이나 그 범위를 잘못 정한다면 비원어민 학습자들이 학습 대상 언어의 원활한 의사소통에 필요한 발음 또는 음운 현상을 배우는 데 실패할 수 있다. 반대로 전사 단위를 정밀화하면 음향 모델이 해당 음소들의 변이를 더 세밀하게 포착할 수 있겠으나 필요 이상의 정밀 전사는 반드시 완벽하게 훈련할 필요가 없는 발음들에 대해 지나치게 엄격한 오류 판정을 내려 학습자의 학습 의욕 저하와 불필요한 학습 시간 소요를 야기할 수 있다.

셋째, 정확하고 일관적인 오류 태깅이 필요하다. 최혜원 외(2022)는 비원어민 학습자 발화 데이터셋 구축에 있어 발음 평가자, 즉 전사자의 객관성과 일관성이 중요하다고 지적하였다. 이 연구는 또한 한국어 비원어민 발화 전사자들을 교육하는 데 상대적으로 시간이 더 소요되는 IPA 기호 체계를 배우는 대신에 한글 자모 체계를 확장한 발음 오류 포함 발화의 전사 방안을 제안하였다. 다만 그와 같은 방안은 다양한 언어를 모어 배경으로 가지는 한국어 학습자의 발화에 존재하는 광범위한 발음 오류를 전사하기 위해서는 계속해서 새로운 기호를 도입해야 하므로 체계의 변동이 심할 것이라는 점, 한국어 음소에 존재하지 않는 발음 오류들을 정확히 전사해야만 하는 전사자들은 기본적으로 기초 수준 이상의 음성학적 지식을 갖추고 있을 것임을 고려하였을 때 그들이 IPA 체계를 이미 알고 있거나 그들에게 그것을 교육하기 어렵지 않을 것이라는 점에서 IPA를 사용한 전사 방안이 더욱 적절할 것이다.

5.3. 의의

5.3.1. 원어민 발화만을 이용한 교차 언어적 발음 오류 검출 모델

많은 양의 비원어민 학습자 발화 데이터를 수집하는 데에 어려움이 따르는 발음 오류 검출 연구의 특성에 의해, 비원어민 학습자 발화 데이터셋과 학습 대상 언어의 원어민 발화 데이터셋을 합쳐 발음 오류 검출 모델을 훈련하는 연구들은 많이 있어 왔다. 그러나 외국어 학습자의 모어와 학습 대상 언어의 원어민 발화 데이터셋만으로 모델을 훈련하여 그 성능을 알아 보고자 하는 연구는 거의 없었으며(Joshi et al., 2015; Lee & Glass, 2015) 특히 심층 신경망을 기반으로 한 발음 오류 검출 모델에 있어서는 본 연구가 처음이라고 할 수 있다.

더불어 본 연구에서는 서로 다른 두 언어의 음소 체계 및 음운 현상을 고려하여 두 언어의 발화 데이터셋을 같은 기호 체계로 전사하는 데이터 가공 방안 또한 제시하였다. 어떤 언어를 외국어로서 학습하는 학습자들의 모어 배경은 매우 다양하기 때문에, 비원어민 외국어 학습자 발화 데이터셋을 구축하는 데 있어서 이들 모어의 영향을 받아 생성되는 잠재적 발음 오류들을 모두 반영하는 것은 사실상 불가능하다. 하지만 세계의 많은 언어의 음소 체계에 대한 연구는 시간이 흐르며 많은 양이 축적되었으며 오늘날에도 아주 활발히 진행되고 있다. 만약 발음 오류 검출 모델을 특정 언어를 모어로 사용하는 학습자들에 특화되도록 훈련하고자 하거나 그들 모어의 특징을 반영하고 있는 비원어민 학습자 발화 데이터가 존재하지 않는다면, 본 연구에서 제시한 방안과 같이 IPA 등의 보편적 표기 단위를 이용하여 각 원어민 발화의 전사 데이터를 가공하고, 이를 통해 발음 오류 검출 모델을 훈련하는 방법을 고려할 수 있다.

이와 같은 시도가 거의 있지 않았던 이유로는 몇 가지를 추측해 볼 수 있다. 첫째, 인식 기반 발음 오류 검출 모델에서는 발음 오류 양상을 직접적으로 모델링하는 것이 학습자 모어의 원어민 발화 데이터셋의 음성적 분포를 모델링하는 것보다 쉬운 문제이다. 둘째, 각 언어마다 발화 음성들을 얼마나 정밀하게 전사할 것인지, 또 다양한 음운 현상을 얼마나 반영할 것인지 등을 고려하여 서로 다른 언어를 일관적인 체계 아래 전사하는 방안에 대해서도 많은 논의가 필요하다. 셋째, 계산 모델링 알고리즘의 발전에 따라 발음 오류 검출 모델의 성능도 발전한다. 본 연구 또한 위와 같은 주제에서 발생하는 문제들을 완전히 해소한

것은 아니지만, 발음 오류 검출 분야에 있어 새로운 방향의 연구 주제들과 그 가능성을 제시했다는 점에서 의의를 가진다고 할 수 있다.

본 연구 또한 비원어민 학습자 발화 음성을 직접 모델링하는 기존의 방식이 아닌, 학습자 모어와 학습 대상 언어의 원어민 발화 음성을 사용하여 간접적으로 모델링 과정을 수행한다는 발음 오류 검출 모델 훈련의 방법론상 한계점을 타파하지는 못하였다. 그러나 이것이 근본적인 한계점이 되지는 않는다. 최근 자연어 처리 분야에서는 프롬프트 엔지니어링(prompt engineering) 방법론을 이용하는 대형 언어 모델의 퓨샷 학습(few-shot learning)을 통해 모델이 거의 경험하지 못한 과제에 대한 답변의 정확도를 높이는 데 성공하였다는 연구가 보고되고 있는데(Brown et al., 2020; Wei et al., 2022), 본 연구에서 사용한 학습 대상 언어와 학습자 모어 원어민 발화 데이터셋이 비원어민 학습자 발화 데이터셋과 완전히 같지는 않지만 유사한 음성적 분포를 보인다는 점을 고려하였을 때 이와 같은 새로운 방법론의 적용 가능성을 충분히 탐구해 볼 필요가 있다.

5.3.2. 모델 성능의 언어학적 분석

대체적으로 발음 오류 검출 모델은 이진 분류 성능을 측정하는 척도 또는 음소 인식 및 발음 오류 여부 진단에 관한 척도를 통해 정량적인 수치를 보고하거나, 자음과 모음으로 경우를 나누어 혼동 행렬을 구하는 것으로 그 성능 평가 및 분석을 진행한다. 그러나 본 연구에서는 파인튜닝된 각 발음 오류 검출 모델이 훈련에 사용된 데이터셋에 따라 어떠한 인식 양상을 보이는지를 언어학적인 지식을 이용해 조금 더 깊이 분석하고자 하였다. 이와 같은 분석을 통해서 파인튜닝 모델들은 중국인 한국어 학습자의 어두 초성 파열음의 기식성과 긴장성 여부를 판단하는 데 어려움을 겪거나, 같은 후설 모음인 [o]와 [ʌ] 모음을 구분하는 성능이 떨어진다는 점 등을 알 수 있었다.

