pISSN 2005-8063
eISSN 2586-5854
2025. 9. 30.
Vol.17 No.3
pp. 83-91

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2025.17.3.083

Automatic Korean dialect identification using wav2vec 2.0 XLS-R*

Jooyoung Lee[1] · Sunhee Kim[2] · Minhwa Chung[1,**]

[1]*Department of Linguistics, Seoul National University, Seoul, Korea*
[2]*Deaprtment of French Language Education, Seoul National University, Seoul, Korea*

Abstract

Previous Korean automatic dialect identification systems have demonstrated limited accuracy, hovering around 60%, and have often relied on traditional machine learning models. A further methodological issue has been the categorization of dialects by administrative districts, a method inconsistent with established linguistic boundaries. This study seeks to overcome these limitations by applying a modern, Transformer-based, end-to-end speech model, wav2vec 2.0 cross-lingual speech representation (XLS-R). The model, pre-trained on approximately 436,000 hours of speech from 128 languages, was fine-tuned for this task. We also re-categorized the dialects into six major linguistic zones: Central, Gyeongnam, Gyeongbuk, Jeonnam, Jeonbuk, and Jeju. Using a dataset of middle-aged and senior male speakers, the XLS-R model was compared against baseline models like Bi-LSTM. The results show a significant performance increase, with the XLS-R model achieving an F1 score of 86.8%—a 14 percentage point improvement over the strongest baseline. While confusion between certain dialects persists, this research validates the effectiveness of applying large-scale, pre-trained models to the nuanced task of dialect identification and underscores the importance of using linguistically-informed categories. Furthermore, the findings contribute to advancing dialect identification technology for applications in speech recognition and forensic science.

Keywords: automatic dialect identification, Korean dialects, wav2vec 2.0, XLS-R

1. 서론

방언 자동 식별(automatic dialect identification)은 음성 신호로부터 화자의 지역 정보를 자동으로 추정하는 기술이다. 이는 언어 식별(language identification)의 하위 개념이지만, 방언 간 음운 체계를 공유(shared phonetic inventories)하고 상호 의사소통이 가능하다는 특성(mutual comprehensibility) 때문에 기술적 난이도가 더 높다(Chambers & Trudgill, 1998; Dobbriner & Jokisch, 2019; Shon et al., 2018).

* This research was supported by the Department of Linguistics at Seoul National University under the project LINGUISTICS SNU 10-10 INITIATIVE.
** mchung@snu.ac.kr, Corresponding author
Received 17 August 2025; Revised 10 September 2025; Accepted 11 September 2025

방언 자동 식별 연구는 세 가지 측면에서 중요한 의의를 지닌다. 첫째, 학술적인 측면에서 방언 자동 식별은 방언 구획 연구에 기여한다. 방언 구획은 어떤 언어가 사용되는 지역의 말을 독립된 언어체계를 가지는 방언으로 구분하는 과정(Choi, 2005:42)이다. 방언 자동 식별은 방언권 경계에 타당성을 부여한다. 왜냐하면 방언 경계에 의미가 있다면 방언 자동 식별에서 방언 간 식별률 또한 높을 것이기 때문이다. 둘째, 기술적인 측면에서 방언 자동 식별은 방언 음성인식에 기여한다(Dobbriner & Jokisch, 2019; Liu et al., 2010). 방언별로 음성인식 모델을 구축하면 특정 방언에 대한 인식률이 향상되지만 입력 음성이 어느 방언인지 파악하는 것이 우선되어야 한다. 방언 자동 식별은 입력 음성의 방언을 추정하여 관련한 방언 모델을 선택할 수 있게 도와준다. 마지막으로, 사회적 측면에서 신원 미확인 화자를 추정하는 데 활용될 수 있다. 구체적으로 과학 수사 단계에서 화자 프로파일링을 통해 용의자의 방언 정보를 추정하는 데 도움을 제공할 수 있다(Lee et al., 2022; Schilling & Marsters, 2015). 이와 같이 방언 자동 식별은 여러 분야에서 중요한 역할을 맡고 있어 높은 식별 정확도가 요구된다.

현재까지 한국어에 대한 방언 자동 식별은 6개 지역(서울/경기, 경상, 전라, 충청, 강원, 제주)에 대해 정확도 및 F1 점수 기준 약 60%대에 머물러 있다(Kim & Kim, 2021; Lee et al., 2019; Lee et al., 2021; Lee et al., 2022). 이는 결코 높은 식별 성능이라 말할 수 없다. 또한 방언 식별에 활용한 학습 모델은 SVM(support vector machine), RF(random forest)를 비롯한 기계 학습 알고리즘과 DNN(deep neural network), CNN(convolutional neural network), RNN(recurrent neural network), LSTM(long short-term memory), Bi(bidirectional)-LSTM 등의 프레임워크에서 벗어나지 못하였다. 한편, 해외 연구는 트랜스포머(Transformer)에 기반한 종단간(end-to-end) 음성 학습 모델을 적용하여 높은 식별 성능을 보였다(Imaizumi et al., 2022; Kakouros & Hiovain-Asikainen, 2023; Wang et al., 2021). 언어에 따라, 그리고 방언의 수에 따라 성능이 다르므로 한국어 방언 식별과의 직접적인 비교는 어려우나 트랜스포머에서 성능 향상을 보인다는 것은 여러 연구 사례를 통해 확인할 수 있다.

이에 본 연구에서는 트랜스포머 기반 음성 모델을 한국어 방언 식별에 적용하여 방언 식별 성능을 높이고자 한다. 이를 위해 자기지도학습(self-supervised learning) 방식으로 학습하는 wav2vec 2.0 XLS-R을 활용하여 한국어 방언 식별의 향상된 성능을 확인하고 방언별 식별률이 어떻게 개선되는지 살펴본다.

