Journal of the Korean Data & Information Science Society
2024, **35**(4), 445–459

http://dx.doi.org/10.7465/jkdi.2024.35.4.445
한국데이터정보과학회지

사전 학습 모델과 앙상블 기법을 통한 음성 감정인식[†]

서재진[1] · 강태인[2] · 곽일엽[3]

[123]중앙대학교 통계데이터사이언스학과

접수 2024년 5월 29일, 수정 2024년 6월 14일, 게재확정 2024년 6월 17일

요 약

음성 감정 인식 연구는 인간-기계 상호작용을 향상시키는 데 중요하고, 의료, 교육, 고객 서비스 등 다양한 분야에서 효율성을 높이고 사용자 경험을 개선할 수 있다. 본 연구에서는 DACON의 '월간 데이콘 음성 감정 인식 AI 경진대회'에 참가하여 여섯 가지 감정을 분류하는 AI 모델을 개발을 목표로 하였다. 전통적인 음성 처리 기술 기반 방법과 사전 학습된 모델을 이용한 방법들의 성능을 비교하였고, 사전 학습된 모델을 통해 음성의 일반화된 특징을 효과적으로 학습한 임베딩 벡터의 추가 학습 가능성을 탐구하였다. 그 결과, WavLM에 1D CNN을 결합한 모델이 79.80%의 성능으로 우수한 결과를 보였고, 사용한 모든 사전 학습된 모델들을 하드 보팅 앙상블하여 5등에 준하는 80.79%까지 성능을 향상시켰다. 본 연구는 음성 감정 분류에서 높은 성능을 달성하여 음성 감정 인식 기술의 적용 가능성을 높임으로써, 다양한 실제 응용 분야에서 감정 인식 모델의 활용을 가능하게 하는 데 기여할 것이다.

주요용어: 딥러닝, 사전 학습된 모델, 앙상블 학습, 음성 감정 인식.

1. 서론

음성 감정 인식 연구는 인간과 기계 간의 상호작용을 개선하는 데 중요한 역할을 한다. 이 기술을 통해 인간의 감정을 정확히 이해하고 반응하게 함으로써 컴퓨터와의 대화를 더욱 자연스럽고 인간다운 것으로 만들 수 있다. 이는 의료, 교육, 고객 서비스 등 다양한 분야에서 효율성을 높이고 사용자 경험을 향상시킬 수 있다. 예를 들어, 정신 건강 관리 분야에서는 우울증이나 불안 장애 같은 정신 질환을 조기 발견 및 모니터링할 수 있게 하며, 교육 분야에서는 학습자의 음성과 감정을 분석하여 학습 효율을 높일 수 있다. 또한, 콜센터에서는 고객의 음성에서 감정을 인식해 직원들이 고객의 감정 상태를 더 잘 이해하고 적절히 대응함으로써 서비스의 질을 높이는 데 기여한다. 따라서, 더욱 자연스러운 컴퓨터와의 상호작용을 위해 다양한 음성 감정 인식 연구가 진행되고 있다 (El Ayadi 등, 2011; Khalil 등, 2019).

본 연구는 DACON에서 주관한 '월간 데이콘 음성 감정 인식 AI 경진대회'에 참가하여, 주어진 음성 데이터셋을 활용해 분노, 두려움, 슬픔, 혐오, 중립, 행복 등 여섯 가지 감정을 분류할 수 있는 AI 모델을 개발하는 것을 목표로 한다. Figure 1.1은 음성을 통해 감정을 인식 및 분류하는 상황을 도식화한 것이다.

[†] 이 성과는 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.RS-2023-00208284). 이 논문은 2022년도 중앙대학교 CAU GRS 지원에 의하여 작성되었음.
[1] (06974) 서울시 동작구 흑석로 84, 중앙대학교 통계데이터사이언스학과, 석사과정.
[2] (06974) 서울시 동작구 흑석로 84, 중앙대학교 통계데이터사이언스학과, 석사과정.
[3] 교신저자: (06974) 서울시 동작구 흑석로 84, 중앙대학교 응용통계학과, 부교수.
E-mail: ikwak2@cau.ac.kr

음성 감정 인식에는 MFCC (Mel-frequency cepstral coefficients; MFCC)와 Mel-spectrogram 등의 음성 특징을 추출하여 Decision tree, XGBoost, CNN (Convolutional neural network; CNN), LSTM (Long short-term memory; LSTM)과 같은 머신러닝 및 딥러닝 모델을 사용하는 방법이 많이 활용되었다. 그러나 최근에는 대규모 오디오 데이터를 통해 사전 학습된 모델을 활용하는 것이 성능 향상에 크게 기여하고 있다.

화자의 음성신호는 1차원 배열 (1 dimensional array) 형태의 데이터로 표현되며, 이는 다음 단계의 Feature extraction 단계를 통해서 MFCC나 주파수-시간 정보를 포함하는 spectrogram 행렬로 변환시킨 후, Decision tree, XGboost, CNN 등등 다양한 기계 학습 방법들과 결합하여 감정을 분류하는 분류 모형 (Classification model)을 구축 할 수 있다 (Pandey 등, 2019). 또한, Wav2vec2.0이나 WavLM같은 사전 학습된 모델들을 활용하여 feature를 추출하여 마찬가지로 사전 학습된 Sequence classification 모델을 이용해 감정을 분류할 수 있다.

최근 연구에서는 사전 학습된 모델들의 예측 결과를 앙상블하여 더 나은 성능을 달성하는 방법이 주목받고 있다 (Deb 등, 2022; Heo 등, 2021; Nimmi 등, 2022). 이는 사전 학습된 모델들이 다양한 데이터를 통해 학습한 일반화된 특징을 효과적으로 활용할 수 있음을 시사한다. 이러한 접근 방식은 서로 다른 데이터셋과 방법으로 사전 학습된 여러 모델들을 결합함으로써 각 모델의 장점을 최대한 활용하고, 모델 간의 상호 보완적인 특징을 통해 보다 견고하고 일반화된 특징을 얻을 수 있다. 결과적으로, 이러한 앙상블 기법은 단일 모델이 놓칠 수 있는 세부 사항을 포착하고, 예측의 정확성을 높이는 데 기여할 수 있다.

또한, 사전 학습된 모델의 임베딩 벡터를 활용하는 연구도 진행되고 있다 (Kang 등, 2023). 본 연구에서도 대규모 사전 학습된 모델들을 사용하여 일반화된 feature를 효과적으로 추출하고, 해당 모델의 임베딩 벡터를 활용한 추가 학습의 가능성을 탐구하고자 한다. 또한, 다양한 사전 학습된 모델들을 비교 분석하여 모델 간의 예측 불일치도를 평가하고, 최종적으로는 이를 기반으로 하드 보팅 앙상블 기법을 적용하여 단일 모델을 사용하는 것보다 향상된 성능을 달성할 수 있는 방법을 알아보고자 한다.