언어학적 지식 기반의 음향 모델 인식 양상 분석은 데이터셋 전체 단위에서의 성능 변화를 이끄는 것에는 한계가 있겠지만, 이러한 분석을 통해 발음 오류 검출 모델의 인식률 또는 발음 오류 진단 성능이 음소들 사이의 어떤 자질에 의해 영향을 받는가에 대한 정보를 알 수 있기 때문에 발음 오류 검출 분야의 연구를 수행하는 데 있어 음향 모델을 훈련할 때 더 주의를 기울여야 할 부분을 제시해 줄 수 있을 것이다.

5.4. 한계와 후속 과제

두 언어의 원어민 발화 데이터만을 이용한 교차 언어적 발음 오류 검출 모델 훈련이라는 초기적 개념을 탐구하는 연구인 만큼 본 연구의 한계점도 다수 존재하는데, 모델의 방법론과 성능 측면에서 논의할 만한 것이 있다.

우선 각 전사 데이터를 가공하고 전사 단위를 결정하는 데 더욱 엄밀한 검토가 요구된다. 본 연구에서는 한국어와 중국어 원어민 발화 데이터셋 및 중국인 한국어 학습자 발화 데이터셋의 전사 텍스트를 같은 IPA 체계를 공유하는 전사 단위로 가공하였는데, 각 언어의 음소 체계와 음운 현상을 고려하였다고 하더라도 이는 완벽하지 않다. 두 언어의 음소 체계는 항상 변화하는 과정에 있으며, 특히 모음 체계에 대해서는 학자마다 의견이 나뉘어 어떤 이론이 발음 오류 검출을 위한 음향 모델 훈련에 최적의 것인지 결정하는 것 또한 많은 실험이 동반되어야 하기 때문이다. 더불어 각 언어의 발화 데이터를 전사할 때 얼마나 정밀한 단위로 전사하고, 두 언어의 자음과 모음 음소 체계를 서로 어떻게 대응시켜야 모델의 성능이 가장 좋을 것인지에 대한 문제의 해답 또한 많은 고민과 실험이 있어야만 얻을 수 있다. 본 연구는 이처럼 다양한 음소 체계 이론을 실험에 빈틈 없이 적용하는 데까지는 나아가지 못했으며, 같은 음소 체계 이론을 적용하였을 때 인식 단위를 얼마나 상세히 나누는 것이 모델의 성능을 가장 높일 수 있는지에 대해서도 후속 연구에서 짚어 볼 필요가 상당하다.

다음으로는 모델 성능 측면에서의 논의 사항이다. 선행 연구 중 본 연구와 그 실험의 형태가 가장 가까운 것으로는 Peng et al.(2021) 및 Wu et al.(2021)을 꼽을 수 있다. 전자의 연구는 훈련 데이터로 L2-ARCTIC과 TIMIT 데이터셋을 함께 이용하였고 후자의 연구는 CU-CHLOE 데이터셋을 이용하였다. 본 연구를 포함한 세 편의 연구에서 사용한 데이터셋이 모두 다르기 때문에 직접적인 성능을 비교하는 것은 어렵겠으나, Peng et al.(2021)은 최대 60.44%, Wu et al.(2021)은 최대 80.98%의 F1-점수를 보고하고 있다. 반면 본 연구의 베이스라인 모델은 50.00%의 F1-점수에 그친다. 이는 짧은 단위로만 이루어진 발화, 적은 양의 데이터셋 분량 등 베이스라인 모델에서 훈련 셋으로 사용한 L2KSC 데이터셋의 특성에서 가장 큰 원인을 찾을 수 있을 것이다. 또한 한국어 발화가 타 언어 발화보다 계산 모델링 난이도가 높다는 추측을 할 수도 있고, 원본 전사 데이터를 IPA 단위로 가공하는

과정에서 모델의 성능 저하 요인이 발생하였을 수도 있다. 이와 같은 문제 역시 후속 연구에서 다룰 만하며, 나아가 심층 신경망 기반 발음 오류 검출 모델 훈련의 성격에 부합하는 새로운 비원어민 학습자 발화 데이터셋을 구축할 필요성 또한 제기할 수 있다.

한편 본 연구에서는 기본적인 wav2vec 2.0 음향 모델의 파인튜닝 설정, 즉 wav2vec 2.0 모델을 이용하여 음소 단위 인식을 진행하는 음성 인식 과제와 같은 방식으로 발음 오류 검출 모델을 훈련하였는데, 선행 연구에서 적용한 모델 구조와 같이 외국어 학습자가 읽은 단어 내지는 문장의 텍스트 정보까지 이용하여 모델 알고리즘을 정의할 필요가 있다. Peng et al.(2022)은 이와 같은 방법을 이용해 학습자 발화에서의 발음 오류 발생 가능성을 주어진 텍스트 정보와 발화 음성 사이의 불일치에서 찾음으로써 발음 오류 검출 모델의 성능을 더 높일 수 있음을 보고하였다. 이는 종전의 발음 오류 검출 모델 연구에서 적용해 온 텍스트 정보를 이용하는 방법론이 wav2vec 기반 알고리즘에서도 적용 가능하다는 것이며, 본 연구와 같은 방법론을 더욱 발전시킬 수 있는 방안이 될 것이다.

6. 결론

본 논문은 음소 단위를 이용하는 인식 기반 발음 오류 검출 모델을 훈련할 때 일반적으로 사용하는 비원어민 외국어 학습자 발화 데이터셋이 아닌 학습자의 모어와 학습 대상 언어의 원어민 발화 데이터를 모두 이용하는 방법론을 제안하고, 심층 신경망 기반의 최신 음향 모델인 wav2vec 2.0을 파인튜닝하는 실험을 진행함으로써 그 가능성을 탐구하였다. 교차 언어적 환경에서 사전 훈련 모델을 파인튜닝하였을 때는 같은 모델을 단일 언어적 환경에서 파인튜닝한 것보다 더 높은 음소 인식과 발음 오류 검출 성능을 얻을 수 있었지만 베이스라인 모델인 비원어민 학습자 발화만으로 파인튜닝한 모델에 비해서는 떨어지는 성능을 보였다. 이러한 결과를 얻는 데 영향을 주는 요인은 각 파인튜닝 모델의 훈련 데이터셋 구성, 적은 양의 레이블링된 데이터로도 비교적 좋은 성능을 얻을 수 있는 wav2vec 2.0 모델의 특성 등 여러 가지가 있을 수 있다.

또한 서로 다른 언어의 발화로 이루어진 데이터셋을 같은 모델로 훈련하기 위해서 각 언어의 음소 체계와 음운 현상을 고려하는 전사 데이터 가공 방식을 이용하였으며, 각 파인튜닝 모델의 인식 양상을 정답 발음과 발음 오류의 경우로 나누고 언어학적 지식을 이용하여 분석하였다.

