The Journal of the Convergence on Culture Technology (JCCT)
Vol. 10, No. 1, pp.617-623, January 31, 2024. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2024.10.1.617

JCCT 2024-1-76

통합 CNN, LSTM, 및 BERT 모델 기반의 음성 및 텍스트 다중 모달 감정 인식 연구

Enhancing Multimodal Emotion Recognition in Speech and Text with Integrated CNN, LSTM, and BERT Models

에드워드 카야디*, 한스 나타니엘 하디 수실로**, 송미화***

Edward Dwijayanto Cahyadi*, Hans Nathaniel Hadi Soesilo, Mi-Hwa Song*****

요 약 언어와 감정 사이의 복잡한 관계의 특징을 보이며, 우리의 말을 통해 감정을 식별하는 것은 중요한 과제로 인식된다. 이 연구는 음성 및 텍스트 데이터를 모두 포함하는 다중 모드 분류 작업을 통해 음성 언어의 감정을 식별하기 위해 속성 엔지니어링을 사용하여 이러한 과제를 해결하는 것을 목표로 한다. CNN(Convolutional Neural Networks)과 LSTM(Long Short-Term Memory)이라는 두 가지 분류기를 BERT 기반 사전 훈련된 모델과 통합하여 평가하였다. 논문에서 평가는 다양한 실험 설정 전반에 걸쳐 다양한 성능 지표(정확도, F-점수, 정밀도 및 재현율)를 다룬다. 이번 연구 결과는 텍스트와 음성 데이터 모두에서 감정을 정확하게 식별하는 두 모델의 뛰어난 능력을 보인다.

주요어 : 음성 감정 인식, CNN, LSTM, BERT, 다중 모달 감정 인식, 딥 러닝

Abstract Identifying emotions through speech poses a significant challenge due to the complex relationship between language and emotions. Our paper aims to take on this challenge by employing feature engineering to identify emotions in speech through a multimodal classification task involving both speech and text data. We evaluated two classifiers—Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM)—both integrated with a BERT-based pre-trained model. Our assessment covers various performance metrics (accuracy, F-score, precision, and recall) across different experimental setups). The findings highlight the impressive proficiency of two models in accurately discerning emotions from both text and speech data.

Key words : Speech Emotion Recognition, CNN, LSTM, BERT, Multimodal Emotion Recognition, Deep Learning

Ⅰ. 서 론

감정은 인간 커뮤니케이션의 핵심 요소로, 우리의 상호작용, 관계 등에 큰 영향을 미친다. 감정을 인식하는 능력은 인간에게만 중요한 것이 아니라, 가상 비서 및 정신 건강 지원 시스템과 같은 AI 기반 기술 개발에도 필수적이다. 최근 딥러닝 알고리즘의 발전으로 이미지 분류, 음성 인식, 텍스트-투-스피치 합성과 같은 다양

* 세명대학교 정보통신학부 학부연구생 (제1저자)
** 서강대학교 컴퓨터공학과 학부연구생 (참여자)
***정회원, 세명대학교 스마트IT학부 부교수 (교신저자)
접수일: 2023년 10월 20일, 수정완료일: 2023년 11월 14일
게재확정일: 2023년 12월 10일

Received: October 20, 2023 / Revised: November 13, 2023
Accepted: December 10, 2023
***Corresponding Author : mhsong@semyung.ac.kr
School of Smart IT, Semyung University, Korea

한 도전과제를 해결하고 있으며, 트랜스포머 기반 시스템도 이미지 및 텍스트 생성 작업에 많이 사용된다. 이러한 기술적 기반을 바탕으로, 우리는 텍스트 및 음성 데이터로부터 감정을 인식하는 다중 모달 머신러닝 시스템을 개발했다. 본 논문에서는 “Multimodal Speech Emotion Recognition and Ambiguity Resolution”[1]에 게재된 연구 방법을 참고하여 음성의 특징을 추출한다. 그 후, TER 모델의 출력과 결합하여 딥 뉴럴 네트워크 기반 모델에 입력하여 학습시킨다. 구체적으로, 감정 인식 작업에 특화되도록 데이터셋을 사용하여 BERT 기반 모델을 먼저 미세 조정한다. 그런 다음 CNN[2] 및 LSTM[3] 분류기를 구축하여 주어진 특징을 바탕으로 감정을 해석한다. 두 분류 모델은 IEMOCAP 데이터셋[4]에서 평가된다.