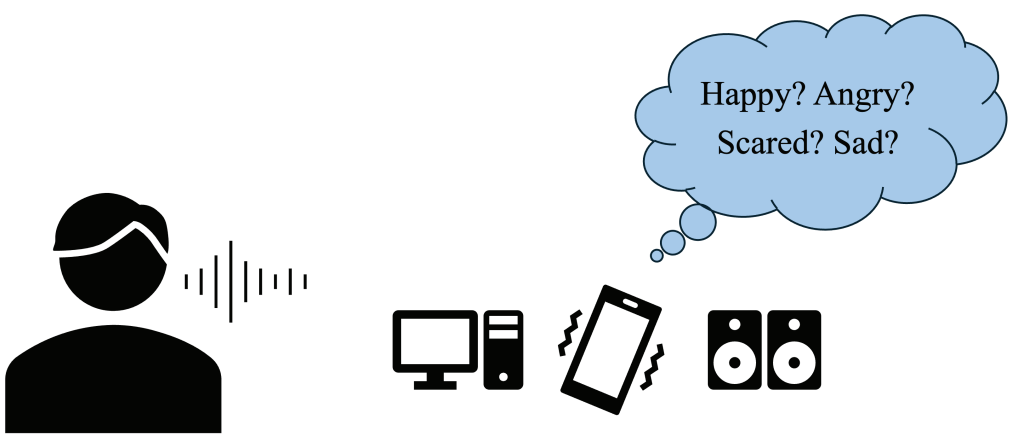

Figure 1.1 Speech emotion recognition scenario

본 논문의 구성은 다음과 같다. 2절에서는 선행연구 기법들을 다루며, 음성 특징 추출 방법과 머신러닝 및 딥러닝 기법에 대해 설명한다. 3절에서는 본 연구에서 제안하는 음성 감정 분류 시스템에 대해 기

술한다. 4절에서는 실험 설정과 연구에 사용된 데이터셋에 대해 소개하고, 5절에서는 제안된 다양한 분류 모형들의 성능을 비교 분석한다. 마지막으로, 6절에서는 본 연구의 결과를 요약하고, 연구의 시사점을 제시한다.

2. 이론적 배경

2.1. 대표적인 음성 특징 추출 방법

2.1.1. Mel-frequency cepstral coefficients (MFCC)

MFCC는 음성 신호 처리에서 가장 널리 사용되는 특징 추출 방법 중 하나이다. 이 방법은 인간의 귀가 특정 주파수 영역에서 음성을 인식하는 방식을 모방하여 개발되었다. MFCC는 음성 신호를 짧은 구간 (프레임)으로 나눈 후, 각 프레임에 대해 멜 스케일로 변환된 스펙트럼을 기반으로 켑스트럼 계수를 계산한다. 이러한 계수들은 음성 신호의 주파수 특성을 잘 반영하며, 음성 인식 등 다양한 오디오 처리 작업에 유용하게 사용된다 (Chakraborty 등, 2014). 연구에 사용된 음성 데이터에서 MFCC를 추출한 예시를 Figure 2.1에서 확인할 수 있다.

2.1.2. Mel-spectrogram

Mel-spectrogram은 음성 신호의 멜 스케일 기반 스펙트럼으로, 인간의 청각 시스템이 실제로 음성을 인식하는 방식에 좀 더 근접한 방식으로 음성 데이터의 특징을 추출한다. Mel-spectrogram은 시간에 따른 신호의 주파수 분포를 시각화하는 방법으로, 각 프레임에서의 주파수 성분들을 멜 스케일로 변환하여 표현한다. 이는 음성의 다양한 특성과 패턴을 효과적으로 포착하며, 특히 음악 장르 분류나 음성 감정 인식 등 고급 오디오 처리 작업에 매우 유용하다. 마찬가지로 Mel-spectrogram의 예시를 Figure 2.1에서 확인할 수 있다.

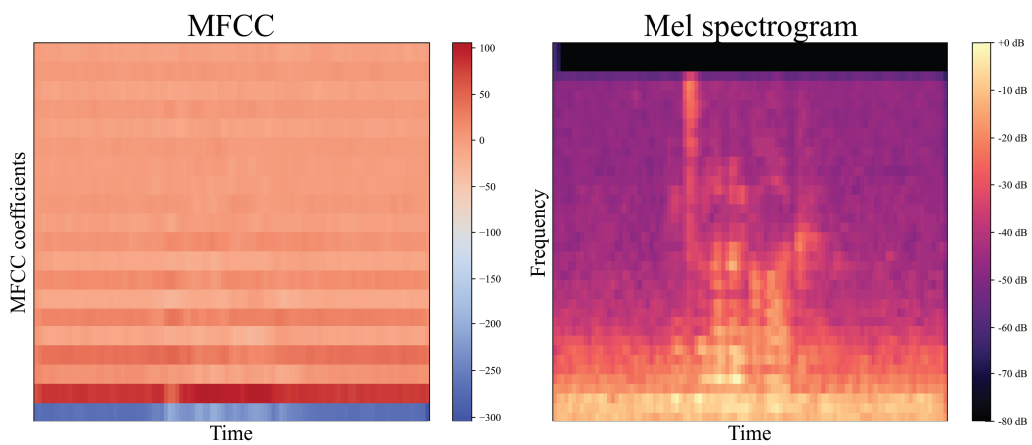

Figure 2.1 Example of MFCC and Mel-Spectrogram

2.2. 머신러닝 및 딥러닝 모형

2.2.1. Decision tree

의사 결정 나무는 데이터를 기반으로 하여 의사 결정 규칙을 학습하고, 이를 통해 새로운 데이터를 분류하는 알고리즘이다 (Song 등, 2015). 이 알고리즘은 분류나 예측 문제에 널리 사용되며, 그 구조는 트리 (Tree) 형태로 표현된다. 트리의 각 내부 노드 (internal node)는 특정 특성 (feature)의 값을 기준으로 데이터를 분할하고, 각 Leaf Node는 최종적으로 분류되는 클래스 레이블 (class label)을 가지게 된다. 이를 통해 모델은 데이터의 패턴을 학습하고, 새로운 데이터에 대해 예측을 수행한다.

2.2.2. XGboost

Chen 등 (2016)이 제안한 XGBoost는 그래디언트 부스팅 알고리즘의 한 종류로, 앙상블 학습 방법을 사용한다. XGBoost는 모델의 복잡도를 조절하기 위한 규제 기법을 사용하여 과적합을 방지하고, 이전 트리의 오차를 보정하는 방식으로 새로운 트리를 학습하여 모델의 성능을 향상시킨다. 또한 각 특성의 중요도를 추정하여 모델의 해석 가능성을 높이고, 불필요한 가지를 자동으로 제거하여 모델을 최적화한다. XGBoost는 트리의 구성이 병렬화되어 있어 대용량 데이터셋에 대해 빠른 학습이 가능하다.