본 연구에서 제안한 방법론과 실험은 별도의 모델 구조 변경이 없는 wav2vec 2.0 음향 모델을 이용한 음소 단위 인식 실험으로, 텍스트 정보를 함께 이용할 수 있는 발음 오류 검출 과제의 특성을 이용하여 추가 모듈 또는 계층을 구성하거나 전사 단위를 더욱 엄밀하게 정의하는 등의 방법을 사용한다면 전체적인 모델의 성능 제고에 도움이 될 수 있다. 이러한 주제들은 추후 과제로 남겨 두고, 본 연구는 두 개의 원어민 발화 데이터셋을 교차 언어적으로 이용하는 발음 오류 검출 모델의 기초적 개념과 방법론을 제안하고 그 초기적 실험 결과를 제시하는 것으로 끝맺는다.

참고 문헌

교육부 국립국제교육원. [웹사이트]. (2020). *한국어능력시험(TOPIK) 지원자 현황_12/31/2020*. URL: https://www.data.go.kr/data/3059526/fileData.do#tab-layer-file

노채환. (2020). 중국인 학습자들의 한국어 이중모음 오류 분석 연구. *한중인문학연구, 68*, 227-247.

박시균, & 최경복. (2014). 중국인 학습자의 겹받침 발음 오류 분석: 겹받침 앞·뒤 자음 선택의 오류를 중심으로. *언어학, 22(4)*, 317-338.

배영환, & 金成蘭. (2019). 구술 발화를 통한 중국인 학습자의 발음 오류 양상 연구 - 종성 오류를 중심으로. *어문론집, 80*, 387-415.

신지영. (2014). *한국어의 말소리.* 박이정.

양승희. (2016). *Corpus-Based Analysis and Pronunciation Modeling of Korean Produced by Chinese Learners* (석사학위, 서울대학교 대학원).

양일. (2015). *중국어권 학습자를 위한 한국어 비음화와 유음화의 발음교육 연구* (석사학위, 서울대학교 대학원).

원다송. (2016). *한국어 학습자 오류 분석 연구의 연구 동향 분석* (석사학위, 연세대학교 대학원).

위국봉. (2014). 중국인 학습자의 /-ㄴ, -ㅇ/ 종성 발음 오류 분석 및 교정 방안 연구. *외국어로서의 한국어교육, 41*, 143-169.

위원징, & 홍미주. (2013). 중국인 학습자의 청취,발음 오류 분석을 통한 한국어 평음,경음,격음 교육 방안 연구. *한국어 교육, 24(4)*, 155-191.

유효려. (2009). 한국어 습득 과정에서의 발음 오류에 대한 분석 - 중국인 학습자를 중심으로. *한국(조선)어교육연구, 6*, 262.

이다연. (2017). *프랑스인 학습자를 위한 한국어 종성 발음 교육 연구* (석사학위, 서울대학교 대학원).

이미향, & 안미애. (2021). 음성 인식 기술(Speech Recognition)을 기반으로 한 한국어 학습자의 발음 진단·평가 시스템 구축 방안 연구. 국제한국언어문화학회 학술대회, 88-98.

이석재, 김정아, & 장재웅. (2006). 외국어로서의 한국어 음성 코퍼스의 설계, 구축 및 응용방안 연구. *한국(조선)어교육연구, 4*, 267-282.

이호영. (1996). *국어 음성학*. 태학사.

장찌엔. (2017). 중국인 학습자 비음 종성 /ㄴ/, /ㅇ/ 음절의 발음 오류 재고 -한·중 음절 유형을 통하여. *한국어 교육, 28(1)*, 251-268.

장향실. (2009). 중국인 학습자의 한국어 음절 오류와 교육 방안. *우리어문연구, 34*, 349.

정효주, & 김영주. (2011). 중국인 학습자의 한국어 비음화 발음 오류 양상. *이중언어학, 47*, 165-187.

정효주, & 김영주. (2012). 중국인 학습자의 유음화 발음 오류 연구. *한말연구, 0(31)*, 283-307.

조문우. (2006). 중국인 학습자의 한국어 단순모음 발음에 대한 연구. *중국어문학지, 21*, 365-388.

조민하. (2017). 중국인 한국어 학습자의 파찰음 발음 오류에 대한 음운·음성학적 연구. *Journal of Korean Culture, 37*, 93-128.

조민하, & 여문리. (2017). 중국인 학습자의 한국어 경음화 오류 분석을 통한 교육 방안 연구 - 인지와 산출의 상관성을 중심으로. *우리말글, 73*, 195-220.

최혜원, 김수연, & 고건희. (2022). 한국어 말하기 자동평가를 위한 발음 오류 분석 -중국어, 베트남어, 태국어 화자를 중심으로. *한국언어문화학, 19(1)*, 221-247.

한국교육개발원. [웹사이트]. (2021). *2021 고등교육기관 외국인 유학생 현황_211214*. URL: https://kess.kedi.re.kr/post/6672839?itemCode=04&menuId=m_02_04_03

한성우. (2008). 중국어권 학습자를 위한 맞춤형 한국어 발음 교육 방안. *우리말글, 44*, 165-194.

황미경, & 조남민. (2021). 중국인 한국어 학습자의 /ㅡ, ㅜ, ㅓ, ㅗ/ 발음에 대한 음향음성학적 연구 -선행 자음에 따른 발음 실현 양상을 중심으로. *이중언어학, 82*, 227-257.

Zheng Yihan, & 원미진. (2021). 중국인 학습자의 한국어 파열음 발음에 대한 조사 연구. *국제어문, 91*, 303.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer,

J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Baevski, A., Zhou, Y., 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.

Bang, J. U., Yun, S., Kim, S. H., Choi, M. Y., Lee, M. K., Kim, Y. J., Park. J., Lee, Y. J., & Kim, S. H. (2020). Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences, 10*(10), 6936.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ..., & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877-1901.

Bu, H., Du, J., Na, X., Wu, B., & Zheng, H. (2017). Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)* (pp. 1-5). IEEE.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Duanmu, S. (2007). *The phonology of standard Chinese*. OUP Oxford.

Fan, A., Grave, E., & Joulin, A. (2019). Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*.

Feng, Y., Fu, G., Chen, Q., & Chen, K. (2020, May). SED-MDD: Towards sentence dependent end-to-end mispronunciation

detection and diagnosis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3492-3496). IEEE.

Franco, H., Neumeyer, L., Kim, Y., & Ronen, O. (1997, April). Automatic pronunciation scoring for language instruction. In *1997 IEEE international conference on acoustics, speech, and signal processing* (Vol. 2, pp. 1471-1474). IEEE.

Gales, M. J., Knill, K. M., Ragni, A., & Rath, S. P. (2014, May). Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)* (pp. 16-23). International Speech Communication Association (ISCA).

Garofolo, J. S. (1993). Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993.*

Gong, Y., Chen, Z., Chu, I., Chang, P., & Glass, J. (2022). Transformer-Based Multi-Aspect Multi-Granularity Non-Native English Speaker Pronunciation Assessment. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7262-7266). IEEE.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning* (pp. 369-376).