본 논문은 다음과 같이 구성한다. 2장에서는 한국어 방언 자동 식별 관련 선행 연구를 방언권, 음향 특징, 학습 모델 종류, 데이터에 따라 탐구하고 개선할 수 있는 부분을 살펴본다. 3장에서는 한국어 방언권에 대해 논하고 한국어 음소 인식으로 적응한 XLS-R 모델의 구성을 설명한다. 4장에서는 데이터셋, 한국어 음소 인식 미세조정 절차, 모델 훈련 설정에 대해 설명한다. 5장에서는 여러 베이스라인 모델과 XLS-R 모델의 방언 식별 성능을 비교하고 최고 성능을 보고한다. 6장에서는 혼동행렬을 제시하면서 XLS-R에서의 방언별 성능 개선과 방언 간 혼동 패턴에 대해 분석하고 그 의미를 논의한다. 마지막으로 7장에서는 본 연구의 기여와 향후 연구 방향을 제시하면서 결론을 맺는다.

2. 선행 연구

이전 연구에서 다룬 방언권은 서울/경기, 경상, 전라, 충청, 강원, 제주의 6개 행정 구역 단위 권역이었다(Kim & Kim, 2021; Lee et al., 2019; Lee at al., 2021; Lee et al., 2022). 하지만 행정 구역 단위 경계는 언어체계에 따른 방언 경계와 다르기 때문에(Lee, 2005) 이러한 구분 방식은 올바르지 않다. 또한, 방언 연구에서는 경상과 전라를 남도와 북도로 구분하여 살펴보기 때문에(경남 – Lee, 1997, 1998; 전남 – Lee, 1998; 전북 – Jang, 2019) 방언 식별 연구에서도 이러한 점을 반영하여야 한다.

음향 특징은 MFCC(Mel-Frequency Cepstral Coefficient)와 F0(기본 주파수; fundamental frequency)를 중심으로 기타 음향 특징을 추가하여 방언 자동 식별의 입력으로 사용하였다. Lee et al.(2019)과 Lee et al.(2021)에서는 MFCC와 F0를, Kim & Kim(2021)은 scaled MFCC를 추출하였다. 한편, Lee et al.(2022)에서는 음소 단위 분절을 통해 음소별 F0와 함께 음소의 길이(duration) 및 세기(intensity), 그리고 모음 포먼트(formant)를 활용하였다. Na & Lee(2023)에서는 MFCC에 더하여 발화 속도, 휴지 길이, 나이, 성별을 입력 특징으로 고려하였다.

학습 모델은 다양한 종류가 사용되었으며, 특히 Kim & Kim(2021)에서 8가지 학습 모델[SVM, RF, DNN, RNN, LSTM, Bi-LSTM, GRU(Gated Recurrent Unit), 1D-CNN]을 사용한 것이 돋보인다. 이밖에도 Lee et al.(2021)는 Bi-LSTM에 attention 층을 더한 구조를, Lee et al.(2022)는 거기에 DNN을 병렬로 학습하여 합치는 혼합(fusion) 형태의 모델을 학습하였다. 그리고 Na & Lee(2023)에서는 SVM, RF, LightGBM(Gradient Boosting Machine)을 사용하였다. 전반적으로 SVM 및 RF의 기계 학습 모델과 RNN 계열의 딥러닝 모델이 활용된 것을 알 수 있다.

방언 데이터는 최신 공개형 코퍼스인 AI Hub를 사용하는 추세이다. Lee et al.(2019)와 Lee et al.(2021)에서는 '한국인 표준 음성 DB'(Shin & Kim 2017; Shin et al., 2015)라는 화자 수가 약 2,500명 수준인 비공개 데이터를 사용하였다. 이후 Kim & Kim(2021)에서는 AI-Hub의 '한국인 대화음성'(AI Hub, 2021a), Lee et al.(2022)에서는 AI-Hub의 '자유대화 음성(일반남여)'(AI Hub, 2021b), Na & Lee(2023)에서는 AI-Hub의 '자유대화 음성(노인남여)'(AI Hub, 2021c)를 사용하여, 공개 코퍼스를 통한 객관적인 연구 비교가 가능하도록 기준을 마련하였다.

3. 제안 방법론

3.1. 한국어 방언권

Lee(2005)에서는 한반도의 방언 분포를 1차 구획부터 4차 구획까지 제시하였다. 1차 구획은 동서 혹은 남북으로 나누는 이분할을, 2차 구획은 중부, 동남(경남), 서남(전라), 제주의 4분류를 의미한다. 3차 구획은 2차 구획의 동남과 서남을 각각 경남과