2.2.3. ResNet

이 연구에서는 CNN (Convolutional neural network; CNN) 알고리즘을 활용하여 음성 데이터를 Mel-Spectrogram으로 변환하여 분류 문제를 다루었다. CNN은 신경망의 한 종류로, 이미지 처리에 특화된 구조를 가지고 있다. CNN은 이미지의 지역적 패턴을 감지하는데 효과적으로 사용되며, 이를 위해 이미지에 대해 필터링과 풀링 등의 작업을 수행한다. 이러한 작업을 통해 이미지의 특징을 추출하고, 이를 기반으로 이미지를 분류하거나 예측한다. 또한 CNN은 합성곱 (Convolution)과 풀링 (Pooling) 층을 여러 겹으로 쌓아 깊은 신경망을 구성할 수 있으며, 이를 통해 복잡한 문제에 대한 효과적인 해결이 가능하다. 특히, 본 연구에서 사용된 CNN 모델은 He 등 (2016) 이 제안한 ResNet (Residual network; ResNet)이다. ResNet은 깊은 신경망에서 발생하는 기울기 소실 문제를 해결하여 더 깊은 네트워크를 효율적으로 학습할 수 있고, 잔차 연결을 통해 정보가 직접적으로 전달되므로 학습이 더욱 안정적으로 이루어지며, 따라서 더 깊은 네트워크를 구성하여 더 높은 성능을 달성할 수 있는 모델이다.

2.3. Bi-LSTM

LSTM (Long short-term memory; LSTM)은 일반적인 RNN이 가지는 기울기 소실 문제를 해결하기 위해 장기 의존성을 효과적으로 학습할 수 있는 모델이다. 이러한 LSTM를 확장하여 시퀀스를 양방향으로 처리하여 양 끝에서 나오는 정보를 모두 활용하는 Bi-LSTM (Bidirectional LSTM; Bi-LSTM)은 더욱 풍부한 문맥 정보를 모델이 학습할 수 있다 (Huang 등, 2015). Bi-LSTM은 입력 시퀀스를 처음부터 끝까지와 끝에서 처음까지 두 번 처리하여 시퀀스의 앞뒤 맥락을 모두 고려할 수 있어, 단방향 LSTM보다 더 효과적으로 문맥 정보를 학습할 수 있게 된다. 또한, 최근에는 복잡한 시계열 데이터를 더 효과적으로 분석하기 위해 CNN과 결합하여 Bi-LSTM을 사용하는 연구도 진행되고 있다 (Yadav, 2020).

2.4. 사전 학습된 모델

2.4.1. Wav2vec 2.0

Baevski 등 (2020) 이 Facebook에서 개발한 Wav2vec 2.0은 자기 지도 학습 방식을 통해 대규모 라벨되지 않은 음성 데이터로부터 원시 음성 파형의 직접적인 음성 인식을 위한 표현을 학습하는 모델이다. 이 모델은 음성 데이터의 일부를 마스킹하고 해당 부분을 예측하도록 하여 음성의 구조를 이해하게 되며, 대비 학습 기법을 통해 정확한 오디오 조각을 식별하고 잘못된 것과 구별하는 능력을 학습한다.

2.4.2. HuBERT

Hsu 등 (2021) 이 Facebook에서 개발한 HuBERT-Base 모델은 16000Hz로 샘플링된 음성 오디오를 입력으로 사용하며, 자기 지도 학습을 통해 음성 데이터의 효과적인 표현을 학습한다. 이 모델은 LibriSpeech 960시간 데이터셋으로 사전 학습되었고, 추가적으로 레이블이 있는 텍스트 데이터를 사용하여 미세 조정을 진행할 수 있다. HuBERT는 음성 인식을 포함하여, 음성에서 텍스트로의 변환, 음성의 감정 분석 등 다양한 음성 기반 작업에 활용될 수 있다. HuBERT의 핵심 아이디어는 음성 데이터의 연속적인 숨겨진 유닛을 예측하도록 학습시킴으로써 더 나은 음성 표현을 학습하는 것이다.

2.4.3. WavLM

Chen 등 (2022) 이 Microsoft에서 개발한 WavLM 모델은 다양한 음성 처리 작업에서 높은 성능을 제공한다. 이 모델도 마찬가지로 자기 지도 학습 방법을 사용하여, 16000Hz로 샘플링된 음성 데이터로부터 효과적인 음성 표현을 학습한다. WavLM은 Transformer 구조에 게이트된 상대 위치 편향을 탑재하여 인식 작업에 대한 처리 능력을 향상시킨다고 한다. 이를 통해 모델은 더 정확하게 음성의 구조를 파악하고, 맥락을 이해할 수 있게 된다. WavLM은 Libri-Light, GigaSpeech, VoxPopuli 등의 대규모 데이터셋으로 사전 학습되었고, 특히, 화자 검증, 화자 식별, 그리고 화자 구분 작업에서 특히 우수한 성능을 보인다고 한다.

2.5. 앙상블 학습

앙상블 학습 (ensemble learning)은 여러 개의 학습 알고리즘을 결합하여 보다 나은 예측 성능을 얻는 기법이다. 머신러닝 및 딥러닝에서 주로 사용하는 앙상블 방법은 여러 모델들의 예측을 결합하여 단일 모델보다 예측 성능을 향상시키고, 과적합 (overfitting)을 방지하며, 일반화 성능을 높이는 데 유용하다 (Dietterich, 2000). 앙상블 학습의 유형으로는 보팅 (Voting), 스태킹 (Stacking), 배깅 (Bagging), 부스팅 (Boosting) 등 다양한 방식이 있다. 보팅과 스태킹은 여러 모델의 결합으로 사용되며, 이들은 각 모델의 예측을 결합하여 최종 예측을 수행한다. 반면, 배깅과 부스팅은 표본의 재추출 (resampling)을 통해 단일 모델을 반복하여 분류기를 만들어 성능을 향상시킨다 (Breiman, 1996; Freund 등, 1996; Parhami, 1994).

본 연구에서 사용한 앙상블 기법은 보팅 방식 중 하드 보팅 (hard voting) 앙상블 기법을 사용한다. 하드 보팅은 다수의 분류기 (모델)가 예측한 결과 중 가장 많은 표를 받은 클래스를 최종 예측 결과로 선택하는 방식이다. 모델의 성능이 서로 비슷할 때, 각 모델의 예측 결과가 서로 다른 클래스를 선택할 가능성이 있으므로, 이때 하드 보팅은 다수결 원칙을 적용하여 가장 많은 모델이 예측한 클래스를 최종 결과로 선택할 수 있다. 따라서 모델 성능이 비슷할 때는 하드 보팅이 소프트 보팅보다 효과적일 수 있으며, 이를 통해 앙상블 모델의 성능을 향상시킬 수 있다.