Hu, W., Qian, Y., & Soong, F. K. (2015, September). An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners' speech. In *SLaTE* (pp. 71-76).

Huang, H., Xu, H., Wang, X., & Silamu, W. (2015). Maximum F1-score discriminative training criterion for automatic mispronunciation detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(4)*, 787-797.

Jang, E., Gu, S., & Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144.*

Joshi, S., Deo, N., & Rao, P. (2015). Vowel mispronunciation detection using DNN acoustic models with cross-lingual training. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Lee, A., & Glass, J. (2015). Mispronunciation detection without nonnative training data. In *Sixteenth Annual Conference of the International Speech Communication Association*.

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

Likic, V. (2008). The Needleman-Wunsch algorithm for sequence alignment. *Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne*, 1-46.

Lin, B., & Wang, L. (2022, May). Phoneme Mispronunciation Detection By Jointly Learning To Align. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6822-6826). IEEE.

Lo, T. H., Weng, S. Y., Chang, H. J., & Chen, B. (2020). An effective end-to-end modeling approach for mispronunciation detection. In *INTERSPEECH* (pp. 3027-3031).

Lopes, C., & Perdigao, F. (2011). Phoneme Recognition on the TIMIT Database. In *(Ed.), Speech Technologies*. IntechOpen. https://doi.org/10.5772/17600

Mathad, V. C., Mahr, T. J., Scherer, N., Chapman, K., Hustad, K. C., Liss, J., & Berisha, V. (2021). The Impact of Forced-Alignment Errors on Automatic Pronunciation Evaluation. In *Interspeech* (pp. 1922-1926).

Meng, H., Lo, Y. Y., Wang, L., & Lau, W. Y. (2007, December). Deriving salient learners' mispronunciations from cross-language phonological comparisons. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)* (pp. 437-442). IEEE.

Meng, H., Harrison, A., & Wang, L. (2009). Developing a Computer-Aided Pronunciation System for Chinese-Speaking Learners of English. In *Bulletin of Advanced Technology Research*.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., & Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015, April). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5206-5210). IEEE.

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Peng, L., Fu, K., Lin, B., Ke, D., & Zhang, J. (2021). A Study on Fine-Tuning wav2vec 2.0 Model for the Task of Mispronunciation Detection and Diagnosis. In *Interspeech* (pp. 4448-4452).

Peng, L., Gao, Y., Lin, B., Ke, D., Xie, Y., & Zhang, J. (2022). Text-Aware End-to-end Mispronunciation Detection and Diagnosis. *arXiv preprint arXiv:2206.07289*.

Qian, X., Meng, H., & Soong, F. (2010, November). Capturing L2 segmental mispronunciations with joint-sequence models in computer-aided pronunciation training (CAPT). In *2010 7th International Symposium on Chinese Spoken Language Processing* (pp. 84-88). IEEE

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint*

arXiv:1807.03748.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Wei, S., Hu, G., Hu, Y., & Wang, R. H. (2009). A new method for mispronunciation detection using support vector machine based on pronunciation space models. *Speech Communication, 51(10)*, 896-905.

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2-3), 95-108.

Wu, M., Li, K., Leung, W. K., & Meng, H. (2021). Transformer Based End-to-End Mispronunciation Detection and Diagnosis. In *Interspeech* (pp. 3954-3958).

Xu, X., Kang, Y., Cao, S., Lin, B., & Ma, L. (2021). Explore wav2vec 2.0 for Mispronunciation Detection. In *Interspeech* (pp. 4428-4432).

Yan, B. C., Wu, M. C., Hung, H. T., & Chen, B. (2020). An End-to-End Mispronunciation Detection System for L2 English Speech Leveraging Novel Anti-Phone Modeling in *Interspeech* (pp. 3032-3036).

Yang, M., Hirschi, K., Looney, S. D., Kang, O., & Hansen, J. H. (2022). Improving Mispronunciation Detection with Wav2vec2-based Momentum Pseudo-Labeling for Accentedness and Intelligibility Assessment. *arXiv preprint arXiv:2203.15937*.

Zhang, F., Huang, C., Soong, F. K., Chu, M., & Wang, R. (2008, March). Automatic mispronunciation detection for Mandarin. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5077-5080). IEEE.

Zhao, G., Sonsaat, S., Silpachai, A., Lucic, I., Chukharev-Hudilainen, E., Levis, J., & Gutierrez-Osuna, R. (2018). L2-ARCTIC: A non-native English speech corpus. In *INTERSPEECH* (pp.

2783-2787).

부록

A. 각 데이터셋에 대한 원 전사 단위 및 IPA 기호 대응 표

A.1. KsponSpeech 원본 전사 한글을 jamo 및 g2pK로 변환한 한글 자모와 IPA 대응 표

한글 자모	IPA	한글 예시	한글 자모	IPA	한글 예시
중성 ㅏ	a	**아**버지	초성 ㄱ	k	**강**
중성 ㅐ	e	**개**미	초성 ㄲ	k=	토**끼**
중성 ㅔ		수**레**	초성 ㄴ	n	**노**래
중성 ㅑ	j a	**약**	초성 ㄷ	t	무**대**
중성 ㅒ	j e	**얘**기	초성 ㄸ	t=	**또**는
중성 ㅖ		**계**산	초성 ㄹ	ɾ	거**리**
중성 ㅓ	ʌ	**어**머니		l	빨**래**
중성 ㅕ	j ʌ	**여**름	초성 ㅁ	m	고**민**
중성 ㅗ	o	**모**래	초성 ㅂ	p	**병**
중성 ㅘ	w a	**과**제	초성 ㅃ	p=	**빨**강
중성 ㅙ	w e	유**쾌**	초성 ㅅ	s	**상**장
중성 ㅚ		**외**국		ɕ	**시**각
중성 ㅞ		**훼**손	초성 ㅆ	s=	**쏘**다
중성 ㅛ	j o	학**교**		ɕ=	글**씨**
중성 ㅜ	u	소**수**	초성 ㅈ	tɕ	**저**녁
중성 ㅝ	w ʌ	**권**리	초성 ㅉ	tɕ=	이**쪽**
중성 ㅟ	w i	**위**임	초성 ㅊ	tɕʰ	**차**
중성 ㅠ	j u	여**유**	초성 ㅋ	kʰ	**커**피
중성 ㅡ	ɯ	**그**림	초성 ㅌ	tʰ	**토**론
중성 ㅢ	ɯ i	**의**사	초성 ㅍ	pʰ	**판**단
중성 ㅣ	i	**기**타	초성 ㅎ	h	**하**루
			종성 ㄱ	k’	기**적**
			종성 ㄴ	n’	**만**두
			종성 ㄷ	t’	**닫**기
			종성 ㄹ	l	법**률**
			종성 ㅁ	m’	공**감**
			종성 ㅂ	p’	**입**
			종성 ㅇ	ŋ	**종**료

