pISSN 2005-8063
eISSN 2586-5854
2023. 6. 30.
Vol.15 No.2
pp. 43-51

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2023.15.2.043

Dialect classification based on the speed and the pause of speech utterances*

Jonghwan Na · Bowon Lee**

Department of Electrical and Computer Engineering, Inha University, Incheon, Korea

Abstract

In this paper, we propose an approach for dialect classification based on the speed and pause of speech utterances as well as the age and gender of the speakers. Dialect classification is one of the important techniques for speech analysis. For example, an accurate dialect classification model can potentially improve the performance of speaker or speech recognition. According to previous studies, research based on deep learning using Mel-Frequency Cepstral Coefficients (MFCC) features has been the dominant approach. We focus on the acoustic differences between regions and conduct dialect classification based on the extracted features derived from the differences. In this paper, we propose an approach of extracting underexplored additional features, namely the speed and the pauses of speech utterances along with the metadata including the age and the gender of the speakers. Experimental results show that our proposed approach results in higher accuracy, especially with the speech rate feature, compared to the method only using the MFCC features. The accuracy improved from 91.02% to 97.02% compared to the previous method that only used MFCC features, by incorporating all the proposed features in this paper.

Keywords: dialect classification, feature extraction, low resource conditions

1. 서론

방언 분류는 음성 인식(Automatic Speech Recognition, ASR)이나 화자 인식/확인과 같은 음성 분석 기술에 적용될 수 있다. 최근 팬데믹으로 인해 비대면 애플리케이션에 대한 수요가 급증하면서 음성 분석 기술이 보다 널리 사용되고 있다. 더욱이, ASR의 상용화 단계에서는 사용자가 표준 언어를 사용하는 것이 보장되지 않기 때문에 방언이 큰 어려움이 될 수 있다. 최근의 연구에 따르면, 입력 데이터에 우편번호 정보를 추가하는 것이 ASR 성능을 향상시킬 수 있다는 결과가 발표되었다(Dheram et al., 2022). 이는 지역 기반의 코호트 정보가 ASR 성능에 영향을 미칠 수 있음을 의미하며, ASR의 성능 향상에 방언 정보가

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018 S1A5A2A03037308) and by the Institute of Information & Communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2022-00155915, Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)].

** bowon.lee@inha.ac.kr, Corresponding author

Received 25 May 2023; Revised 14 June 2023; Accepted 14 June 2023

영향을 미친다고 볼 수 있다.

이전 연구에 따르면, Mel-Frequency Cepstral Coefficients(MFCC) 특성을 사용한 딥러닝 기반의 방법이 주류를 이루었다(Chowdhury et al., 2020; Khurana et al., 2017; Mukherjee et al., 2020; Tawaqal & Suyanto, 2021; Wan et al., 2022; Wang et al., 2021; Zhang & Hansen, 2018). 최근에는 MFCC 특성에 운율학적(Bhattacharjee & Sarmah, 2013; Lee et al., 2021, 2022; Mehrabani & Hansen, 2015), 음성학적(Chittaragi & Koolagudi, 2019; Najafian et al., 2018) 혹은 어휘적(Michon et al., 2018; Ying et al., 2020) 특성을 결합하여 방언 분류 모델을 훈련시키는 연구가 많이 소개되고 있다. 최근의 연구에서는(Kim & Kim, 2021), 한국 방언 분류에서 MFCC 특성만을 사용한 랜덤포레스트(Random Forest, RF) 분류기의 정확도가 63.93%를 기록했다. 본 논문에서는 전통적인 MFCC 특성에 더하여 음성 발화의 속도와 휴지 길이를 추출함으로써 방언 분류 성능을 향상시키는 방법론을 제안한다.

본 논문에서 제안하는 접근법은 기존 음향 특성에 추가 정보를 포함시켜 성능을 향상시키는 것이다. 방언 데이터셋의 크기가 제한적이므로, 딥러닝 모델을 적용하는 대신에 서포트벡터머신(Support Vector Machine, SVM), 랜덤포레스트 그리고 라이트지비엠(Light Gradient Boosting Machine, LightGBM)과 같은 전통적인 분류 알고리즘을 사용하여 이러한 음향적 특성 사용의 효과성을 검증했다. 본 논문에서 제안하는 특성은 발화 속도와 휴지 구간 길이로 해당 특징을 사용하는 경우 전통적인 MFCC 특성을 사용하여 얻은 것들에 비해 우수한 성능을 보이는 것을 실험을 통해 확인했다.

2. 선행 연구

2.1. 기계 학습 모델

음성을 사용하여 성별, 나이 또는 지역 분류를 수행하는 전형적인 방법은 머신러닝 기반 분류기를 사용하는 것이다. 최근 연구에 따르면, 다양한 머신러닝 및 딥러닝 모델이 방언 분류에 사용되었으며, 그 결과 한국어 방언의 경우 랜덤포레스트가 가장 우수한 성능을 보였다(Kim & Kim, 2021). 본 논문에서 사용된 데이터셋의 크기가 딥러닝 모델을 학습하기에 충분하지 않았기 때문에, 본 연구에서는 세 가지 머신러닝 모델 서포트벡터머신, 랜덤포레스트 그리고 라이트지비엠을 고려하였다.

