사물인터넷융복합논문지 Vol. 12, No. 2, pp. 173-181, 2026 https://doi.org/10.20465/KIOTS.2026.12.2.173

멀티모달 LLM 기반 심리상담 인공지능 시스템 구현

박준용[1], 김균호[1], 조강래[1], 김태국[2*]
[1]국립부경대학교 컴퓨터·인공지능공학부 학생, [2]국립부경대학교 컴퓨터·인공지능공학부 교수

Implementation of a Multimodal LLM-based Psychological Counseling AI System

Jun-Yong Park[1], Gyun-Ho Kim[1], Kang-Rae Jo[1], Tae-Kook Kim[2*]
[1]Student, Division of Computer and Artificial Intelligence Engineering, Pukyong National University
[2]Professor, Division of Computer and Artificial Intelligence Engineering, Pukyong National University

요 약 본 연구는 현대 사회의 정신건강 서비스 수요 급증과 기존 텍스트 중심 상담 챗봇이 가진 비언어적 맥락 파악의 한계를 극복하기 위해, 음성·텍스트·얼굴 표정을 통합한 멀티모달 감정 인식 기반 AI 심리상담 시스템을 제안하고 그 효과를 검증하였다. 기술적으로는 한국어 음성 인식을 위한 Wav2Vec2, 텍스트 문맥 분석을 위한 KoBERT, 실시간 표정 인식을 위한 ResNet18 모델을 개별 모달리티의 백본으로 채택하였으며, 각 모델에서 산출된 감정 확률 분포를 결정 수준 후기 융합(Decision-level Late Fusion) 방식으로 통합하였다. 실험 결과, 제안한 시스템은 단일 모달리티 기반 접근 방식에 비해 정서 상태 인식의 일관성과 정확도를 향상시키고, 상담 맥락에 대한 이해 수준을 유의미하게 개선하는 것으로 나타났다. 특히 음성, 텍스트, 얼굴 표정 정보를 결합한 통합 감정 분석은 상담 과정에서 사용자 정서 변화를 보다 정밀하게 반영함으로써 상담자의 판단을 보조하는 도구로서의 활용 가능성을 확인하였다.

주제어 : 멀티모달, 감정 인식, 디지털 헬스케어, 심리상담, 대규모 언어 모델, 인공지능

Abstract This study proposes and empirically evaluates a multimodal emotion recognition-based AI psychological counseling system that integrates speech, text, and facial expressions, aiming to address the rapidly increasing demand for mental health services in modern society and the limitations of conventional text-based counseling chatbots in capturing nonverbal cues. From a technical perspective, Wav2Vec2 was employed for Korean speech recognition, KoBERT for textual context analysis, and ResNet18 for real-time facial expression recognition, each serving as the backbone model for its respective modality. The emotion probability distributions generated from each model were integrated using a decision-level late fusion approach. Experimental results demonstrate that the proposed system improves the consistency and accuracy of emotional state recognition compared to single-modality approaches, while also significantly enhancing the understanding of counseling context. In particular, the integrated emotion analysis combining speech, text, and facial expression data enables more precise reflection of users' emotional changes during the counseling process, thereby confirming its potential as a supportive tool for assisting counselors in decision-making.

Key Words : Multimodal, Emotion recognition, Digital healthcare, Large Language Model (LLM), Affective computing, Artificial intelligence

이 논문은 국립부경대학교 자율창의학술연구비(2024년)에 의하여 연구되었음
*교신저자 : 김태국(king@pknu.ac.kr)
접수일 2026년 02월 01일 수정일 2026년 02월 14일 심사완료일 2026년 02월 23일

1. 서론

최근 디지털 미디어 기술의 급격한 발달과 현대 사회의 스트레스 증가로 인해 정신 건강 관리에 대한 사회적 수요가 급증하고 있다. 특히 고령화 사회 진입과 1인 가구의 증가로 정서적 고립 문제가 대두되면서, 시공간의 제약 없이 접근 가능한 디지털 헬스케어 서비스의 중요성이 강조되고 있다[1,2].

이와 함께 사물인터넷(Internet of Things, IoT) 기술의 발전은 웨어러블 디바이스, 스마트폰, 환경 센서 등 다양한 장치를 통해 사용자의 생체 신호 및 행동 데이터를 실시간으로 수집·분석할 수 있는 기반을 제공하고 있다. 이러한 IoT 기반 데이터는 사용자의 심박수, 활동량, 수면 패턴 등과 같은 정량적 정보뿐 아니라, 일상 생활 속 정서 상태를 간접적으로 추정할 수 있는 중요한 단서를 제공한다[3,4]. 더불어 인공지능(Artificial Intelligence, AI) 기술의 발전은 이러한 다차원 데이터를 통합적으로 해석하고 의미 있는 패턴을 도출하는 데 핵심적인 역할을 수행하고 있으며, 특히 대규모 언어모델(LLM)과 딥러닝 기반 감정 인식 기술은 인간의 언어적·비언어적 표현을 이해하는 수준으로 발전하고 있다[5-7].

기존의 심리상담 서비스는 주로 대면 방식이나 텍스트 중심의 챗봇 형태로 제공되어 왔다. 그러나 대면 상담은 물리적 거리와 비용의 한계가 있으며, 단순 텍스트 기반 챗봇은 사용자의 비언어적 표현을 포착하지 못해 심도 있는 정서적 교감이 어렵다는 한계가 있다. 이러한 제약을 보완하기 위한 대안으로, IoT를 통해 수집된 다양한 정서·행동 데이터와 AI 기반 분석 기술을 결합한 지능형 심리상담 서비스가 주목받고 있으며, 그 학술적 타당성 또한 점차 검증되고 있다[8-10].

