사물인터넷융복합논문지 Vol. 12, No. 4, pp. 195-206, 2026 https://doi.org/10.20465/KIOTS.2026.12.4.195

CNN-WavLM 기반 환경음·음성 감정 인식 플랫폼 구현

윤혜림[1], 이현승[1], 김태국[2*]
[1]국립부경대학교 컴퓨터·인공지능공학부 학생, [2]국립부경대학교 컴퓨터·인공지능공학부 교수

Implementation of CNN-WavLM-based Environmental Sound and Speech Emotion Recognition Platform

Hye-Rim Yoon[1], Hyun-Seung Lee[1], Tae-Kook Kim[2*]
[1]Student, School of Computer and Artificial Intelligence Engineering, Pukyong National University
[2]Professor, School of Computer and Artificial Intelligence Engineering, Pukyong National University

요 약 본 연구는 음성과 환경음을 활용하여 사용자의 감정을 기록하고 회상할 수 있는 청각 중심 감정 기록 플랫폼을 구현하였다. 기존 사진, 영상, 텍스트 중심 기록 방식은 일상 소리가 지니는 감정적 분위기와 상황적 맥락을 충분히 반영하기 어렵다. 이에 본 연구에서는 사용자가 직접 녹음한 소리를 입력으로 활용하고, 인공지능 기반 감정 분석 기술을 적용하여 자동 감정 태깅 기능을 구현하였다. 제안한 시스템은 입력 오디오의 음성 포함 여부를 판별한 뒤, 음성이 포함된 경우 WavLM 기반 음성 감정 인식 모델을 적용하고, 음성이 포함되지 않은 경우 멜-스펙트로그램 기반 CNN 환경음 분석 모델을 적용하는 분기 구조로 설계하였다. CNN 모델은 환경음을 장소(scene) 단위로 분류한 후, 사전에 정의한 scene-emotion mapping 규칙을 통해 감정 태그로 변환한다. 또한 플랫폼은 감정 태그와 함께 녹음 시점의 시간, 위치, 사용자 기록을 저장하여 사용자가 감정별, 날짜별, 위치별로 기록을 탐색하고 회상할 수 있도록 구성하였다. 이를 통해 본 연구는 음성 감정 인식과 환경음 분석 모델을 실제 서비스 흐름에 통합하고, 청각 정보를 활용한 감정 아카이빙 플랫폼의 구현 가능성을 제시하였다.

주제어 : 감정 인식, 사물인터넷, 음성 감정 분석, 환경음 분석, WavLM, Convolutional Neural Network (CNN), 인공지능

Abstract This study implements an auditory-centered emotion recording platform that enables users to record and recall their emotions using speech and environmental sounds. Conventional recording methods based on photos, videos, and text have limitations in capturing the emotional atmosphere and contextual information conveyed by everyday sounds. To address this issue, this study uses sounds directly recorded by users as input data and applies artificial intelligence-based emotion analysis techniques to implement an automatic emotion-tagging function. The proposed system is designed as a branching structure that first determines whether the input audio contains speech. If speech is detected, a WavLM-based speech emotion recognition model is applied. If speech is not detected, a Mel-spectrogram-based CNN model is used to analyze environmental sounds. The CNN model classifies environmental sounds at the scene level and then converts the classification results into emotion tags based on predefined scene-emotion mapping rules. In addition, the platform stores the emotion tags together with the time, location, and user records at the moment of recording, allowing users to explore and recall their records by emotion, date, and location. Through this implementation, this study integrates speech emotion recognition and environmental sound analysis models into an actual service flow and demonstrates the feasibility of an emotion archiving platform that utilizes auditory information.