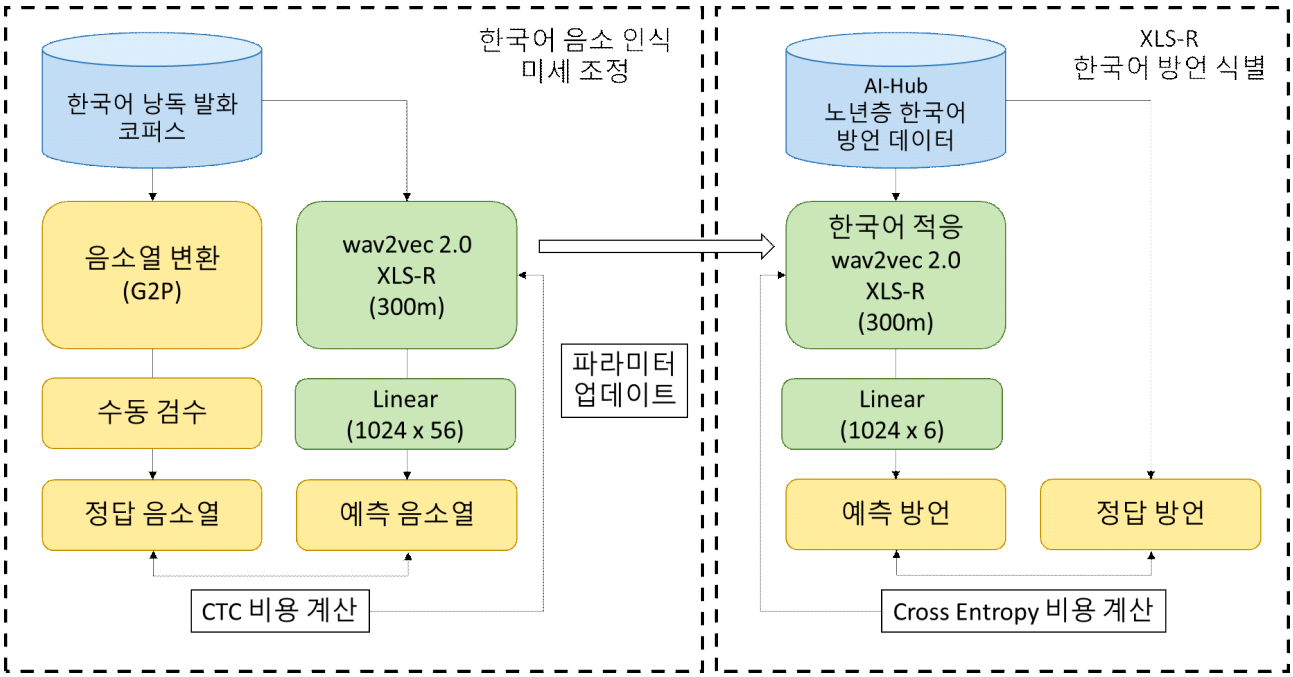

그림 1. 한국어 음소 인식 미세조정 XLS-R(cross-lingual speech representation) 모델의 방언 자동 식별
Figure 1. Automatic dialect identification of a cross-lingual speech representation (XLS-R) model fine-tuned with Korean phoneme recognition

경북, 전남과 전북으로 나눈 6분류 구획이고, 4차 구획은 시군 단위 구획으로 정의하였다.

본 연구에서는 3차 구획에 해당하는 6개의 중방언권(중부, 경남, 경북, 전남, 전북, 제주)을 방언 분석 대상으로 삼는다. 1차와 2차 구획은 각 방언권의 범위가 넓어 방언 식별에서의 유의미한 분석을 하기 어려운 한편, 4차 구획은 분석해야 할 방언의 수가 지나치게 많아지게 된다. 방언 식별 대상의 수가 6개라는 점에서는 선행 연구와 일치한다. 그러나 행정 구역의 기준에 맞춰 서울/경기, 경상, 전라, 충청, 강원, 제주로 분류한 선행 연구와 달리 국어 방언학의 관점에 따라 경상과 전라는 남북 방언권으로, 서울/경기, 충청 및 강원 영서 일부 지역은 중부 방언권으로 재구성한 점에서 차별성을 두고 있다.

3.2. 한국어 적응 wav2vec 2.0 XLS-R

wav2vec 2.0 XLS-R(cross-lingual speech representation)은 Meta AI에서 개발한 대규모 다국어 사전 훈련(pre-trained) 음성 모델로, 128개 언어로 구성된 약 436,000시간의 음성 데이터를 활용하여 자기지도학습 방식으로 학습되었다(Babu et al., 2022). XLS-R은 음성인식을 비롯한 음성 번역, 언어 식별 등 다양한 음성 처리 태스크에서 높은 수준의 성능을 달성하였으며, 본 연구에서는 한국어 방언 식별 영역에 적용하고자 한다.

XLS-R의 구조는 크게 CNN 특징 추출 블록과 트랜스포머 인코더 블록으로 이루어져 있다. CNN 특징 추출 블록은 여러 7개 층의 1D CNN으로 구성되어 있고, 원시 음성으로부터 1,024차원의 저차원 특징을 추출한다. 트랜스포머 인코더 블록은 multi-head self-attention과 feed-forward 네트워크로 구성되어 있으며, 중간층마다 layer norm과 residual connection을 적용한다.

XLS-R은 여러 언어의 다양한 음향 패턴을 이미 학습한 상태이나 여기서는 같은 음소 카테고리에 속하더라도 방언에 따라 음성적 실현이 다른, 한국어 방언의 미세한 음성적 차이를 구분하기 위해 음소 인식(phoneme recognition) 태스크로 한국어 음성에 적응시킨다. 음소 인식을 통해 한국어 음운에 익숙해진 식별 모델이 그렇지 않은 모델보다 방언 식별 성능이 더 좋을 것으로 기대한다. 음소 인식 태스크는 한국어 문장 낭독 발화의 음소열을 학습하는 방식으로 진행한다.

그림 1은 XLS-R 모델의 음소 인식 훈련 단계와 이후 한국어 음소를 학습한 XLS-R 모델로 한국어 방언 식별을 학습하는 단계를 도식화한 것이다. XLS-R은 우선 음소 인식에 맞게 56개 음소로 이루어진 선형 층을 디코더(decoder) 층으로 사용하고, 이후 방언 식별에서는 방언권의 수에 맞게 6차원의 선형 층을 디코더 층으로 사용한다. 또한, 음소 인식에서는 정답과 예측 음소열 사이에서 CTC(connectionist temporal classification) loss를 사용하는 한편, 방언 식별 태스크에서는 cross entropy loss를 통해 정답 방언권을 학습한다.

4. 실험 설정

4.1. 데이터셋

Lee(2007)에서는 방언 자료제공인을 선정할 때 'NORM'이라는 기준을 제시하였다. 'NORM'은 Non-mobile, Old, Rural, Male의 약자로 지역 이동이 적고 지방에 거주하는 노년층 남성을 의미한다. 이러한 기준을 반영하여 본 연구에서는 다음의 조건을 충족하는 코퍼스를 탐색하였다: 첫째, 지역 메타 정보가 경상과 전라 지역이 남도와 북도까지 구분되어 있어야 한다. 둘째, 화자의 성별은 남성이다. 셋째, 화자의 연령대는 50대 이상이다. 이러한 조건을 종합한 결과, AI-Hub의 '중 · 노년층 한국어 방언 데

표 1. 방언 식별 실험 데이터 분포
Table 1. Data distribution for the dialect identification experiments

방언	학습 데이터		테스트 데이터	
	화자 수(명)	샘플 수(개)	화자 수(명)	샘플 수(개)
중부	1,065	108,197	30	3,000
경남	320	32,494	30	3,000
경북	652	67,728	30	3,000
전남	670	63,721	30	3,000
전북	305	39,906	30	3,000
제주	114	12,911	30	3,000
전체	**3,126**	**324,957**	**180**	**18,000**

이터'(AI Hub, 2024a, 2024b)가 적합한 것으로 확인되었다. '중·노년층 한국어 방언 데이터'는 방언 화자의 시군 지역 정보까지 제공되어 경상과 전라를 남도와 북도로 나눌 수 있을 뿐만 아니라 강원 지역에서는 춘천과 원주 화자만 선별할 수 있었다.

