멀티모달 특징 결합을 통한 감정인식 연구

김성식[1], 양진환[2], 최혁순[2], 고준혁[2], 문남미[3]
[1]호서대학교 컴퓨터공학부 학부생
[2]호서대학교 컴퓨터공학과 석사과정
[3]호서대학교 컴퓨터공학부 교수
sungsik001004@gmail.com, yjh970706@naver.com, hyuksoon2001@gmail.com,
junhyeok970306@gmail.com, nammee.moon@gmail.com

The Research on Emotion Recognition through Multimodal Feature Combination

Sung-Sik Kim[1], Jin-Hwan Yang[1], Hyuk-Soon Choi[1],
Jun-Heok Go[1], Nammee Moon[1]
[1]Dept. of Computer Science, Hoseo University

요 약

본 연구에서는 음성과 텍스트라는 두 가지 모달리티의 데이터를 효과적으로 결합함으로써, 감정 분류의 정확도를 향상시키는 새로운 멀티모달 모델 학습 방법을 제안한다. 이를 위해 음성 데이터로부터 HuBERT 및 MFCC(Mel-Frequency Cepstral Coefficients)기법을 통해 추출한 특징 벡터와 텍스트 데이터로부터 RoBERTa를 통해 추출한 특징 벡터를 결합하여 감정을 분류한다. 실험 결과, 제안한 멀티모달 모델은 F1-Score 92.30으로 유니모달 접근 방식에 비해 우수한 성능 향상을 보였다.

1. 서론

최근 인간-컴퓨터 상호작용 기술의 발전에 따라 사용자의 감정 상태를 인지하고 적절한 피드백을 제공하는 분야인 감성 컴퓨팅이 다양한 산업 분야에서 주목받으며 그 연구가 활발히 이루어지고 있다. 사람의 의사 표현은 음성, 언어, 얼굴 표현 등의 다양한 모달리티로 나타난다[1].

본 연구는 음성 데이터로부터 HuBERT 및 MFCC 기법을 통해 추출한 특징 벡터와 텍스트 데이터로부터 RoBERTa를 통해 추출한 특징 벡터를 결합하여 분류하는 멀티모달 모델 학습 방법을 제안한다. 이러한 접근 방법은 각 모달리티에서 추출된 특징 정보를 상호 보완적으로 활용하여, 감정 인식 모델의 성능을 향상시키는 것을 목표로 한다.

2. 특징추출

2.1 HuBERT 특징 추출

HuBERT는 음성의 원시 파형 데이터에서 복잡한 언어적 및 비언어적 특성을 추출하는 자기지도학습 모델이다[2]. 이 모델은 음성 데이터의 복잡한 특성을 이해하고 분석하기 위해 발화의 리듬, 강세 및 톤과 같은 비언어적 요소도 포착한다. 본 연구에서는 HuBERT를 사용하여 감정인식에 필요한 음성 특징을 추출하였으며, 각각의 음성 요소가 감정 표현에 미치는 영향을 포괄적으로 포착하여, 감정 분류의 정확도를 향상시키는데 기여한다.

2.2 MFCC 특징 추출

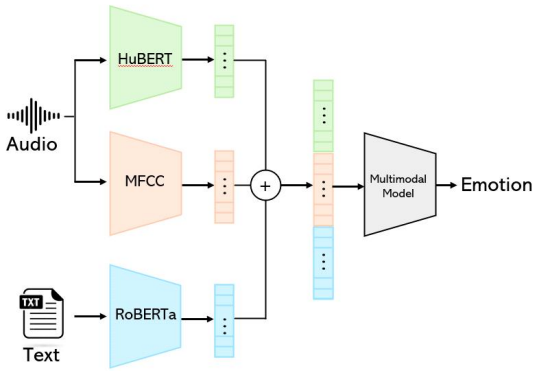
MFCC는 음성 인식 분야에서 널리 사용되는 특징 추출 방법 중 하나로, 인간의 청각 특성을 반영하여 설계된 주파수 스케일을 사용한다[3]. 특히, MFCC는 인간의 청각 특성을 반영하는 Mel 스케일을 사용하여 주파수 정보를 변환하며, 이를 통해 음성의 특성을 더 정확하게 추출할 수 있다. 본 연구에서는 MFCC를 사용하여 음성의 음색, 강도와 같은 음향학적 특성을 추출한다.