A.2. AISHELL-1 원본 전사 한자와 pypinyin으로 변환한 한어병음 및 IPA 대응 표

한어병음	IPA	한자 예시	한어병음 예시	한어병음	IPA	한자 예시	한어병음 예시
b	p	冰	bing	a	ɑ	马路	malu
p	pʰ	气泡	qipao	o	o	放送	fangsong
m	m	慢	man	e	ɤ	特别	tebie
f	f	父母	fumu	er	ɚ	而且	erqie
d	t	期待	qidai	ai	a i	招牌	zhaopai
t	tʰ	太阳	taiyang	ei	ə i	送给	songgei
n	n	努力	nuli	ao	a u	电脑	diannao
l	l	离开	likai	ou	ə u	口罩	kouzhao
g	k	光	guang	an	a nˀ	上班	shangban
k	kʰ	裤子	kuzi	en	ə nˀ	请问	qingwen
h	x	海浪	hailang	ang	a ŋ	健康	jiankang
z	ts	兔子	tuzi	eng	ə ŋ	梦	meng
c	tsʰ	从来	conglai	ong	u ŋ	分钟	fenzhong
s	s	意思	yisi	ia	j ɑ	下去	xiaqu
zh	ʈʂ	主张	zhuzhang	iao	j a u	桥	qiao
ch	ʈʂʰ	出发	chufa	ie	j e	节日	jieri
sh	ʂ	树木	shumu	iou	j ə u	由于	youyu
r	ʐ	当然	dangran	ian	j a nˀ	钱	qian
j	tɕ	根据	genju	in	i nˀ	心理	xinli
q	tɕʰ	清明	qingming	iang	j a ŋ	方向	fangxiang
x	ɕ	小心	xiaoxin	ing	i ə ŋ	聪明	congming
				iong	j u ŋ	雄	xiong
				u	u	目前	muqian
				ua	w ɑ	花	hua
				uo	w o	脱出	tuochu
				uai	w a i	快	kuai
				uei	w ə i	喂	wei
				uan	w a nˀ	万里	wanly
				uen	w ə nˀ	问题	wenti
				uang	w a ŋ	窗户	chuanghu
				ueng	w ə ŋ	老翁	laoweng
				ü	y	必须	bixu
				üe	ɥ e	却	que
i	i	比率	birü	üan	ɥ a nˀ	选择	xuanze
	z	司机	siji	ün	ɥ i nˀ	裙子	qunzi
	ʐ	好吃	haochi	io	j o	哟	yo

A.3. L2KSC 원본 전사 PLU(양승희, 2016)와 IPA 대응 표

PLU	한글 자모	IPA	한글 예시	PLU	한글 자모	IPA	한글 예시
AA	중성 ㅏ	a	**아**버지	K	초성 ㄱ	k	**강**
EH	중성 ㅐ	e	**개**미	KK	초성 ㄲ	k=	토**끼**
EY	중성 ㅔ		수**레**	N	초성 ㄴ	n	**노**래
JA	중성 ㅑ	j a	**약**	T	초성 ㄷ	t	무**대**
JH	중성 ㅒ	j e	**얘**기	TT	초성 ㄸ	t=	**또**는
JE	중성 ㅖ		**계**산	R	초성 ㄹ	ɾ	거**리**
AX	중성 ㅓ	ʌ	**어머**니	L	초성 ㄹ	l	빨**래**
JX	중성 ㅕ	j ʌ	**여**름		종성 ㄹ		고**민**
OW	중성 ㅗ	o	**모**래	M	초성 ㅁ	m	**병**
WA	중성 ㅘ	w a	**과**제	P	초성 ㅂ	p	**빨**강
WH	중성 ㅙ	w e	유**쾌**	PP	초성 ㅃ	p=	**상**장
WE	중성 ㅞ		**훼**손	S	초성 ㅅ	s	**시**각
JO	중성 ㅛ	j o	학**교**			ɕ	**쏘**다
UW	중성 ㅜ	u	소**수**	SS	초성 ㅆ	s=	글**씨**
WX	중성 ㅝ	w ʌ	**권**리			ɕ=	**저**녁
UI	중성 ㅟ	w i	**위**임	Z	초성 ㅈ	tɕ	이**쪽**
JU	중성 ㅠ	j u	여**유**	ZZ	초성 ㅉ	tɕ=	**차**
WW	중성 ㅡ	ɯ	**그**림	CH	초성 ㅊ	tɕʰ	**커**피
WI	중성 ㅢ	ɯ i	**의**사	KH	초성 ㅋ	kʰ	**토**론
IY	중성 ㅣ	i	**기**타	TH	초성 ㅌ	tʰ	**판**단
				PH	초성 ㅍ	pʰ	**하**루
				HH	초성 ㅎ	h	기**적**
				KQ	종성 ㄱ	k’	**만**두
				NN	종성 ㄴ	n’	**닫**기
				TQ	종성 ㄷ	t’	법**률**
				MM	종성 ㅁ	m’	공**감**
				PQ	종성 ㅂ	p’	**입**
				NX	종성 ㅇ	ŋ	**종**료

B. wav2vec 2.0 Large 사전 훈련 모델 기반 각 파인튜닝 모델의 L2KSC 시험 셋에 대한 혼동 행렬

B.1. L2KSC

	-	a	e	ʌ	o	u	ɯ	i	j	w	k	kʰ	kˀ	n	nˀ	t	tʰ	tˀ	ɾ	l	m	mˀ	p	pʰ	pˀ	s	sˀ	ɛ	ɛˀ	tɕ	tɕˀ	tɕʰ	ŋ	h	ts	ʂ	ɚ	ɥ	ɑ	ə	ɤ	y	ɥ	tsʰ	ʃʂ	ʃʂʰ	z	ʐ	f	x
-	27	2	1	3	5	2	1	1	22	3	2			13	1	8	2			8	2	5	37	3	1			2	2		1				1	3															
a	1	514									1					1						1																														
e	1		77		1	1		3	1					1																																					
ʌ	3	6	1	64	37	3	5				6					2	1					5			2																										
o	2			24	127	31		1		1																																									
u	12			5	5	6	7	9									1								1										2																
ɯ				2	3	6	8																																												
i	8			3				135														3	1													1																
j	38		1		1			4	132		2							1	1			1			2		1																									
w	3				1	2		1		59																			18	3	1										1											
k	2			1							15	2	14																																						
kˀ	1										36	52	3																																						
kʰ	7					1			1		34		54																													4										
kˋ	22													4						1																1																
n	2														46																																					
nˋ	2	1	1													5				1		1		1																												
t																	111	32	2			1										3		1																		
tˀ				1													14	24															1																			
tʰ	1	1			3										1		16	3	66															2																		
tˋ	8													39		1				25																																
ɾ	1																				5	9																														
l	6							1		1				1	1	2					1	75																														
m	4							1							2								52																													
mˋ	3		1		2		4	1								2								45											11																
p	1																7						1		23	15	1								2																
pˀ																										1																										
pʰ	7										4					1									28		51																									
pˋ																												1																								
s																													3	2																						
sˀ																																																				
ɛ	17				2	1																							7	2	6	3				1															
ɛˀ	3			1																									4		18	19			1																
tɕ																																	1																		
tɕˀ													1																			8	25	2			2														
tɕʰ																																	1	2																	
ŋ	4																							2												5															
h	1									1			1									1															9															
ts																																1	4				1														
ʂ	1																												2		1																				
ɚ	4																					3																													