3. 연구 방법

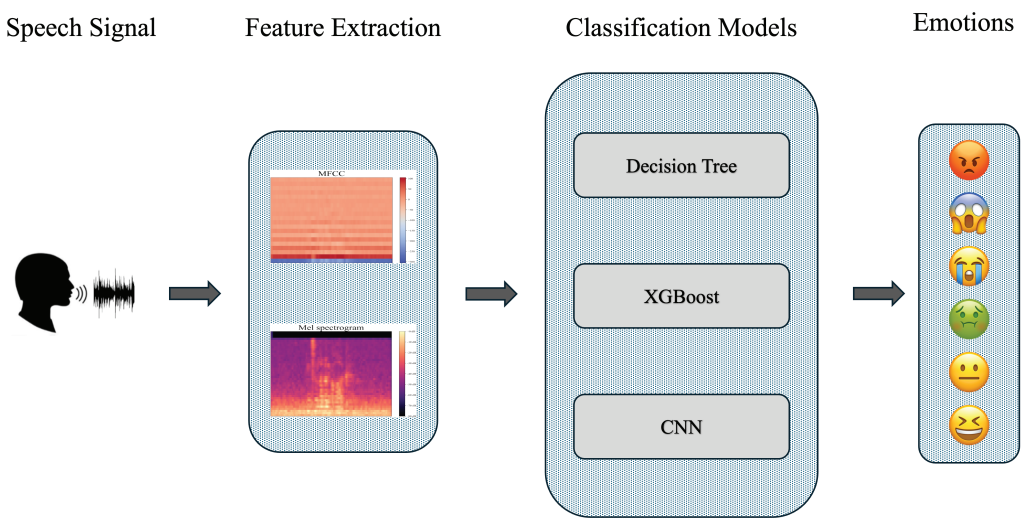

Figure 3.1 Emotion classification model using speech data

본 연구에서는 음성 데이터를 활용해 화자의 감정을 분류하는 다양한 모형을 고려하였다. 먼저, 음성 처리에 널리 사용되는 기법인 MFCC와 Mel-spectrogram을 활용하여 음성신호를 전처리하고, 여기에 기본적인 머신러닝 및 딥러닝 모델들을 적용한 후 성능을 평가하였다. 해당 과정은 Figure 3.1를 통해 확인할 수 있다. 음성 데이터를 Python의 librosa 라이브러리를 통해 쉽게 MFCC를 추출할 수 있다. 추출된 MFCC를 시간 방향에 따라 평균 값을 계산하여 하나의 음성 데이터를 32개의 Coefficient로 표현할 수 있고, 이를 Decision Tree와 XGBoost를 통해 6가지 감정으로 분류하였다. Decision Tree는 모델에 초기 설정된 hyperparameter를 통해 실험을 진행하였고, XGBoost 모델의 주요 hyperparameter로는 생성할 트리의 개수인 n_estimators, 각 트리의 학습률을 조절하는 learning_rate, 트리의 최대 깊이를 설정하는 max_depth 등이 있고, 각각 1000, 0.1, 3으로 설정하였다. 마찬가지로, Mel-spectrogram도 librosa 라이브러리를 통해 추출할 수 있다. 추출된 Mel-spectrogram를 통해 음성 데이터를 이미지로 변형시켜 이를 ResNet의 입력으로 사용하여 모델을 학습하고 감정을 분류하였다.

또한, 대규모 사전 학습된 모델을 활용함으로써, 대규모 데이터셋에서 학습된 일반화된 특징을 음성 감정 인식 작업에 적용한다. Figure 3.2은 사전 학습 모형 활용시의 모형에 대해 기술하였다.

사전 학습된 모델들은 허깅페이스 (Hugging face)라는 플랫폼을 통해 쉽게 접근하고 사용할 수 있다 (Wolf 등, 2019). 허깅페이스는 다양한 자연어 처리 (NLP) 및 음성 인식 모델을 호스팅하는 오픈 소스 라이브러리 및 커뮤니티로, 연구자들과 개발자들에게 사전 학습된 모델을 쉽게 다운로드하여 특정 task에 맞게 미세 조정할 수 있는 환경을 제공한다. 이를 통해 연구자들은 복잡한 모델 학습 과정 없이도 고성능의 음성 인식 및 처리 모델을 빠르게 개발하고 적용할 수 있게 된다.

허깅페이스의 AutoFeatureExtractor와 AutoModelForAudioClassification 클래스를 사용하면, 선택한 사전 학습 모델에 맞는 feature 추출기와 분류 모델을 자동으로 로드할 수 있다. feature 추출기는 음성 데이터를 입력으로 받아 모델이 이해할 수 있는 형식으로 변환하며, 분류 모델은 사전 학습된 가

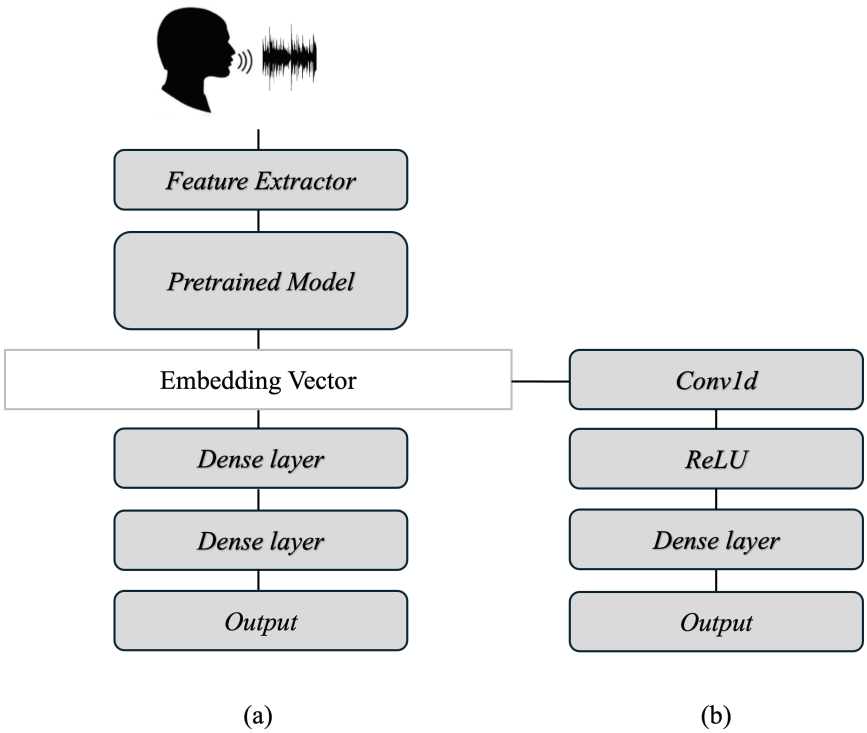

Figure 3.2 The process of emotion classification of speech data through pre-trained models

중치를 사용하여 높은 성능을 보장한다. Figure 3.2의 (*a*) 절차에 따라 연구자는 허깅페이스에서 사전 학습된 모델을 선택하여 오디오 분류 작업에 사용할 수 있으며, 목표로 하는 task에 맞게 마지막 Dense layer를 수정하여 6개의 감정 카테고리 중 하나로 출력할 수 있도록 한다.