2017년 무작위 대조군 연구를 통해 자동화된 대화형 에이전트인 Woebot이 인지행동치료(Cognitive Behavioral Therapy, CBT)를 효과적으로 전달하여 사용자들의 우울 증상을 유의미하게 감소시킬 수 있음을 입증하였다[11]. 이후 연구들에서는 AI 기반 심리상담 챗봇이 인간 상담자에 대한 두려움이나 낙인(stigma) 없이 이용할 수 있는 '비판받지 않는 안전한 공간'을 제공함으로써, 사용자의 심리적 장벽을 낮추고 자기개방(self-disclosure)을 촉진하는 정성적 가치 또한 확인되었다[12,13]. 이러한 결과는 AI 심리상담이 단순한 보조 도구를 넘어, 실제 임상적 효용을 지닌 디지털 치료 개입으로 발전할 수 있음을 시사한다.

본 연구의 멀티모달 LLM은 단순히 다중 양식 데이터를 처리하는 개별 모델의 나열을 의미하는 것이 아니라, 각 모달리티에서 독립적으로 산출된 감정 확률 분포를 결정 수준 후기 융합(Late Fusion)을 통해 LLM의 시스템 프롬프트 및 추론 컨텍스트에 직접 반영하는 시스템 레벨의 통합적 아키텍처를 의미한다.이러한 구조적 특징을 통해 LLM은 사용자의 언어적 발화 내용뿐만 아니라 비언어적 정서 신호가 통합된 '멀티모달 문맥'을 바탕으로 공감적 상담 응답을 생성하게 된다.

본 시스템은 임상 진단이나 치료를 대체하지 않으며, 전문가의 판단을 보조하기 위한 지원 도구로 설계되었다. 특히 자해·타해 등 위기 신호가 감지되는 경우 시스템은 즉시 전문기관/긴급 연락 안내를 우선 제공하도록 안전 정책을 적용한다.

2. 구현

본 장에서는 음성(Voice), 텍스트(Text), 얼굴 표정(Face) 신호를 실시간으로 수집·분석하고, 결정 수준 후기 융합(Late Fusion)을 통해 통합 감정 확률 분포를 산출한 뒤, 이를 멀티모달 LLM 기반 상담 응답 생성에 활용하는 전체 구현 구조를 설명한다. Late Fusion은 텍스트, 음성, 이미지와 같은 서로 이질적인 멀티모달 데이터를 각 모달리티에 최적화된 개별 모델에서 독립적으로 처리한 후, 각 모델이 산출한 예측 결과(예: 클래스 확률 분포 또는 결정 값)를 결합하여 최종 의사결정을 수행하는 방식이다. 이 방법은 모달리티 간 특징 표현을 직접 결합하는 조기 융합(Early Fusion)과 달리, 각 모달리티의 고유한 특성을 충분히 반영할 수 있으며, 모델 간 독립적인 학습과 유연한 확장성이 가능하다는 장점이 있다. 또한 가중 평균, 투표 기반 결합, 확률 정규화 등 다양한 결합 전략을 적용함으로써 상황에 따라 성능을 최적화할 수 있다[14].

제안 시스템은 (1) 입력 수집 및 전처리, (2) 모달별 감정 분류, (3) 시간 구간 기반 누적 및 결정 수준 후기 융합, (4) 실시간 스트리밍/시각화, (5) 상담 대화 생성 및 서비스 제공의 단계로 구성된다.

먼저 입력 단계에서는 사용자로부터 마이크 음성, 카메라 영상(얼굴 표정), 그리고 텍스트(직접 입력 또는 음성 인식 결과)를 수집한다. 음성은 발화 이벤트를 기준으로 일정 길이의 오디오 구간을 추출하여 분석하며, 필요 시 무음 구간 제거(VAD) 및 입력 품질 관리(클리핑/저품질 샘플 제거 등)를 통해 모델 입력 신호의 안정성을 확

보한다. 영상은 웹캠 프레임을 일정 주기로 수집한 후 프레임 내 얼굴을 검출하여 얼굴 영역(ROI)에 대해 리사이즈 및 정규화를 수행하고, 이를 표정 분류 모델에 전달한다. 텍스트는 문장 단위로 토큰화 및 길이 정규화를 수행하여 텍스트 감정 분류 모델 입력으로 변환한다.

다음으로 모달별 감정 분류 단계에서는 음성 SER 모델(2.1), 텍스트 감정 분류 모델(2.2), 얼굴 표정 기반 감정 분류 모델(2.3)이 각각 동일한 감정 클래스 순서의 확률 분포를 출력한다. 각 모달의 출력 포맷을 통일함으로써, 입력 형태가 서로 다르더라도 후기 융합 단계에서 직접 결합이 가능하도록 설계하였다.

후기 융합 단계(2.4)에서는 일정 시간 동안 누적된 모달별 확률 벡터를 모달리티별 평균으로 요약한 뒤, 모달 가중치를 적용한 가중 평균으로 통합 확률 분포를 계산한다. 모달 가중치는 각 모달 모델의 성능(정확도)을 반영한 성능 비례 가중치와 균등 가중치를 결합하는 하이브리드 전략으로 설정하여, 특정 모달에 편향되지 않으면서도 신뢰도 높은 모달의 기여를 반영하도록 하였다. 최종 결과는 단일 단계(0~4)로 정량화하지 않고, 5개 감정 클래스에 대한 확률 분포로 유지하여 복합 정서를 보존한다.

마지막으로 통합 감정 분포와 주요 감정(최대 확률 클래스)은 실시간 스트리밍(SSE) 기반으로 시각화되며, 동시에 상담 LLM의 입력 컨텍스트로 제공되어 응답의 톤, 공감 강도, 제안 전략이 사용자의 정서 신호를 반영하도록 한다. 이하 절에서는 각 모달 감정 분류 모델의 구현(2.1~2.3)과 결정 수준 후기 융합 기반 통합 감정 확률 산출 방법(2.4)을 순차적으로 상세히 기술한다.

