ISSN 2287-5026 (Print)
ISSN 2288-159X (Online)

Journal of the Institute of Electronics and Information Engineers

2026 4 제63권 4호

Vol.63, No.4 April 2026

www.theieie.org Vol.63, No.4 April 2026
The Institute of Electronics and Information Engineers (IEIE)
Room #907, The Korea Science Technology Center The first building, 22,
Teheran-ro 7 Gil, Gangnam-gu, Seoul, Republic of Korea

전자공학회 논문지
2026 4
제63권 4호
IEIE
사단법인 대한전자공학회

전자공학회논문지
•이 논문집은 한국연구재단 **우수등재학술지**임.

차 례

2026년 4월 제63권 제4호

SD / 반도체

[SoC 설계]

[RF 집적회로기술]

CI / 컴퓨터

[인공지능 및 보안]

AISP / 인공지능 신호처리

[영상 신호처리]

[음향 및 신호처리]

IE / 산업전자

[신호처리 및 시스템]

Journal of The Institute of Electronics and Information Engineers Vol.63, NO.4, April 2026 https://doi.org/10.5573/ieie.2026.63.4.110
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2026-63-4-13

오디오-비주얼 멀티모달 감정 인식을 위한 Cross-Attention Fusion

(Cross-Attention Fusion for Audio-Visual Multimodal Emotion Recognition)

김 정 윤*, 이 승 호**

(Jeong-Yoon Kim and Seung-Ho Lee©)

요 약

본 논문에서는 얼굴 영상과 음성 신호를 함께 활용한 오디오-비주얼 멀티모달 감정 인식을 위해 cross-attention 기반 융합 구조를 제안한다. 시각 정보는 RetinaFace를 이용한 얼굴 검출 및 정렬 과정을 거쳐 224×224×3 크기로 정규화되며, 음성 정보는 wav2vec2.0-large-robust 사전학습 모델을 통해 시간 의존적 임베딩 시퀀스(Batch, T, 1024)로 변환된다. 두 모달리티는 각각 transformer 인코더를 통해 시퀀스 수준 특징을 학습하고, 이후 cross-attention 모듈을 통해 상호 보완적 정보를 선택적으로 결합함으로써 단순 병합 방식보다 더 정교한 멀티모달 표현을 생성한다. 제안한 방법의 성능을 검증하기 위해 CREMA-D을 활용하여 실험을 수행하였다. 전체 데이터는 80%와 20% 비율로 학습·테스트 세트로 분할하였으며, 감정 데이터의 클래스 불균형 특성을 고려하여 accuracy와 weighted F1-score를 주요 평가 지표로 채택하였다. Weighted F1-score는 precision과 recall의 조화 평균 및 해당 클래스의 개수를 비율로 곱하여 더하는 것으로, 특정 감정의 등장 비율이 낮은 상황에서도 분류 성능을 균형 있게 평가할 수 있는 장점이 있다. 실험 결과, 제안하는 cross-attention 기반 멀티모달 모델은 정확도 88.3%, weighted F1-score 0.883의 성능을 기록하며 단일 모달 기반 모델 또는 단순 early/late fusion 방식 대비 유의미하게 향상된 결과를 보였다.

Abstract

In this paper, we propose a cross-attention fusion architecture for audio-visual multimodal emotion recognition utilizing both face images and audio signals. The visual information is normalized to 224×224×3 through face detection and alignment using RetinaFace, and the audio information is converted into a time-dependent embedding sequence (Batch, T, 1024) using the wav2vec2.0-large-robust pre-trained model. Each modality learns sequence-level features through a transformer encoder, and the cross-attention module selectively combines complementary information to generate more sophisticated multimodal representations than simple merging methods. To validate the performance of the proposed method, we conducted experiments using the CREMA-D. The entire data set was split into training and test sets at a ratio of 80% and 20%, respectively. Accuracy and weighted F1-score were adopted as the main evaluation metrics considering the class imbalance of the emotion data. Weighted F1-score is calculated by multiplying the harmonic mean of precision and recall by the number of corresponding classes and adding them together. It has the advantage of being able to evaluate classification performance in a balanced manner even in situations where the appearance rate of a specific emotion class has lower quantity. Experimental results show that the proposed cross-attention-based multimodal model achieved an accuracy of 88.3% and an F1-score of 0.883, demonstrating significant improvements over single-modality models or simple early/late fusion methods.