기존의 분류 모델은 1024개의 임베딩 벡터를 두 개의 Dense layer를 통과시켜 6가지 감정으로 분류하는데, 본 연구에서는 임베딩 벡터를 추가로 활용하고자 1D CNN을 거쳐서 분류하는 방법을 사용했다. Figure 3.2의 (*b*) 절차대로 추출된 1024개의 feature를 입력으로 받아 kernel size가 5, stride가 1, padding이 2인 Conv1d layer를 통과하여 256개의 feature 맵을 생성하고 ReLu 활성화 함수를 적용하여 비선형성을 추가한 후, 6개의 감정 클래스로 분류하는 Dense layer를 거치도록 하였다. 이를 통해 대규모 사전 학습된 모델을 통한 고차원 벡터에서 추가적으로 중요한 정보를 추출하고 이를 최종 분류 작업에 반영하여 성능 향상을 기대할 수 있다.

마지막으로, 사전 학습된 모델들 간 예측 성능이 유사한 경우, 예측값 불일치 비율을 분석하여 모델들이 각각 서로 다른 특성을 학습하고 있는지 확인한다. 예측값 불일치 비율이 높을 경우, 이를 바탕으로 하드 보팅 앙상블 기법을 적용하여 모델 간의 보완적 특성을 활용하고, 이를 통해 전체 모델의 성능 향상을 기대할 수 있는 방법론을 제안한다.

4. 실험

4.1. 실험 설정

본 연구에 사용된 모든 모델은 NVIDIA Quadro RTX 8000 (48GB) GPU와 INTEL i9-9900K CPU를 기반으로 하여, PyTorch 1.13.1 및 CUDA 11.7을 사용하여 훈련되었다. 실험에 사용된 모든 오디오 데이터는 16000Hz의 샘플링 레이트로 불러와서 전처리되었고, 최적화를 위해 Adam 옵티마이저를 적용하였으며, 학습 과정에서 배치 사이즈를 16으로 설정하였다. 총 50회의 Epoch 동안 가장 높은 성능을 보인 모델을 최종 모델 선정하였고, 모델 성능 평가에는 정확도 (Accuracy) 지표를 사용하여 모델이 음성 데이터 속 화자의 감정을 얼마나 정확하게 분류하는지를 측정하였다. 연구에 사용된 사전 학습된 모델은 'wav2vec2-large-xlsr-53-english', 'wavlm-large', 'hubert-large-ls960-ft'이다.

4.2. 데이터셋

본 연구에 사용된 데이터셋은 '음성 감정 인식 AI 경진대회 월간 데이콘'에서 제공된 것으로, 주로 1초에서 5초 사이의 길이를 가진 영어 음성 데이터로 구성되어 있다. 이 데이터셋은 다양한 환경에서 녹음된 음성 샘플들을 포함하고 있으며, 각 음성 샘플은 분노, 두려움, 슬픔, 싫증, 중립, 행복 등 6가지 감정 유형으로 분류된다. 총 5001개의 라벨이 지정된 학습 데이터를 통해 학습을 한 후, 1881개의 라벨이 없는 테스트 데이터을 올바르게 분류하는 것이 해당 대회의 목표였다. 본 연구를 통해 이러한 다양한 감정 상태를 포함한 음성 데이터를 통해 모델이 얼마나 정확하게 감정을 인식하여 분류할 수 있는지를 평가하고자 하였다.

4.3. 시간 및 메모리 사용량 비교

각 모델의 학습 시간과 메모리 사용량을 비교하여, 각 방법론이 요구하는 자원이 어느 정도인지 평가하였다.

Table 4.1 Training time and memory usage comparison

Model	Training Time	Memory Usage
MFCC + DT	10 minutes	Very low
MFCC + XGBoost	20 minutes	Very low
Mel-Spec + CNN	50 minutes	2GB
Mel-Spec + CNN + LSTM	55 minutes	2GB
HuBERT	226 minutes	17GB
Wav2Vec 2.0	239 minutes	17GB
WavLM	237.5 minutes	17GB

Table 4.1을 보면, MFCC와 Decision Tree, XGBoost와 같은 머신러닝 모델 조합은 처리 시간이 매우 짧고 메모리 사용량도 매우 낮아 적은 자원을 필요로 한다. Mel-Spectrogram을 사용한 ResNet 모델과 LSTM까지 결합한 모델은 50 에폭 기준 약 1시간 소요되며, 2GB의 메모리를 사용한다. 반면에, 세 가지 사전 학습된 모델은 각각 학습 시간이 에폭 당 약 5배 정도 더 걸려 50 에폭 기준 약 4시간 동안 학습한다. 즉, 앞선 방법들에 비해 길고, 메모리 사용량도 크게 요구된다. 이러한 모델들은 대규모 데이터셋을 이용하여 미리 학습되었기에 많은 수의 파라미터를 가지고 있으며, 이러한 파라미터들을 처리하고 최적화하기 위해서는 보다 많은 계산 자원을 필요하다. 따라서, 자원이 제한된 상황에서는 전통적인 모델들이 더 적합할 수 있으며, 자원 사용에 여유가 있는 경우에는 사전 학습된 모델들을 사용하는 것이 성능 면에서 유리할 수 있다.