2.1 음성 기반 감정 분류 모델 구현 및 고도화 방안

Wav2Vec2를 백본(Backbone)으로 사용하는 한국어 음성 감정 인식(SER) 모듈의 성능 최적화 방안을 제안한다. 제안 모델은 MFCC(Mel-Frequency Cepstral Coefficients, 멜-주파수 켑스트럼 계수)나 로그멜 필터뱅크(Log-Mel Filterbank)와 같은 수작업 특징 추출(Hand-crafted Feature Extraction) 과정 없이 원시 음성 파형(Raw Waveform)으로부터 직접 음향 특징을 학습하는 종단간(End-to-End) 접근법을 채택한다. 이를 통해 한국어 고유의 복잡한 운율적 특성(Prosodic Features)과 발화 환경의 다양성에 효과적으로 대응하고자 한다. 특히, 레이블링 비용이 높은 한국어 자유 대화 데이터의 한계를 극복하기 위해, 다국어 데이터로 사전 학습된 wav2vec2-large-xlsr-53 모델을 이용한 전이 학습(Transfer Learning)을 적용하여 소량의 데이터만으로도 높은 일반화 성능 확보를 목표로 한다.

모델 학습에는 AIHub에서 제공하는 '감정이 태깅된 자유대화 (성인) 데이터셋'을 활용하였다[15]. 모델의 아키텍처는 wav2vec2-large-xlsr-53 백본 위에 4~6 계층의 Transformer Encoder를 추가하여 시간적 맥락 정보를 보강하는 구조를 가진다. 입력 음성은 3초 길이의 창(Window)과 1초의 스트라이드(Stride)로 분할되어 장기 의존성(Long-term Dependency)을 학습하며, 최종 출력 시퀀스는 Global Average Pooling을 통해 고정된 차원의 z_audio 임베딩으로 요약된다. 데이터 전처리 단계에서는 스테레오 채널 분리 후 모노 채널을 선택하고, VAD(Voice Activity Detection)를 통해 무음 구간을 제거한다. 또한, 피크 클리핑(Peak Clipping) 및 저품질 샘플 제거를 통해 데이터의 질을 관리하며, 7-클래스 감정 레이블을 5-클래스로 매핑하여 모델의 변별력을 높인다. 학습 안정화를 위해 AdamW 옵티마이저, OneCycleLR 스케줄러, FP16 혼합 정밀도 학습, 레이블 스무딩(Label Smoothing), 그리고 그래디언트 클리핑(Gradient Clipping)을 적용한다. 추론 시에는 z_audio, 로짓(logits), 확률(probabilities)과 함께 시간 구간 (t_start, t_end) 정보를 출력하여 타 모달리티와의 동기화를 지원하며, Temperature Scaling을 통해 모델 출력 확률의 신뢰도를 보정한다.

성능 최적화는 정규화, 데이터 증강, 학습 전략 고도화, 모델 규모 확장 등 네 가지 방향으로 체계적으로 접근한다. 첫째, 과적합 제어를 위해 Dropout(0.5~0.6), Weight Decay(0.03~0.05), Label Smoothing(0.15~0.2) 등의 정규화 기법을 적용하여 검증 정확도를 62-65%까지 확보한다. 둘째, 실제 환경의 가변성에 대한 모델 강건성을 높이기 위해 배경 소음 믹싱, 속도 및 피치 조절, 볼륨 변조, SpecAugment와 같은 데이터 증강 기법을 적극 활용하고 VAD 정밀화, 노이즈 제거, 50% 중첩 윈도잉 등 전처리 과정을 강화하여 65-70%의 정확도를 목표로 한다. 셋째, 클래스 불균형 문제를 완화하는 Focal Loss, 안정적인 수렴을 유도하는 Cosine Annealing 스케줄러, 그리고 조기 종료(Early Stopping, patience=5)와 같은 고도화된 학습 전략과 필요시 소규모 앙상블을 적용하여 정확도를 68-75% 수준까지 끌어올린다. 마지막으로, 가용 자원이 허용하는 범위 내에서 백본을 wav2vec2-large-xlsr-53으로 상향하고 상단 Transformer 및 분류기 계층을 확장하여 모델의 표현력을 극대화한다.

최종적으로 성능과 자원 효율의 균형을 고려하여, 정

규화(Dropout 0.5, Weight Decay 0.03)와 균형 잡힌 데이터 증강, Focal Loss 및 Cosine Annealing을 기본 구성으로 채택하는 것을 권장한다. 이 구성을 통해 검증 정확도 75-82%, 훈련 정확도 85-88%에 도달하고, 과적합 갭은 5-10%p 범위 내로 수렴할 것으로 기대된다. 추론 지연 시간은 3초 클립 기준 약 0.15초, 메모리 사용량은 약 12GB 수준으로 유지된다. 나아가 텍스트 및 비전 신호와의 후기 융합(Late-fusion)을 통해 3-5%p의 추가적인 성능 향상을 기대할 수 있다. 다만, 과도한 증강이나 모델 확장은 성능 저하를 유발할 수 있으므로, 하이퍼파라미터 탐색과 조기 종료를 통해 위험을 관리하는 것이 필수적이다.

2.2 텍스트 기반 감정 분류 모델 구현 및 고도화 방안

본 연구에서는 한국어 문장의 감정 분류를 위해 사전학습 언어모델인 SKT의 KoBERT(skt/kobert-base-v1)를 기반으로 fine-tuning을 수행하였다. KoBERT는 한국어 코퍼스를 기반으로 사전학습된 BERT 구조의 모델로, 문맥 정보를 효과적으로 반영할 수 있어 감정 분석과 같은 자연어 이해 과제에 적합하다.모델 학습에는 'KEMDy20 한국어 멀티모달 감정 데이터셋[16]'과 AIHub에서 제공하는 '한국어 감정 정보가 포함된 연속적 대화 데이터셋[17]'을 활용하였다. 두 데이터셋은 다양한 감정 상태와 실제 대화 맥락을 포함하고 있어, 모델이 보다 현실적인 상담 상황에서의 정서 표현을 학습할 수 있도록 구성하였다.전체 학습 파이프라인은 데이터 전처리, 모델 구성, 학습 및 검증 단계로 이루어진다.