이후 논문은 다음과 같이 구성된다. Ⅱ장에서는 관련 연구를 파악하고, 그 방법론에 대해 확장한다. Ⅲ장에서는 이 연구에 사용된 데이터 세트를 소개한다. Ⅳ장에서는 연구 방법을 탐구하며, 데이터 세트에 특징 추출 전에 적용된 구현 및 전처리 단계의 세부 사항을 살펴본다. 이어서 Ⅴ장에서는 실험 결과를 세심하게 제시한다. 마지막으로, 논문은 연구 결과를 요약하고 향후 연구 가능성에 대한 제안으로 결론을 맺는다.

Ⅱ. 관련연구

인간 감정을 인식하는 것은 감정이 표현될 수 있는 다양한 방식 때문에 도전적인 문제이다[5-10]. 최근에는 다중 모달 딥러닝 구조를 사용하여 감정을 분별하는 데 있어서 유망한 결과를 보여주고 있다. [11]에서 연구팀은 음성 및 텍스트 처리 방법을 설명하였으며, 이는 오디오 순환 인코더를 구축하는 것을 첫 단계로 포함한다. 다음 단계는 텍스트 순환 인코더(TRE)와 융합 모델을 구축하는 것이며, 각각 멀티모달 이중 순환 인코더(MDRE)와 멀티모달 이중 순환 인코더 주의 모델(MDREA)을 선택한다. CNN은 많은 연속적인 계층으로 구성된 체계적인 신경망의 한 유형이다[2]. 이 순차적 네트워크는 특징 추출 방법을 사용하여 입력의 추상적 모델을 생성한다. LSTM은 데이터의 순차적 패턴을 포착하고 장기 의존성을 처리하기 위해 설계된 순환 신경망(RNN) 아키텍처의 한 유형이다[3]. BERT는 자연어 처리에 특화된 머신러닝을 위한 오픈소스 도구이다[12]. 위키백과의 방대한 텍스트로 훈련을 시작하여 특정 질문 및 답변 세트를 사용하여 추가로 맞춤화할 수 있다. BERT는 'Bidirectional Encoder Representations from Transformers'의 약자이며, 트랜스포머 모델에 기반을 두고 있다. 시스템은 이러한 연결을 동적으로 미세 조정하는데, 이 과정을 자연어 처리에서 'Attention'라고 한다.

Ⅲ. 연구방법

1. 데이터 집합

본 연구에서는 남부 캘리포니아 대학교(USC)의 연구자들이 2008년에 처음 제공한 IEMOCAP 데이터셋[4]을 사용했다. 이 데이터셋은 다른 연사 10명 사이의 대화를 포함한 5개의 녹음 세션으로 구성되어 있으며, 총 12시간에 가까운 오디오-비주얼 데이터와 전사본을 포함한다. 데이터셋은 분노, 행복, 슬픔, 중립, 놀람, 두려움, 좌절, 흥분을 포함한 8가지 감정 범주로 풍부하게 주석이 달려 있다. 또한, 1에서 5까지의 척도로 지배, 활성화, 가치 수준을 나타내는 차원적 라벨을 제공한다. 데이터셋은 각 세션별로 개별 발화로 분할된다. 데이터셋을 분석하는 준비 과정에서, 발화 파일을 해당 시작 및 종료 타임스탬프에 따라 더 세분화하여 약 7,000개의 고유 오디오 파일을 생성했다. 이 오디오 파일들은 연구에서 특징 추출에 이용되었다. 또한, 미세 조정된 BERT 모델을 위해 Hugging Face를 통해 수집된 6가지 감정 라벨이 있는 14,696개의 영어 텍스트를 사용했다.

