방언 별 화자 식별 알고리즘 분석

김수림

성신여자대학교 정보시스템공학과 20201071

[요 약]

음성인식 기술에서 화자의 방언 사용은 컴퓨터와의 상호작용을 방해하는 요소 중 하나이다. 방언은 음의 높낮이, 음소, 음절, 단어, 문장 등 다양한 요소를 통해 구분할 수 있지만 이를 모두 식별하여 구분해 내기에 어려움이 있다. 본 논문에서는 한국인 대화 음성 데이터를 통해 MFCC를 활용하여, 머신 러닝 및 딥러닝 분류 모델에서 한국어 방언 분류 성능을 비교 및 분석한다. 5가지의 머신 러닝 및 딥러닝 분류 모델에서 강원, 전라, 경상, 충청, 제주 5개 지역의 한국어 방언 분류 성능을 비교한다. 실험 결과 XGB 분류 모델이 정확도 62.73%로 가장 높은 성능을 기록하였다.

▶주제어: 머신 러닝, 딥러닝, MFCC, 음성 분석, 방언 분류, 성능 비교

Ⅰ. 서론

최근 스마트폰, 인공지능 기술이 탑재된 다양한 가전제품에서 음성 서비스를 제공하고 있다. 또한 코로나19(Covid-19)로 인해 비대면 회의[1]와 같은 비대면 플랫폼에 대한 수요가 급증하였다. 이에 따라 음성 인식 기술이 지속적으로 발전하고 있다. 음성 인식 기술은 사용자가 발성하는 음성을 이해하여 컴퓨터가 다룰 수 있는 문자(코드) 정보를 변화하는 기술[2]을 의미한다. 이때 화자의 음성이 표준어가 아니라 방언인 경우 컴퓨터가 제대로 이해하지 못하여 성능이 저하되는 문제가 발생한다.

방언은 특정 지역이나 공동체에서 사용하는 특별한 언어 형태로, 억양, 어휘, 문법 등에서 독특한 특징을 가지고 있다. 이러한 방언의 특징은 기존 음성인식 모델에서 처리하기 어려운 점을 만든다. 특정 방언의 발음이 다른 지역의 발음과 매우 유사하거나, 특이한 억양이 존재할 수 있기 때문에 음성 명령 혹은 질의를 정확하게 이해하기 어렵다. 또한, 방언은 특정 지역 또는 공동체에서 사용되기 때문에 해당 방언에 대한 충분한 데이터가 부족하여 모델의 학습과 일반화에 어려움을 겪는다.

따라서, 이러한 어려움을 극복하고 정확한 방언 인식 및 구분을 위한 최적의 알고리즘을 연구하는 것은 중요한 과제이다. 다양한 방언을 인식하고 구별하는 기능은 사용자의 경험을 향상시키는 데 도움을 주고, 지역적 또는 문화적으로 다양한 사용자들에게 맞춤화된 서비스를 제공하는 데 기여한다. 또한, 음성 데이터 분석을 통해 사용자의 출신 지역이나 언어적 배경을 식별할 수 있다.

이는 다양한 분야에서 응용될 수 있다. 예를 들어, 음성 비서 제품에서는 사용자의 방언을 인식하여 정확하게 명령을 이해하고 응답할 수 있다.

---

[1] ZOOM(줌) 또는 Google Meet(구글미트) 등

[2] IT사전, 한국정보통신기술협회, 1992.

이를 통해 사용자가 자연스럽게 음성 명령을 내릴 수 있으며, 사용자와의 상호작용이 원활하게 이루어지는데 도움을 준다. 또한, 방언 정보를 기반으로 사용자에게 맞춤형 광고를 제공하거나 문화적인 차이를 고려한 번역 서비스를 제공하는 등의 응용이 가능하다.