B.2. KsponSpeech

	-	a	e	ʌ	o	u	ɯ	i	j	w	k	kʷ	kʰ	k˞	n	n˞	t	tʷ	tʰ	t˞	ɾ	l	m	m˞	p	pʷ	pʰ	p˞	s	sʷ	ɕ	ɕʷ	tɕ	tɕʷ	tɕʰ	ŋ	h	ts	ʂ	ɚ	ɯ̞	ɑ	ə	ɤ	y	ɥ	tsʰ	ʈʂ	ʈʂʰ	z	ʐ	f	x
-	29	11	3	28	8	6	8	9	34	15	26	1		2	1	7	1	1	1	19	7	5	38	9	2		1	1	1	1		1	1		3	1	3				1												
a		515		1				1			1																																										
e		1	76	1				3			1			1								1		1																														
ʌ	4	4	1	7	4		8	3			1						1					2			1																												
o	4	9		45	13	5	3	1			11			1		1				3																																		
u	8	1		9	27	41	2	2		7		1		1																	2				1																		
ɯ	1	1		1	2	2	81											1				1											1																				
i	1		3	1	1			145																																													
j	9		5	3	1	3		25	13		5						1												1								1																
w	4	1		1	5	8	8	7	2	31	3	1																	1	1			4				3																
k	3	1									91	12	12		1		1	1																			2																
kʷ	11						7				31	42	1																																								
kʰ	1					1	4				32	1	54						1		2									1							4																
k˞	18	1		3	1	1								35						1								3								1																
n								1							37								9													1																	
n˞	7	1	1	1												41				1		3		1												1																
t	2			1													11	28	5		1	1											2																			
tʷ																	1	2	1																																	
tʰ	1	1	1		1										1		13	2	67														1		5																
t˞	22	3	1	13										13		1				17								3																								
ɾ															3						6	15																														
l	9		1		2		1	3						1	6		1				4	59																				1											
m	3							1															53						2																								
m˞	17		1					2								5							1	6												1																	
p	1					1		1							1		1						6		33	5	1																										
pʷ																										1																											
pʰ	3							1																	14		71				2																						
p˞																								1																													
s																													2	1	1		1																				
sʷ																																																						
ɕ	9			5	3																								4	1	17																						
ɕʷ				1																									4	1	39	1																					
tɕ																																		1																			
tɕʷ													1																	1			1	34	1																		
tɕʰ																													1					1	1																		
ŋ	5													1										2												48																	
h	3		2					3		2	7		1								1						1		1								73																
ts																																		6																		
ʂ								2																							2																						
ɚ	3			1																		3																															

B.3. AISHELL−1

	-	a	e	ʌ	o	u	ɯ	i	j	w	k	kʷ	kʰ	kˇ	n	nˇ	t	tʷ	tʰ	tˇ	ɾ	l	m	mˇ	p	pʷ	pʰ	pˇ	s	sʷ	ɕ	ɕʷ	tɕ	tɕʷ	tɕʰ	ŋ	h	ts	ʂ	ɚ	ɥ	ɑ	ə	ɤ	y	ɥ	tsʰ	ʈʂ	ʈʂʰ	z	ʐ	f	x
-	17	24			3	43		43	12	12	15		5			12	5		1			1	4		2		1				7				1	11		1	3	3		16	186	13				1			2		7
a	28	222	1			1		1	11				5			5	2						6				5				2							2	1	1		141	25	55			1	2					1
e		1	16		2			55								1							1						1														7		1							
ʌ	5	23	1		5	47		1								3																				6							21	16	2						4		1
o	2	4			22	12		1			6						1																			17							19	1									2
u					6	91		2																																				1	1						1		
ɯ	31					4		7	1		1		13			1	2					1																1					4	19				1	1	3			1
i	8		1			1		125	4													3									1												4		4								
j	5	2			1	2		1	147	1												1									2												16	2		2			1				1
w	5								2	58																																	1	1		13							
k	24					1					6		31						2																							5										1
kʷ	8	1				1				12	65		2												1										1									1										
kʰ											4		9						3																1								1										2
kˇ	11					1		2					1			1																	4							1		4	1	28							1		
n															18							7	2													2							1										
nˇ	1	1						1								48																				5						1										
t						1					1						141																1					1				2	2						1			
tʷ	1	2				2				1							22		1																							1	7	1				1				1
tʰ						1											8		74								1								1												1	1	6				
tˇ	19		3			2										2	2		2												1					2				3		13	2	4								
ɾ	13																					9																					1	1								
l	2					1		22	1							6						44																	2		3	2	1							4		
m	11							1															45		1																									1		
mˇ	2					4		4								24	1					1	8													36				1		2	2	2								
p	7																					1			11		31																									
pʷ																									1																												
pʰ										2							1		2						3		8								1	1											1						
pˇ	1																																																				
s	1	1																													1												1				1						
sʷ																																																					
ɕ	1																												13		14				4			1	6														
ɕʷ	4																												6		31								1							4							
tɕ																																															1						
tɕʷ	1								4				1																1								1										29	1					
tɕʰ																	1																		1														1				
ŋ	2					1		1								5																				46						1											
h	14							2	5				1		1		1		6								2		1		1				2				5				4		2				1				37
ts																																															6						
ʂ																															4																						
ɚ	1					1		1								2																							1		1												

B.4. Cross-lingual

	-	a	e	ʌ	o	u	ɯ	i	j	w	k	kʷ	kʰ	kˀ	n	nˀ	t	tʷ	tʰ	tˀ	ɾ	l	m	mˀ	p	pʷ	pʰ	pˀ	s	sʷ	ɕ	ɕʷ	tɕ	tɕʷ	tɕʰ	ŋ	h	ts	ʂ	ɚ	ɥ	ɑ	ə	ɤ	y	ɥ	tsʰ	ʈʂ	ʈʂʰ	z	ʐ	f	x
-	23	10	4	20	5	11	10	7	24	12	16		1	3	1	5	2				20	4	6	38	7	2			1			1	1	2		2	3	4					1	1									
a	2	509		1				1	2																																		3										
e		1	74	2			2	3						1									2																														
ʌ	5	25	1	85	2	1	8	2			2					1	1						2																														
o	18	7		50	83	9	1	1		1	6			2		1	1		1	4																1																	
u	5			9	22	52	2			12																																											
ɯ	1	1		1	1	3	80																1											2																	1			
i	1		3	1				145	1																																												
j	5		4	2		2		22	142		3											1	1			2																												
w	3	1		1	2	4	4	6	1	40	4																			17	2			3														1						
k	1				1						86	22	12				1																					1																
kʷ	11			1			2				23	53	2																																									
kʰ	2					1	3		1		28	7	52	1			1		1		1																	3																
kˀ	22			3		1								34		1												3																										
n								1							34									13																														
nˀ	5	1	1	1												45				1			3																															
t	2			1							1						104	35	4				1											1					1															
tʷ																	7	27	6																																			
tʰ		1			2										1		10	2	71															1		5																		
tˀ	28	2	1	17				1						16		1				5								1															1											
ɾ	3														3						3	15																																
l	9		1		1			1							4						3	67														2																		
m	5							1							1								51						1																										
mˀ	11		1					1							1	8							1	61												3																			
p	2																1				1		6		31	5	4																											
pʷ																										1																												
pʰ	3																				1				11		74		1		1																								
pˀ																								1																														
s																													3	1	1																							
sʷ																																																						
ɕ	10			8	1		1																						6	1	9	3																						
ɕʷ				3																									3		32	8																						
tɕ																																		1																			
tɕʷ																			1											1				35	1																		
tɕʰ																		1											1						1																			
ŋ	7															1								2												46																	
h	1			1				2		1	7		1									1							2								78																
ts																																		6																			
ʂ																															4																							
ɚ	2			2																			3																															

