Smart Media Journal / Vol.15, No.1 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2026.15.1.100

한국어 감정 인식을 위한 다중 특징 기반 순차적 정제 모델

(A Multi-feature-based Sequential Refinement Model for Korean Emotion Recognition)

이명호*, 임명진*, 신주현*

(Moung Ho Yi, Myung Jin Lim, Ju Hyun Shin)

요 약

인간-컴퓨터 상호작용(HCI)에서 복잡한 감정 이해의 중요성이 커지면서 멀티모달 감정 인식 연구가 활발히 진행되고 있다. 최근에는 인간의 복잡한 감정을 인식하기 위해 다양한 모달리티를 사용하여 연구되고 있다. 그러나 기존 연구는 한국어 특유의 복합적인 감정 표현에서 나타나는 모달리티 간 상호작용을 충분히 포착하는 데 한계가 있다. 본 연구는 이를 극복하기 위해 텍스트, 음성, 자모 정보를 결합하는 순차적 정제 기반 멀티모달 융합 모델을 제안한다. 먼저 KLUE-BERT, HuBERT, 자모 기반 Transformer Encoder로 각 모달리티의 특징을 추출하고, Cross-Modal Attention을 통해 특징을 순차적으로 정제하며 융합한다. 이후 Conformer 블록과 Attention Pooling을 적용해 핵심 표현을 도출한다. 제안 모델은 정확도 0.8462를 달성하였다. 이는 모달리티 간 상관관계를 정밀하게 모델링하여 한국어 감정 인식 성능을 유의미하게 향상시킴을 보여준다.

■ 중심어 : 멀티모달 감정 인식 ; 순차적 정제 ; 컨포머 ; 자모 특징 ; 어텐션 풀링

Abstract

As the importance of understanding complex emotions in human - computer interaction (HCI) has grown, research on multimodal emotion recognition has become increasingly active. Recently, diverse modalities have been leveraged to recognize nuanced human emotions. However, prior studies have limitations in sufficiently capturing inter-modal interactions that arise in Korean-specific, compound emotional expressions. To address this gap, we propose a sequential refinement-based multimodal fusion model that integrates text, speech, and jamo (Korean grapheme) information. Specifically, we extract modality-specific representations using KLUE-BERT, HuBERT, and a jamo-based Transformer encoder, and then sequentially refine and fuse these features via cross-modal attention. Subsequently, we apply Conformer blocks and attention pooling to derive salient expressive cues. The proposed model achieves an accuracy of 0.8462, demonstrating that precisely modeling cross-modal correlations can meaningfully improve Korean emotion recognition performance.

■ keywords : Multimodal Emotion Recognition ; Sequential Refinement ; Conformer ; Jamo Features ; Attention Pooling

Ⅰ. 서 론

인간과 컴퓨터 간의 상호작용 기술은 디지털 환경의 급격한 변화와 함께 그 중요성이 날로 증가하고 있다. 최근에는 단순한 명령 수행을 넘어, 사용자의 경험을 극대화하기 위해 인간의 복잡한 감정과 내재된 의도를 깊이 있게 이해하고 이에 공감하며 반응하는 방향으로 발전하고 있다[1-3]. 감정은 텍스트, 음성 등 다양한 모달리티

* 정회원, 조선대학교 미래융합학부
본 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(No. 2023R1A2C1006149)이며, 조선대학교 학술연구비의 지원을 받아 연구되었음(2025년)

접수일자 : 2025년 12월 31일
수정일자 : 2026년 01월 21일
게재확정일 : 2026년 01월 27일
교신저자 : 신주현 e-mail : jhshinkr@chosun.ac.kr