입력 데이터는 'Sentence'와 'Emotion' 필드로 구성된다. 문자열 형태의 감정 레이블은 정수 인덱스로 변환(Label Encoding)된다. 각 문장은 KoBERTTokenizer를 통해 토큰화되며, 문장의 시작과 끝에 특수 토큰 [CLS]와 [SEP]가 추가된다. 이후 모든 시퀀스는 고정 길이(max_len)에 맞춰 패딩(Padding) 또는 절단(Truncation) 과정을 거쳐 input_ids와 attention_mask 텐서로 변환되어 모델의 입력으로 사용된다.

사전 학습된 KoBERT 모델의 마지막 트랜스포머 블록 출력에서 BERT 구조에서 제안된 [CLS] 토큰에 해당하는 768차원 은닉 상태 벡터를 문장 전체의 대표 임베딩으로 사용한다[18]. 이 벡터 위에 과적합 방지를 위한 Dropout 레이어와 최종 감정 클래스 수만큼의 출력을 갖는 단일 Linear Layer로 구성된 분류 헤드(ClassificationHead)를 추가하여 최종 모델을 구성했다.

손실 함수로는 다중 클래스 분류에 표준적인 Cross-Entropy Loss를, 평가 지표로는 정확도(Accuracy)를 사용하였다. 옵티마이저는 AdamW를 채택하였으며 학습률은 5e-5로 설정하고, 선형 웜업(비율 0.1) 이후 코사인 스케줄러를 적용하였다. 배치 크기는 64, 에폭 수는 5로 두었고, 드롭아웃 비율은 0.5로 설정하였다. 또한 그래디언트 클리핑은 최대 노름 1.0으로 제한하였다.

학습 절차는 데이터를 학습, 검증, 테스트 셋으로 분할하고, DataLoader를 통해 미니배치를 구성하는 것으로 시작한다. 매 에폭마다 학습 데이터로 모델 가중치를 업데이트(역전파)하고, 검증 데이터로 성능을 평가하여 정확도가 가장 높은 최적의 모델을 최종 선택 및 저장한다.

KoBERT 기반 감정 분류 모델은 추론 단계에서 입력 문장에 대한 분석 결과를 출력한다. 모델의 최종 Linear Layer는 각 감정 클래스에 대한 로짓(Logit)을 산출하며 이는 소프트맥스 함수를 통해 개별 감정에 대한 값으로 변환된다.

최종 출력은 두 가지 형태로 제공된다. 첫째, 가장 높은 확률을 가진 클래스를 단일 예측 레이블로 반환한다. 둘째, 모든 감정 클래스에 대한 전체 확률 분포({'happy': 0.92, 'depressed': 0.03, ...})를 출력할 수 있다. 이 확률 분포는 모델 예측의 신뢰도를 나타내며, 어떤 감정들 사이에서 모델이 혼동하는지에 대한 분석 정보를 제공한다.

2.3 얼굴 표정 기반 감정 분류 모델 구현 및 고도화 방안

비전 처리 모듈은 사용자 표정의 세부 감정 상태를 실시간으로 정밀 인식하기 위해 잔차 네트워크(Residual Network) 기반의 ResNet18을 적용하였다. 모델 학습에는 Kaggle의 FER-2013(Facial Expression Recognition 2013)[19] 데이터셋을 활용했으며, 서비스 적용 시 혼동이 큰 클래스를 정리하고 분별력을 높이기 위해 7개 기본 감정(angry, disgust, fear, happy, sad, surprise, neutral) 중 핵심 5개 클래스(angry, happy, depressed, surprised, neutral)로 재구성하였다. 이때 FER-2013의 'sad' 라벨은 서비스 및 융합 단계에서 사용하는 감정 체계(EMOTIONS = ['happy', 'depressed', 'surprised', 'angry', 'neutral'])와 일관성을 맞추기 위해 'depressed'로 매핑하여 사용하였다.

FER-2013은 48×48 크기의 흑백(face-cropped) 이미지로 구성되어 있으므로, ResNet18의 입력 규격(224×224)에 맞추기 위해 Bicubic interpolation 기반의 업스케일링을 수행하였다. 전처리 단계에서는 transforms.Resize

((224, 224), interpolation=InterpolationMode.BICUBIC)로 입력 크기를 정규화한 뒤, ToTensor() 변환과 Normalize(mean=[0.5]*3, std=[0.5]*3) 정규화를 적용하여 모델 입력 분포를 안정화하였다. 또한 학습 과정에서는 클래스 불균형의 영향을 완화하고 특히 'depressed'와 같이 상대적으로 성능이 낮은 클래스의 인식률을 개선하기 위해, 해당 클래스 중심의 데이터 증강(Data Augmentation)을 강화하였다(좌우 반전, 회전 등). 이를 통해 데이터 편향을 줄이고 실제 환경에서의 표정 변화에 대한 강건성을 높였다.

모델 평가 단계에서는 학습된 체크포인트를 로드한 후(model.load_state_dict), 테스트 데이터셋을 배치 단위로 추론하여 소프트맥스 확률 분포를 산출하고(torch.softmax), 예측 결과를 CSV로 저장하여 모달 융합 단계에서 활용 가능하도록 하였다. 이러한 구현은 서비스 관점에서 "표정 입력, 감정 확률 출력, 결과 기록/전달" 흐름을 명확히 제공하며, 추론 결과를 후속 의사결정(후기 융합 및 우울 관련 신호 반영)에 직접 연결할 수 있다는 장점이 있다.

실험 결과, 전체 정확도와 Macro F1-score 관점에서 안정적인 성능을 확보하였으며, 특히 데이터 증강을 통해 기존 FER-2013에서 상대적으로 취약하게 나타나기 쉬운 'depressed' 클래스의 분류 성능을 유의미하게 개선하였다. 그 결과, 전체 분류 성능의 균형이 향상되었고, 멀티모달 후기 융합에서 중요한 우울 관련 신호의 신뢰도를 높여 최종 감정 분포 산출의 안정성에도 긍정적으로 기여하였다.