서포트벡터머신(Hearst et al., 1998)은 데이터 분석에 사용되는 지도학습 알고리즘이다. 이 모델은 입력 데이터를 더 높은 차원의 공간으로 매핑한 다음, 훈련 데이터셋에 기반한 비확률적 이진 선형 분류를 사용하여 테스트 데이터가 어느 카테고리로 분류될지 결정하는 방식을 사용한다.

랜덤포레스트(Breiman, 2001)는 전체 특성 중에서 무작위로 특성을 선택하여 단일 결정 트리를 만드는 앙상블 학습 방법이다. 다양한 결정 트리가 학습될 때까지 여러 번 반복되며, 이러한 트리들의 예측 결과를 조합하여 최종 예측을 수행한다. 이런 종류의 알고리즘은 다양성을 보장하며 과적합을 방지하고 일반화 성능을 향상시키는 장점을 가진다.

라이트지비엠(Ke et al., 2017)은 트리 기반 학습 알고리즘을 사용하여 강력한 머신러닝 모델을 구축하는 그래디언트 부스팅 모델이다. 이는 전통적인 단계별 트리 성장 전략 대신에 잎사귀 중심(leaf-wise) 트리 성장 전략을 사용한다. 라이트지비엠의 핵심 아이디어는 그래디언트 기반 일방 표본 추출(gradient-based one-side sampling, GOSS)과 배타적 특성 꾸러미화(feature bundling)이라는 새로운 기법을 채택함으로써 훈련 속도를 더욱 향상시키고 메모리 사용량을 줄일 수 있다.

2.2. Mel-Frequency Cepstral Coefficients(MFCC)

MFCC(Davis & Mermelstein, 1980)는 저주파 대역에서의 오디오 신호를 세밀하게 표현하고 고주파 대역에서는 상대적으로 소략하게 표현하는 특성을 가진다. MFCC는 방언 분류(Chowdhury et al., 2020; Khurana et al., 2017; Mukherjee et al., 2020; Tawaqal & Suyanto, 2021; Wan et al., 2022; Wang et al., 2021; Zhang & Hansen, 2018)뿐만 아니라 음성인식(Shahnawazuddin et al., 2016; Tüske et al., 2014; Wallington et al., 2021), 화자 인식(Fenu et al., 2020; Garcia-Romero et al., 2019; Lin & Mak, 2020; Pappagari et al., 2020) 그리고 감정 인식(Keesing et al., 2021; Likitha et al., 2017; Sarma et al., 2018; Saste & Jagdale, 2017; Seo & Lee, 2022)과 같은 음성 분석의 다양한 분야에서도 활용된다.

2.3. 음성 구간 탐지(Voice Activity Detection, VAD)

Voice Activity Detection(VAD; Sohn et al., 1999)은 음성 발화에서 활성 음성 세그먼트를 결정하는 데 유용한 도구로, 음성에서의 휴지 구간을 VAD 결과에서 추출할 수 있다. 본 실험에 따르면, VAD에 의해 결정된 활성 음성 세그먼트에서만 MFCC 특성을 추출하는 것이 VAD를 사용하지 않는 것보다 더 높은 정확도를 보였다. 따라서 본 논문에서는 VAD에 의해 결정된 음성 세그먼트에 대한 MFCC 특성만을 추출하여 사용하였다. 우리는 실시간, 온라인 그리고 비지도학습을 사용한 세그먼트 기반의 VAD 알고리즘인 rVAD(Tan et al., 2020)를 사용하였다. rVAD는 소음에 강인한 VAD 방법으로 오디오 프레임이 음성을 포함하고 있는지에 대한 이진 결정을 출력한다. 본 실험에 rVAD를 사용한 이유는 실험 진행 당시 SOTA에 가까운 성능을 보이는 비지도 학습 방법의 알고리즘이었기에 사용하였다.

3. 제안하는 접근 방법

그림 1과 그림 2는 각각 경상도(Gyeongsang-do)와 충청도(Chungcheong-do) 방언을 사용하는 발화자가 말하는 동일한 예문인 “오늘 달이 참 예뻐”를 발화한 오디오 파일의 멜 스펙트로그램을 나타낸다. 본 논문에서 사용된 그림에서 경상도와 충청도의 표기는 편의를 위해 각각 gs와 cc로 사용하였다. 그림 1과 그림 2의 노란색으로 표시된 영역은 휴지 구간을 표시한 영역이며, 하늘색으로 그려진 선은 기본 주파수(fundametal frequency, F0)를 나타낸 것이다. 그림 1과 그림 2에서 볼 수 있듯이, 경상도 방언은 충청도 방언에 비해 상대적으로 짧은 휴지 구간과 더 빠

른 발화 속도를 가지고 있는 것을 알 수 있다. 또한, 두 예시의 F0값을 살펴보면, 경상도 방언이 충청도 방언보다 더 높은 F0 변동성을 가지고 있음을 알 수 있다. 따라서 본 실험에서는 상대적인 휴지 길이, 발화 속도, 그리고 F0 분산을 특징으로 추출하였다. 앞서 언급했듯이, 이전 연구에서 널리 사용되는 MFCC 특성과의 성능을 비교하기 위해, 우리는 추가적으로 13차 MFCC 특성을 추출했다. 각각의 특성을 추출하는 방법은 다음과 같다.