표 1은 본 연구에서 사용하는 데이터 분포이다. 테스트 데이터는 방언별로 같은 수의 화자와 샘플이 되도록 조정하였는데, 원본에는 동일한 화자가 학습 데이터와 테스트 데이터에 모두 포함되어 있어 이를 피하기 위해 모든 데이터를 모은 후 일부를 테스트 데이터로 다시 구성하였다. 방언 간 학습 데이터의 양이 차이가 크지만, 주어진 원본 코퍼스를 있는 그대로 최대한 사용하기 위해 규모가 가장 작은 제주 방언에 맞추어 나머지 방언의 데이터를 줄이거나 규모가 가장 큰 중부 방언에 맞추어 데이터 증강을 하지 않았다. 이러한 한계는 4.3장에서 다루는 class weight를 통해 극복하고자 하였다. 본 데이터는 50대 이상 남성 화자의 발화만 사용하고 여성 화자는 제외되었다. 그리고 원주와 춘천을 제외한 나머지 강원 지역 역시 제외하였다.

4.2. 한국어 음소 인식 미세조정 및 결과

XLS-R의 음소 인식 태스크는 서울대학교 언어학과에서 보유한 한국어 낭독 발화 음성으로 진행한다. 해당 코퍼스는 표준어 문장 텍스트를 읽은 것이지만 다양한 지역 출신의 화자로 구성되어 있어 음소 단위에서는 음성적 실현이 표준 발음과 다를 수 있다. 따라서 방언 발음이 일부 포함된 준 표준어 낭독 발화라 할 수 있다. 정답 음소열은 음소열 변환을 통해 표준 음소열(canonical phoneme sequence)을 자동 생성한 후 직접 들으면서 실제 발음이 다른 음소는 수동으로 수정하는 과정을 거쳐 마련하였다. 예를 들어, '얼음'이라는 단어는 'EO R EU M'으로 음소열이 자동 생성되지만 실제 발음이 [으름]이면 'EU R EU M'으로 고친 것이다.

본 코퍼스는 60,000개 발화로 이루어진 120시간 규모의 음성 데이터이다. 이 중 90%에 해당하는 54,000개 발화(108시간)를 음소 인식 훈련에, 나머지 10%인 6,000개 발화(12시간)는 모델 평가에 사용한다. 음소열 변환은 Montreal Forced Aligner(MFA; McAuliffe et al., 2017) 툴킷을 사용한다. MFA 툴킷은 한국어 음향 모델을 지원하기 때문에 한국어 음소열 생성을 바로 수행할 수 있다. 본 연구에서 다루는 음소열의 종류 수는 MFA에서 지원하는 56개이다(그림 1 참고).

음소 인식 훈련 결과, 음소오류율(phone error rate)은 3.88%였다. 이는 낮은 오류율로 XLS-R이 한국어 음소를 잘 학습하여 한국어에 적응하였음을 확인할 수 있다.

4.3. 방언 식별 학습 설정

XLS-R과의 성능 비교를 위한 베이스라인으로 Kim & Kim(2021)에서 다룬 DNN, LSTM, Bi-LSTM을 사용하며 설정값 또한 동일한 값을 따른다. DNN은 5개의 은닉층으로 구성되며, 각 층의 뉴런 수는 순서대로 256, 256, 128, 128, 64개로 한다. LSTM과 Bi-LSTM에서는 128개 차원 수를 가진 RNN 층과 128, 64, 32개 차원 수로 이루어진 선형 층으로 구성한다. Kim & Kim(2021)의 모델에 더하여 attention 층을 추가한 Bi-LSTM의 성능도 함께 확인한다.

음향 특징은 두 가지를 사용한다. 첫째는 Kim & Kim(2021)에서 제안한 scaled MFCC이다. DNN의 입력 형태인 특징 행렬(feature matrix)의 경우 13차원의 MFCC 계수의 평균을 사용하고, RNN 계열의 입력인 시계열 데이터(time-series data)는 첫 5초의 음성에 대한 MFCC를 사용한다. 이후 scaled MFCC는 MinMax 스케일링을 통해 얻는다. 두 번째는 eGeMAPS 특징 세트이다. eGeMAPS는 extended Geneva Minimalistic Acoustic Parameter Set로 연구물 간 객관적인 비교 평가를 위해 제안된 음향 특징 세트이다(Eyben et al., 2016). eGeMAPS는 데이터의 형태에 따라 LLDs, low-level descriptors와 functionals로 나뉜다. LLDs는 순환 신경망의 입력 형태인 프레임 단위 시퀀스 데이터로서 25개의 음향 특징으로 구성되어 있다. functionals는 LLDs의 특징에 대한 통계치로 88개의 특징 행렬 형태로 이루어져 있다. functionals는 DNN의 입력으로 사용한다.

XLS-R 모델 학습에서는 첫 5초 구간의 음성을 입력으로 사용한다. 5초의 음성이면 한 문장 이상의 발화가 포함되며 대용량 샘플을 동시에 학습하는 배치 학습을 효과적으로 수행할 수 있다. 그리고 선형 출력층에서 나온 6차원의 출력값 중 가장 높은 값의 뉴런에 해당하는 방언이 예측 방언이 된다. XLS-R 사전 훈련 모델은 HuggingFace에서 파라미터 수가 3억 개인 모델(wav2vec2-xls-r-300m)을 다운로드하여 사용한다. 또한 XLS-R 모델의 입력인 원시 음성(raw audio)의 형식은 모노 채널, 16-bit precision, 16 KHz 샘플링레이트이다.