본 논문에서는 음성 인식 분야에서 주로 사용하는 MFCC(Mel-Frequency Cepstral Coefficient)를 통해 음성 특징 벡터를 추출하여 5가지 분류 모델 로지스틱 회귀(Logistic Regression), XGBoost(Extreme Gradient Boost), SVM(Support Vector Machine), RF(Random forest), CNN(Convolutional neural network)에 대해 성능을 비교하고 최적의 분류 모델을 제안한다.

본 논문은 다음과 같이 구성된다. 2장에서는 음성 특징을 추출하는 MFCC와 머신 러닝 및 딥러닝 모델을 설명한다. 3장에서는 실험 과정 및 결과를 요약하고, 4장에서 결론 및 향후 연구에 관하여 기술한다.

Ⅱ. 관련 연구

2.1 선행연구

본 논문에서는 사람의 음성을 분석하여 지역의 방언을 구분하는 연구를 수행할 것이다. 화자 인식을 위해 MFCC 음성 특징을 사용한다. MFCC는 음성 데이터 전체를 대상으로 하지 않고, 일정 구간으로 나누어 구간에 대한 스펙트럼을 분석하여 특징을 추출한다. 음성을 사용하여 지역 분류를 수행하는 전형적인 방법은 머신 러닝 기반 분류기를 사용하는 것이다. 최근 연구에 따르면, 다양한 머신 러닝 및 딥러닝 모델이 방언 분류에 사용되었으며, 그 결과 한국어 방언의 경우 RF 모델이 가장 우수한 성능을 보였다[3]고 한다.

2.2 오디오 데이터의 특징

2.2.1 waveform

소리는 시계열 데이터로, 음압의 변화를 내포한다. 오디오 데이터는 기본파(fundamental frequency)와 배음(harmonics)로 구성되어 있는 여러 주파수가 결합되어 있는 형태[4]이기 때문에 오디오 고유의 waveform에서 특징을 추출하기 어렵다. 그림 1은 train 데이터 셋에 포함된 방언 음성에 대한 waveform이다.

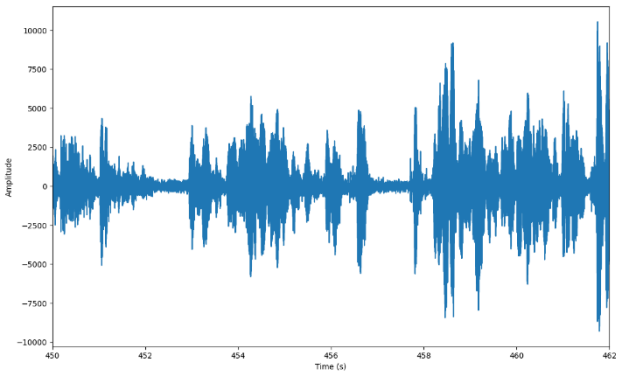

그림 1. 방언 음성에 대한 waveform

2.2.2 Mel spectrogram

사람은 높은 주파수의 소리보다는 낮은 주파수의 소리에 민감하다. 하지만 신경망을 통해 음성에서 추출한 특징을 학습하면 광범위한 주파수의 소리를 구별할 수 있다. 오디오의 waveform에

[3] 나종환, 이보원.(2023).발화 속도와 휴지 구간 길이를 사용한 방언 분류.말소리와 음성과학,15(2),43-51.

[4] 박대서, 방준일, 김화종, 고영준.(2018).CNN을 이용한 음성 데이터 성별 및 연령 분류 기술 연구.한국정보기술학회논문지,16(11),11-21.

STFT(Short Time Fourier Transform)를 수행하면 주파수를 x축으로 하는 스펙트럼이 생성되고, 이때 y축인 magnitude를 제곱하면 Power spectrum이 생성된다. Magnitude에 log scale을 적용하여 데시벨 단위로 변환한 것을 Log spectrum이라고 한다. Log spectrum을 세로로 세워서 프레임 마다 쌓으면, 푸리에 변환으로 사라졌던 time domain을 복원할 수 있고, 이를 Spectrogram이라고 한다.