5. 실험결과

5.1. 단일 모델 성능 비교

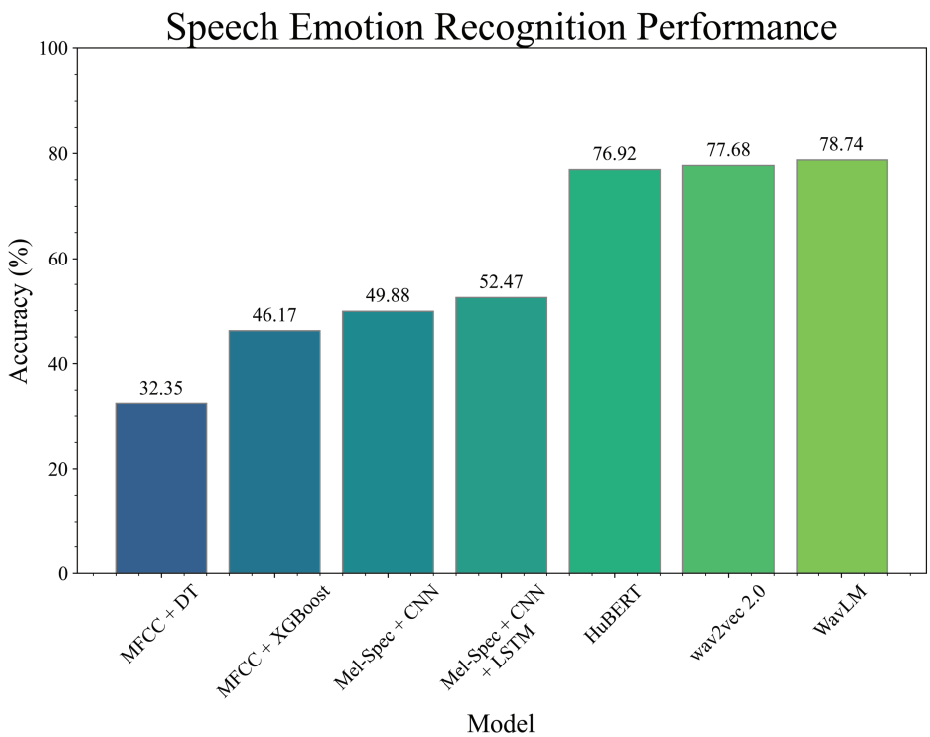

Figure 5.1 Comparison of single model performance

Figure 5.1은 기본적인 머신러닝 및 딥러닝 모델 (MFCC + DT, MFCC + XGBoost, Mel-spectrogram + ResNet, Mel-spectrogram + ResNet + Bi-LSTM)과 사전학습 모델의 성능 비교 결과를 보여준다. MFCC feature와 결정 트리 알고리즘을 결합한 방식의 Accuracy는 32.35%로 나타났다. 결정 트리는 비교적 간단한 머신러닝 모델로, 복잡한 음성 데이터를 처리하에는 한계가 있다. 마찬가지로 MFCC feature를 활용하여 더 발전된 머신러닝 모델인 XGBoost를 적용한 결과, 46.17%의 정확도를 달성했다. XGBoost를 사용한 것이 결정 트리를 사용한 것보다 성능이 개선되었지만, 여전히 절반에 미치지 못하는 Accuracy를 달성했다. 더 나아가, Mel-spectrogram feature를 활용한 ResNet의 경우 49.88%의 정확도를 기록했고, Bi-LSTM까지 결합한 모델은 52.47%의 성능을 기록하였다. 이는 Bi-LSTM의 결합을 통해 문맥 정보를 모델이 추가로 학습하여 3% 정도 성능 향상을 이끌어냈고, 효과가 있는 결합임을 확인할 수 있었다. 그러나, 이전의 머신러닝 기법들보다는 성능이 개선되었지만, 여전히 아쉬운 성능을 달성하였다.

사전 학습 모델을 활용한 결과를 보면 성능이 크게 향상되었다. 'hubert-large-ls960-ft' 모델은 76.92%의 정확도를 달성했고, 'wav2vec2-large-xlsr-53-english' 모델은 77.68%의 정확도를 기록했으며, 'wavlm-large' 모델은 78.74%로 가장 높은 성능을 보였다. 본 대회의 경우 라벨이 지정된 학습 데이터의 개수가 5001개로 매우 적은 편이다. 따라서 사전학습모형의 사용 없이 모형을 데이터 만으로

학습시키면 좋은 성능을 달성할 수 없었고, 대량의 음성 데이터로 사전학습된 HuBERT, Wav2vec 2.0, WavLM를 기반으로 하는 모델들을 활용했을 때 큰 폭의 성능 향상을 보였다.

5.2. 추가 레이어 결합 성능 비교

Table 5.1 Performance comparison with additional 1D CNN

Number	Models	Accuracy (%)
1	HuBERT	76.92
2	Wav2vec 2.0	77.68
3	WavLM	78.74
4	HuBERT + 1D CNN	78.06
5	Wav2vec2.0 + 1D CNN	76.39
6	WavLM + 1D CNN	**79.80**

사전 학습된 모델을 활용하여 충분한 성능을 확인하였지만, 1D CNN을 추가한 모델의 성능을 Table 5.1에 제시하였다. 본 실험에서는 사전 학습된 모델만을 사용했을 때와 1D CNN을 통합하여 사용했을 때의 성능 차이를 비교하였다. 이를 통해 감정 분류 작업에서 1D CNN의 추가적인 성능 향상을 기대하였다.

1D CNN을 추가한 모델들은 사전 학습된 모델만을 사용한 경우에 비해 성능이 일부 향상되거나, 약간의 성능 저하를 보였다. Hubert와 WavLM 모델의 경우, 성능이 약 1% 정도 상승하였고, 특히 WavLM 모델에 1D CNN 레이어를 추가했을 때 본 연구에서의 단일 모델로는 가장 높은 성능인 79.80%의 정확도를 기록하였다. 이는 1D CNN 레이어가 특정 모델과의 조합에서 성능 향상에 기여할 수 있음을 시사한다. 반면, Wav2vec 2.0 모델의 경우 1D CNN 레이어를 추가했을 때 성능이 약간 감소하는 모습을 보였다. 이러한 결과는 1D CNN을 추가하는 것이 모든 모델에서 동일한 성능 향상을 보장하지는 않지만, 모델에 따라 긍정적인 영향을 미칠 수 있음을 보여준다. 따라서, 모델의 특성과 작업에 따라 최적의 결합을 찾는 것이 중요하다.

5.3. 앙상블 학습 성능 비교

Table 5.1을 확인해보면, 모델들의 예측 정확도가 대체로 76%에서 80% 사이로 나타나, 모든 모델이 비슷한 수준의 성능을 보이는 것으로 관찰되었다. 그러나, Figure 5.2를 확인해보면 모델 간 예측값 불일치 비율이 상당히 높게 나타났다. 이러한 불일치는 모델들이 서로 다른 특성을 학습하고 있고, 이로 인해 특정 상황에서는 한 모델이 다른 모델보다 더 정확한 예측을 제공할 수 있다고 보여진다.