를 통해 복합적으로 표출되는 특성을 가진다. 따라서 단일 모달리티만 활용하는 방식을 벗어나 모달리티를 상호보완하여 분석하는 한국어 멀티모달 감정 인식 연구가 활발히 되고 있다[4,5]. 기존의 한국어 멀티모달 감정 인식 연구는 각 모달리티를 독립적인 네트워크로 처리한 후, 최종 예측 단계에서 결과를 병합하는 Late Fusion 방식을 채택한다. 예를 들어, 텍스트 데이터는 KoBERT와 같은 언어 모델로 처리하고 음성 데이터는 CNN 기반의 모델로 처리한 뒤, 도출된 각 확률값을 가중 평균하거나[6], 다수의 사전학습 음성 모델들의 결과를 다수결로 통합하는 방식이 주를 이루었다[7]. 이러한 접근법은 단일 모달리티의 특성을 보존할 수 있다는 장점이 있으나, 서로 다른 모달리티 간의 상호작용과 상관관계를 세밀하게 포착하지 못해 감정을 정밀하게 모델링하는 데 한계가 있다. 특히, 음향 모델링 시 발화 내용(음운 구조)을 간과하는 것은 모델의 일반화 성능을 저하시키는 주요 원인이 된다. 실제로 텍스트의 음운 정보와 음향 신호를 정밀하게 정렬하지 않을 경우 평음과 격음을 혼동하는 등 결정적인 인식 오류가 발생함이 보고된 바 있다[8]. 반면, 문장 전체를 단순 병합하는 대신 감정 정보가 집약된 문미 억양과 같은 특정 음운 구간을 선별하여 집중했을 때는 인식 성능이 유의미하게 향상됨이 입증되었다[9]. 이는 단순한 정보의 결합을 넘어, 텍스트의 음운 구조와 미세한 음향 특징을 정밀하게 동기화하는 과정이 필수적임을 시사한다.

이러한 문제를 해결하기 위해 특징 벡터를 단순히 연결한 후 Conformer 등에 입력하는 중간 융합 방식이 제안되기도 한다[12]. 하지만 단순히 벡터를 결합하는 방식만으로는 각 정보가 서로의 의미를 보완하거나 강화하는 고차원적인 전이 과정을 충분히 반영하기 어렵다. 특히 한국어는 언어적 의미 외에도 말소리의 높낮이와 같은 음향적 요소, 그리고 음소의 배열과 조음 특성인 음운적 요소가 결합되어 감정을 전달하는 특성이 강하다. 따라서 이러한 특징들이 상호작용하며 감정 인식의 성능을 향상시키는 연구가 필수적이다. 따라서 본 논문에서는 기존 연구의 한계를 극복하고 한국어 감정 인식의 정확도를 향상시키기 위해 다중 특징 기반 순차적 정제 모델을 제안한다. 본 모델은 하나의 발화 데이터로부터 의미, 음향, 음운이라는 세 가지 관점의 독립적인 특징을 추출한다. 이후 제안하는 핵심 메커니즘인 Cross-Modal Attention을 통해 모달리티 간 상호작용을 유도하고 정보를 정제한다. 이처럼 순차적 정제 과정을 통해 생성된 고차원 벡터는 Conformer 블록을 통과한다. 마지막으로 Attention Pooling 기법을 적용하여 발화 내에서 감정 표현이 집중된 구간에 가중치를 부여함으로써 최종 감정을 인식한다. 본 연구의 기여는 한국어의 언어적 특성을 고려한 다중 특징 추출 체계를 구축하고, 모달리티 간의 순차적 정제 메커니즘을 통해 멀티모달 융합 성능을 입증한다는 점에 있다.

본 논문의 구성은 다음과 같다. 2장에서는 멀티모달 및 텍스트 기반 감정 인식과 관련된 기존 연구를 고찰한다. 3장에서는 본 논문이 제안하는 다중 특징 기반 순차적 정제 모델의 상세 구조을 설명한다. 4장에서는 실험 및 결과, 5장에서 결론 및 제언에 관해 기술하고 마무리한다.

Ⅱ. 관련 연구

1. 멀티모달 감정 인식

기존 멀티모달 감정 인식 연구는 각 모달리티를 독립적인 네트워크로 처리한 후, 최종 추론 단계에서 결과를 통합하는 후기 융합 방식에 집중한다. 박혜민[6]의 연구에서는 텍스트와 오디오 데이터를 각각 KoBERT와 CNN 모델로 개별 학습하고, 각 모델이 산출한 예측 확률값에 최적의 가중치를 부여하여 결합하는 가중 평균 앙상블 방식을 제안한다. 서재진[7]의 연구에서는 Wav2vec 2.0, HuBERT, WavLM 등 다양한 자기지도학습 기반의 사전 학습 모델들을 활용하