방언 식별은 분류 과제(classification task)에 해당하기 때문에 cross entropy loss로 정답 방언과 예측 방언 간 비용을 계산한다. 그리고 과적합(overfitting) 방지를 위한 dropout 값은 0.3으로, 최적화를 위한 학습률은 1×10^{-4}로 설정한다. 훈련 반복 수(epoch)는 최대 100으로 하되 patience를 5로 설정한 early stopping을 추가하여 F1 점수가 5회 동안 개선되지 않으면 훈련을 중단하도록 한다.

또한 방언 간 샘플 수 차이에 의한 과적합을 극복하기 위해 class weight를 적용하여 비용을 계산한다. 이는 학습 샘플 수가 적은 제주 방언에 대해서는 높은 가중치를 부여한다는 것을 의미한다. 모델 훈련 과정에 class weight를 적용하기 위해 Python

라이브러리인 scikit-learn(Pedregosa et al., 2011)의 'compute_class_weight' 클래스를 이용한다. 이는 다음의 수식 (1)로 방언별 가중치를 계산한다:

$$w_j = \frac{n_{samples}}{n_{classes} \times n_j} \quad (1)$$

w_j는 j번째 방언에 대한 가중치, $n_{samples}$는 전체 샘플 수(여기서는 324,957), $n_{classes}$는 방언의 수(여기서는 6), n_j는 j번째 방언의 샘플 수를 뜻한다. 위 식에 따라 방언별 가중치는 중부, 경남, 경북, 전남, 전북, 제주 순서대로 0.501, 1.667, 0.800, 0.850, 1.357, 4.195로 나타난다.

한편, 본 연구에서는 자세하게 다루지 않으나 class weight를 적용하지 않고 모델 훈련을 한 경우 훈련 방향이 샘플 수가 많은 중부 방언으로 편향되는 과적합이 발생하여 한국어 미세조정 XLS-R 기준 F1 점수가 0.45-0.55 수준으로 나타났다. 이는 표 2에서의 결과와 비교할 때 class weight를 적용하는 것은 필수임을 알 수 있다.

4.4. 평가 방법

방언 자동 식별은 전체 음성 샘플 중 정확하게 방언을 예측한 샘플의 비율로 평가한다. 따라서 평가 지표로 precision(정밀도), recall(재현율), 그리고 F1 점수를 사용한다. precision은 모델이 예측한 방언 중 실제 정답의 비율을 뜻하고, recall은 정답 방언 중 모델이 정확하게 예측한 수의 비율을 의미한다. 그리고 F1 점수는 precision과 recall의 조화평균으로 산출한다. F1 점수의 산출식은 다음과 같다.

$$F1 = \frac{\mathrm{Pr}ecision \times Recall}{\mathrm{Pr}ecision + Recall} \quad (2)$$

5. 실험 결과

표 2는 한국어 방언 식별 성능을 나타낸 것이다. MFCC와 eGeMAPS를 비교한 결과, 모든 모델에 대해 eGeMAPS의 성능이 더 우수한 것으로 나타났다(DNN F1: 0.438<0.626; Bi-LSTM F1: 0.639<0.738; Attention Bi-LSTM F1: 0.647<0.728).

모델 간 비교에서는 DNN보다 LSTM 계열이 더 높은 식별률을 보였다. MFCC를 기준으로는 DNN(0.438)<LSTM(0.565)<Bi-LSTM(0.639) <Attention Bi-LSTM(0.647)의 순으로, eGeMAPS를 기준으로는 DNN(0.626) LSTM(0.715)<Attention Bi-LSTM (0.728)<Bi-LSTM(0.738) 순으로 F1 점수가 관찰되었다. 한편, attention 기반 Bi-LSTM과 기본 Bi-LSTM 사이에서는 MFCC의 경우 attention 기반 Bi-LSTM의 성능이 더 우수하지만 eGeMAPS의 경우 기본 Bi-LSTM의 성능이 소폭 더 높게 나타났다.

XLS-R과의 비교에서는 XLS-R 모델이 Bi-LSTM에 비해 F1 점수가 14%p나 더 높은 식별률을 보여 86.8%를 달성하였다. 두 XLS-R 모델 사이에서는 미세조정(fine-tuned) 모델이 사전 훈련(pre-trained) 모델보다 조금 더 우수한 성능을 보이나 그 차이가 거의 없었다(F1 기준 0.003 차이).

6. 논의

6.1. 방언별 식별 성능 개선

그림 2는 scaled MFCC 및 eGeMAPS의 음향 특징과 학습 모델의 조합 중 가장 높은 식별률을 보였던 eGeMAPS Bi-LSTM 모델과 두 XLS-R 모델의 방언별 식별률을 혼동행렬로 나타낸 것이다. 혼동행렬의 행과 열은 각각 정답 방언과 모델의 예측 방언을 나타내며, 셀 안의 숫자는 정답 개수(각 행의 합) 중 해당 방언으로 예측한 개수의 비율(재현율; recall)을 의미한다. 또한, 괄호 속 숫자에는 예측한 음성 샘플 수를 표시한 것이다.

eGeMAPS Bi-LSTM과 미세조정 XLS-R 모델 사이에서 성능 개선을 보인 방언은 중부 방언(0.70→0.91), 경남 방언(0.69→0.74), 경북 방언(0.66→0.88), 전남 방언(0.67→0.89), 그리고 제주 방언(0.85→0.93)이다. 전북 방언은 두 모델에서 모두 0.86의 재현율을 보였으며, 올바르게 맞힌 샘플 수로는 오히려 eGeMAPS 기반 Bi-LSTM에서 더 많았다(2,582 vs 2,569).