Keywords : Multimodal emotion recognition, Transformer, Convolutional neural network, Self-supervised learning, Affective computing

*학생회원, **평생회원, 국립한밭대학교 전자공학과(Dept. Electronic Engineering, Hanbat National University)
© Corresponding Author(E-mail : shyolee@hanbat.ac.kr)
※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 학·석사연계ICT핵심인재양성사업의 연구결과로 수행되었음(IITP-2026-RS-2022-00156212).
Received ; December 19, 2025 Revised ; January 16, 2026 Accepted ; January 22, 2026

Ⅰ. 서 론

사람의 감정은 음성의 억양, 얼굴 표정, 구강 움직임 등 다양한 비언어적 표현을 통해 전달된다. 단일 모달 기반 감정 인식[1]은 특정 환경에서는 강건한 성능을 보이나, 실제 상황에서 발생하는 잡음, 조명 변화, 모달리티 결손 등의 문제로 인해 정확도가 크게 저하되는 경향이 있다. 이러한 한계를 보완하기 위해 오디오와 비주얼 모달리티를 동시에 활용하는 멀티모달 감정 인식 연구[2]가 활발히 진행되고 있다.

기존 연구들은 주로 입력 단계에서 두 모달리티를 결합하는 early fusion[3], feature 추출 이후에 두 모달리티를 결합하는 late fusion[4]을 사용하여 감정 인식을 수행했다. 이러한 방법들은 단일 모달 감정 인식의 한계를 넘어 오디오-비주얼 정보를 결합하여 감정 인식의 성능을 크게 향상시켰다. 그러나 여전히 한계가 존재했다. 기존의 방법들은 두 모달리티 간의 독립성이 지나치게 강해서 상호 정보 참조가 어려워 학습의 불안정을 야기하거나, 초기 단계에서부터 지나친 정보의 결합으로 인해 훨씬 크고 복잡한 딥러닝 모델을 필요로 하여 멀티모달을 사용함으로써 얻는 득보다 더 큰 실이 있었다.

대표적인 기존 연구인 AuxFormer[5]는 각 모달리티 간에 독립성을 유지하기 위하여 멀티모달을 분석하는 transformer[6, 7] 외에 보조적으로 오디오, 비주얼 단일 모달리티를 분석하는 auxiliary transformer를 사용하였다. 이러한 독립성 유지 전략은 효과적이었으나 멀티모달 분석, 오디오 분석, 비주얼 분석을 각각 수행하기 위한 여러 개의 transformer가 있어야하며, 따라서 많은 양의 컴퓨팅 파워를 필요로 하는 문제가 있다.

본 연구에서는 이러한 한계를 극복하기 위해 사전 학습된 feature extractor (RetinaFace[8], wav2vec2.0[9])를 이용하여 오디오 feature, 비주얼 feature를 추출하고 cross-attention을 통해 오디오 feature와 비주얼 feature의 각 요소 간의 가중치를 적응적으로 조절하는 오디오-비주얼 멀티모달 감정 인식을 위한 cross-attention fusion을 제안한다. 본 연구의 기여는 다음과 같다.

i) Cross-attention 기반 오디오-비주얼 fusion 제안: 오디오/비주얼 feature의 각 요소 간 상호작용을 cross-attention으로 모델링하고, 중요도를 적응적으로 조절하는 결합 방식을 제안한다.

ii) Feature-level fusion 설계로 fusion 복잡도 감소: feature 단계에서 결합을 수행해 fusion 단계의 데이터 복잡도를 낮추고, 효율적인 멀티모달 결합을 가능하게 한다. 또한 유사한 방식인 late fusion 보다 멀티모달 감정인식에서 높은 정확도를 달성할 수 있음을 제시한다.