2. 모델 설계 및 구축

1) 음성 특징 추출

모델을 구축하는 첫 단계는 데이터 전처리를 수행하는 것으로, 데이터 세트에서 오디오 및 텍스트 특징을 추출한다. 사운드 특징 추출 방법은 'Multimodal Speech Emotion Recognition and Ambiguity Resolution' [1] 연구에서 제안한 방법을 활용한다. 오디오 특징의 경우, 피치(pitch), 하모닉스(harmonics), RMSE 정지(RMSE pause), 중심 모멘트(central moments)와 같은 주요 특징들을 추출한다. 피치(pitch)는 우리의 의사소통에서 중요한 요소로, 감정에 의해 영향을 받을 수 있는 성대에서 생성된 파형의 변화를 반영한다. 피치 신호를 평가하고 결정하기 위한 다양한

알고리즘이 개발되었다. 본 논문에서는 중심 클립된 프레임의 자기 상관에 기반한 가장 일반적인 방법을 사용한다. 식(1)에서 입력 신호 $y[n]$는 중심 클립되어 결과 신호 $y_{clipped}[n]$를 생성한다 :

$$y_{clipped}[n] = \begin{cases} y[n] - C_1, & \text{if } y[n] \geq C_1 \\ 0, & \text{if } |y[n]| < C_1 \\ y[n] + C_1, & \text{if } y[n] \leq C_1 \end{cases} \quad (1)$$

c_1로 표시된 값은 입력 신호 평균의 대략 절반이며 입력 신호는 본질적으로 이산적이다. 그 후, 이 결과 신호에 대해 자동 상관이 계산되며, 이는 정규화 목적을 위해 추가로 조정된다. 이 자기 상관 신호에서 가장 높은 값은 입력 신호 $y[n]$의 피치에 해당한다. 입력 신호에 중심 클리핑을 적용하면 자기 상관에서 더 뚜렷한 피크가 관찰된다. 다른 오디오 특징으로 넘어가서, 우리는 하모닉스에 집중한다. 분노 또는 스트레스를 받은 음성과 같이 감정이 고조된 경우, 피치와 함께 추가적인 여기 신호가 존재한다. 이러한 추가 신호들은 스펙트럼에서 하모닉스와 교차 하모닉스로 나타난다. 이러한 하모닉스를 식별하고 계산하기 위해, 중앙값 기반 필터를 사용한다. 초기에는 중앙값 필터는 식(2)와 같이 결정되는 l로 표시된 특정 창 크기를 사용하여 생성된다.

$$y[n] = median9x[n-k : n+k]k = (l-1/2) \quad (2)$$

값 l이 홀수인 경우, 정렬된 목록에서 중간값을 사용하여 중앙값을 계산한다. 그러나 l이 짝수인 경우에는 정렬된 목록의 중간에 위치한 두 값의 평균으로 중앙값을 계산한다. 이후, 이 중앙값 필터는 주어진 스펙트로그램의 h번째 주파수 조각 S_h 을 처리하는 데 사용되어 향상된 스펙트로그램 주파수 조각 H_i를 생성한다 (식3).

$$H_i = M(S_h, l_{harm}) \quad (3)$$

이러한 맥락에서, M은 미디어 필터를 나타내고, i는 i번째 시간 단계를 의미하며, l_{harm}은 하모닉 필터의 길이를 나타낸다. 음성 신호에서 에너지가 그 볼륨과 상관 관계가 있기 때문에, 이 특징은 특정 감정을 식별하는 데 활용될 수 있다. 평균 제곱근 에너지(Root Mean Square Energy)의 공식은 식 (4)와 같다.

$$E = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y[i]^2} \quad (4)$$

프레임별로 Root Mean Square Error (RMSE)를 계산하고, 평균과 표준편차를 특징으로 추출한다. 마지막으로, 정지(pause)와 중심 모멘트라는 두 가지 추가 특징을 포함한다. 정지(pause)의 개념은 오디오 신호 내의 침묵의 순간을 나타내는 데 사용된다. 이 지표는 우리의 감정 상태와 밀접하게 연결되어 있다; 예를 들어, 우리는 흥분할 때 빠르게 말하는 경향이 있다. 이 특징 값은 다음 식(5)와 같이 결정된다.

$$Pause = \Pr(y[n] < t) \quad (5)$$

여기에서 t는 신중하게 선택된 임계값으로, RMSE (Root Mean Square Error)의 0.4배에 해당한다. 여기서 E는 RMSE 자체를 나타낸다. 마지막으로, 신호의 진폭의 평균과 표준편차를 활용하여 입력 정보의 요약된 표현을 캡슐화한다.