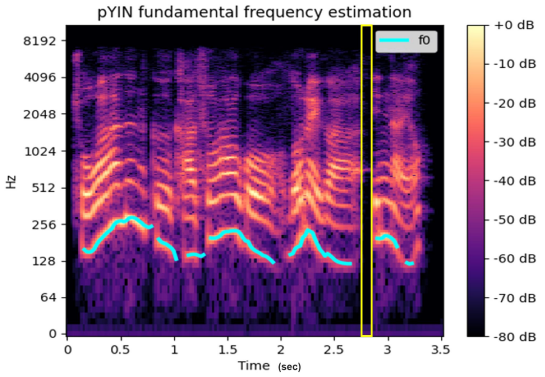


그림 1. 경상도 방언의 멜 스펙트로그램 (“오늘 달이 참 예뻐”)
Figure 1. Mel-spectrogram of Gyeongsang-do dialect (“Today’s moon is so beautiful”)

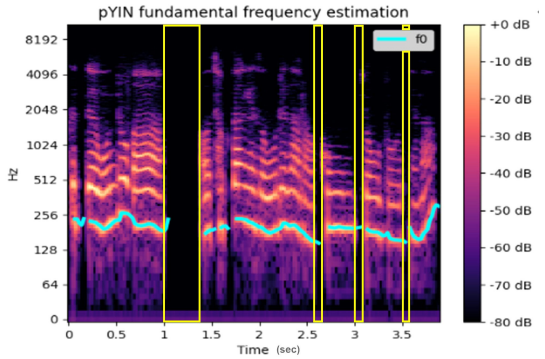


그림 2. 충청도 방언의 멜 스펙트로그램(“오늘 달이 참 예뻐”)
Figure 2. Mel-spectrogram of Chungcheong-do dialect (“Today’s moon is so beautiful”)

3.1. 휴지 길이 특성

휴지 길이 특성을 얻는 과정은 아래의 수식으로 나타낼 수 있다. 아래 수식에서 상대적 휴지 길이 특성(pause feature)은 줄여서 *pf*로 표기하였다.

$$y_t = r\,VAD(x_t) \tag{1}$$

$$z = \sum_{t=0}^{T} y_t \tag{2}$$

$$pf = 1 - \frac{z \cdot \Delta}{l} \tag{3}$$

입력 오디오 x는 time-step Δ로 세분화되며, 이는 $x = [x_0, x_1, x_2, \cdots, x_t, \cdots, x_T]$로 표현된다. 오디오 파일의 time-step으로 잘려진 세그먼트들(x_t)이 rVAD 모델의 입력으로 들어가게 되면 출력값(y_t)이 나오게 된다. 이 출력값은 발화 구간에서는 1의 값을 가지며, 휴지 구간은 0의 값을 가진다. 1의 값을 가지는 활성 음성 세그먼트들을 모두 더하여 얻은 값을 z로 지정하며 z와 time-step의 크기(Δ)와 곱하면 오디오 파일의 발화 구간 길이를 얻을 수 있다. 오디오 길이(l)에 대한 발화 구간 길이의 비율을 1에서 빼서 상대적인 휴지 길이 특성(*pf*)을 얻었다.

3.2. 발화 속도 특성

발화 속도(speech rate) 특성과 발화된 문자의 수(number of characters), 오디오 길이(audio length)는 각각 *sr, nc, l*로 아래 수식에서 표현하였다. 발화 속도(*sr*) 특성은 발화된 문자의 수(*nc*)를 오디오 길이(*l*)로 나눔으로써 얻을 수 있다. 이때 오디오 길이(*l*)는 VAD 결과가 아닌 원본 오디오의 처음과 끝의 휴지 구간만을 제거한 길이를 말한다. 이 두 휴지 구간을 제거한 이유는 오디오 파일의 길이에 해당 특성이 영향을 받지 않기 위함이다. 이는 아래 수식과 같이 표현할 수 있다.

$$sr = \frac{nc}{l} \tag{4}$$

발화된 문자의 수(*nc*)는 레이블된 데이터의 스크립트에서 얻을 수 있으며, 해당 정보는 ASR 모델의 출력에서도 얻을 수 있다. 영어와 달리, 한글은 문장의 문자수가 문장의 음절수와 같다는 특성을 가지고 있기에 이와 같은 방법으로 발화 길이를 구하였다.

그림 3은 충청도와 경상도의 발화 속도(*sr*)와 휴지 길이 특성(*pf*)의 분포를 보여준다. 그림에서 확인할 수 있듯이 두 그룹이 겹치는 구간이 존재하지만 쉽게 분리되는 것을 확인할 수 있다. 따라서 이 두 특성은 두 방언의 분류에 유의미한 영향을 줄 것을 알 수 있다.