특히, 중부, 경북, 전남 방언에서 20%p 이상의 F1 점수 향상을 보였다. 중부 방언에서는 경북, 전남, 제주로 혼동되었던 점이 개선되었고, 경북 방언에서는 경남, 전남과의 혼동이 줄었으며, 전남 방언에서는 중부, 경남, 전북과의 혼동률이 낮아졌다.

이처럼 XLS-R 모델이 더 우수한 성능을 보인 데는 여러 요인을 생각할 수 있다. 첫째, XLS-R 모델의 사전 훈련에 사용된 데이터의 규모 및 다양성이다. 앞서 언급하였듯이 XLS-R의 사전 훈련 데이터는 한국어를 포함한 128개 언어로 구성된 약 436,000시간의 음성이다. 사전 훈련을 통해 다양한 말소리 패턴에 익숙한 상태에서 한국어 방언 식별하는 것이 그렇지 않은 상태보다 더 유리했을 것이다. 둘째, 학습 모델 구조의 차이다. Bi-LSTM 모델보다 XLS-R 모델의 구조가 더 복잡하며, 이러한 복잡한 구조로 인해 한국어 방언 간 미세한 차이를 더 잘 구분하여 학습한 것일 수도 있다. 셋째, 입력 특징의 차이다. Bi-LSTM이 학습한 음향 특징은 88개의 음향 특징인데 반해 XLS-R의 CNN 특징 추출 블록은 1,024차원의 임베딩 특징을 추출하여 사용한다. 입력 음성과 목표 방언 간의 연결성을 신경망으로 나타내었을 뿐만 아니라 이를 표현하는 차원 수도 훨씬 더 풍부하기 때문에 방언 식별에 더 적합하였을 것이다.

6.2. 방언 간 혼동 양상

XLS-R 모델이 eGeMAPS Bi-LSTM 모델보다 대부분의 방언에서 더 높은 식별률을 보이나 방언 간 혼동은 여전히 존재한다. 첫째, 모든 방언이 중부 방언으로 혼동하는 비율이 다른 방언에 비해 높다. 이는 중부 방언으로 정의된 넓은 지역적 범위가 요인일 것이다. 중부 방언권은 서울/경기, 충청, 그리고 강원 영서를 아우르는 광범위한 지역 방언이다(Lee, 2005). 범위가 넓은 만큼 다른 방언의 특징과도 겹칠 것이고, 이는 방언 간 혼동을 일으켰을 것이다. 둘째, 중부 방언과 전남 방언이 서로 혼동하

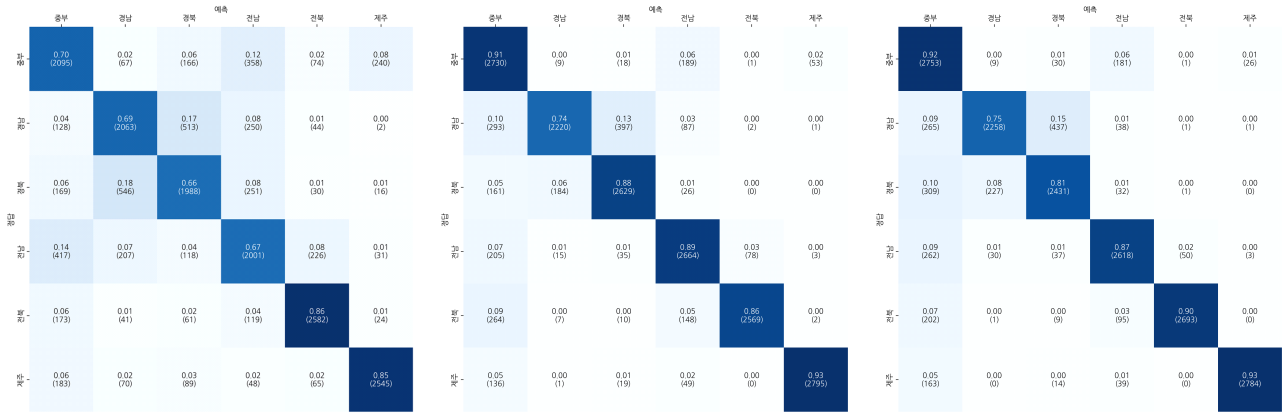

그림 2. 방언 식별 모델 혼동행렬(좌: eGeMAPS Bi-LSTM, 가운데: 미세조정 XLS-R, 우: 사전 훈련 XLS-R)
Figure 2. Confusion matrices of dialect identification (Left: eGeMAPS Bi-LSTM, Middle: Fine-tuned XLS-R, Right: Pre-trained XLS-R)

는 비율이 높다. 두 방언권이 비슷한 양상을 보이는 경우는 억양에서다. Lee(1998)에서는 중부, 경남, 전남 방언의 억양을 비교하였다. 여기서 경남 방언과 대립하는 측면에서 중부 방언과 전남 방언 간 비슷한 점이 많다고 언급한다. XLS-R 모델이 구체적으로 방언 음성의 어떠한 특성에 주목하였는지 현재로서는 알기 어려우나 1D CNN 특징 추출 블록으로 만들어지는 저차원 특징이 음성의 시간 축도 담고 있기 때문에 억양 패턴도 함께 학습에 반영될 수 있다고 예상한다.

6.3. 한국어 음소 인식 미세조정의 효과

표 2에서 한국어 음소 인식으로 미세조정한 XLS-R 모델이 그렇지 않은 모델보다 미미한 차이로 성능이 높았다. 사실상 두 모델의 성능 차이가 없는 것이라 할 수 있는데, 그 이유는 사전 훈련 모델에 한국어가 반영되어 있기 때문일 뿐만 아니라 다양한 언어와 대규모 사전 훈련 데이터로 이미 풍부한 음성 패턴을 경험하였기 때문이다. 한국어 음소에 대한 추가 학습이 이루어졌더라도 새로운 패턴을 학습한 것은 아니라는 것이다.