Table 5.2 F1 score comparison of individual models and ensemble model on validation set

Emotion	WavLM	Wav2Vec 2.0	HuBERT	Ensemble
Anger	0.84	0.86	0.86	0.86
Fear	0.70	0.69	0.69	0.73
Sadness	0.68	0.69	0.69	0.70
Disgust	0.73	0.74	0.74	0.77
Neutral	0.85	0.85	0.85	0.86
Happiness	0.85	0.82	0.82	0.84
Overall Accuracy	0.78	0.78	0.78	0.79
Macro Average	0.78	0.78	0.78	0.79
Weighted Average	0.78	0.77	0.77	0.79

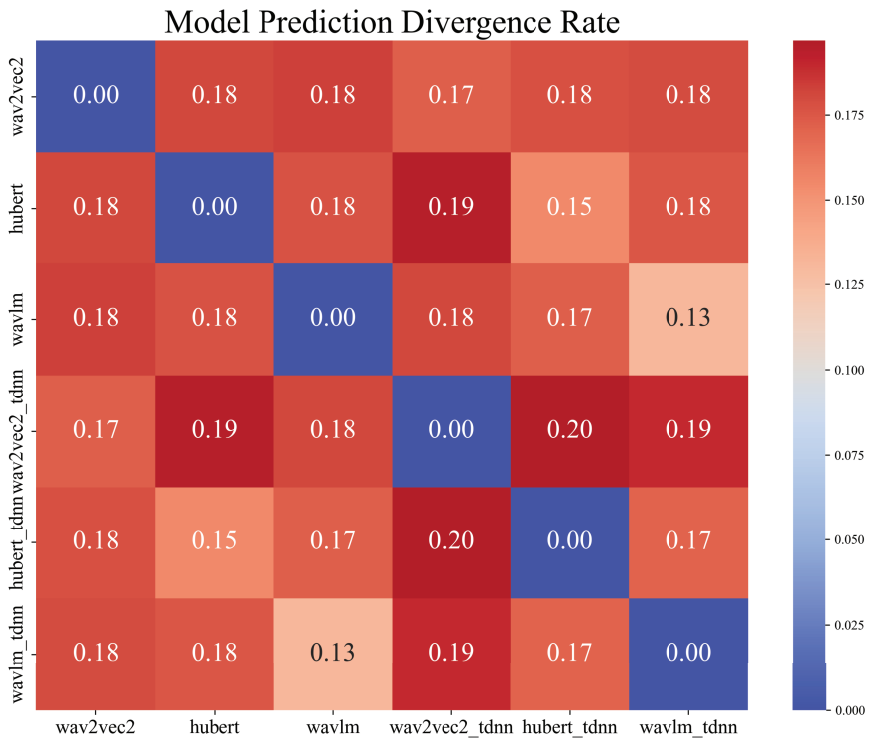

Figure 5.2 Model prediction divergence rate

Table 5.3 Performance comparison of ensemble learning

Number	Models	Accuracy (%)
1	Wav2vec 2.0 + HuBERT + WavLM	79.80
2	Wav2vec 2.0 + HuBERT + WavLM combined with 1D CNN	80.41
3	All models	**80.79**

모델별로 감정을 어떻게 예측했는지 확인하기 위해서 학습에 사용되지 않은 1001개의 검증 데이터를을 이용하여 세 개의 사전 학습된 모델에 대해 각각 F1 score와 하드 보팅 앙상블 기법 사용시의 F1 score를 Table 5.2에 기술하였다. 각 모델별로 감정 예측 성능에서 차이가 나타났는데, 예를 들어, 분노와 두려움 감정에 대해서는 Wav2Vec 2.0과 HuBERT가 더 높은 F1 score를 보였고, 반면에 중립 감정에 대해서는 세 모델 모두 유사한 성능을 보였다. 이를 앙상블한 결과가 개별 모델보다 더 좋은 성능을 보였으며, 특히 두려움과 혐오 감정에서의 F1 score가 눈에 띄게 향상되었다. 이를 통해, 다양한 모델의 예측을 결합하는 앙상블 기법이 모델 간의 상호 보완적인 특성을 활용하여 예측 성능을 향상시킬 수 있음을 확인하였다.

Table 5.3을 확인해보면, 모든 실험에서 단일 모델을 사용했을 때보다 하드 보팅 앙상블 기법을 사용했을 때 정확도가 향상되었다. 이는 다양한 모델들의 예측을 결합함으로써 각 모델의 한계를 보완하고, 더 강력한 예측 성능을 달성할 수 있음을 시사한다. 1D CNN을 결합하지 않은 모델끼리 하드 보팅을 진행한 결과, 79.80%의 정확도를 얻을 수 있었다. 이는 WavLM에 1D CNN을 결합한 모델의 성

능과 동일하다. 사전 학습된 모델에 1D CNN를 결합한 모델들끼리 하드 보팅을 진행한 결과, 성능이 80.41%로 1D CNN이 없는 모델들보다 예측 성능이 좋았고, 1D CNN를 추가한 것이 의미가 있었음을 나타낸다. 마지막으로, 1D CNN을 포함한 모델과 포함하지 않은 모델을 모두 결합하여 하드 보팅 앙상블 학습을 한 결과, 본 연구를 통해 달성한 최고 성능인 80.79%의 정확도를 달성하였다. 이는 다른 특성을 가진 모델들을 조합함으로써, 각 모델의 장점을 최대한 활용한 것으로 보인다. 따라서, 사전 학습 모델마다 음성의 특성을 다양하게 학습하는 것으로 보이므로 다양한 모델을 선택하여 앙상블하는 것이 예측 성능을 향상시키는 효과적인 방법일 수 있다. 이는 음성 감정 분류와 같은 작업에서 모델의 다양성을 활용하여 높은 성능을 달성하는데 중요한 역할을 할 수 있다.

5.4. 대회 상위 성능 방법론과 비교

Table 5.4 Performance comparison of top 4 methods

Rank	Model	Method	Accuracy (%)
1	Hubert_emotion	Classification Layer Modification	87.85
2	10 pretrained models	Stacking	82.92
3	hubert-large-ll60k	F1 score	81.85
4	wav2vec2-large-xlsr-53-english	Data augmentation	81.78
5	**Ours**	**Ours**	**80.79**

Table 5.4를 통해 확인해보면, 대회의 상위 랭킹을 달성한 참가자들 모두 대규모 오디오 데이터를 사전 학습한 모델들을 활용하여 높은 성능을 달성하였다. 1위를 달성한 참가자는 'Hubert_emotion' 모델을 활용하여 Classification layer 수정을 통해 87.85%의 정확도를 달성하였고, 2위를 차지한 참가자는 10가지 대규모 사전 학습 모델을 활용해 앙상블 기법인 Stacking을 통해 82.92%의 성능을 달성하는 등 최신 기술인 사전 학습 모델의 도입은 음성 감정 인식 및 분류의 성능을 매우 향상시켰다. 대회에서 상위 랭킹을 달성한 방법론과 비교하여, 본 연구에서도 후속 연구를 사전 학습된 모델을 활용하여 유사한 수준의 성능을 달성할 수 있었다. 이러한 결과를 통해 기존의 전통적인 음성 처리 기법보다 사전 학습된 모델을 활용한 접근 방식이 효과적임을 확인할 수 있었다.

6. 결론

본 연구에서는 음성 데이터를 통해 화자의 감정을 분류하는 문제에 대해 다루었다. 연구를 위해 초기 단계에서는 MFCC와 Mel-Spectrogram과 같은 음성에서 추출한 feature를 바탕으로 기본적인 머신러닝 및 딥러닝 모델들을 적용하여 성능을 평가하였다. 다음으로, 다양한 사전 학습된 모델을 선택하여 대회의 목표에 맞게 출력 레이어를 수정하여 연구에 사용하였고, 그 중 WavLM 기반의 'wavlm-large' 모델이 사용한 모델이 78.74%로 가장 높은 성능을 보였다. 사전 학습된 모델들이 상대적으로 높은 정확도를 달성함으로써, 대규모 데이터셋을 통한 사전 학습의 중요성을 확인할 수 있었다. 위의 과정을 통해 '월간 데이콘 음성 감정 인식 AI 경진대회'에 참여하는 동안, 총 431팀 중 24등을 차지하였다.