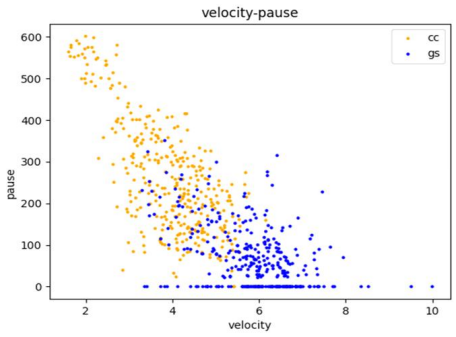
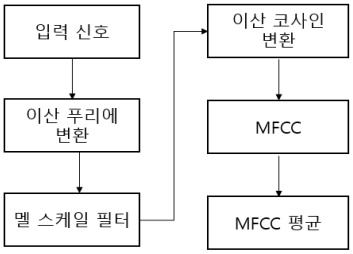


그림 3. 발화 속도와 휴지 구간 길이 특성 분포
Figure 3. Velocity and pause feature distributions

3.3. Mel-Frequency Cepstral Coefficients(MFCC) 특성

본 실험에서는 rVAD에 의해 결정된 활성 음성 세그먼트의 13차 MFCC를 사용하였다. 또한 MFCC값들은 선행 연구(Kim & Kim, 2021)와 마찬가지로 각각의 빈(bin)에 대해 평균값을 구하여 사용하였다. 그림 4는 본 실험에서 사용한 MFCC 평균 특성을 구하는 과정을 도식화한 그림이다. 평균화되지 않은 MFCC 특성들로 딥러닝 모델을 사용하면 더 오래 걸리는 반면, MFCC의 값을 평균화하여 특성으로 사용할 경우 훈련과 평가 모두에 있어서 빠르다는 장점을 가질 수 있다. 또한 해당 방법은 발화 길이에 영향을 받지 않고 머신 러닝 모델의 입력으로 넣을 수 있다는 장점을 가지고 있다.



MFCC, Mel-Frequency Cepstral Coefficients.

그림 4. MFCC 평균 특성 추출 과정
Figure 4. Process of extracting mean MFCC feature

3.4. Fundametal frequency(F0) 분산 특성

F0의 분산을 계산하기 위해 먼저 YIN 알고리즘(de Cheveigné & Kawahara, 2002)을 사용하는 librosa[1] 라이브러리를 통하여 오디오 파일의 F0를 추출하였다. 결과적으로 얻어진 시계열 F0값들은 F0 변동 속도의 특성으로 분산을 계산하는 데 사용하였다. 본 실험에서는 전체 오디오 구간 길이에 대한 분산 및 10, 15, 20, 25 ms의 네 가지 다른 구간 길이에 대해 F0 분산을 계산하였다. F0의 분산의 특성을 추출하는 식은 아래 수식과 같다.

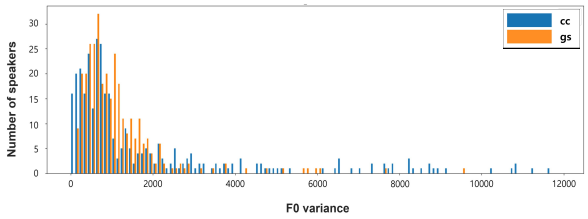
$$F0_t = YIN(x_t) \tag{5}$$

$$F0_{array} = [F0_0, F0_1, \ldots, F0_t, \ldots, F0_T] \tag{6}$$

$$F0_{variance} = variance(F0_{array}) \tag{7}$$

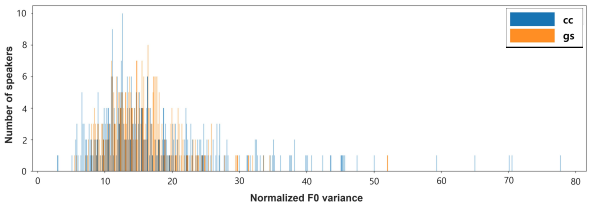
먼저 YIN 알고리즘에 입력으로 오디오 파일(x)의 time-step으로 잘려진 세그먼트(x_t)가 들어가며, 이를 F0 배열($F0_{array}$)에 저장한다. 이렇게 구해진 F0 배열의 분산을 구하면 전체 오디오 구간 길이에 대한 F0 분산 특성을 얻을 수 있다.

그림 5는 F0 분산 특성의 두 지역의 분포를 나타낸다. 그림 4에 비해 그림 5는 두 지역 사이의 명확한 경계를 찾기가 다소 어렵다. 따라서 F0의 전체적인 분산을 계산하는 대신, 우리는 F0 분산을 고정된 시간 단계 내에서 최대값과 최소값의 차이로 나눔으로써 정규화된 F0 분산을 추출해보았다. 그림 6은 정규화된 F0 분산의 분포를 보여주며, 여전히 두 그룹 사이에 유의미한 차이를 확인할 수 없었다. 이는 주어진 데이터셋에서 F0의 변화가 지역적 특성이라기보다는 개인 차이에 더 크게 영향을 받는 것이라고 판단되었다. 따라서 F0 기반 특성은 이후 실험에서 제외되었다.



F0, fundametal frequency.

그림 5. F0 분산 분포
Figure 5. F0 variance distributions



F0, fundametal frequency.