고, 각 모델의 개별 예측 결과를 다수결 원칙으로 통합하는 기법을 적용해 단일 모델 대비 성능 향상을 입증한다. 이러한 Late Fusion 방식은 구현이 용이하고 각 모달리티의 특성에 최적화된 독립적인 모델을 유연하게 활용할 수 있다는 장점이 있다. 그러나 모달리티 간의 복잡한 상관관계나 동적인 상호작용을 모델의 학습 과정에서 직접적으로 포착하기 어렵다는 근본적인 한계를 가진다. 이주환[10]의 연구에서는 이를 개선하기 위해 EEG와 음성에서 추출한 특징 벡터를 연결한 후, 이를 Conformer 블록의 입력으로 사용하는 중간 융합 방식을 제안한다. 이 방식은 후기 융합보다 상대적으로 낮은 레벨에서 정보 교환이 이루어지지만, 단순히 특징 벡터를 물리적으로 결합하는 것만으로는 두 모달리티 간의 깊은 비선형적 의존성을 모델링하기에 여전히 부족하다. 따라서 본 논문에서는 각 모달리티 정보가 서로를 단계적으로 참조하고 정교화하는 순차적 정제 메커니즘을 제안하여 기존 융합 방식의 한계를 극복하고자 한다.

2. 텍스트 기반 감정 인식

텍스트 기반 감정 인식 연구는 주로 KoBERT, KoBART, KCELECTRA와 같은 한국어 사전 학습 언어 모델을 특정 도메인의 데이터로 미세 조정하여 문장의 문맥적 의미를 고차원 벡터로 표현하는 방식에 주력해 왔다. 최근에는 단일 문장의 해석을 넘어 대화 전체의 흐름과 문맥을 모델링에 반영하려는 시도가 활발히 이루어지고 있다. 임명진[11]의 연구에서는 대화의 전개에 따른 문장 간의 감정 연관성을 상관계수로 분석하고, 이를 어텐션 가중치로 활용하여 복합적인 감정 변화를 인식하는 모델을 제안한다. 김향경[12]의 연구에서는 이전 발화에 대한 청자의 감정 반응을 현재 발화자의 상태를 예측하기 위한 추가적인 특징으로 활용하여 모델의 판별 성능을 높였다. 이처럼 기존 연구들은 텍스트의 고수준 의미 정보를 깊이 있게 이해하고 대화의 동적인 흐름을 포착하는 방향으로 발전한다. 그러나 이러한 접근법들은 텍스트를 오직 추상적인 의미 정보를 담은 매체로만 간주했을 뿐, 텍스트가 내포하고 있는 저수준의 음운론적 특성을 감정 인식에 직접적으로 투영하려는 시도는 미흡하다. 한국어의 경우 동일한 단어라도 음운의 배열이나 조음 방식에 따라 감정의 강도가 다르게 표현될 수 있다. 이에 본 논문에서는 텍스트를 의미와 음운이라는 다중 관점으로 분리하여 분석하고, 음운 정보를 실제 음향 데이터와 정밀하게 정렬함으로써 감정 인식의 정확도를 향상시키는 방법을 제안한다.

III. 한국어 감정 인식을 위한 다중 특징 기반 순차적 정제 모델

1. 연구 구성도

본 절에서는 텍스트, 음성 그리고 한국어 자모 정보를 결합한 다중 특징 기반 순차적 정제 모델을 제안한다. KLUE-BERT, HuBERT, Transformer Encoder를 통해 의미, 음향, 음운 특징을 각각 추출한 뒤, 순차적 Cross-Modal Attention 메커니즘을 통해 이를 융합한다. 이 후 Conformer 블록과 Attention Pooling을 거쳐 시퀀스 내 중요 정보를 추출하고 7가지 감정을 인식한다. 그림 1은 본 논문에서 제안하는 모델의 연구 구성도를 나타낸다.