향후에는 AffectNet·RAF-DB 등 대규모 고해상도 데이터셋을 추가한 전이 학습(Transfer Learning)을 적용하여 표정 표현의 다양성과 실제성을 더 폭넓게 학습시키고, 클래스별 성능 편차를 추가로 줄임으로써 비전 모듈 및 최종 멀티모달 모델의 성능을 지속적으로 강화할 수 있다[20].

2.4 후기 융합 기반 통합 감정 확률 분포 산출

마지막으로, 음성·텍스트·표정 모달리티에서 독립적으로 산출된 감정 분석 결과는 결정 레벨의 후기 융합(Decision-level Late Fusion) 방식으로 통합된다. 본 연구의 융합 단계는 '우울 단계(0~4)'와 같은 이산 등급으로 정량화하지 않으며, 최종 결과를 5개 감정 클래스에 대한 확률 분포 형태로 유지한다. 이를 통해 사용자 정서가 단일 범주로 과도하게 단정되는 것을 방지하고, 복합 감정(혼재된 정서) 상태를 상담 응답 생성 단계에서 유연하게 반영할 수 있도록 설계하였다.

구현 측면에서 Late Fusion 모듈은 각 모달리티로부터 출력되는 감정 확률 벡터를 일정 시간 구간 동안 버퍼(buffer)에 누적한 후, 설정된 간격(interval)마다 이를 통합하는 방식으로 동작한다. 감정 클래스는 EMOTIONS = ['happy', 'depressed', 'surprised', 'angry', 'neutral']의 5개로 범주로 정의되며, 각 모달리티 모델은 동일한 클래스 순서에 대응하는 확률 벡터를 출력한다. 얼굴 모달리티는 웹캠 프레임 기반으로 비교적 높은 빈도(약 10 FPS)로 지속적으로 추론되어 버퍼에 누적되며, 음성과 텍스트는 사용자의 발화가 발생했을 때만 이벤트 기반으로 추론되어 버퍼에 추가된다. 융합 시점에는 각 모달 버퍼 내 결과들을 평균(mean)으로 요약한 뒤, 사전에 정의된 모달 가중치를 적용해 통합 확률을 계산한다.

모달 가중치는 임의값이 아니라, 각 모달 모델의 성능(정확도)을 반영하도록 설계하였다. 본 시스템은 얼굴(74%), 음성(65%), 텍스트(66%)의 성능 지표를 사용하여 성능 비례 가중치(performance-proportional weights)를 계산하고, 특정 모달에 편향되지 않도록 균등 가중치(equal weights)를 함께 섞는 하이브리드 전략을 적용하였다. 구체적으로 α=0.6으로 설정하여 "성능 기반 60% + 균등 40%"를 결합한 가중치를 사용하였으며, 최종적으로 얼굴≈0.350, 음성≈0.323, 텍스트≈0.326의 가중치가 산출된다. 융합은 다음과 같이 수행된다. (1) 모달별 버퍼 평균 확률을 계산하고, (2) 해당 모달 가중치를 곱해 누적합을 구하며, (3) 사용된 모달 가중치 총합으로 정규화하여 확률 합이 1이 되도록 보정한다. 이후 최종 감정 라벨은 통합 확률 분포에서 가장 큰 값을 갖는 클래스(argmax)로 결정하되, 상담 단계에서는 상위 감정뿐 아니라 전체 확률 분포를 함께 활용하도록 설계하였다.

또한 Late Fusion 모듈은 융합 결과와 함께 실제로 융합에 사용된 모달리티 및 각 모달에서 누적된 결과 개수를 함께 기록한다(얼굴 N개, 음성 M개, 텍스트 K개). 이는 실시간 서비스 환경에서 특정 모달 입력이 부족한 상황(음성 미입력, 얼굴 인식 실패 등)을 구분하고, LLM 상담 응답 생성 시 "근거 모달의 존재 여부"를 함께 제공하여 해석 가능성과 신뢰성을 높이는 데 기여한다. 융합이 완료되면 버퍼를 초기화하고 다음 주기에서 새롭게 누적된 관측값으로 다시 통합을 수행한다.

3. 시스템 설계 및 서비스 구현

본 연구에서는 멀티모달 감정 인식 결과를 기반으로 사용자 맞춤형 상담 응답을 생성하기 위해 대형 언어 모델(LLM)을 활용하였다. LLM은 멀티모달 데이터를 직접 처리하는 구조가 아니라, 감정 인식 모듈(Wav2Vec2, KoBERT, ResNet18)에서 도출된 최종 감정 상태를 입력으로 받아 자연어 상담 응답을 생성하는 역할을 수행한다.

구체적으로, 시스템은 음성, 텍스트, 얼굴 표정으로부터 각각 추출된 감정 확률을 Late Fusion 방식으로 통합하여 최종 감정 레이블을 결정하고, 이를 상담 컨텍스트 정보와 함께 LLM의 입력 프롬프트로 구성한다. 이때 프롬프트는 단순 질의 형태가 아니라, 상담 상황을 반영한 구조화된 템플릿 형태로 설계되었다.

3.1 시스템 아키텍처 및 데이터 파이프라인

본 연구에서 제안하는 시스템은 웹 기반 실시간 AI 심리상담 애플리케이션으로, 사용자의 얼굴 표정(Face), 음성(Voice), 텍스트(Text) 신호를 통합적으로 활용하여 감정 상태를 추정하고, 이를 기반으로 공감적 상담 응답을 생성·제공한다. 기존 텍스트 중심 상담 챗봇은 사용자의 비언어적 정서 신호(표정, 음성 톤)를 충분히 반영하기 어렵다는 한계가 있다. 이에 본 애플리케이션은 멀티모달 감정 분석 결과를 결정 수준 후기 융합(Late Fusion) 방식으로 통합하고, 그 결과를 실시간으로 시각화함과 동시에 상담 대화 생성 과정에 직접 반영하도록 설계하였다. 특히 정서 상태를 단일 점수나 단계로 단정하지 않고, 5개 감정 클래스(happy, depressed, surprised, angry, neutral)에 대한 확률 분포 형태로 유지함으로써 복합적인 감정 상태를 유연하게 표현한다.