그림 6. 정규화된 F0 분산의 분포
Figure 6. Normalized F0 variance distributions

화자의 나이와 성별은 화자 인식 분야에서도 주요 특성으로 사용된다(Fenu et al., 2020; Pappagari et al., 2020). 방언 특색의 차이가 연령대와 성별에 따라 미세하게 존재할 것이라고 생각되어 이 두 메타 정보도 사용하였다. 특히 나이 정보는 10년 단위의 나이 그룹으로 변환하여 50, 60, 70대의 세 그룹으로 분류하여 사용하였다. 따라서 본 논문에서는 발화 속도, 휴지 길이 특성, 나이, 성별, 그리고 13차 MFCC 특성들을 사용하여 실험을 진행하였다.

그림 7은 제안된 알고리즘의 의사코드를 나타낸다. rVAD(Tan et al., 2020) 알고리즘을 사용하여 입력 오디오 파일의 발화와 휴지 구간을 판별한다. VAD 결과에 기반하여, 휴지 길이 특성은 휴지 구간을 사용하여, MFCC 특성은 발화 구간을 사용하여 각각 구하였다. 레이블 데이터에서는 성별, 나이 그리고 문자수를 추출하였다. 그리고 마지막으로, 추출된 특성들을 입력 데이터로 사용하여 기계 학습 모델을 이용하여 방언 분류를 수행하였다.

1 https://librosa.org/doc/latest/index.html

```
함수 입력(오디오_파일, 라벨_파일):
    만약 오디오_파일이라면:
        시작_끝_휴지_구간_제거(오디오_파일)
        오디오_길이 = 오디오_파일_길이(오디오_파일)
        rVAD_출력 = rVAD_분석(오디오_파일)

        만약 rVAD_출력이 1이라면:
            MFCC_특성_추출(오디오_파일)
        만약 rVAD_출력이 0이라면:
            휴지_구간_길이_특성_추출(오디오_파일)

    만약 라벨_파일이라면:
        나이, 성별 = 나이_성별_특성_추출(라벨_파일)
        스크립트_글자_수 = 스크립트_글자_수_추출(라벨_파일)
        발화_속도_특성_추출(오디오_길이, 스크립트_글자_수)

    추출한_특성들 = [나이, 성별, 휴지_구간_길이, 발화_속도, MFCC_특성]

    방언_분류기_입력(추출한_특성들)
```

그림 7. 제안된 알고리즘의 의사코드
Figure 7. Pseudocode of the proposed algorithm

4. 실험 결과

4.1. 하드웨어 환경

본 실험에서 머신 러닝 모델을 훈련시키기 위해 12 GB 메모리를 가진 4개의 NVIDIA TITAN Xp GPU, 64 GB RAM 그리고 6 코어로 3.6 GHz에서 실행되는 Intel Xeon E5-1650 v4 CPU가 장착된 환경을 사용하였다.

4.2. 모델 및 데이터셋

데이터셋의 크기가 딥러닝 모델을 사용하기에 충분히 크지 않았기에 본 실험에서는 앞서 언급한 대로 서포트벡터머신, 랜덤포레스트 그리고 라이트지비엠 - 세 가지 머신러닝 모델을 실험에 사용하였다. 서포트벡터머신의 경우 선형 커널과 C=10의 페널티 파라미터를 가진 모델로 훈련시켰다. 랜덤포레스트의 경우 엔트로피, 부트스트랩 집합, 150개의 결정 트리, 42의 랜덤 상태 그리고 최대 깊이 12의 하이퍼 파라미터를 가진 모델을 사용했다. 마지막으로, 라이트지비엠은 400개의 추정기를 가진 모델을 사용하였고 로그로스(logloss)를 평가 메트릭으로 사용하였으며 early stopping은 100 epoch로 설정한 모델을 사용하였다.

본 실험에서는 AI-hub에 공개된 '자유대화 음성(노인남녀)[2]' 데이터셋을 사용하였다. 방언 데이터가 아닌 노인 데이터를 사용한 이유는 노인의 경우가 방언색이 더욱 짙은 경향을 보였기에 해당 데이터셋을 선택하였다. 오디오 데이터의 샘플링 레이트는 16 kHz이고, 각 발화의 길이는 8초에서 15초 사이였다. 추가적으로, 각 오디오 데이터는 나이, 지역, 성별, 화자 ID 그리고 대본을 포함하는 해당 레이블 데이터도 제공되었다. 이 메타 데이터들 중 스크립트, 성별 그리고 나이 레이블을 사용하였다. 전체 데이터셋은 약 3,000시간의 오디오를 포함하며, 이는 1,000명의 노인들에 의해 녹음되었다. 또한, 데이터셋은 한국의 여섯 지역에서의 방언을 포함하고 있고, 그중에서 우리는 충청도와 경상도 지역의 방언을 사용하였다. 실험에서 사용된 데이터셋에서 이 두 지역에 해당되는 300여 명의 발화자들 중 오디오의 음질이 좋지 않은 환경에서 녹음한 발화자와 발음이 부정확한 발화자는 실험에서 제외시켜 최종적으로 발화자 134명의 오디오를 실험에 사용하였고 각 발화자당 5개의 발화를 무작위로 추출하였다. 실험을 위하여 학습 데이터와 시험 데이터의 비율은 8:2로 구성하였다.