Ⅱ. 본 론

1. 제안하는 연구의 개요도

본 연구에서는 오디오와 비주얼 모달리티를 입력으로 받아 이들 간의 상호 연관성을 효과적으로 학습할 수 있는 cross-attention 기반 멀티모달 감정 인식 네트워크를 제안한다. 제안하는 네트워크는 사전 학습된 feature extractor를 이용해 각 모달리티의 고품질 feature를 추출한 후, cross-attention fusion 모듈을 통해 오디오 feature와 비주얼 feature 간의 상호 정보를 선택적으로 교환할 수 있도록 설계되었다. 그림 1은 제안한 멀티모달 감정 인식 모델의 전체 개요도를 나타낸다.

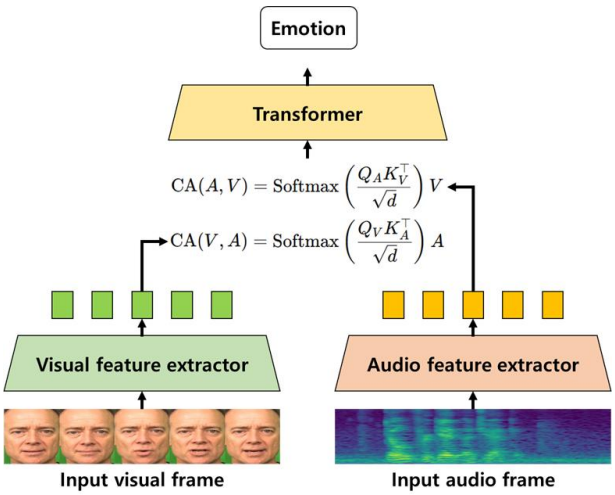

그림 1. 제안하는 방법의 개요도
Fig. 1. Overview of the proposed method.

2. 네트워크 구조

제안하는 네트워크는 사전 학습된 feature extractor, cross-attention 그리고 transformer 로 구성된다. 이는 AuxFormer 같이 auxiliary transformer가 필요한 방식이 아니라, 융합 이후의 시퀀스 하나만을 처리하는 효율적인 단일 인코더 구조로 설계하였다. 그림 2는 본 논문에 사용한 네트워크 구조도를 나타낸다.

가. 비주얼 특징 추출

비주얼 모달리티는 얼굴 영역을 안정적으로 획득하

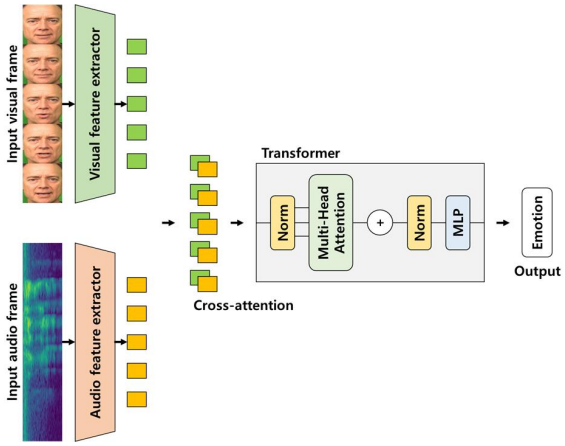

그림 2. 본 논문에 사용한 네트워크 구조도
Fig. 2. Network structure used in this paper.

기 위해 RetinaFace 기반의 얼굴 검출 및 정렬 과정을 거친다. RetinaFace는 Wider Face 데이터셋[10]을 이용하여 사전 학습된 얼굴 검출 및 bounding box 정렬에 특화된 모델이다. 입력 영상의 얼굴은 RetinaFace를 통해 정확하게 위치가 추정되며, 이후 얼굴 정렬을 수행하여 표준화된 얼굴 패치를 생성한다. 이렇게 정규화된 얼굴 이미지는 224×224×3 크기로 변환되어 시각적 일관성을 확보한다.