시스템의 전체 구성은 (1) 백엔드 실시간 감정 분석 서비스, (2) 감정 데이터 실시간 스트리밍, (3) 상담 대화 API 및 응답 생성, (4) 프론트엔드 UI/UX의 네 부분으로 이루어진다. 백엔드는 Python 기반의 Flask 프레임워크로 구현되었으며, 감정 분석은 백그라운드 스레드에서 지속적으로 수행된다. 얼굴 모달리티의 경우 웹캠 입력을 약 10 FPS로 수집하여 표정 기반 감정 확률을 추론하고 이를 Late Fusion 버퍼에 누적한다. 음성과 텍스트 모달리티는 사용자의 발화 이벤트(음성 인식 결과 확정 또는 텍스트 전송) 발생 시 분석이 수행되며, 각각 음성 감정 모델과 텍스트 감정 모델의 출력 확률을 동일한 클래스 순서의 확률 벡터로 정규화하여 버퍼에 추가한다.

시스템은 설정된 융합 간격(본 구현에서는 2초)마다 버퍼에 누적된 관측값을 모달리티별 평균으로 요약한 후, 모달 가중치를 적용한 가중 평균을 통해 최종 감정 확률 분포를 산출한다. 이때 최종 결과에는 확률 분포뿐만 아니라 융합에 사용된 모달리티의 종류와 누적 개수(face N개, voice M개, text K개)를 함께 기록하여, 실시간 환경에서 특정 모달리티 입력이 부재한 상황을 구분하고 결과 해석의 투명성을 높였다.

실험 환경은 Ubuntu Linux 운영체제, Framework는 PyTorch, Flask (Backend), React (Frontend), 라이브러리는 Transformers (KoBERT, Wav2Vec2), OpenCV (ResNet18), API는 OpenAI GPT-4o (Chat), Web Speech API (STT)를 사용하였다.

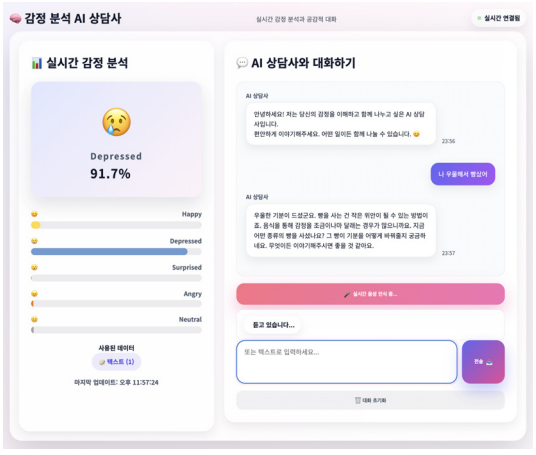
실시간 서비스 제공을 위해 감정 결과 전달에는 SSE(Server-Sent Events) 방식을 적용하였다. 백엔드는 Late Fusion이 수행될 때마다 최신 감정 결과를 큐에 저장하고, 스트리밍 엔드포인트(/api/emotions/stream)를 통해 클라이언트로 지속적으로 전송한다. 프론트엔드는 EventSource를 통해 스트림을 구독하여 감정 데이터를 수신하며, 수신된 확률 값에 따라 주요 감정 표시와 막대 그래프를 실시간으로 갱신한다. 이 과정에서 별도의 새로고침 없이 감정 분포가 지속적으로 반영되므로, 사용자는 상담 대화 중에도 자신의 정서 변화가 어떻게 추정되고 있는지를 직관적으로 확인할 수 있다.

3.2 멀티모달 융합 기반 상담 응답 생성

상담 대화 기능은 /api/chat 엔드포인트를 중심으로 구성된다. 텍스트 입력의 경우 해당 발화에 대한 텍스트 감정 분석 결과를 Late Fusion 버퍼에 누적하고, 최신 융합 결과를 기반으로 상담 응답을 생성한다. 음성 입력의 경우에는 음성 인식 결과 텍스트를 텍스트 감정 모델에 입력하여 의미 기반 감정 신호를 확보하고, 오디오 데이터가 제공될 경우 음성 감정 모델을 통해 음성 톤 기반 감정 신호를 추가로 추론한다. 또한 실시간으로 갱신되는 얼굴 표정 기반 감정 결과를 함께 활용하여 통합 감정 확률 분포를 산출한다. 이후 시스템은 Late Fusion 결과와 모달리티별 근거 요약을 LLM의 시스템 프롬프트에 포함시켜, 사용자의 발화 내용과 정서 상태를 동시에 고려한 공감적 상담 응답을 생성한다. 이때 Late Fusion 결과는 결론이 아닌 정서적 단서로 활용되며, 최근 관측이 없는 모달리티가 존재할 경우 이를 명시하도록 프롬프트 지침을 구성하여 과도한 확신을 방지하였다.

3.3 사용자 인터페이스 및 경험 (UI/UX)

프론트엔드 UI는 단일 페이지 구조로 설계되었으며, 화면을 좌측의 감정 대시보드 영역과 우측의 상담 채팅 영역으로 분할하였다.