4.3. 실험 결과

표 1. MFCC 추출 방법에 따른 모델별 정확도
Table 1. Performance comparison based on MFCC extraction methods

모델	rVAD 미사용	rVAD 사용
LightGBM	0.873	**0.903**
RF	0.896	**0.910**
SVM	0.657	**0.672**

MFCC, Mel-Frequency Cepstral Coefficients; VAD, Voice Activity Detection; GBM, Gradient Boosting Machine; RF, Random Forest; SVM, Support Vector Machine.

표 1은 MFCC 추출 방법에 따른 모델별 정확도를 보인다. 위의 표 1에서 알 수 있듯이 rVAD를 사용하여 휴지 구간을 제외한 부분의 MFCC 평균값을 사용한 경우와 rVAD를 사용하지 않고 모든 구간을 사용하여 MFCC 평균값을 사용한 경우의 정확도를 측정하였다. 실험 결과 세 모델 모두 rVAD를 사용하여 휴지 구간을 제외한 부분의 MFCC 평균값을 사용한 결과 가장 성능이 좋은 것을 알 수 있다. 따라서 이후의 실험에서 MFCC 특성을 추출할 때 rVAD를 사용하여 휴지 구간을 제외한 실제 발화 구간만을 사용하여 MFCC값을 구하였고 이 평균값을 사용하였다.

표 2에서 볼 수 있는 것처럼 표 2는 사용 특성별 모델들의 정확도를 나타낸다. MFCC는 앞서 언급한 바와 같이 13차원으로 13개의 특성을 사용하였으며, 본 논문에서 제안한 New 특성의 경우 발화속도, 휴지 구간의 길이, 나이, 성별 네 개의 특성을 사용했다. 마지막으로 모든 특성을 사용한 경우는 앞서 언급한 두 경우를 함께 사용한 총 17개의 특성을 사용한 결과이다. 표 2에서 각 특성의 옆에 표기한 괄호 안의 숫자는 각각의 경우에 사용된 특성의 개수를 나타낸다. 표 2에서 알 수 있듯이 모델의 경우 랜덤포레스트가 가장 높은 정확도를 보인다. 또한 본 논문에서 제안한 특성과 MFCC 특성을 함께 사용하는 것이 다른 경우들에 비해 성능이 가장 좋은 것을 확인할 수 있다. 또한 제안된 특성만 사용하는 경우가 MFCC 특성만을 사용하는 것보다 더

2 https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=107

높은 정확도를 가지는 것을 알 수 있다.

표 2. 사용 특성별 모델의 정확도
Table 2. Accuracy of models based on used featuers

모델	MFCC (13)	New (4)	New+MFCC (17)
LightGBM	0.903	0.933	0.955
RF	0.910	0.948	**0.970**
SVM	0.672	0.881	0.933

MFCC, Mel-Frequency Cepstral Coefficients; GBM, Gradient Boosting Machine; RF, Random Forest; SVM, Support Vector Machine.

표 3은 어떤 특성이 정확도에 중요한 영향을 미치는지에 대한 실험 결과를 제시한다. 이를 확인하기 위해 모든 특성을 사용한 경우에서 각각의 특성들을 하나씩 제외했을 때의 정확도를 확인하였다. 이는 각 실험에서 제외된 특성이 정확도에 얼마나 영향을 미치는가를 알 수 있는 실험이라고 생각된다. 해당 실험에서 사용한 모델은 앞선 실험에서 가장 성능이 뛰어났던 랜덤포레스트를 사용하였다. 표 3에서 볼 수 있듯이 발화 속도 특성만을 제외하고 다른 모든 특성을 사용한 경우 91.8%로 가장 낮은 정확도를 보인다. 이는 본 실험에서 발화 속도 특성이 성능에 있어 가장 중요하다는 것을 나타낸다. MFCC 특성도 다른 특성들에 비해 상대적으로 높은 영향력을 보였다. 마지막으로, 성별과 나이 특성을 제외한 경우 모든 특성을 다 사용했을 때 대비 0.7% 낮아져 96.3%의 정확도를 보였다. 이는 해당 실험에서 가장 작은 감소폭으로 나이와 성별 특성이 정확도에 가장 덜 영향을 미친다는 것을 알 수 있다.

표 3. 다양한 특성 조합에 따른 정확도
Table 3. Performance of different feature combinations

사용 특성	랜덤포레스트
New+MFCC 특성	0.970
- 성별, 나이 특성	**0.963**
- 휴지 구간 길이 특성	0.955
- MFCC 특성	0.948
- 발화 속도 특성	**0.918**

MFCC, Mel-Frequency Cepstral Coefficients.

5. 결론

본 논문에서는 방언 분류에 있어서 서로 다른 집단 간의 음향학적 차이에 의해 결정된 발화 속도와 휴지 구간 길이 같은 음성 특성을 추출하여 사용하는 새로운 접근법을 제안한다. 또한 제안된 특성과 선행 연구들의 주요 방법인 MFCC 특성을 함께 사용하는 것이 다른 경우들에 비해 가장 높은 성능을 보이는 것을 확인할 수 있었다. 각 특성이 정확도에 얼마나 영향을 미치는지 확인한 실험에서는 사용한 특성들 중에서 발화 속도 특성이 성능에 가장 강한 영향을 미친다는 것을 알 수 있었다. 또한 나이와 성별 특성이 가장 적은 영향력을 미치는 것도 확인할 수 있었다.