Mel filter는 저음의 주파수보다 고음의 주파수에 덜 민감한 사람의 청력에 기반하여 1kHz까지는 선형적으로, 그 이상의 주파수는 log scale로 변환한다. 이와 같은 특성을 가지고 있는 Mel filter를 Spectrogram에 적용시키면 주파수는 Mel frequency로, Power는 log로 mapping된다. 이와 같이 생성되는 것이 Mel spectrogram이다. 즉. Mel spectrogram은 오디오 신호에 STFT를 가해 얻은 Spectrogram에 Mel filter를 적용하여 얻을 수 있는 특징이다. 그림 2는 방언 음성에서 추출한 Mel spectrogram이다[5].

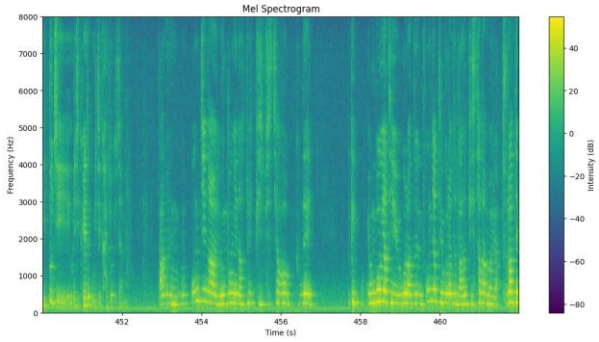

그림 2. 방언 음성에 대한 Mel spectrogram

2.2.3 Mel-Frequency Cepstral Coefficients (MFCC)

MFCC는 오디오 신호에서 추출할 수 있는 특징으로, 소리의 고유한 특징을 나타내는 수치이다. 주로 음성 인식, 화자 인식, 음성 합성, 음악 장르 분류 등 오디오 도메인의 문제를 해결하는데 사용된다.

MFCC는 Mel spectrogram을 구하는 과정에서 log을 취한 뒤 이산 코사인 변환(DCT; Discrete Cosine Transform)을 수행한 것이다. 그리고 주파수가 낮고 정보와 에너지가 몰려있는 12개의 계수(cepstrum coefficient)와 이들을 더한 값을 feature로 사용한다. 즉, MFCC는 Mel spectrogram에 log 및 DCT를 취하고 12개의 계수와 이로부터 구해진 에너지를 더한 값으로 프레임 마다 13개의 값을 feature로 가지는 특징이다.[5]

MFCC는 저주파 대역에서의 오디오 신호를 세밀하게 표현하고 고주파 대역에서는 상대적으로 간략하게 표현하는 특징[3]을 가지기 때문이다. 그림 3은 방언 음성으로부터 추출한 MFCC이다.

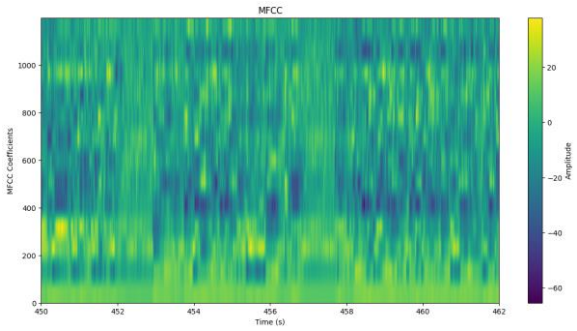

그림 3. 방언 음성에 대한 MFCC

[5]정윤아, 김동회.(2022).MFCC 기반 환경음 분류 CNN에서 커널 사이즈와 풀링 레이어에 의한 성능분석.한국디지털콘텐츠학회 논문지,23(5),913-920.

2.3 머신 러닝 및 딥러닝 기술

2.3.1로지스틱 회귀(Logistic Regression)

로지스틱 회귀는 회귀를 사용하여 데이터가 어떤 범주에 속할 확률을 0에서 1 사이의 값으로 예측하고 그 확률에 따라 가능성이 더 높은 범주에 속하는 것으로 분류해주는 지도 학습이다.