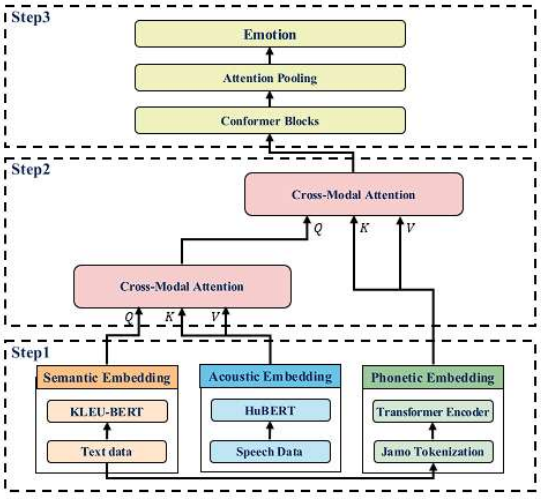

그림 1. 연구 구성도

2. 입력 및 특징 추출 단계

본 절에서는 하나의 발화 데이터로부터 세 가지의 독립적이면서도 상호 보완적인 특징 표현을 추출하는 과정을 기술한다. 첫째, 발화의 의미론적 정보를 추출하기 위해 원본 텍스트를 한국어 자연어 처리에 특화된 사전 학습 모델인 KLUE-BERT 모델을 사용한다. 이를 통해 문장의 문맥적 의미를 반영한 고차원의 Semantic Embedding 벡터를 생성한다. 둘째, 발화의 음향적 특징을 도출하기 위해 원본 음성 데이터를 자기지도학습 기반의 모델인 HuBERT 모델을 사용한다. 이는 음성 파형의 물리적 특성을 효과적으로 학습하여 음성의 톤, 속도, 강세 등 감정적 변이를 담은 Acoustic Embedding 벡터를 생성한다. 마지막으로 발화의 음운론적 정보를 모델링하고자 텍스트를 자모 단위로 분해한 뒤, 이를 Transformer Encoder에 통과시켰다. 해당 인코더는 자모의 순차적 패턴을 학습하여 실제 발음과 운율의 변화를 암시하는 Phonetic Embedding 벡터를 생성한다. 추출된 세 가지 임베딩 벡터는 딥러닝 모델의 입력으로 사용되기 위해 Batch, Sequence Length, Dimension으로 구성된 3차원 Tensor 형태로 생성한다.

3. 순차적 정제 단계

본 절에서는 입력 및 특징 추출 단계에서 추출된 세 가지 특징 벡터를 Cross-Modal Attention 메커니즘을 활용하여 단계적으로 융합하는 순차적 정제 과정을 기술한다. 의미 및 음운 특징은 최대 시퀀스 길이를 고정하여 입력을 생성했으며, 음향 특징은 발화마다 상이한 길이를 보존하기 위해 배치의 최대 길이에 맞춰 제로 패딩을 적용한다. 서로 다른 모달리티의 상관관계를 연산하여 감정 인식에 유의미한 정보를 추출하는 것을 목적으로 한다. 기본적인 Cross-Modal Attention의 연산 구조는 특정 모달리티의 정보를 질의(Query)로 하고, 참조하고자 하는 모달리티의 정보를 키(Key)와 값(Value)으로 설정하며, 식 1과 같이 정의한다.

$$CMA(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

식의 d_k는 특징 벡터의 차원 수 이며, Softmax 함수를 통해 각 정보 간의 연관성 가중치를 산출한다. 본 모델의 순차적 정제 과정은 먼저 가장 풍부한 문맥 정보를 보유한 Semantic Embedding E_s를 중심축인 Query로 설정하여 정제를 시작한다. 첫 번째 블록에서는 E_s를 Query로 Acoustic Embedding E_a를 Key와 Value로 사용하여 식 2와 같이 음향 정보가 반영된 1차 정제 벡터 F_{sa}를 생성한다.

$$F_{sa} = CMA(E_s, E_a, E_a) \quad (2)$$

이 과정을 통해 텍스트의 각 의미 단위에 대응하는 음성의 뉘앙스와 물리적 특성이 결합되어, 의미론적 정보 위에 음향적 특징이 정제된 형태로 중첩된다. 다음은 생성된 F_{sa}를 새로운 Query로 삼아 Phonetic Embedding E_p와 다시 융합한다. 두 번째 블록은 E_p를 Key와 Value로 참조하여 식 3과 같이 정제 벡터 F_{final}을 도출한다.

$$F_{final} = CMA(F_{sa}, E_p, E_p) \quad (3)$$

이러한 Sequential 정제 방식은 음향 정보가 반영된 벡터에 한국어 발음의 세부적인 패턴 정보까지 유기적으로 결합하게 한다. 이는 각 모달리티가 감정 인식에 미치는 영향력을 계층적으로 모델링함으로써 단일 융합 방식보다 더욱 안정적이고 표현력 높은 특징 표현을 가능하게 한다.