정규화된 영상은 transformer 기반의 비주얼 인코더로 입력되며, 인코더는 이미지 패치를 토큰 단위 시퀀스로 분해하여 얼굴 표정 변화, 근육 움직임, 구강 형태 등 감정 인식에 중요한 시각적 단서를 시퀀스 표현으로 학습한다. 이를 통해 최종적으로 감정 분류에 활용 가능한 고차원 시각 특징 벡터를 생성한다.

나. 오디오 특징 추출

오디오 모달리티는 감정 표현의 핵심 요소인 음성 높낮이, 속도, 음색, 발성 강도 등의 정보를 반영하기 위해 wav2vec2.0-large-robust 사전학습 모델을 활용한다. 본 논문에서 오디오 feature extractor로 사용한 wav2vec2.0 모델은 LibriSpeech 데이터셋[11]을 이용하여 사전 학습된 음성 특징 추출 모델로 robust한 특징 추출이 가능하다. 원본 음성 신호는 wav2vec2.0을 통해 프레임 단위로 처리되며, 시간 의존적 특성을 포함한 임베딩 시퀀스로 변환된다. 변환된 오디오 feature는 (Batch, T, 1024) 형태를 가지며, 여기서 T는 음성 길이에 따라 동적으로 변하는 시간축 시퀀스 길이를 의미한다.

이 오디오 임베딩 시퀀스는 transformer 기반의 오디오 인코더로 전달되어 말소리의 억양, 강세, 리듬과 같은 장기·단기 패턴을 효과적으로 학습한다. 이를 통해 감정 표현에 중요한 음향적 특징을 고차원 시퀀스 표현으로 정제하여 fusion 단계에서 시각적 특징과 상호 보완적으로 활용할 수 있도록 한다.

다. Cross-Attention Fusion

추출된 비주얼·오디오 특징은 서로 다른 시간 해상도와 표현 방식으로 구성되어 있기 때문에, 두 정보를 단순히 병합하는 것만으로는 감정 표현의 상관성을 충분히 반영하기 어렵다. 이러한 문제를 해결하기 위해 본 연구에서는 cross-attention fusion 전략을 도입한다.

Cross-attention fusion은 한 모달리티의 시퀀스를 Query로, 다른 모달리티의 시퀀스를 Key와 Value로 활용하여, 두 시퀀스 간의 상호작용을 비선형적으로 모델링한다. 즉, 오디오가 비주얼 정보를 참조하도록 하거나, 반대로 비주얼이 오디오 정보를 참조하도록 하는 방식으로 양방향 의미 결합을 수행한다. 이러한 구조는 단순 연결 방식에서는 얻기 어려운 모달리티 간의 영향도 조절, 시점 간 상호 연관성, 상보적 정보의 선택적 반영을 가능하게 한다. 식 (1)은 비주얼 기준 cross-attention을, 식 (2)는 오디오 기준 cross-attention을 나타낸다.

$$\mathrm{CA}(V, A) = \mathrm{Softmax}\left(\frac{Q_V K_A^\top}{\sqrt{d}}\right) A \tag{1}$$

$$\mathrm{CA}(A, V) = \mathrm{Softmax}\left(\frac{Q_A K_V^\top}{\sqrt{d}}\right) V \tag{2}$$

여기서 Q_V는 비주얼 특징에서 생성된 query, K_A는 오디오 특징에서 생성된 key, A는 오디오 value를 의미한다. 반대로 Q_A는 오디오 특징에서 생성된 query, K_V는 비디오 특징에서 생성된 key, V는 비디오 value를 의미한다. 이러한 cross-attention을 통해 각 모달의 요소를 감정인식에 얼마나 이용할지 적응적으로 조정할 수 있다.