2.3.2 XGBoost(Extreme Gradient Boost)

XGBoost는 기존 그래디언트 부스팅[6](Gradient Tree Boosting)알고리즘에 과적합 방지를 위한 기법이 추가된 지도 학습 알고리즘이다. XGBoost는 기본 학습기(Base Learner)를 의사결정나무로 하며 Gradient Boosting과 같이 Gradient(잔차)를 이용하여 이전 모형의 약점을 보완하는 방식으로 학습한다.

2.3.3 SVM(Support Vector Machine)

SVM은 기계 학습의 분야 중 하나로 지도 학습 모델이며, 주로 분류와 회귀 분석을 위해 사용한다. 두 카테고리 중 어느 하나에 속한 데이터의 집합이 주어졌을 때, SVM 알고리즘은 주어진 데이터 집합을 바탕으로 하여 새로운 데이터가 어느 카테고리에 속할지 판단하는 결정 경계(Decision Boundary)이다.

2.3.4 RF(Random forest)

RF는 기계 학습에서의 분류, 회귀 분석 등에 사용되는 앙상블 학습 방법의 일종이다. 훈련 과정에서 구성한 다수의 결정 트리(decision tree)로부터 분류 또는 평균 예측치를 출력함으로써 동작한다. 많은 수의 결정 트리를 생성 하므로 일부 결정 트리가 오버피팅(overfitting)되어도 최종 분류에 영향을 미치지 않는다.

2.3.5 CNN(Convolutional neural network)

CNN은 딥러닝에서 시각적 이미지를 분석하는 데 가장 일반적으로 적용되는 인공 신경망의 한 종류이다. 기본적으로 CNN은 입력 계층, 여러 은닉 계층, 출력 계층으로 구성되어 있다. CNN은 레이어 중 적어도 하나에서 일반 행렬 곱셈 대신 컨볼루션이라는 수학적 연산을 사용한다. 픽셀 데이터를 처리하도록 설계되어 있어 이미지 인식 및 처리에 사용된다. 필터는 간단한 특징부터 객체를 고유하게 정의하는 특징으로 복잡도를 늘릴 수 있다.

Ⅲ. 실험

3.1 데이터 셋

본 논문에서는 AI Hub에서 제공하는 방언 별 한국인 대화 음성 데이터를 이용하여 모델 학습 및 테스트를 수행한다. 한국인 대화 음성 데이터 세트는 한국인의 일당 대화를 인식하고 음성을 문자로 실시간 변환하는 인공지능 기술 개발을 위해 춘천 MBC와 EBS의 방송콘텐츠에서 음원을

[6] 그래디언트 부스팅은 회귀 및 분류 작업에 사용되는 기계학습 기술이다.

추출하고 화자의 성별, 나이, 방언 등의 정보를 라벨링 한 데이터 세트이다[7]. 강원, 경상, 전라, 충청, 제주로 분류된 5개의 데이터에서 7:3의 비율에 맞춰 각 1000개씩 총 5000개의 데이터를 훈련 데이터로, 각 300개씩 총 1500개의 데이터를 테스트 데이터로 사용하였다.

3.2 데이터 전처리

MFCC를 이용하여 음성 데이터의 특징을 추출한다. 수집한 음성 데이터의 12초 길이를 선정하여 실험을 진행하였다. 음성의 초당 샘플링 수는 인간의 목소리에 가장 적합한 16000Hz로 설정하였고, 구간 당 40개의 특징을 추출하였다.

3.3 실험 결과

각 모델의 성능 비교 지표로는 정확도(accuracy), 정밀도 (precision), 민감도(sensitivity), F1-점수 (F1-score)를 사용하였다. 정확도(accuracy)는 전체 데이터 셋 중 올바르게 분류한 데이터 수의 비율을 말한다. 정밀도(precision)는 분류기가 정답이라고 분류한 대상 중 실제 정답인 경우를 말한다. 민감도(sensitivity)는 재현율(recall)이라고도 하며 정답 중 분류기가 올바르게 정답으로 분류한 비율을 말한다. F-점수는 정밀도와 재현율의 조화평균으로 계산된 성능 지표로 모델 평가지표로 많이 활용되며 여러 지표가 아닌 하나의 수로 평가를 내리는데 용이하다[4].