4. 최종 감정 인식

본 절에서는 순차적 정제 단계를 통해 도출된 멀티모달 융합 정보를 바탕으로 최종 감정을 인식하는 과정을 기술한다. 이 단계에서는 시퀀스

데이터 내에 존재하는 전역적 문맥과 지역적 패턴을 동시에 추출하고 감정 인식에 유의미한 시점에 집중하여 정보를 요약하도록 설계한다. 먼저 식 3에서 얻은 최종 벡터 F_{final}은 후처리 모듈인 Conformer 블록의 입력으로 사용한다. Conformer는 Transformer의 Self-Attention과 CNN의 연산을 결합한 구조로 전역적인 감정의 흐름과 지역적인 특징 패턴을 동시에 학습한다. 이를 통해 도출된 벡터는 식 4와 같이 한 차원 더 고차원적인 특징 표현 X로 정제된다.

$$X = Conformer(F_{final}) \tag{4}$$

이후, 정제된 특징 시퀀스 X에서 감정 인식에 결정적인 기여하는 시간대에 높은 가중치를 부여하기 위해 Attention Pooling 기법을 적용한다. 단순히 시간축에 대해 평균을 내는 방식보다 감정의 핵심 정보를 효과적으로 포착할 수 있게 한다. 각 시점 t에 대한 Attention weighting a_t와 최종 요약 벡터 v는 식 5, 6과 같이 표현된다.

$$a_t = \frac{\exp(w^T X_t)}{\sum_{t=1}^{T}\exp(w^T X_t)} \tag{5}$$

$$v = \sum_{t=1}^{T} a_t X_t \tag{6}$$

식에서 w는 학습 가능한 가중치 벡터이다. 마지막으로 고정된 크기의 단일 벡터 v를 Softmax 함수가 포함된 Classifier에 통과시켜 최종적인 감정 인식 결과 $\hat{y}$를 도출한다.

$$\hat{y} = Softmax(W_c v + b_c) \tag{7}$$

본 단계는 다중 특징이 순차적으로 정제된 결과물을 바탕으로 핵심적인 시간 정보를 압축하여 복합적인 감정 상태를 정확하게 식별하게 한다.

Ⅳ. 실험 및 결과

1. 데이터 셋 및 전처리

본 연구의 데이터 셋 및 전처리 단계에서는 AI-Hub의 감정 분류를 위한 대화 음성 데이터셋(4, 5차년도 및 5차년도 2차)을 사용했으며, 초기 데이터셋은 총 43,976개의 발화 샘플로 구성되었다. 데이터의 엄정성을 위해 먼저 전역적인 중복 제거를 수행한다. 특수문자와 공백을 제거하는 정규화 과정을 거친 후, 텍스트 내용이 완전 동일한 데이터 6,779개를 제외함으로써 Train과 Test 간의 정보 중첩으로 인한 Data Leakage 문제를 원천 차단한다. 다음은 감정 레이블의 신뢰성을 높이기 위해 전문가 5인의 평가 결과를 처리한다. 다수결 투표를 통해 3표 이상의 합의를 얻은 데이터만을 유효한 것으로 간주했으며, 감정이 명확하지 않아 2표 이하를 획득한 모호한 샘플 6,421개를 실험 대상에서 제외한다. 이러한 단계적 정제를 거쳐 최종적으로 30,776개의 고유한 발화 데이터를 실험에 확보한다. 다만, 정제 과정에서 모호한 데이터가 제외됨에 따라 실제 환경의 복합적 감정 표현보다 데이터의 판별 난이도가 낮아졌을 가능성이 있으며, 본 결과는 이러한 정제된 데이터 환경을 기반으로 도출된 성능임을 명시한다. 실험의 효율적인 수행과 검증을 위해 전체 데이터셋은 8:2의 비율로 분할하여 사용한다. 표 1은 감정별 데이터 개수이다.