본 논문에서는 경상도와 충청도, 두 지역에 대한 방언 분류 실험을 진행하였지만, 더 많은 방언을 분류하는 데 대한 추가 실험은 추후 수행할 계획이다. 또한 더 큰 데이터셋을 확보하여 해당 기법을 딥러닝 모델에도 적용해볼 계획이다. 본 논문에서 제안한 접근 방법은 다른 국가의 방언을 분류하는 데에도 유의미한 결과를 보일 수 있을 것으로 생각된다. 제안한 방법을 통해 높은 정확도의 방언 정보를 추출할 수 있다면, 해당 정보를 화자 인식이나 음성인식과 같은 음성 분석의 다른 분야에서도 충분히 활용될 수 있을 것으로 기대된다.

References

Bhattacharjee, U., & Sarmah, K. (2013, March). Language identification system using MFCC and prosodic features. *Proceedings of the 2013 International Conference on Intelligent Systems and Signal Processing (ISSP)* (pp. 194-197). Vallabh Vidyanagar, India.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Chittaragi, N. B., & Koolagudi, S. G. (2019). Acoustic-phonetic feature based kannada dialect identification from vowel sounds. *International Journal of Speech Technology*, *22*(4), 1099-1113.

Chowdhury, S. A., Ali, A., Shon, S., & Glass, J. (2020, October). What does an end-to-end dialect identification model learn about non-dialectal information? *Proceedings of the INTERSPEECH 2020* (pp. 462-466). Shanghai, China.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 28*(4), 357-366.

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America, 111*(4), 1917-1930.

Dheram, P., Ramakrishnan, M., Raju, A., Chen, I. F., King, B., Powell, K., & Stolcke, A. (2022, September). Toward fairness in speech recognition: Discovery and mitigation of performance disparities. *Proceedings of the INTERSPEECH 2022* (pp. 1268-1272). Incheon, Korea.

Fenu, G., Medda, G., Marras, M., & Meloni, G. (2020, November). Improving fairness in speaker recognition. *Proceedings of the 2020 European Symposium on Software Engineering* (pp. 129-136). Rome, Italy.

Garcia-Romero, D., Snyder, D., Watanabe, S., Sell, G., McCree, A., Povey, D., & Khudanpur, S. (2019, September). Speaker recognition benchmark using the chime-5 corpus. *Proceedings of the INTERSPEECH 2019* (pp. 1506-1510). Graz, Austria.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Schol-kopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, *13*(4), 18-28.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q, ...

Liu, T. Y. (2017, December). LightGBM: A highly efficient gradient boosting decision tree. *Proceedings of the 31st Conference on Neural Information Processing Systems*. Long Beach, CA.

Keesing, A., Koh, Y. S., & Witbrock, M. (2021, August). Acoustic features and neural representations for categorical emotion recognition from speech. *Proceedings of the INTERSPEECH 2021* (pp. 3415-3419). Brno, Czechia.

Khurana, S., Najafian, M., Ali, A., Hanai, T. A., Belinkov, Y., & Glass, J. (2017, August). QMDIS: QCRI-MIT advanced dialect identification system. *Proceedings of the INTERSPEECH 2017* (pp. 2591-2595). Stockholm, Sweden.

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, J., Kim, K., & Chung, M. (2021, November). Korean dialect identification based on intonation modeling. *Proceedings of the 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (O-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 2022 25th Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)* (pp. 1-6). Hanoi, Vietnam.

Likitha, M. S., Gupta, S. R. R., Hasitha, K., & Upendra Raju, A. (2017, March). Speech based human emotion recognition using MFCC. *Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)* (pp. 2257-2260). Chennai, India.

Lin, W., & Mak, M. W. (2020, October). Wav2spk: A simple DNN architecture for learning speaker embeddings from waveforms. *Proceedings of the INTERSPEECH 2020* (pp. 3211-3215). Shanghai, China.

Mehrabani, M., & Hansen, J. H. L. (2015). Automatic analysis of dialect/language sets. *International Journal of Speech Technology*, *18*(3), 277-286.

Michon, E., Pham, M. Q., Crego, J., & Senellart, J. (2018, August). Neural network architectures for arabic dialect identification. *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)* (pp. 128-136). Santa Fe, NM.

Mukherjee, H., Obaidullah, S. M., Santosh, K. C., Phadikar, S., & Roy, K. (2020). A lazy learning-based language identification from speech using MFCC-2 features. *International Journal of Machine Learning and Cybernetics, 11*(1), 1-14.

Najafian, M., Khurana, S., Shan, S., Ali, A., & Glass, J. (2018, April). Exploiting convolutional neural networks for phonotactic based dialect identification. *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5174-5178). Calgary, AB.

Pappagari, R., Cho, J., Moro-Velazquez, L., & Dehak, N. (2020, October). Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer's disease and assess its severity. *Proceedings of the INTERSPEECH 2020* (pp. 2177-2181). Shanghai, China.