2.3 RoBERTa 특징 추출

RoBERTa는 BERT의 하이퍼파라미터 등을 조절하여 BERT의 성능보다 훨씬 더 나을 수 있음을 보여준 모델이다[4]. 본 연구에서는 RoBERTa 모델을 사용하여 텍스트 데이터로부터 언어적 의미와 문맥상의 정보를 추출한다. 이를 통해 감정 분류에 필요한 언어적 특징을 정밀하게 제공하며, 감정적으로 중요한 어휘와 문맥의 뉘앙스를 파악하는데 중요한 역할을 한다. 특히, 문장의 구성요소가 감정 표현에 미치는 영향을 세밀하게 분석함으로써 보다 정교한 감정 분류를 가능하게 한다.

3. 멀티모달 감정인식 모델

본 연구에서는 텍스트 데이터와 음성 데이터를 이용한 멀티모달 감정 인식 모델을 제안한다. 이 모델은 음성 데이터로부터 HuBERT와 MFCC 기법을 통해 특징 벡터를 추출하며, 텍스트 데이터로부터 RoBERTa를 통해 특징 벡터를 추출한다. 추출된 모든 특징은 Z-Score를 통해 표준화되며, 이후 concatenate되어 하나의 특징 벡터로 통합한다. 이 통합된 벡터는 멀티모달 모델을 거쳐 감정을 분류하는데 사용된다. 제안한 모델 구조는 (그림 1)과 같다.



(그림 1) 멀티모달 감정인식 모델 구조

4. 실험

4.1 실험환경

실험을 위해 사용된 데이터는 AI Hub에서 제공하는 '감정 분류를 위한 대화 음성 데이터셋'의 4, 5차년도 데이터를 병합한 총 43,989개의 데이터로 <표 1>과 같다.

<표 1> 데이터 셋

감정상태	개수
Sad	13,999
Angry	11,635
Disgust	4,659
Happy	4,548
Fear	4,131
Neutral	3,262
Surprise	1,755
합계	43,989

전체 데이터셋을 8:1:1의 비율로 설정하여 Train, Validation, Test 데이터셋으로 구성한다. 실험은 50epoch를 학습시키고, 학습 중 Validation 성능이 가장 우수한 모델을 최적 모델로 선정한다. 선정된 최적 모델을 사용하여 Test 데이터셋의 성능을 평가하였으며, 평가 지표로는 Accuracy와 F1-Score를 사용한다.

4.2 실험 결과

실험 결과 Test 데이터셋에 대한 Accuracy와 F1-Score는 <표 2>와 같다.

<표 2> 멀티모달 감정인식 모델 학습 결과

Model	Accuracy	F1-Score
HuBERT	87.56	87.24
RoBERTa	91.88	90.93
HuBERT+MFCC +RoBERTa	**92.90**	**92.30**

실험 결과, 유니모달 모델에 비해 HuBERT, MFCC, 그리고 RoBERTa로 특징을 추출하여 결합한 멀티모달 모델이 F1-Score 92.30으로 우수한 성능 향상을 보였다.

5. 결론

본 연구에서는 HuBERT와 RoBERTa로부터 얻은 언어적, 비언어적 특성과 음성 데이터의 MFCC 특성을 결합하여 감정을 분류하는 모델을 제작하였다. 실험 결과 F1-Score 92.30으로 유니모달 접근 방식에 비해 우수한 성능 향상을 보였다. 이 결과는 멀티모달 데이터의 효과적인 통합이 모델의 성능을 개선하는 데 기여할 수 있음을 보여준다. 향후 연구에서는 멀티모달 데이터 통합 방법의 최적화를 통해, 감정 분류 모델의 성능을 더욱 향상시키는 방안을 모색할 것이다.

ACKNOWLEDGMENT

본 연구는 과학기술정부통신부와 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음 (No. 2019-0-01834).

참고문헌

[1] 유지현. (2022). 중간 융합 모듈을 사용한 트랜스포머 기반의 멀티모달 감정인식 네트워크.

[2] Wei-Ning Hsu, et al. “HuBERT: Self-Supervised Speech Representation Learning by MaskedPrediction of Hidden Units”, arXiv preprint, arXiv:2106.07447, 2021.

[3]박정현, et al. "음성데이터 증강을 통한 3D 특징 벡터 기반 신생아 울음소리 분류." 한국컴퓨터정보학회논문지 28.9 (2023): 47-54.

[4] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692, 2019.