표 1. 감정별 데이터 개수

감정	개수
Sadness	12,908
Angry	6,175
Neutral	4,767
Happiness	2,745
Fear	1,933
Disgust	1,732
surprise	516

2. 실험 환경 및 구현

본 연구의 제안 모델은 PyTorch 프레임워크를 사용하여 구현한다. 실험의 재현 가능성을 보장하기 위해 Seed는 42로 고정한다. 모델 학습 시

최적화 알고리즘은 Adam을 사용했으며, 학습률은 0.0001로 설정한다. Epoch은 10으로 설정하고, 검증 데이터에 대한 정확도가 가장 높은 시점의 모델을 최종 선택한다. 모델의 주요 하이퍼파라미터 설정값은 표 2와 같다.

표 2.. 주요 하이퍼파라미터 설정

항목	설정값
Optimizer	Adam
Learning Rate	0.0001
Batch Size	32
Epochs	10
Random Seed	42
Hidden Dimension	768
Dropout Rate	0.1

3. 실험 결과 및 성능 분석

본 절에서는 제안한 모델의 성능을 검증하기 위해 Accuracy, Precision, Recall, Weighted F1-Score를 지표로 사용하여 성능 분석을 수행한다. 본 연구에서 F1-Score는 감정 범주별 데이터의 불균형을 고려하여 각 클래스의 샘플 수에 비례한 가중치를 부여하는 Weighted 방식을 적용한다. 이는 모든 클래스를 동일 비중으로 처리하는 Macro 방식보다 실제 데이터 셋의 분포를 충실히 반영하여 모델의 전반적인 신뢰성을 종합적으로 평가하는 데 적합하기 때문이다. 또한, 본 모델의 각 모듈이 성능 향상에 미치는 기여도를 확인하고자 단일 모달리티 모델 및 소거 실험 결과와 비교한다. 표 3은 제안 모델과 비교 대상 모델들의 실험 결과를 정리한 것이다.

표 3. 비교실험 결과

모델	Accuracy	Precision	Recall	F1-Score
단일 텍스트 (LSTM)	0.8114	0.8107	0.8114	0.8104
단일 음성 (CNN)	0.7630	0.7629	0.7630	0.7582
텍스트+음성 평균 앙상블	0.8354	0.8324	0.8354	0.8325
제안 모델 (Con, Atten 제외)	0.8207	0.8177	0.8207	0.8175
제안 모델 (자모 제외)	0.8356	0.8353	0.8356	0.8349
제안 모델	0.8462	0.8456	0.8462	0.8451

실험 결과, 제안한 모델은 모든 평가 지표에서 우수한 성적을 기록하며 가장 높은 성능을 보였다. 본 실험에서 비교 대상으로 설정한 단일 모달리티 모델(LSTM, CNN)은 제안 모델의 구조적 기여도를 검증하기 위해 KLUE-BERT와 HuBERT를 동일한 특징 추출기로 활용하고, 그 위에 시퀀스 모델링 레이어를 결합한 모델이다. 특히 단순한 Average Ensemble 방식보다 본 연구의 순차적 정제 방식이 더 높은 성능 향상을 보였는데, 이는 텍스트, 음성, 자모 정보를 단계적으로 결합하여 정보의 손실을 최소화하고 상호 보완적인 특징을 효과적으로 추출했기 때문이라고 분석한다. 추가로 Conformer 블록과 Attention Pooling의 기여도를 확인하기 위해 이를 제외한 모델과 비교 실험을 진행한다. 두 모듈을 제외했을 때 성능이 하락한 것을 확인할 수 있다. 이를 통해 전역적 문맥과 지역적 패턴을 동시에 파악하는 Conformer의 재정제 기능과 감정의 핵심 시점에 집중하는 Attention Pooling 기법이 최종 인식 정확도 향상에 결정적인 역할을 수행했음을 입증한다. 또한, 본 연구의 독창적 구성 요소인 Phonetic Embedding인 자모 임베딩이 전체 성능에 미치는 기여도를 분석하기 위해 이를 제외한 모델과의 비교 실험을 수행한다. 자모 임베딩을 제거하고 텍스트와 음성 정보만을 활용했을 때의 정확도는 0.8356, Weighted F1-Score는 0.8349로 최종 제안 모델보다 낮았다. 특히 자모 특징 유무에 따른 클래스별 성능을 분석한 결과, 중립 감정의 F1-Score는 0.70에서 0.72로 상승했으며 정밀도는 0.71에서 0.75로 향상되었다. 이는 자모 단위의 음운 정보가 제공하는 미세한 운율 단서가 사라질 경우, 음향적으로 유사한 타 감정들을 중립으로 오분류하는 경향이 강해짐을 시사한다. 다만, 제안 모델은 대규모 사전학습 모델을 복합적으로 활용하기 때문에 안정적인 구동을 위해 연산 자원이 충분한 환경에서 진행할 것을 제안한다. 결론적으로, 본 연구에서 제안하는 모델은 한국어의 의미적, 음향