2) 텍스트 처리

텍스트 특징에 대해서는, 원시 텍스트를 모델이 이해할 수 있는 수치 형태로 변환하기 위해 BERT 토큰화를 사용했다. 구글에 의해 개발된 BERT는 문장에서 단어의 맥락을 이해하기 위해 그 앞뒤의 단어들을 고려하는 딥러닝 모델이다. 텍스트를 수치 데이터로 변환하는 과정에는 토큰화, 임베딩, 풀링, 출력이 포함된다. 토큰화는 원시 텍스트를 토큰으로 분할하는 과정이다. 각 토큰은 하위 단어가 될 수 있으며, 이는 알려지지 않은 단어를 처리하고 어휘 크기를 줄이는 데 도움이 된다. BERT는 Word Piece라는 토크나이저를 사용한다. 예를 들어, "BERT helps in NLP"라는 문장은 ["BERT", "helps", "in", "N", "##LP"]와 같이 토큰화된다. 이어서 임베딩 과정에서, 각 토큰은 벡터로 변환된다. BERT는 각 토큰에 대한 초기 임베딩을 찾고, 단어의 순서를 인코딩하기 위해 위치 임베딩을 추가한다. 이 임베딩의 합은 BERT 모델의 입력을 나타낸다. BERT는 이러한 임베딩을 트랜스포머 인코더의 여러 계층을 통해 처리한다. 각 계층은 주의 메커니즘을 적용하여, 모델이 각 단어를 처리할 때 문장의 다른 부분에 집중할 수 있도록 한다. BERT의 Attention 메커니즘은 다음 공식으로 표현될 수 있다.

$$Atn(Q,K,V) = softmax(\frac{QK^t}{\sqrt{d_k}})V \quad (6)$$

식 (6)에서 Atn은 Attention을 나타내고, Q는 질의 행렬, K는 키 행렬, V는 값 행렬, 그리고 d_k는 키의 차원을 나타낸다. 이 공식은 모델이 출력을 생성할 때 맥락 내 다른 단어들의 영향력을 가중치를 두고 고려하도록 한다. 다음은 풀링 과정을 설명한다. 입력 텍스트가 BERT 계층을 통과한 후, 특징을 추출할 수 있다. 일반적인 접근 방식 중 하나는 각 입력 시퀀스의 시작에 추가되는 "[CLS]" 토큰에 해당하는 출력을 취하는 것이며, 이는 전체 시퀀스의 본질을 포착하도록 의도된다. 또 다른 접근 방식은 모든 토큰의 출력을 평균하는 것이다. BERT 모델의 출력은 벡터 집합이며, 이 벡터들은 입력 텍스트의 언어적 특성을 포착하는 "특징"이다.

3) 모델 통합

이 프로젝트에서는 두 종류의 딥 뉴럴 네트워크 알고리즘과 하나의 트랜스포머 기반 모델을 사용한다. 첫 번째는 순차적 데이터를 처리하기 위해 설계된 전통적인 CNN의 변형인 1차원 컨볼루션 뉴럴 네트워크이다. 이미지에서 일반적으로 발견되는 2D 데이터 대신, 1D CNN은 시간에 따라 한 차원으로 전개되는 시계열 데이터, 오디오 신호 또는 그와 유사한 데이터를 분석하기에 이상적이다. 1D CNN에서 컨볼루션 필터는 데이터를 따라 한 차원으로 이동한다. 1D CNN의 구성 요소는 컨볼루션 레이어, 활성화 함수, 풀링 레이어, 완전 연결 레이어, 출력 레이어로 구성된다고 자세히 설명될 수 있다. Conv 1d 레이어에 대한 컨볼루션 연산은 다음 식 (7) 으로 정의될 수 있다.