라. Transformer를 통한 분석

Cross-attention fusion을 통해 생성된 멀티모달 결합 시퀀스는 감정 상태의 시간적 패턴을 정밀하게 분석하기 위해 transformer 구조로 입력된다. 본 연구에서는 감정 인식 문제의 특성을 고려하여 transformer 인코더를 구성하였다. 이는 기존 연구처럼 여러 개의 transformer를 병렬적으로 두는 방식이 아니라, 융합 이후의 시퀀스 하나만을 처리하는 단일 인코더 구조로 설계하였다.

본 논문에서 활용한 transformer 인코더는 총 3개의

블록으로 구성되며, 각 블록은 8개의 헤드를 갖는 multi-head self-attention과 두 층으로 이루어진 feed-forward network로 이루어진다. 모든 어텐션 헤드는 동일하게 512 차원의 임베딩을 사용하며, feed-forward network의 내부 은닉 차원은 1024로 설정하였다. Fusion 된 멀티모달 시퀀스는 이 transformer 인코더를 통과하면서 시간적 의존성과 감정 관련 특징이 정제되고, 마지막으로 출력된 시퀀스는 평균 풀링을 통해 하나의 벡터로 요약된다. 최종 단계에서는 선형 퍼셉트론을 통해 감정 클래스에 대한 확률을 산출하며, 이를 통해 오디오-비주얼 통합 표현을 기반으로 감정 인식을 수행한다.

그림 3. 본 논문에서 사용한 CREMA-D 예시
Fig. 3. Example of CREMA-D used in this paper.

Ⅲ. 실 험

1. 실험 환경

모든 실험은 Windows 10 워크스테이션을 사용하여 수행되었으며 사양은 다음과 같다. Intel(R) Core i7-10700K 8-Core Processor (3.80 GHz), 16 GB RAM, Nvidia RTX 3070 GPU (8GB VRAM, CUDA Cores: 5,888). 모든 워크플로우는 CUDA 12.8 버전, cuDNN 9.7.1 버전, PyTorch 2.7.1 버전이 사용되었다. 모든 모델은 cross entropy loss를 사용하여 50 epoch 동안 학습을 진행하였다.

2. 데이터 셋

본 논문에서는 CREMA-D[12]를 사용하여 제안하는 방법을 학습하고 평가한다. CREMA-D는 다양한 인종 및 국적 배경을 가진 91명의 배우(남성 48명, 여성 43명)가 특정 감정적 의도를 지닌 문장을 연기하는 고품질 녹음을 특징으로 하는 시청각 데이터셋이다. 이 데이터셋은 7,438개의 비디오클립을 포함하며, 각 클립은 평균 7명의 평가자가 평가하여 총 5.26시간 분량의 데이터를 제공한다. CREMA-D의 이러한 감정 평가 레이블링 작업은 오디오 전용, 비디오 전용, 시청각의 세 가지 조건에서 지각 평가를 포함하므로 본 연구에 특히 적합하다. CREMA-D는 화자와 독립적인 데이터 분할을 통해 분노, 혐오, 두려움, 행복, 슬픔, 그리고 중립 상태를 포함하는 6개 클래스의 다중 레이블 분류 작업으로 분류된다. 그림 3은 본 논문에서 사용한 CREMA-D 예시를 나타낸다.

3. 실험 및 고찰

본 연구에서는 CREMA-D를 사용하여 제안하는 방법의 감정 인식 성능을 검증하였다. 학습 데이터와 테스트 데이터는 80%:20% 비율로 클래스 균등하게 임의로 분리하였다.