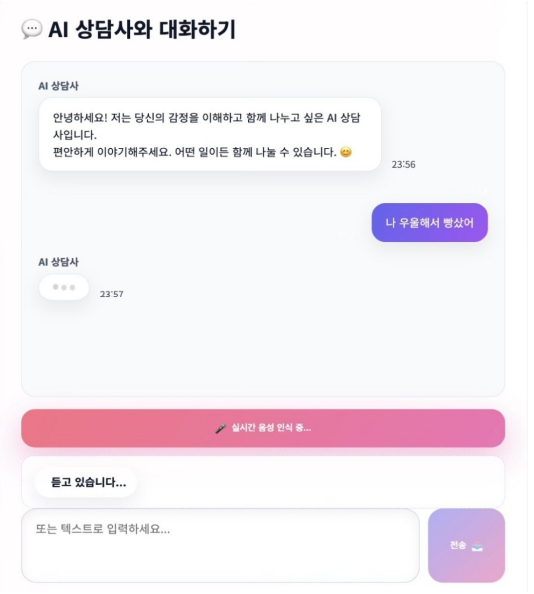
[Fig. 1] Single-page UI: emotion dashboard (left), chat (right).

감정 대시보드는 현재 가장 높은 확률을 가지는 감정을 강조하는 Primary Emotion 카드와 함께, 5개 감정 클래스에 대한 확률 막대 그래프를 제공한다. 이러한 시각화 방식은 정서 상태를 단일 라벨로 단정하지 않고 확률 분포로 제시함으로써, 사용자가 자신의 복합적인 감정 상태를 스스로 해석할 수 있도록 돕는다. 또한 최근 융합에 포함된 모달리티와 누적 횟수를 함께 표시하여 결과의 근거와 최신성을 사용자에게 제공한다.

상담 채팅 영역은 텍스트 입력과 더불어 브라우저의 Web Speech API를 활용한 연속 음성 인식을 지원한다. 음성 인식 과정에서 중간 인식 결과(interim transcript)를 실시간으로 UI에 표시하고, 일정 시간 침묵이 감지되면 발화를 자동으로 확정하여 서버로 전송함으로써 자연스러운 대화 흐름을 제공한다. AI 응답 생성 대기 시간에는 전체 화면 로딩 대신 채팅창 내부에 타이핑 인디케이터를 표시하고, 응답 수신 후에는 문자 단위로 출력되는 타이핑 효과를 적용하였다.

이는 상담사가 답변을 작성 중인 것과 유사한 사회적 신호를 제공하여 대화 몰입감을 유지하기 위한 UI 설계이다.

마지막으로, 본 애플리케이션은 민감한 데이터를 취급하므로 보안과 개인정보 보호를 고려하였다. OpenAI API 키는 코드에 직접 포함하지 않고 환경변수(.env)를 통해 관리하였으며, 웹캠과 마이크 입력은 브라우저 권한 기반으로 사용자 동의 하에 수집된다. 또한 실시간 환경에서 일부 모달리티 입력이 부족하거나 인식 실패가 발생할 수 있음을 전제로, 모달리티 부재 상태를 명시적으로 처리하여 UI와 프롬프트에 반영하였다.



[Fig. 2] Chat: voice dictation + typing indicator.

요약하면, 본 연구의 웹 애플리케이션은 얼굴·음성·텍스트 기반 감정 확률 산출, Late Fusion을 통한 통합 감정 분포 생성, SSE 기반 실시간 시각화, 멀티모달 정서 문맥을 반영한 LLM 상담 응답 생성, 그리고 사용자 경험을 고려한 인터페이스 설계를 통합함으로써 기존 텍스트 중심 AI 상담의 한계를 보완하는 실시간 멀티모달 심리 상담 서비스를 구현하였다.

4. 결론

본 연구는 음성, 텍스트, 얼굴 표정이라는 상이한 정서 신호를 통합한 멀티모달 감정 인식과 이를 기반으로 한 실시간 상호작용을 결합하여, AI 심리 상담 대화 서비스의 기술적·서비스적 가능성을 제시하였다. 기술적으로는 각 모달리티의 특성을 반영한 감정 분석 결과를 결정 수준에서 통합함으로써, 단일 입력 채널에 의존하는 기존 대화형 시스템에서 발생할 수 있는 정서 정보의 불완전성과 맥락 인식의 한계를 보완할 수 있음을 시사한다.

이러한 접근은 사용자 대화 맥락 속에서 보다 안정적이고 일관된 정서 추론을 가능하게 하며, 감정 상태 변화에 반응하는 대화 흐름을 구성하는 데 기여할 수 있다.

서비스 관점에서 본 연구는 AI 기반 심리 상담 대화 서비스가 정신건강 지원의 접근성과 지속성을 확장할 수 있음을 보여준다. 실시간 대화형 인터페이스는 사용자가 일상적인 환경에서 정서적 상태를 표현하고 반응을 받을 수 있는 저부담 상호작용을 가능하게 하며, 초기 정서 점검이나 정서적 지지 제공 수단으로 활용될 잠재력이 있다. 특히 시간적·공간적 제약으로 인해 기존 상담 서비스 이용이 어려운 사용자에게, 이러한 대화 서비스는 심리적 지원의 진입 장벽을 낮추는 보완적 도구로 기능할 수 있다. 또한 멀티모달 구조는 사용 환경에 따라 입력 채널을 유연하게 활용할 수 있어 다양한 서비스 시나리오로의 확장 가능성을 시사한다.

한편, AI 심리 상담 대화 서비스의 도입은 윤리적·사회적 고려를 필수적으로 수반한다. 상담 대화 과정에서 수집되는 음성, 텍스트, 영상 데이터는 민감한 개인 정보를 포함할 수 있으므로, 개인정보 보호와 데이터 보안, 명확한 사전 동의 절차가 전제되어야 한다. 또한 문화적 배경과 정서 표현 방식의 다양성에 따라 서비스 경험이 달라질 수 있으므로, 공정성과 신뢰성을 지속적으로 점검할 필요가 있다. 향후 연구에서는 설문 기반 만족도 평가 등을 통해 상담 품질 및 사용자 경험을 보완할 예정이다.