대회 기간이 지나고, 추가로 성능 향상을 이루고자 사전 학습된 분류 모델의 마지막 Dense layer 전에 1D CNN 레이어를 추가한 모델의 성능을 평가하여, 특정 모델과의 조합에서 성능 향상을 관찰할 수 있었다. 예를 들어, 'wavlm-large' 모델에 1D CNN을 추가한 모델의 성능이 기존의 성능보다 약 1% 높은 79.80%를 달성하였다. 이는 새로 추가된 레이어가 음성 데이터의 시간적 특성을 더 잘 포착하여 성능 향상에 기여할 수 있음을 시사한다. 그러나 모든 모델에서 일관된 성능 향상을 보장하지는 않아, 모델의 특성과 고려한 최적의 조합을 찾는 것이 중요함을 알 수 있었다.

또한, 사전 학습된 모델을 사용한 모델 간의 예측 정확도가 비슷한 수준임에도 불구하고, 예측한 레이블의 불일치 비율이 높은 문제점을 발견하였다. 이를 해결하기 위해, 여러 모델의 예측 결과를 결합하는 앙상블 기법 중 하나인 하드 보팅을 적용하여 성능 향상을 이끌어냈고, 이는 각 모델의 장점을 통합하고 한계를 보완함으로써, 더욱 뛰어난 예측 성능을 달성할 수 있음을 보여준다. 연구에 사용된 모든 사전 학습된 모델을 하드 보팅한 결과, 정확도 80.79%를 달성하여 해당 대회에서 5등에 준하는 정확도를 달성할 수 있었다.

결론적으로, 본 연구는 음성 데이터의 감정 분류 작업에 있어 사전 학습된 모델의 중요성을 강조하며, 사전 학습된 모델을 통해 추출된 임베딩 벡터를 활용하여 새로운 레이어를 추가한 것이 특정 조건에서 성능 향상에 기여할 수 있음을 보여주었다. 또한, 다양한 모델을 조합하는 앙상블 기법이 예측 성능을 높이는 유효한 방법임을 확인하였다. 이러한 발견은 음성 데이터를 활용하여 감정 분류 및 인식 뿐만 아니라 다양한 응용 분야에서 모델 선택 및 설계 방법에 중요한 지침을 제공할 것으로 기대된다.

References

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **24**, 123-140.

Baevski, A., Zhou, Y., Mohamed, A. and Auli, M. (2020). Wav2vec2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, **33**, 12449-12460.

Chakraborty, K., Talele, A. and Upadhya, S. (2014). Voice recognition using MFCC algorithm. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, **1**, 2349-2163.

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

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785-794.

Deb, S. D., Jha, R. K., Jha, K. and Tripathi, P. S. (2022). A multi model ensemble based deep convolution neural network structure for detection of COVID19. *Biomedical Signal Processing and Control*, **71**, 103126.

Dietterich, T. G. (2000). Ensemble methods in machine learning, *International Workshop on Multiple Classifier Systems*, 1-15. Springer, Berlin, Heidelberg.

El Ayadi, M., Kamel, M. S. and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, **44**, 572-587.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Icml*, **96**, 148-156.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778.

Heo, J. H., I, S. B., Yang, W. H. and Lim, D. H. (2021). Transfer learning-based ensemble deep learning for image classification of COVID-19 patients. *Journal of the Korean Data & Information Science Society*, **32**, 1219-1235.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R. and 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.

Huang, Z., Xu, W. and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv Preprint arXiv*:1508.01991.

Kang, T. and Kwak, I. (2023). A two-stage training approach for voice spoofing detection. *Journal of the Korean Data & Information Science Society*, **34**, 203-214.

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H. and Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, **7**, 117327-117345.

Nimmi, K., Janet, B., Selvan, A. K. and Sivakumaran, N. (2022). Pre-trained ensemble model for identification of emotion during COVID-19 based on emergency response support system dataset. *Applied Soft Computing*, **122**, 108842.

Pandey, S. K., Shekhawat, H. S. and Prasanna, S. M. (2019, April). Deep learning techniques for speech emotion recognition: A review. *In 2019 29th International Conference Radioelektronika (RADIOELEK-*

TRONIKA), 1-6. IEEE.

Parhami, B. (1994). Voting algorithms. *IEEE Transactions on Reliability*, **43**, 617-629.

Song, Y. Y. and Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, **27**, 130.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... and Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv Preprint arXiv*:1910.03771.

Yadav, A. and Vishwakarma, D. K. (2020). A multilingual framework of CNN and bi-LSTM for emotion classification. *In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1-6. IEEE.

Journal of the Korean Data & Information Science Society
2024, **35**(4), 445–459

http://dx.doi.org/10.7465/jkdi.2024.35.4.445
한국데이터정보과학회지

Pre-trained models and ensemble technique for speech emotion recognition[†]

Jaejin Seo[1] · Taein Kang[2] · Il-Youp Kwak[3]

[123]Department of Statistics and Data Science, Chung-Ang University

Received 29 May 2024, revised 14 June 2024, accepted 17 June 2024

Abstract

Research on speech emotion recognition plays a crucial role in enhancing human-machine interaction and improving efficiency and user experience in various fields such as healthcare, education, and customer service. In this study, we aimed to develop an AI model to classify six emotions by participating in DACON's 'Monthly DACON Speech Emotion Recognition AI Competition'. We compared the performance using traditional speech processing techniques and pretrained models, and investigated the potential for additional learning using embedding vectors effectively learned through pretrained models. As a result, a model combining WavLM with 1D CNN demonstrated superior performance at 79.80%, and by ensembling all pretrained models using hard voting, we further improved performance to 80.79%, achieving a ranking equivalent to 5th place in the competition. This research is expected to contribute to the application potential of speech emotion recognition technology, enabling the utilization of emotion recognition models in various real-world applications.

Keywords: Deep learning, ensemble learning, pre-trained models, speech emotion recognition.

[†] This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (RS-2023-00208284). This research was supported by the Chung-Ang University Graduate Research Scholarship in 2022.

[1] Graduate student, Department of Statistics and Data Science, Chung-Ang University, Seoul 06974, Korea.

[2] Graduate student, Department of Statistics and Data Science, Chung-Ang University, Seoul 06974, Korea.

[3] Corresponding author: Associate professor, Department of Statistics and Data Science, Chung-Ang University, Seoul 06974, Korea. E-mail: ikwak2@cau.ac.kr