감정 인식 문제는 클래스 간 데이터 분포가 균일하지 않거나, 특정 감정이 다른 감정보다 더 뚜렷하게 표현되는 경향이 존재하기 때문에 단순 정확도만으로는 모델의 성능을 판단하기 어렵다. 이에 따라 본 논문에서는 전체 예측의 정확성을 나타내는 accuracy와 더불어, 클래스 불균형을 반영할 수 있는 weighted F1-score를 주요 평가 지표로 채택하였다. 클래스 불균형을 반영할 수 있는 weighted F1-score를 주요 평가 지표로 채택하였다 weighted F1-score는 precision과 recall의 조화 평균 및 해당되는 감정 클래스 개수의 비율을 곱하여 더하는 것으로 정의되며, 특정 감정 클래스가 상대적으로 적게 등장하는 경우에도 모델의 분류 성능을 균형 있게 확인할 수 있다는 장점이 있다.

CREMA-D를 대상으로 실험을 수행한 결과, 제안하는 Cross-Attention 기반 멀티모달 모델은 정확도 88.3%, weighted F1-score 0.883의 성능을 보였다. 이는 단일 모달 기반 모델 또는 단순 결합 방식의 멀티모달 모델보다 높은 성능을 나타낸 것으로, 오디오·비주얼 두 모달리티 간의 상호 정보 참조가 감정 인식 성능 향상에 효과적임을 보여준다. 특히 cross-attention을 통해 각 모달리티의 시퀀스 요소 간 관계를 동적으로 조정함으로써, 특정 구간에서 음성 정보가 부족하거나 얼굴 표정 변화가 미약한 경우에도 상대 모달리티가 이를 보완하여 안정적인 예측을 수행하는 것이 확인되었다. 표 1은 CREMA-D를 대상으로 실험을 수행한 결

과를 나타낸다.

이러한 결과는 사전 학습된 feature extractor 모델로부터 추출된 고품질 특징과, 제안한 교차-주의 융합 방식이 결합된 구조가 멀티모달 감정 인식에서 효과적으로 작동함을 시사한다. 또한 비교적 경량화된 Transformer 인코더 구조를 사용하였음에도 높은 weighted F1-score를 기록한 것은, 본 연구의 융합 전략이 복잡한 구조 없이도 모달리티 간 시너지 효과를 충분히 끌어낼 수 있음을 의미한다. 그림 4는 CREMA-D 감정별 confusion matrix를 나타낸다.

표 1. CREMA-D를 대상으로 실험을 수행한 결과
Table 1. Result of conducting experiments targeting CREMA-D.

Method	Accuracy(%) ↑	Weighted F1-score ↑
Visual-only	74.6	0.738
Audio-only	78.3	0.779
Late Fusion	82.1	0.816
Early Fusion	81.5	0.808
The Proposed Method	**88.3**	**0.883**

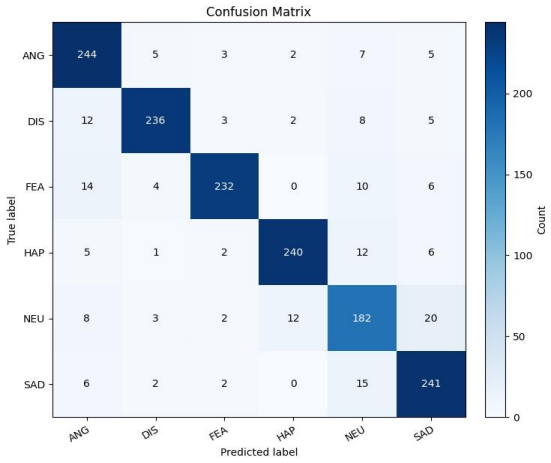

그림 4. CREMA-D 감정 별 confusion matrix
Fig. 4. Confusion matrix per emotion in CREMA-D.

또한 실험 결과 제안하는 방법의 training loss가 local minimum에 빠지지 않고 더욱 낮은 값으로 수렴하는 것을 확인하였다. 그림 5는 early fusion, late fusion 방법과 본 논문에서 제안하는 방법 간의 학습 과정에서의 cross entropy loss 그래프를 나타낸다.