* 세로 축이 정답 음소, 가로 축이 모델 예측 음소이며, 세로 축의 '-'은 학습자의 삭제 오류, 가로 축의 '-'은 모델의 삭제 오류를 의미함. 단 세로와 가로축 모두 '-'인 칸은 모델이 학습자의 삭제 오류를 올바르게 검출했음을 의미함.

C. XLSR-53 사전 훈련 모델 기반 각 파인튜닝 모델의 L2KSC 시험 셋에 대한 혼동 행렬

C.1. L2KSC

	-	a	e	ʌ	o	u	ɯ	i	j	w	k	kʷ	kʰ	kˀ	n	nˀ	t	tʷ	tʰ	tˀ	ɾ	l	m	mˀ	p	pʷ	pʰ	pˀ	s	sʷ	ɕ	ɕʷ	tɕ	tɕʷ	tɕʰ	ŋ	h	ts	ʂ	ɚ	ɯ̞	ɑ	ə	ɤ	y	ɥ	tsʰ	ʈʂ	ʈʂʰ	z	ʐ	f	x
-	35	2	1	6	4	2	2	2	11	1	3			7	1	6	1			3	4	2	37	4	2			2	1	1	1		1		2	1																
a		513	1	1			1				1					1																																				
e			79		1			3	1					1																																						
ʌ	4	1	1	5	48	2	8	2			1					2						7																														
o	2			23	145	13		1												2																																	
u	3		1	3	7	81	1	1									1			2															2																		
ɯ					3	2	86																																														
i	7			2				139	1							1							1																														
j	7				1			4	17	1												1																															
w	1					1		1		69																			12	2	1							1			1												
k				1							16	13	4																																								
kʷ											15	75	2																																								
kʰ	1								1	1	35		58				1																				4																
kˀ	13													39		1				1								1																										
n				1											46	1																																					
nˀ	2	1	1													49				1		1		1												1																	
t																	11	42	2			1											2	1	1																		
tʷ																	8	31	1																																		
tʰ	1				2		1	1							1		6	2	78																1																		
tˀ	12													21		1				38								1																										
ɾ	7								1												6	1																																
l	5		1				1			1				1	1	1					2	75																																
m	4							1							2								52																															
mˀ	2					1		1								4				1				74											4																		
p	1																1								24	9	15																										
pʷ																										1																											
pʰ	4										2														29		56																										
pˀ																												1																									
s	1																												2	2																							
sʷ																																																					
ɕ	15					2																							1	1	8	2			1																		
ɕʷ																													3	3	27	12					1																
tɕ																																		1																			
tɕʷ													1																					36	1																		
tɕʰ																																		1	2																		
ŋ	6			1																				2												47																	
h	2								1																												91																
ts																																		6																			
ʂ																													1		3																						
ɚ	3																					3																	1														

C.2. KsponSpeech

	-	a	e	ʌ	o	u	ɯ	i	j	w	k	kʷ	kʰ	k˭	n	n˭	t	tʷ	tʰ	t˭	ɾ	l	m	m˭	p	pʷ	pʰ	p˭	s	sʷ	ɕ	ɕʷ	tɕ	tɕʷ	tɕʰ	ŋ	h	ts	ʂ	ɚ	ɯ̞	ɑ	ə	ɤ	y	ɥ	tsʰ	ʈʂ	ʈʂʰ	z	ʐ	f	x
-	26	11	5	18	1	14	15	15	27	8	26	1	2	5	3	8	1	2		18	7	4	36	16			1	1	1			1	2		3	1	8															
a	1	513		1			2	1																																													
e		1	76	1				4						1		1								1																													
ʌ	5	33	2	76	2		8	4			1					1	1					2																															
o	2	7		6	84	12	3	1			11	1				1				4																																	
u	1			7	26	51	3			3				1																	1																						
ɯ	1			1	2	2	8							2								1											2																				
i	1		3	1	1			145																																													
j	7		8		1	3	1	36	119	3	3						1																			1	1															
w	4	1		1	5	11	18	6	1	25	1										1			1					1	1		1	2																				
k	2								1		16	9	2		1		1																1				1																
kʷ	6						6		1	1	5	27	1																																								
kʰ	1					1	1		1		35	3	54						1		1																3																
k˭	17			3	1	1								37		1				1								3																									
n								1							35								12																														
n˭	4	2	2													44				1		4																															
t	1			2	1												13	34	3		1	1											3	1																			
tʷ																	8	25	7																																		
tʰ		1	1		2										1		11	2	67														1		7																		
t˭	19			12			1							13		1				24								2	1																									
ɾ															3						7	13															1																	
l	1				1		1	1							5		2				4	62														2																	
m	3			1				1															53						1																									
m˭	13		1					1								11							5	55												1																		
p						1				1					1		2						5		33	6	1																										
pʷ																										1																												
pʰ	1																								1		79				1																						
p˭																								1																													
s																													3	1	1																						
sʷ																																																					
ɕ	8			2	4	1																							9		12	2	1																			
ɕʷ						1																							5		33	7																					
tɕ																																		1																		
tɕʷ													1																					35	2																	
tɕʰ																																		1	2																	
ŋ	8															2								1												45																	
h	1			1	1					1	5		1																								84															
tʂ																																		6																		
ʂ								1																							3																						
ɚ	3			1																		3																															