또한, 음소 인식 미세조정으로 한국어 음소 음향에 익숙해진 XLS-R 인코더가 방언 식별에도 영향을 미칠 것이라는 기대와 달리, 선형 층의 디코더가 바뀌면서 학습 효과가 사라진 것 또한 요인일 수 있다. 음소 인식 태스크의 기대 효과가 미미한 것을 고려하여 contrastive loss로 한국어 음성을 추가로 학습하는 방식으로 한국어 적응을 수행해야 할 것으로 판단된다.

표 2. 한국어 방언 자동 식별 결과
Table 2. Korean Dialect Identification Performance

음향 특징	모델	precision	recall	F1
scaled mean MFCC	DNN	0.440	0.440	0.438
frame-level scaled MFCC	LSTM	0.595	0.559	0.565
	Bi-LSTM	0.651	0.638	0.639
	Attention Bi-LSTM	0.658	0.644	0.647
eGeMAPS func-tionals	DNN	0.632	0.626	0.626
eGeMAPS LLDs	LSTM	0.717	0.715	0.715
	Bi-LSTM	0.739	0.737	0.738
	Attention Bi-LSTM	0.734	0.727	0.728
16KHz mono raw audio	XLS-R (pre-trained)	0.876	0.863	0.865
	XLS-R (fine-tuned)	**0.878**	**0.867**	**0.868**

XLS-R, cross-lingual speech representation.

7. 결론

본 연구에서는 한국어 방언 식별 문제에 트랜스포머 기반 종단간 음성 모델인 wav2vec 2.0 XLS-R을 새롭게 적용해 봄으로써 LSTM 중심의 기존 방법론과의 성능 비교를 통해 방언별로 개선된 식별률을 살펴보았다. 그리고 행정 구역 단위 방언권 대신 Lee(2005)에 따라 방언권을 새로 구성하여 방언 간 혼동 패턴을 설명하는 데 방언학을 적용해 보려는 가능성을 제시하였다.

본 연구에서는 XLS-R 모델이 식별 성능 면에서 우수하다는 점을 보였으나 모델이 방언 식별 훈련 과정에서 무엇을 학습하며 그것을 어떻게 해석해야 하는지에 대한 설명은 부재했다. 향후 연구에서는 XLS-R 모델 내 multi-head self-attention 층에서의 attention weight 분포를 살피어 음운 측면에서 이해할 수 있는지 살펴본다. 그리고 Whisper(Radford et al., 2023), HuBERT(Hsu et al., 2021), WavLM(Chen et al., 2022) 등과 같은 다른 형태의 트랜스포머 기반 음성 모델과의 성능 비교도 추가로 진행할 예정이다. 한편, NORM의 기준에 따른 화자 집단을 대상으로 방언 식별 실험을 수행한 본 연구를 확장하여 다른 집단에서의 식별 성능과 비교하여 화자 집단의 특성이 방언 식별에 미치는 영향에 대해 살펴볼 수 있다.

References

AI Hub. (2021a). Korean conversational speech (Version 1.2) [Data set]. Retrieved from https://aihub.or.kr/aihubdata/data/view.do?&dataSetSn=130

AI Hub. (2021b). Spontaneous conversational speech (General) (Version 1.2) [Data set]. Retrieved from https://aihub.or.kr/aihubdata/data/view.do?&dataSetSn=109

AI Hub. (2021c). Spontaneous conversational speech (Senior) (Version 1.2) [Data set]. https://aihub.or.kr/aihubdata/data/view.do?&dataSetSn=107

AI Hub. (2024a). Middle-aged/senior dialect data (Ganwon-do, Gyeongsang-do) (Version 1.2) [Data set]. https://aihub.or.kr/aihubdata/data/view.do?&dataSetSn=71517

AI Hub. (2024b). Middle-aged/senior dialect data (Chungcheong-do, Jeolla-do, Jeju-do) (Version 1.2) [Data set]. https://aihub.or.kr/aihubdata/data/view.do?&dataSetSn=71558

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., ... Auli, M. (2022, September). XLS-R: Self-supervised cross-lingual speech representation learning at scale. *Proceedings of Interspeech 2022* (pp. 2278-2282). Incheon, Korea.

Chambers, J. K., & Trudgill, P. (1998). *Dialectology*. Cambridge, UK: Cambridge University Press.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., ... Wei, F. (2022). WavLm: Large-Scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1505-1518.

Choi, M. (2005). The System of Korean Dialectology. *The Journal of Korean Dialectology*, *1*, 35-72.

Dobbriner, J., & Jokisch, O. (2019, August). Towards a dialect classification in German speech samples. *Proceedings of the 21st International Conference on Speech and Computer* (pp. 64-74). Istanbul, Turkey.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., ... Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2). 190-202.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451-3460.

Imaizumi, R., Masumura, R., Shiota, S., & Kiya, H. (2022). End-to-end Japanese multi-dialect speech recognition and dialect identification with multi-task learning. *APSIPA Transactions on Signal and Information Processing*, *11*(1), e4.

Jang, S. (2019). A study on the orthography of the Jeollabuk-do Dialect Dictionary. *Korean Language and Literature*, *71*, 97-120.

Kakouros, S., & Hiovain-Asikainen, K. (2023, August). North sami dialect identification with self-supervised speech models. *Proceedings of Interspeech 2023* (pp. 5306-5310). Dublin, Ireland.

Kim, Y. K., & Kim, M. H. (2021). Performance comparison of Korean dialect classification models based on acoustic features. *Journal of The Korea Society of Computer and Information*, *26*(10), 37-43.

Lee, B. (1997). A study on the intonation of the Gyeongnam dialect. *The Woorimal Journal of Korean Language*, *7*, 79-103.

Lee, B. (1998). A comparative study of intonation in the central, Gyeongnam, and Jeonnam dialects. *The Woorimal Journal of Korean Language*, *8*, 1-62.

Lee, J., Kim, K., & Chung, M. (2021, November). Korean dialect identification based on intonation modeling. *Proceedings of the 24th Conference of the Oriental COCOSDA* (pp. 168-173). Singapore, Singapore.