또한 기존 논문들과 정확도 및 클래스 별 불균형은 반영되지 않는 unweighted F1-score를 CREMA-D를

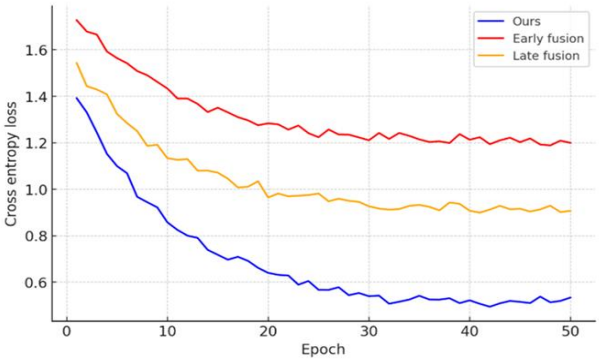

그림 5. 학습 과정에서의 cross entropy loss 감소 그래프
Fig. 5. Cross entropy loss graph during the training.

표 2. CREMA-D에 대해 기존 논문들과의 비교 결과
Table 2. Comparison of CREMA-D with existing papers.

Method	Accuracy(%) ↑	F1-score ↑
AuxFormer[5]	71.7	0.713
Goncalves et al.[13]	77.3	0.772
The Proposed Method	**88.3**	**0.881**

대상으로 비교하였다. 표 2는 CREMA-D에 대해 기존 논문들과의 비교 결과를 나타낸다.

그럼에도 본 실험은 실험 설정과 데이터 특성으로 인해 몇 가지 한계점이 존재한다. 우선, CREMA-D는 통제된 환경에서 수집된 배우 기반 데이터셋으로, 실제 환경에서 나타나는 자발적 감정 표현이나 복잡한 배경 요소가 충분히 반영되어 있지 않다. 따라서 제안된 모델이 실제 상황에서 발생하는 잡음, 비정형 표정, 다양한 조명 조건 등과 같은 비정규적 변동성에 대해 얼마나 강건한 성능을 유지할 수 있는지는 추가 검증이 필요하다.

모달리티 자체의 제약 역시 존재한다. RetinaFace 기반 얼굴 정렬 과정은 비정상적 포즈나 심한 가림이 포함된 영상에서는 오탐지 가능성이 있으며, wav2vec2.0 기반 오디오 특징 역시 심한 잡음 환경에서 성능이 저하될 수 있다. 이러한 입력 품질 저하가 모델 전체 성능에 미치는 영향을 본 연구에서는 깊이 분석하지 못했다. 마지막으로, 제안한 모델은 멀티모달 융합 구조로서 비교적 가벼운 편이지만, transformer와 cross-attention의 조합 특성상 여전히 시퀀스 길이가 증가하면 계산 비용이 크게 증가하는 구조적 제약이 존재한다. 이는 실시간 감정 인식이나 모바일 환경 적용 시 추가적인 최적화가 필요함을 의미한다.

Ⅳ. 결 론

본 논문에서는 오디오와 비주얼 정보를 결합하여 감정 인식을 수행하는 새로운 멀티모달 구조를 제안하였다. 기존 연구들은 두 모달리티의 독립성을 유지하기 어렵거나, 반대로 지나치게 많은 정보를 조기에 결합하여 복잡도가 크게 증가하는 문제가 있었다. 이를 해결하기 위해 본 연구는 사전 학습된 feature extractor를 활용하여 고품질의 비주얼·오디오 특징을 효율적으로 추출하고, 이후 cross-attention을 적용하여 두 모달리티 간의 상호 보완적 정보를 선택적으로 통합할 수 있는 구조를 설계하였다.

실험 결과, 제안된 모델은 CREMA-D에서 높은 accuracy와 weighted F1-score를 기록하며 기존의 단일 모달 기반 접근 또는 단순 융합 방식보다 우수한 성능을 보였다. 이러한 성능 향상은 cross-attention이 각 모달리티의 시점별 특징을 정교하게 조정하고, 불완전한 감정 표현을 반대 모달리티가 보완해 주는 구조적 장점에서 비롯된 것으로 해석된다. 또한 멀티모달 분석을 위한 복수의 transformer를 사용하는 방식을 지양하고, 단일 transformer 인코더를 활용하여 효율성과 성능을 동시에 확보했다는 점에서도 실용적 의미가 크다.