Key Words : Emotion Recognition, Internet of Things (IoT), Speech Emotion Analysis, Environmental Sound Analysis, WavLM, Convolutional Neural Network (CNN), Artificial Intelligence (AI)

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2023-00242528).
*교신저자 : 김태국(king@pknu.ac.kr)
접수일 2025년 12월 24일 수정일 2026년 05월 25일 심사완료일 2026년 06월 22일

1. 서론

현대 사회에서 개인의 일상과 감정을 기록하는 방식은 디지털 기술의 발전과 함께 지속적으로 변화해 왔다. 스마트폰과 소셜 미디어의 보급으로 사진과 영상은 가장 보편적인 기록 수단이 되었으며, 개인의 경험은 시각적 콘텐츠를 중심으로 저장되고 공유되고 있다. 이러한 기록 환경 속에서 청각 기반 기록은 시각적 기록을 보완하여 개인의 감정과 경험을 보다 다층적으로 표현할 수 있는 기록 방식으로 제시될 수 있다[1].

인지과학 및 심리학 연구에 따르면 인간의 기억은 단일 감각에 의해 형성되지 않으며, 다양한 감각 자극이 결합된 형태로 저장된다. 그중에서도 청각 자극은 특정한 시간과 공간, 그리고 감정 상태를 즉각적으로 떠올리게 하는 특성을 지닌다[2,3]. 일상에서 무심코 지나치는 목소리, 주변 환경음, 배경 소리는 과거의 경험과 감정을 자연스럽게 상기시키는 강력한 기억 단서로 작용할 수 있다[4]. 그럼에도 불구하고 이러한 청각 기반 기억을 체계적으로 기록하고 활용할 수 있는 서비스는 상대적으로 제한적이다.

한편 최근 인공지능 기술의 발전으로 음성 데이터를 활용한 감정 인식 연구가 활발히 진행되고 있다[5-8]. 그러나 대부분의 연구는 사람 간의 대화 음성을 중심으로 감정을 분류하는데 초점을 맞추고 있으며, 주변 환경에서 발생하는 비언어적 소리를 감정 분석의 대상으로 확장한 사례는 충분히 축적되지 않았다. 이는 실제 일상 환경에서 발생하는 다양한 청각 자극을 감정 기록에 활용하는 데 있어 한계를 남긴다.

이에 본 연구는 사용자가 직접 녹음한 음성과 환경음을 기반으로 감정을 분석하고 저장할 수 있는 청각 중심의 감정 기록 플랫폼을 제안한다. 인공지능 모델을 활용하여 소리에 포함된 감정을 자동으로 태깅하고, 녹음 시점의 시간과 위치 정보, 사용자 메모를 함께 저장함으로써 감정을 기준으로 한 기억 탐색을 가능하게 한다.

2. 관련 연구 및 기술

본 연구는 청각 자극을 기반으로 감정을 기록하고 회상하는 플랫폼을 제안한다는 점에서, 관련 연구 역시 두 가지 측면에서 검토될 필요가 있다. 하나는 소리와 감정, 기억 간의 관계를 다룬 선행 연구이며, 다른 하나는 음성 및 환경음을 입력으로 감정을 분류하는 인공지능 기술에 관한 연구이다. 본 장에서는 이러한 연구 흐름을 바탕으로, 청각 자극의 기억 환기 효과에 대한 기존 연구를 정리하고 본 연구에서 적용한 합성곱 신경망 (Convolutional Neural Network, CNN) 기반 환경음 분석 기술과 WavLM 기반 음성 감정 인식 모델을 중심으로 관련 기술 동향을 살펴본다.

2.1 소리와 기억의 연관성

청각 자극은 인간의 기억 형성과 회상 과정에서 중요한 역할을 수행하는 감각 요소로 알려져 있다. 인지과학 및 신경과학 연구에 따르면, 소리와 음악과 같은 청각 자극은 감정 처리와 기억 형성에 관여하는 뇌 영역을 동시에 활성화함으로써, 감정과 결합된 기억을 효과적으로 환기시키는 특성을 지닌다.