적, 음운적 특성을 순차적으로 정제함으로써 복합적인 감정 상태를 정밀하게 인식한다.

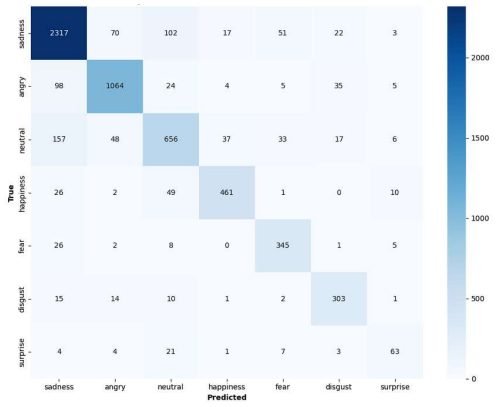

그림 2. 제안한 모델의 혼동 행렬

최종적으로 제안한 모델의 상세 인식 성능과 클래스별 오분류 경향을 파악하기 위해 그림 2의 Confusion Matrix를 분석한다. 분석 결과, 슬픔에서 2,317건, 분노에서 1,064건의 높은 정답 수를 기록하며 감정의 색채가 뚜렷한 영역에서 탁월한 식별력을 보였다. 이는 모델이 강한 에너지를 내포한 분노와 특유의 침전된 특징을 가진 슬픔의 핵심 신호를 안정적으로 학습했음을 시사한다. 특히 주목할 점은 중립과 슬픔간의 관계다. 중립 데이터 중 157건이 슬픔으로, 슬픔 데이터 중 102건이 중립으로 교차 오분류된 양상을 보였다. 이는 두 감정 모두 Low-Arousal 특성을 공유하며 음성학적으로 낮은 피치와 느린 발화 속도를 보이기 때문에 발생하는 Acoustic Overlap 현상으로 풀이된다. 그런데도 본 모델이 중립 데이터의 상당수(656건)를 정확히 식별해 낸 것은, 자모 단위의 미세한 운율 변화를 정제하여 유사한 저각성 감정 간의 경계를 성공적으로 구분했음을 입증하는 결과다. 또한, 행복과 놀람 간의 혼동은 두 감정이 공통적으로 지니는 높은 Arousal과 긍정적인 Valence에 기인한 것으로 분석한다. 결론적으로 본 연구의 순차적 정제 과정은 각 모달리티의 정보를 단계적으로 다듬음으로써, 음향적으로 유사한 감정들 사이에서도 유의미한 변별 지점을 찾아내어 한국어 감정 인식의 정밀도를 제고한다.

V. 결론 및 제언

본 연구에서는 한국어 감정 인식 성능을 향상시키기 위해 텍스트, 음성, 한국어 고유의 음운 정보를 결합한 다중 특징 기반 순차적 정제 모델을 제안한다. 먼저 KLUE-BERT, HuBERT, 자모 단위의 Transformer Encoder를 활용하여 각 모달리티의 특징을 추출한다. 추출된 특징들은 Cross-Modal Attention 메커니즘을 통해 텍스트 정보를 중심으로 음향과 음운 정보가 단계적으로 결합되는 순차적 정제 과정을 거쳤다. 이후 전역적 문맥과 지역적 패턴을 동시에 포착하는 Conformer 블록과 핵심 시점의 정보에 집중하는 Attention Pooling을 적용한 결과, 향상된 정확도와 F1-Score를 달성하며 제안 모델의 우수한 성능을 입증한다. 결론적으로 본 연구는 한국어의 의미적, 음향적 특성에 음운적 특성을 순차적으로 정제하여 결합하는 방식이 복합적인 감정 상태를 인식하는 데 매우 유효함을 확인한다. 다만, 본 연구는 정제된 데이터를 사용한다는 한계가 있어, 향후 연구에서는 실제 환경의 소음이나 다중 화자의 발화가 섞인 상황에서도 강건하게 작동할 수 있는 데이터셋의 다양성을 확보할 필요가 있다. 또한, 대규모 사전 학습 모델을 사용함에 따른 연산 비용 문제를 해결하기 위해 모델 경량화 연구가 병행되어야 하며, 감정의 범주 분류를 넘어 감정의 강도를 수치화하는 모델로 확장을 제언한다. 후속 연구를 통해 본 모델은 향후 지능형 상담 시스템이나 감성형 AI 서비스 등 다양한 실무 분야에서 핵심적인 기술로 활용될 수 있을 것으로 기대한다.