그럼에도 제안하는 방법에는 몇 가지 한계가 남아 있다. 실험이 단일 데이터셋에 국한되어 있어 다양한 상황에서의 일반화 성능을 충분히 검증하지 못했으며, 실제 환경에서 흔히 발생하는 배경 소음·얼굴 가림·발화 강도 변화와 같은 문제들에 대해 추가적인 평가가 필요하다. 또한 cross-attention 기반 융합은 효과적이지만, 모달리티 간 동적 의존성이 더 큰 시나리오에서는 보다 정교한 시점 정렬 기법이나 self-supervised 사전학습 전략이 요구될 수 있다.

향후 연구에서는 다양한 멀티모달 데이터셋을 포함한 광범위한 실험을 수행하고, 실제 환경에서 발생하는 변동성에 강건한 모델을 개발하는 방향으로 확장할 예정이다. 또한 기존의 cross-attention을 개선하여, 모달리티의 품질을 실시간으로 판단하고 신뢰도 기반 가중치를 조정하는 적응형 융합 방법, 또는 경량화를 위한 knowledge distillation 기반 구조도 함께 고려할 수 있다.

REFERENCES

[1] Khalil, Ruhul Amin, et al. "Speech emotion recognition using deep learning techniques: A review." IEEE access 7 (2019): 117327-117345.
[2] Ahmed, Naveed, Zaher Al Aghbari, and Shini Girija. "A systematic survey on multimodal emotion recognition using learning algorithms." Intelligent Systems with Applications 17 (2023): 200171.
[3] Bucur, Beniamin, et al. "An early fusion approach for multimodal emotion recognition using deep recurrent networks." *2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2018.
[4] Dixit, Chhavi, and Shashank Mouli Satapathy. "Deep CNN with late fusion for real time multimodal emotion recognition." *Expert Systems with Applications* 240 (2024): 122579.
[5] Goncalves, Lucas, and Carlos Busso. "AuxFormer: Robust approach to audiovisual emotion recognition." ICASSP 2022-2022 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.
[6] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
[7] Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint* arXiv:2010.11929 (2020).
[8] Deng, Jiankang, et al. "Retinaface: Single-shot multi-level face localisation in the wild." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[9] Hsu, Wei-Ning, et al. "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training." arXiv preprint arXiv:2104.01027 (2021).
[10] Yang, Shuo, et al. Wider face: A face detection benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 5525-5533.
[11] Panayotov, Vassil, et al. Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015. p. 5206-5210.
[12] Cao, Houwei, et al. "Crema-d: Crowd-sourced emotional multimodal actors dataset." IEEE transactions on affective

computing 5.4 (2014): 377-390.
[13] L. Goncalves and C. Busso, "Robust Audiovisual Emotion Recognition: Aligning Modalities, Capturing Temporal Information, and Handling Missing Features," in IEEE Transactions on Affective Computing, 2022, vol. 13, no. 4, pp. 2156-2170

저 자 소 개

김 정 윤(학생회원)
2020년 국립한밭대학교 전자공학과 학사 졸업.
2022년 국립한밭대학교 전자공학과 석사 졸업.
2022년~현재 국립한밭대학교 전자공학과 박사과정.
<주관심분야 : 감정인식, 딥러닝>

이 승 호(평생회원) - 교신저자
1986년 한양대학교 전자공학과 학사 졸업.
1989년 한양대학교 전자공학과 석사 졸업.
1994년 한양대학교 전자공학과 박사 졸업.
1994년~현재 국립한밭대학교 전자공학과 교수
<주관심분야 : 영상신호처리, 딥러닝, AR>