$$y(t) = (x*w) + b = \sum_{a=0}^{M-1} x(t+a).w(a) + b \quad (7)$$

이 공식에서 x는 입력 데이터를, w는 커널을, y는 출력 맵을, t는 입력의 컨볼루션 작업의 현재 위치를, M은 커널의 크기를, b는 바이어스 항을 나타내고, *는 컨볼루션 작업을 나타낸다. 또한, 순환 신경망(RNN) 아키텍처의 변형인 장단기 기억(LSTM)도 구현했다. LSTM은 표준 RNN이 직면하는 장기 의존성 관리와 기울기 소실 문제를 해결하기 위해 특별히 고안되었다. LSTM의 핵심 구성 요소에는 입력 게이트가 포함되어 있으며, 이는 셀로의 정보 흐름을 제어한다. 입력 게이트(input gate)는 시그모이드 활성화 함수를 사용하고, 삭제 게이트(forget gate)는 이전 셀 상태의 정보를 유지하거나 잊어야 할 것을 결정한다. 마지막으로 숨겨진 상태(hidden state)는 LSTM 셀의 출력으로, 현재 시간 단계에 대한 정보를 담고 있으며 다음 시간 단계에서 사용된다. 텍스트 분류기로는 BERT 기반의 대문자/소문자를 구별하는 모델을 사용했다. NLP에서 "cased"는 모델이 단어의 원래 대소문자를 유지한다는 것을 의미한다.

IV. 실험 및 결과

이 장에서는 실험 방법과 상세 내용을 기술한다. 머신러닝 모델을 만들기 위해 파이썬을 프로그래밍 언어로, TensorFlow를 딥러닝 라이브러리로, Librosa를 음성 특징 추출 라이브러리로 사용했다. 모든 모델은 더 빠른 처리를 위해 NVIDIA RTX 3090 GPU에서 훈련되었다. 과정은 행복, 슬픔, 분노, 두려움, 놀람, 중립의 6가지 감정으로 라벨링된 14,696개의 영어 문장 데이터 셋으로 시작한다. 이 데이터 셋을 사용하여, 실험에 특화된 BERT 모델을 훈련하였다.

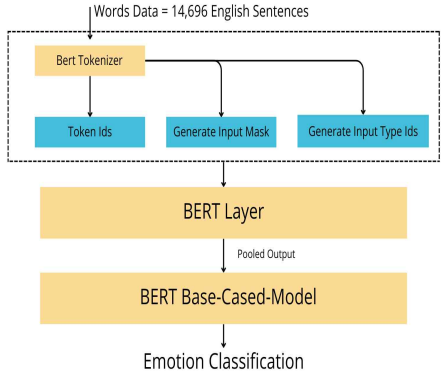

그림1. 미세 조정 BERT 시스템 구조
Figure 1. Fine tuned BERT system architecture

그림 1에서 보이듯이, 먼저 BERT 토크나이저를 사용하여 모든 문장을 토큰화하고, 그것들을 3부분(토큰, 입력 마스크, 타입 ID)으로 분할한다. 그 다음 데이터를 BERT 기반 대문자/소문자 모델에 입력하고 결과 출력을 출력한다.

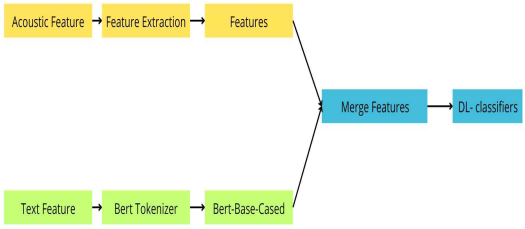

그림2. 멀티모달 SER 시스템 아키텍처
Figure 2. Multimodal SER system architecture

연구 방법 챕터에서 설명된 대로 음성 특징을 추출한 후, 미세 조정된 BERT 기반 대문자/소문자 모델의 출력과 결합한다. 마지막으로, 이 복합 데이터를 CNN 및 LSTM 분류기에 입력하여 성능 평가를 진행한다. 표1과 표2에서 각각 CNN, LSTM 모델의 파라미터를 기술하였다.