REFERENCES

[1] Z. Xie and L. Guan, "Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools," *Proc. of 2013 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1-6, 2013.

[2] 임명진, 이명호, 신주현, "상담 챗봇의 다차원 감정 인식 모델," *스마트미디어저널,* 제10권, 제4호, 21-27쪽, 2021년 12월

[3] 이명호, 임명진, 신주현, "텍스트와 음성의 앙상블을 통한 다중 감정인식 모델," *스마트미디어저널,* 제11권, 제8호, 65-72쪽, 2022년

[4] 조찬영, 정현준, "얼굴 영상과 다차원 감정 기반의 텍스트를 이용한 멀티모달 감정인식 시스템," *한국정보기술학회논문지,* 제21권, 제5호, 39-48쪽, 2023년

[5] 김선희, "감정인식을 위한 PK-CCA 기반의 다중 모달 생체신호 융합 모델," *스마트미디어저널,* 제14권, 제4호, 39-47쪽, 2025년

[6] 박혜민, "KoBART 모델과 CNN 모델의 가중평균 앙상블을 이용한 멀티모달 감정인식 분류 모델 개선," *한국정보과학회 학술발표논문집,* 2157-2159쪽, 2023년

[7] 서재진, 강태인, 곽일엽, "사전 학습 모델과 앙상블 기법을 통한 음성 감정인식," *한국데이터정보과학회지,* 제35권, 제4호, 445-459쪽, 2024년

[8] 최승호, 이희영, "Silver Mate 개발을 위한 음성기반 감정을 이용한 인간-로봇 상호작용," *한국로봇학회,* 제2권, 제2호, 16-22쪽, 2025년

[9] 김아름, "한국어 자동 음성 인식의 오류 유형에 대한 음운론적 연구," *서울대학교, 박사학위논문,* 2022년

[10] 이주환, 김형국, "뇌전도와 해당 오디오 신호의 시공간적 특징 융합을 이용한 감정인식," *한국음향학회지,* 제43권, 제6호, 630-636쪽, 2024년

[11] 임명진, 신주현, "대화문 감정 연관성 기반 복합 감정인식 모델," *멀티미디어학회논문지,* 제28권, 제2호, 366-377쪽, 2025년

[12] 김향경, 우민선, 박선정, 김법민, 김용민, "대화에서 감정반응을 고려한 멀티모달 기반 발화자 감정인식 모델 개발," *한국정보과학회 학술발표논문집,* 2115-2117쪽, 2023년

저 자 소 개

이명호(정회원)

2018년 조선대학교 제어계측로봇공학과 학사 졸업
2020년 조선대학교 소프트웨어융합공학과 석사 졸업
2025년 조선대학교 전자공학과 박사 졸업

<주관심분야 : 빅데이터, 머신러닝, 텍스트마이닝, 인공지능, 딥러닝 등>

임명진(정회원)

2022년 조선대학교 컴퓨터공학과 박사 졸업
2022년~현재 조선대학교 신산업융합학부 겸임교수

<주관심분야 : 빅데이터 처리, 데이터마이닝, 자연어처리, 머신러닝, 감정인식 등>

신주현(정회원)

2007년 조선대학교 전자계산학과 박사 졸업
2018년~현재 조선대학교 신산업융합학부 부교수

<주관심분야 : 데이터베이스, 데이터마이닝, 자연어처리, 인공지능 등>