Sarma, M., Ghahremani, P., Povey, D., Goel, N. K., Sarma, K. K., & Dehak, N. (2018, September). Emotion identification from raw speech signals using DNNs. *Proceedings of the INTERSPEECH 2018* (pp. 3097-3101). Hyderabad, India.

Saste, S. T., & Jagdale, S. M. (2017, April). Emotion recognition from speech using MFCC and DWT for security system. *Proceedings of the 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA)* (pp. 701-704). Coimbatore, India.

Seo, J., & Lee, B. (2022). Multi-task conformer with multi-feature combination for speech emotion recognition. *Symmetry*, *14*(7), 1428.

Shahnawazuddin, S., Dey, A., & Sinha, R. (2016, September). Pitch-adaptive front-end features for robust children's ASR. *Proceedings of the INTERSPEECH 2016* (pp. 3459-3463). San Francisco, CA.

Sohn, J., Kim, N. S., & Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, *6*(1), 1-3.

Tan, Z. H., Sarkar, A. K., & Dehak, N. (2020). rVAD: An unsupervised segment-based robust voice activity detection method. *Computer Speech & Language,* 59, 1-21.

Tawaqal, B., & Suyanto, S. (2021). Recognizing five major dialects in Indonesia based on MFCC and DRNN. *Journal of Physics: Conference Series, 1844*, 012003.

Tüske, Z., Golik, P., Nolden, D., Schlüter, R., & Ney, H. (2014, September). Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages. *Proceedings of the INTERSPEECH 2014* (pp. 1420-1424). Singapore, Singapore.

Wallington, E., Kershenbaum, B., Klejch, O., & Bell, P. (2021, August-September). On the learning dynamics of semi-supervised training for ASR. *Proceedings of the INTERSPEECH 2021* (pp. 716-720). Brno, Czechia.

Wan, M., Ren, J., Ma, M., Li, Z., Cao, R., & Gao, Q. (2022, March). Deep neural network based Chinese dialect classification. *Proceedings of the 2021 Ninth International Conference on Advan-*

ced Cloud and Big Data (CBD) (pp. 207-212). Xi'an, China.

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 the INTERSPEECH – 2021* (pp. 3266-3270). Brno, Czechia.

Ying, W., Zhang, L., & Deng, H. (2020). Sichuan dialect speech recognition with deep LSTM network. *Frontiers of Computer Science*, *14*(2), 378-387.

Zhang, Q., & Hansen, J. H. L. (2018). Language/dialect recognition based on unsupervised deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(5), 873-882.

• **나종환 (Jonghwan Na)**
인하대학교 전기컴퓨터공학과 석사과정
인천광역시 서구 가좌동 건지로 404
Tel: 032-860-7411
Email: jhna@dsp.inha.ac.kr
관심분야: 음성인식, 방언분석, 음성분석

• **이보원 (Bowon Lee)** 교신저자
인하대학교 전기컴퓨터공학과 교수
인천광역시 미추홀구 인하로 100
Tel: 032-860-7423
Email: bowon.lee@inha.ac.kr
관심분야: 음성인식, 음성분석, 음성감정분석, 방언분석

발화 속도와 휴지 구간 길이를 사용한 방언 분류*

나 종 환 · 이 보 원

인하대학교 전기컴퓨터공학과

국문초록

본 논문에서는 음성의 발화 속도와 휴지 구간의 길이 그리고 화자의 연령과 성별에 기반한 방언 분류 접근 방법을 제안한다. 방언 분류는 음성 분석을 위한 중요한 기술 중 하나이다. 예를 들어 정확한 방언 분류 모델은 화자 인식 또는 음성 인식의 성능을 향상시킬 수 있는 잠재력을 가질 수 있다. 선행 연구에 따르면, Mel-Frequency Cepstral Coefficients(MFCC) 특징을 사용한 딥러닝 기반의 연구가 주류를 이루었다. 우리는 지역 간의 음향적 차이에 주목하여 그 차이를 바탕으로 추출한 특징을 사용하여 방언 분류를 진행하였다. 본 논문에서는 음성의 발화 속도, 휴지 구간의 길이 특성을 추출하여 사용하며 이와 함께 화자의 연령과 성별과 같은 메타데이터를 추가로 사용하는 새로운 접근 방법을 제안한다. 실험 결과 제안된 접근 방법이 더 높은 정확도를 보이는 것을 확인하였으며 특히 음성의 발화 속도 특성을 사용하는 것이 기존 MFCC만을 사용하는 방법보다 향상된 성능을 보여준다는 것을 확인할 수 있었다. MFCC 특성만을 사용한 방법과 비교했을 때 본 논문에서 제안한 특성들을 모두 사용하였을 때의 정확도는 91.02%에서 97.02%로 향상되었다.

핵심어: 방언 분류, 특성 추출, 적은 데이터셋 환경

참고문헌

김영국, 김명호(2021). 음향 특성에 따른 한국어 방언 분류 모델의 성능 비교. *한국컴퓨터정보학회논문지*, *26*(10), 37-43.

* 이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원(NRF-2018S1A5A2A03037308) 및 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임[RS-2022-00155915, 인공지능융합혁신인재양성사업(인하대학교)].