표1. CNN 모델 파라미터
Table 1. CNN Model Parameter

Parameter	
Batch size	65
Activation	Relu
Optimizer	Adam
Epoch	20
Dropout	No
Loss	Mean_Squared_Error

표2. LSTM 모델 파라미터
Table 2. LSTM Model Parameter

Parameter	
Pooling Method	Relu
Pooling in second layer	Maxpooling
Epoch	20
Activation	50
Batch size	65
Optimizer	Rmsprop
Loss	Categorical_crossentropy
Droput	No

표3. 모델 성능 비교
Table 3. Model Comparison

Metrics	LSTM-BERT	CNN-BERT	BERT
Precision	0.9126	0.9120	0.7177
Recall	0.9126	0.9120	0.7219
F1_score	0.9129	0.9131	0.719

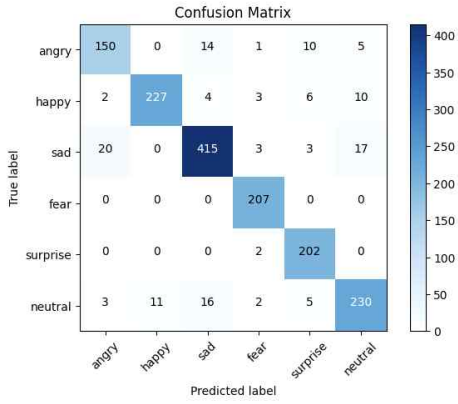

그림3. CNN 모델 혼동행렬
Figure 3 CNN Model Confusion Matrix

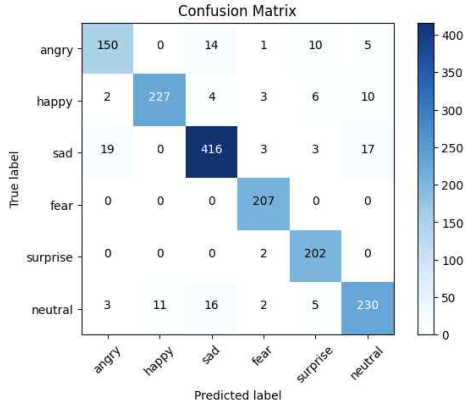

그림4. LSTM 모델 혼동행렬
Figure 4 LSTM Model Confusion Matrix

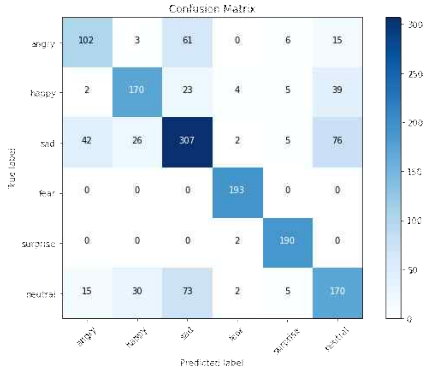

그림5. BERT 모델 혼동행렬
Figure 5 BERT Model Confusion Matrix

이제 연구 방법 섹션에서 설명된 모델들의 성능을 평가해보자. 표3, 그림3, 4, 5에서 볼 수 있듯이, CNN과 LSTM 모델은 감정 인식 측면에서 비슷한 효과를 보인다. 반면 BERT 모델은 만족스러운 결과를 얻기는 했지만, 이는 사용된 데이터셋이 텍스트 데이터만이었기 때문에 발생한 것으로 보인다. 오디오와 텍스트 특징을 결합하는 것이 성능을 향상에 기여하였음을 명확히 보여주며, 텍스트와 음성 특징 사이의 강한 상관관계를 증명한다. 이 두 모달리티를 통합함으로써, 각 모달리티의 약점을 보완할 수 있다. 예를 들어, 오디오에만 의존하는 모델은 많은 소음이 포함된 데이터에서 어려움을 겪을 수 있으며, 텍스트만 사용하는 모델은 문장에서 비꼬는 말투를 감지하지 못할 수 있다.

V. 결 론

인간의 감정은 복잡하고 다양하며, 우리의 일상생활에서 중요한 역할을 하며 행동, 의사결정, 상호작용에 영향을 미친다. 음성 감정 인식은 말하는 언어에서 표현된 감정을 식별하고 분류하려는 전산 준언어학(computational paralinguistics) 및 음성 처리의 과제이다. 목적은 운율, 음조, 리듬을 포함한 화자의 음성 패턴에서 기쁨, 분노, 슬픔, 좌절과 같은 화자의 감정 상태를 추론하는 것이다. 성별, 성격, 기분, 의도 및 정신 상태는 전통적으로 자동 음성 인식 알고리즘에서 간과되어왔던 중요한 보조 언어 정보이다. 인간의 마음은 효율적인 대응을 돕기 위해 단어 뒤에 숨은 의미를 해석하고 이를 위해 모든 음성 및 준언어적 사실을 활용한다. 준언어적 요소에 대한 이해가 부족하면 의사소통의 의미가 손상될 수 있다.