C.3. AISHELL-1

	-	a	e	ʌ	o	u	ɯ	i	j	w	k	kʷ	kʰ	kˀ	n	nˀ	t	tʷ	tʰ	tˀ	ɾ	l	m	mˀ	p	pʷ	pʰ	pˀ	s	sʷ	ɕ	ɕʷ	tɕ	tɕʷ	tɕʰ	ŋ	h	ts	ʂ	ɚ	ɯ̩	ɑ	ə	ɤ	y	ɥ	tsʰ	ʈʂ	ʈʂʰ	z	ʐ	f	x
-	14	27			1	21		36	16	5	6		3			14	2						33				1								1	5			2	4		15	226	5					3		2		4
a	21	231				2		2	15				3			2			1				5				7												1			199	8	17				1			1		2
e	1		1		1	1		64								1							1																				6										
ʌ	4	24	1			41		2								5	1																			6			1					38	1							2		
o		5			15	115		1			3						11																			3								32	1									
u						96		1																			1		1							1									1							1		
ɯ	32				1	2					1		14			1	1																					1						4	29				1		3			1
i	2	1				1		142																							3					1								1									
j	4							2	164													1																						7	1		2			1		1		1
w						1			1	72			1																														5		1	8							
k	2				1						6		38														1															3	1										
kʷ	4	1								3	78		4												1		1																										
kʰ											3		92																																1								5
kˀ	4					12							3																				2			1						4		37							1		
n															29							2	17																														
nˀ						3		1								48																				4							1										
t																	144		1			1																				2	1					1					
tʷ	1	1				1				2							24		1																								8	1				1					
tʰ															1		5		81																								1						5				
tˀ	15		1			3											2																			1				2		19		3									
ɾ	17																2					4																					1										
l	6					2		22		1					1	2						42																		1		4		3							4		
m	4							1															5		4																												
mˀ	7					1		1		1						31																				44							1	1									
p	8																		1						14		25																1								1		
pʷ																									1																												
pʰ																			1								85								1	1								1	1				1					
pˀ																																												1									
s																													3		1												1										
sʷ																																																					
ɕ																													13		12				1			1	9				1				2						
ɕʷ	3																												5		36								2														
tɕ																																																1					
tɕʷ									2				1																				2					1						1				3	1				
tɕʰ																	1																		1								1										
ŋ						1										1																				53						1											
h	8								2				1				1					1									26								1					1						1				52
ts																																						4										2					
ʂ																															4																						
ɚ	2					1		1																																1		2											

C.4. Cross-lingual

	-	a	e	ʌ	o	u	ɯ	i	j	w	k	kʷ	kʰ	k˴	n	n˴	t	tʷ	tʰ	t˴	ɾ	l	m	m˴	p	pʷ	pʰ	p˴	s	sʷ	ɕ	ɕʷ	tɕ	tɕʷ	tɕʰ	ŋ	h	ts	ʂ	ɚ	ɯ̞	ɑ	ə	ɤ	y	ɥ	tsʰ	ʈʂ	ʈʂʰ	z	ʐ	f	x
-	27	1	5	15	9	11	17	9	26	9	28	1		3	2	6	1			19	5	7	39	11	1			1			2	1	1		1	2	4																
a	1	515				1	1																																														
e		2	74	1			4							1										3																														
ʌ	6	26	1	85	2	1	9	2			1					1						1																															
o	4	8		59	86	12	1	1			11			1		1		1		1																																	
u	1			3	3	53	1			1	1			1													1				1																							
ɯ	1				3	2	81												1			1											2																				
i	1		2	1	1			146																																													
j	6		6	1	1	1	1	25	137	2					1		1																			1	1																
w	2			1	6	13	12	7		31	1																		12				1				3																	
k	1								1		17	1	3		1		1																																				
kʷ	9	1		2			6		1		37	35	1																																								
kʰ	1						2		1		35	2	55								1													1		3																	
k˴	17			1	1	1								4		1				1								2																									
n								1							34								13																														
n˴	3	1	1	1												43				1		5		1												1																	
t	2			1													13	36	4		1	1											2																				
tʷ	1																8	25	6																																		
tʰ					1										1		13	2	68														1		7																		
t˴	24		1	1	1		2							15						18								2																									
ɾ	1														2						7	14																															
l	8				1		1	1							3		3				4	67																															
m	3			1																			55																														
m˴	14			1				1								2	1						6	6												2																	
p										1							1						5		35	6	2																										
pʷ																										1																											
pʰ	1		1					1																	12		75				1																						
p˴	1																																																				
s																													3	1	1																							
sʷ																																																					
ɕ	8				5																								9	1	14	2																					
ɕʷ																													5	1	36	4																					
tɕ																																		1																			
tɕʷ													1																				1	35	1																		
tɕʰ																																		1	2																		
ŋ	7															1								2												46																	
h	3		1	1			1			1	3		1																								83																
ts																																		6																			
ʂ	1							2																							1																							
ɚ	3			1																		3																															

* 세로 축이 정답 음소, 가로 축이 모델 예측 음소이며, 세로 축의 '–'은 학습자의 삭제 오류, 가로 축의 '–'은 모델의 삭제 오류를 의미함. 단 세로와 가로축 모두 '–'인 칸은 모델이 학습자의 삭제 오류를 올바르게 검출했음을 의미함.

Abstract

wav2vec 2.0-based Mispronunciation Detection Using Cross-lingual Fine-tuning

Seunghun Ji
Department of Linguistics
The Graduate School
Seoul National University

Mispronunciation detection (MD), which is a subtask of computer-aided pronunciation training (CAPT), aims to construct a computational model which, given an utterance of a foreign language learner, can exactly judge whether each phoneme of the utterance is mispronounced. Utilizing the output of a well-trained mispronunciation detection model, a foreign language instructor can offer corrective feedback on the learner's pronunciation, whether manually or automatically. The current main research trend in MD model is to conduct a phoneme-level speech recognition using deep neural network-based acoustic models, with non-native, foreign language learners' speech dataset of sentences or words as training data. However, it is highly difficult to gather a big amount of this type of dataset which has an enough amount to train such deep neural network-based models. To explore the possibility of solving this problem, this study suggests a method of training an acoustic model only with native data of the learners' background L1 and native data of the target L2, instead of non-native L2 learners' data.

We utilize wav2vec 2.0, which performs well in much of recent audio-related tasks. Target L2 is Korean, learners' background L1 is Mandarin Chinese. The training of wav2vec 2.0 is divided into pre-training and fine-tuning stages, and we investigate the performances and analyze the recognition results of pre-trained models which are fine-tuned with Chinese learners' data of Korean, monolingual native data of Korean or Mandarin Chinese, or cross-lingual native data of Korean and Mandarin Chinese, respectively. Besides, the transcription and recognition units are from IPA system, in order to uniformly represent the phoneme-level transcription text of both languages. To do so, we also suggest a method of converting Korean and Chinese text to IPA sequences, considering the phonemic systems and phonological variations.

The cross-lingually fine-tuned models with native speech data of both languages showed 74.37% correct rate in phoneme recognition, 39.90% F1-score in mispronunciation detection, at their highest. These are 1.42%p and 1.44%p higher than those of the monolingually fine-tuned models, respectively. However, it could not reach the baseline performance, which are 80.11% correct rate, 50.00% F1-score. But when the fine-tuned model are investigated in terms of false rejection, common errors which Chinese learners of Korean make, and so on, each model showed a mixed recognition aspects.

Keywords : mispronunciation detection, wav2vec 2.0, cross-lingual training, acoustic models, Chinese learners of Korean, non-native speech dataset
Student Number : 2020-25664