본 연구는 제한된 데이터와 실험 환경을 기반으로 수행되어 실제 상담 대화의 장기적 상호작용과 임상적 효과를 충분히 검증하지 못했다는 한계를 지닌다. 향후 연구에서는 멀티모달 감정 인식 모델의 성능을 정량적으로 평가하기 위해 단일 모달 대비 성능 비교 실험 및 다양한 공개 데이터셋 기반 검증을 수행할 예정이다. 또한 LLM 기반 상담 응답의 품질을 평가하기 위해 사용자 설문, 전문가 평가, 그리고 실제 상담 시나리오 기반 실험을 포함한 다각적인 평가 체계를 구축할 계획이다. 아울러 개인별 정서 표현 특성과 대화 맥락을 반영하는 개인화 전략과 공공 정신건강 서비스와의 연계를 모색함으로써, AI 심리 상담 대화 서비스가 책임 있는 디지털 정신건강 지원 도구로 발전할 수 있을 것으로 기대한다.

REFERENCES

[1] J.I.Leee, "Analysis of Health Behaviors and Sociodemographic Factors Influencing Depression among Adults," Journal of Internet of Things Convergence, Vol.11, No.5, pp.163-169, 2025.

[2] G.J.Kim, "Intelligent Health Pattern Analysis System Using u-Health Data," Journal of Internet of Things Convergence, Vol.11, No.4, pp.43-48, 2025.

[3] E.S.Oh, S.R.Gwon, J.M.Oh, B.Peng, T.K.Kim, "Implementation of a real-time public transportation monitoring system," Journal of Internet of Things Convergence, Vol.10, No.4, pp.9-19, 2024.

[4] I.R.Numonov, T.K.Kim, "IoT-based Smart alarm system," Journal of Internet of Things Convergence, Vol.10, No.4, pp.35-41, 2024.

[5] S.C.Kwon, D.H.Lee, B.C.Jang, "Zero-shot Korean Sentiment Analysis with Large Language Models: Comparison with Pre-trained Language Models," Journal of The Korea Societyof Computer and Information, Vol.29, No.2, pp.43-50, 2024.

[6] J.M.Choi, "Quiz Generation System Using Google Forms and ChatGPT," Journal of Internet of Things Convergence, Vol.10, No.6, pp.105-110, 2024.

[7] J.S.Ryu, S.K.Jung, "Empirical Study on 70B-Level AI Interview Evaluation Performance of Small LLaMA 3.1 8B Using Generative AI and LLM Techniques," Journal of Internet of Things Convergence, Vol.11, No.6, pp.249-256, 2025.

[8] M.S.Ann, "Social Media Big Data Analysis on Artificial Intelligence (AI) Psychological Counseling," The Society of Convergence Knowledge Transactions, Vol.7, No.3, pp.65-75, 2019.

[9] S.J.Kang, "Applications of Large Language Models in Counseling and Psychotherapy: Opportunities, Limitations, and Ethical Considerations," Journal of Digital Contents Society, Vol.25, No.12, pp.3751-3759, 2024.

[10] S.J.Lee, "An autoethnography on the Development and counseling Experience of Non-face-to-face Mental Health Mobile Service Platform of Mental Health Social Welfare expert," Journal of Internet of Things Convergence, Vol.8, No.5, pp.63-70, 2022.

[11] K.K.Fitzpatrick, A.Darcy, and M.Vierhile, "Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial," JMIR Mental Health, Vol.4, No.2, e19, 2017.

[12] R.Fulmer, A.Joerin, B.Gentile, L.Lakerink, and M.Rauws, "Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial," JMIR Mental Health, Vol.5, No.4, e64, 2018.

[13] V.Ta, C.Griffith, C.Boatfield, X.Wang, M.Civitello, H.Bader, E.DeCero, A.Loggarakis, "User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis," Journal of Medical Internet Research, Vol.22, No.3, e16235, 2020.

[14] K.Gadzicki, R.Khamsehashari, C.Zetzsche, "Early vs

Late Fusion in Multimodal Convolutional Neural Networks," 2020 IEEE 23rd International Conference on Information Fusion (FUSION), pp.1-6, 2022.

[15] AI Hub, Emotion-tagged free conversation (adult) [Internet], https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71631

[16] e·preTX, Korean Multimodal Sentiment Dataset 2020 (KEMDy20)[Internet], https://epretx.etri.re.kr/dataDetail?lang=ko&id=318

[17] AI Hub, Continuous dialogue dataset containing Korean sentiment information[Internet], https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71631

[18] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K., "BERT: Pre-training of deep bidirectional transformers for language understanding", NAACL-HLT, pp.4171-4186, 2019.

[19] kaggle, FER-2013[Internet], https://www.kaggle.com/datasets/msambare/fer2013

[20] Mollahosseini, A., Hasani, B., & Mahoor, M.H. (2017). "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild." IEEE Transactions on Affective Computing, Vol.10, No.1, pp.18-31.

박 준 용(Jun-Yong Park) [준회원]

• 2020년 3월 ~ 2026년 2월 : 국립부경대학교 컴퓨터·인공지능공학부

〈관심분야〉
인공지능(AI)

김 균 호(Gyun-Ho Kim) [준회원]

• 2020년 3월 ~ 2026년 2월 : 국립부경대학교 컴퓨터·인공지능공학부

〈관심분야〉
인공지능(AI)

조 강 래(Kang-Rae Jo) [준회원]

• 2020년 3월 ~ 2026년 2월 : 국립부경대학교 컴퓨터·인공지능공학부

〈관심분야〉
인공지능(AI)

김 태 국(Tae-Kook Kim) [종신회원]

■ 2004년 8월 : 고려대학교 전기전자전파공학부(공학사)
■ 2006년 8월 : 고려대학교 메카트로닉스학과(공학석사)
■ 2014년 8월 : 고려대학교 모바일솔루션학과(공학박사)
■ 2016년 3월 ~ 2022년 2월 : 동명대학교 AI학부 교수
■ 2022년 3월 ~ 현재 : 국립부경대학교 컴퓨터·인공지능공학부 교수

〈관심분야〉
사물인터넷(IoT), 콘텐츠 전송 네트워크(CDN), 이동성, 인공지능(AI), 빅데이터(big data), 모바일 서비스