본 연구에서는 IEMOCAP 데이터셋을 활용하여 다중 모달 음성 감정 인식의 도전을 다룬다. BERT-LSTM 기반 모델과 BERT-CNN 기반 모델을 비교하고, 두 모델 모두 감정 분류에 잘 수행됨을 보여주었다. 향후 연구로는 Sepctral Roll-off 및 Zero Crossing Rate 와 같은 더 풍부한 특징과 더 진보된 융합 방법을 적용하여 다중 모달 음성 감정 인식 모델의 품질과 성능을 향상시킬 계획이다.

References

[1] S. Gaurav, “Multimodal speech emotion recognition and ambiguity resolution”, *arXiv preprint arXiv:1904.06022*, 2019. doi.org/10.48550/arXiv.1904.06022

[2] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... & Farhan, L, “Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions”, *Journal of big Data*, 8, pp 1-74, 2021. doi.org/10.1186/s40537-021-00444-8

[3] YY. Yu, X. Si, C. Hu and J. Zhang, “A review of recurrent neural networks: LSTM cells and network architectures”, *Neural computation*, Vol 31, No. 7, pp. 1235-1270, 2019. doi: 10.1162/neco_a_01199.

[4] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S., “IEMOCAP: Interactive emotional dyadic motion capture database” *Language resources and evaluation*, 42, pp. 335-359, 2008. https://doi.org/10.1007/s10579-008-9076-6

[5] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S., “End-to-End Multimodal Emotion Recognition Using Deep Neural Networks”, *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301-1309, Dec. 2017, doi: 10.1109/JSTSP.2017.2764438.

[6] Kim, J. H. & Lee, S. P., “Multi-Modal Emotion Recognition Using Speech Features and Text Embedding”, *Trans. Korean Inst. Electr. Eng*, 70, pp. 108 - 113, 2021. doi:10.5370/kiee.2021.70.1.108.

[7] Ranganathan, H., Chakraborty, S., & Panchanathan, S., “Multimodal emotion recognition using deep learning architectures” *2016 IEEE winter conference on applications of computer vision (WACV). IEEE*, pp. 1-9, 2016. DOI: 10.1109/WACV.2016.7477679

[8] Liu, W., Qiu, J. L., Zheng, W. L., & Lu, B. L.. “Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition”, *IEEE Transactions on Cognitive and Developmental Systems*, Vol. 14, No. 2, pp.715-729, 2021. DOI: 10.1109/TCDS.2021.3071170

[9] Jo, C.Y. & Jung, H.J., “Multimodal Emotion Recognition System using Face Images and Multidimensional Emotion-based Text”, *The Journal of Korean Institute of Information Technology*, vol. 21, no. 5, pp. 39-47, 2023, doi: 10.14801/jkiit.2023.21.5.39

[10] Lee, S.J., Seo, J.Y. & Choi, J.H., “The Effect of

Interjection in Conversational Interaction with the AI Agent: In the Context of Self-Driving Car", *The Journal of the Convergence on Culture Technology*, vol. 8, no. 1, pp. 551-563, 2022. doi:10.17703/JCCT.2022.8.1.551.

[11] Yoon, S., Byun, S. & Jung, K., "Multimodal Speech Emotion Recognition Using Audio and Text", *2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece,* pp 112-118, 2018, doi: 10.1109/SLT.2018.8639583.

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *In Proceedings of naacL-HLT*, Vol. 1, p. 2, pp 4171-4186, 2019. DOI: 10.18653/V1/N19-1423

※ 이 논문은 2023년 ASK 2023에서 "Speech and Textual Data Fusion for Emotion Detection: A Multimodal Deep Learning Approach"의 제목으로 발표된 논문을 확장한 것임.
※ "이 논문은 2023학년도 세명대학교 대학 혁신지원사업에 의한 연구임"