Lee, J., Kim, K., & Chung, M. (2022, November). Korean dialect identification based on an ensemble of prosodic and segmental feature learning for forensic speaker profiling. *Proceedings of the 25th Conference of the Oriental COCOSDA* (pp. 1-6). Hanoi, Viet Nam.

Lee, J., Kim, K., Lee, K., & Chung, M. (2019, October). Gender, age, and dialect identification for speaker profiling. *Proceedings of the 22nd Conference of the Oriental COCOSDA*. Cebu, Philippines.

Lee, K. (2005). On the dialectal variation and the division of dialect areas. *The Journal of Korean Dialectology*, *1*, 103-123.

Lee, S. K. (2007). *Korean dialectology*. Seoul, Korea: Hakyeonsa.

Liu, G., Lei, Y., & Hansen, J. H. L. (2010, August). Dialect identification: Impact of differences between read versus spontaneous speech. *Proceedings of the 18th European Signal Processing Conference* (pp. 2003-2006). Aalborg, Denmark.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017, August). Montreal forced aligner: Trainable text-speech alignment using kaldi. *Proceedings of Interspeech 2017* (pp. 498-502). Stockholm, Sweden.

Na, J., & Lee, B. (2023). Dialect classification based on the speed and the pause of speech utterances. *Phonetics and Speech Sciences*, *15*(2), 43-51.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., ... Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

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

Schilling, N., & Marsters, A. (2015). Unmasking identity: Speaker profiling for forensic linguistic purposes. *Annual Review of Applied Linguistics*, *35*, 195-214.

Shin, J., Jang, H., Kang, Y., & Kim, K. W. (2015). Developing a Korean standard speech DB. *Phonetics and Speech Sciences*, *7*(1), 139-150.

Shin, J., & Kim, K. W. (2017). Developing a Korean standard speech DB (II). *Phonetics and Speech Sciences*, *9*(2), 9-22.

Shon, S., Ali, A., & Glass, J. (2018, June). Convolutional neural networks and language embeddings for end-to-end dialect recognition. *Proceedings of Odyssey 2018 The Speaker and Language Recognition Workshop* (pp. 98-104). Les Sables d'Olonne, France.

Wang, D., Ye, S., Hu, X., Li, S., & Xu, X. (2021, August). An end-to-end dialect identification system with transfer learning from a multilingual automatic speech recognition model. *Proceedings of Interspeech 2021* (pp. 3266-3270). Brno, Czechia.

• **이주영 (Jooyoung Lee)**
서울대학교 언어학과 박사과정
서울시 관악구 관악로 1
Tel: 02-880-9039
Email: excalibur12@snu.ac.kr
관심분야: 한국어 방언학, 방언 식별, 음성인식

• **김선희 (Sunhee Kim)**
서울대학교 불어교육과 교수
서울시 관악구 관악로 1
Tel: 02-880-7693
Email: sunhkim@snu.ac.kr
관심분야: 불어 음성학, 음성 언어 처리

• **정민화 (Minhwa Chung)** 교신저자
서울대학교 언어학과 교수
서울시 관악구 관악로 1
Tel: 02-880-9195
Email: mchung@snu.ac.kr
관심분야: 음성인식, 외국어 발음 교육, 음성 바이오마커

wav2vec 2.0 XLS-R을 활용한 한국어 방언 자동 식별*

이 주 영[1] · 김 선 희[2] · 정 민 화[1]

[1]서울대학교 언어학과, [2]서울대학교 불어교육과

국문초록

기존의 한국어 자동 방언 식별 시스템은 약 60% 수준의 낮은 정확도를 보였으며, 주로 전통적인 기계 학습 모델에 의존해 왔다. 또한, 행정 구역에 따라 방언을 분류하는 방법론적 문제는 기존의 언어학적 경계와 일치하지 않는 한계가 있었다. 본 연구는 이러한 한계를 극복하기 위해 최신 트랜스포머 기반의 종단간 음성 모델인 wav2vec 2.0 XLS-R(cross-lingual speech representation)을 적용하고자 한다. 128개 언어의 약 436,000시간에 달하는 음성 데이터로 사전 훈련된 이 모델을 본 과제에 맞게 미세조정하였다. 또한, 방언을 중부, 경남, 경북, 전남, 전북, 제주의 6개 주요 언어권으로 재분류하였다. AI Hub '중 · 노년층 한국어 방언 데이터'를 사용하여 XLS-R 모델을 Bi-LSTM과 같은 베이스라인 모델과 비교하였다. 그 결과, XLS-R 모델이 F1 점수 86.8%를 달성하며 가장 성능이 좋은 베이스라인보다 14%p 높은 성능 향상을 보였다. 특정 방언 간의 혼동은 여전히 존재하지만, 본 연구는 대규모 사전 학습 모델을 한국어 방언 식별이라는 정교한 과제에 적용하는 것의 효과를 입증하고, 언어학적 정보에 기반한 범주 사용의 중요성을 강조한다. 연구 결과는 음성인식 및 법과학 분야의 방언 식별 기술 발전에 기여할 것이다.

핵심어: 방언 자동 식별, 한국어 방언, wav2vec 2.0, XLS-R

참고문헌

신지영, 김경화(2017). 한국인 표준 음성 DB 구축(II). *말소리와 음성과학*, *9*(2), 9-22.

신지영, 장혜진, 강연민, 김경화(2015). 한국인 표준 음성 DB 구축. *말소리와 음성과학*, 7(1), 139-150.

이기갑(2005). 방언 분화와 방언 구획. *한국방언학회*, *1*, 103-123.

이병운(1997). 경남 방언의 억양 연구. *우리말연구*, *7*, 79-103.

이병운(1998). 중부방언, 경남방언, 전남방언의 억양에 대한 비교 연구. *우리말연구*, *8*, 1-62.

이상규(2007). 한국어 *방언학*. 서울: 학연사.

최명옥(2018). 국어방언학의 체계. *한국방언학회*, *1*, 35-72.

* 본 연구는 서울대학교 언어학과의 'LINGUISTICS SNU 10-10 INITIATIVE' 사업의 지원을 받아 수행되었습니다.