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN}$$

$$Precision = \frac{TP}{FP + TP}$$

$$Sensitivity(Recall) = \frac{TP}{TP + FN}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

	정확도	정밀도	민감도	F1-점수
LR	47.40	50.04	47.40	47.27
SVM	55.06	59.77	55.06	54.98
RF	50.13	51.18	50.13	49.50
XGB	**62.73**	**64.84**	**62.73**	**62.36**
CNN	59.60	62.04	59.60	59.56

표 1. 분류 모델 성능 비교 지표

표 1은 머신 러닝과 딥러닝 모델을 학습하고 성능을 비교한 결과이다. 5가지 분류 모델 중 XGB가 정확도 62.73%, F1-score 62.36%를 기록하고, 정밀도와 민감도 또한 가장 높은 성능을 기록하였다.

Ⅳ. 결론

음성 인식 및 화자 인식에서 사용자의 방언 사용은 성능을 저하시키는 원인이다. 본 논문은 지역 방언 분류 시스템을 구현하여 음성 인식 시스템의 성능을 향상시키는 방법에 대해 연구하였다.

[7] Young Kook Kim, Myung Ho Kim.(2021).Performance Comparison of Korean Dialect Classification Models Based on Acoustic Features.한국컴퓨터정보학회논문지 ,26(10),37-43.

본 논문에서는 5개 지역(충청, 전라, 경상, 강원, 제주)의 한국어 방언 음성 데이터에서 12초 동안의 MFCC 특징의 평균값을 정규화하여 추출하고, 이 특징을 입력으로 하는 머신 러닝 및 딥러닝 중 효과적으로 분류하는 최적의 모델을 제안한다.

5가지의 분류 모델을 학습하고 성능을 비교한 결과로 XGB 모델이 정확도 62.73%, F1-score 62.36%로 가장 높은 성능을 기록하였다. 분류 모델 중 RF는 다수의 결정 트리를 앙상블 형태로 학습하기 때문에 가장 성능이 좋을 것이라고 예상했다. 그러나 본 실험에서 XGB의 훈련 세트의 정확도는 99%, RF의 훈련 세트의 정확도는 65%로 XGB 모델이 과대 적합이 되어 가장 높은 정확도를 보인 것으로 판단된다.

본 실험에서 사용한 방언 음성 데이터는 남녀 대화 형식이므로 성별과 나이에 의해 나타나는 오차를 감안해야 한다. 향후 성별과 나이의 특징을 추가로 분석하여 확실한 데이터를 확보하고, 지역 방언 분류의 정확도를 높이는 연구가 필요하다.

참고문헌

[1] 나종환, 이보원(2023). 발화 속도와 휴지 구간 길이를 사용한 방언 분류. 말소리와 음성과학,15(2),43-51.

[2] 박대서, 방준일, 김화종, 고영준.(2018).CNN을 이용한 음성 데이터 성별 및 연령 분류 기술 연구.한국정보기술학회논문지,16(11),11-21.

[3] 음성인식기술 ,IT사전, 한국정보통신기술협회, 1992.

[4] 정윤아, 김동회.(2022).MFCC 기반 환경음 분류 CNN에서 커널 사이즈와 풀링 레이어에 의한 성능분석.한국디지털콘텐츠학회논문지,23(5),913-920.

[5] 한국민족문화대백과사전, 방언.

[6] AI Hub, 한국어 대화 음성, https://aihub.or.kr/aidata /7968.

[7] Young Kook Kim, Myung Ho Kim.(2021).Performance Comparison of Korean Dialect Classification Models Based on Acoustic Features.한국컴퓨터정보학회논문지,26(10),37-43.