Jancke의 연구는 음악이 청각 피질뿐만 아니라 편도체와 해마를 자극하여 감정과 기억을 밀접하게 연결한다는 점을 보여주었으며, 이는 기억이 단일 감각이 아닌 감정적·인지적 요소가 결합된 형태로 저장된다는 점을 시사한다[2]. 또한 Janata는 음악과 같은 청각 단서가 개인의 자서전적 기억을 유발하는 신경학적 메커니즘을 분석함으로써, 특정 소리가 과거의 경험과 감정을 즉각적으로 떠올리게 하는 강력한 기억 단서로 작용할 수 있음을 실증적으로 제시하였다[3].

이러한 이론적 배경을 바탕으로, Oleksik 등의 Sonic Gems 연구는 소리를 기억 매체로 활용한 대표적인 사례를 제시한다. 해당 연구는 오디오 기록이 단순한 음성 저장 수단을 넘어, 개인의 감정과 추억을 보존하고 나아가 집단의 문화적 기억까지 담아낼 수 있음을 보여준다. 연구 결과에 따르면 참가자들은 특정 소리를 단순한 정보로 인식하기보다, 자신의 정서적 경험과 결합된 기억의 일부로 받아들이는 경향을 보였다. 특히 짧고 의미 있는 소리는 사진이나 영상과 달리 명확한 시각 정보를 제공하지 않음에도 불구하고, 오히려 사용자의 상상력과 감정적 몰입을 활성화하여 깊은 인상을 남기는 것으로 나타났다[4].

Sonic Gems 연구는 시각 정보의 부재가 기억 회상에 부정적으로 작용하기보다는, 개인의 경험과 감정을 기반으로 한 능동적인 기억 재구성을 유도할 수 있음을 시사한다. 이는 소리 기록이 감정 중심의 회상과 몰입을 촉진하며, 사진이나 영상과는 차별화된 방식으로 기억을 구성할 수 있음을 의미한다. 이러한 특성은 소리가 개인의 내면적 경험과 결합되어 기억을 형성하는 매체로서 높은 잠재력을 지닌다는 점을 보여주며, 청각 기반 감정 기록

시스템의 필요성을 뒷받침하는 중요한 근거가 된다.

2.2 음성 및 환경음 감정 인식 기술

2.2.1 CNN

기존의 환경음 분류 연구에서는 음향 신호를 시간-주파수 영역으로 변환한 후 이를 입력으로 활용하는 접근이 일반적으로 사용되어 왔다. 특히 멜-스펙트로그램(Mel-spectrogram)은 인간의 청각 특성을 반영한 주파수 스케일을 제공함으로써 환경음에 포함된 에너지 분포와 음색 변화를 효과적으로 표현할 수 있는 특징으로 널리 활용된다. 이러한 특성으로 인해 멜-스펙트로그램을 입력으로 하는 CNN 구조는 환경음 분류 분야에서 대표적인 분석 방법으로 자리 잡아왔다[9]. Piczak은 ESC-50과 UrbanSound8K 데이터셋을 이용한 연구를 통해, CNN이 환경음에 내재된 시간-주파수 패턴을 효과적으로 학습할 수 있음을 보였다[10,11]. 해당 연구에서는 비교적 단순한 CNN 구조만으로도 기존의 수작업 특징 기반 분류 기법보다 향상된 분류 성능을 달성하였다. 이를 통해 CNN이 환경음 분류 문제에 적합한 모델 구조임을 입증하였다. 이후 멜-스펙트로그램 기반 CNN은 다양한 환경음 및 음향 장면 분류 연구에서 기준선 모델(Baseline Model)로 활용되며 표준적인 접근 방식으로 확립되었다.

다만 이러한 CNN 기반 환경음 분류 방식은 입력 특징이 사전에 정의된 스펙트로그램 변환 과정에 의존한다는 한계를 지닌다. 주파수 분해 방식이나 시간 해상도와 같은 요소가 전처리 단계에서 고정되기 때문에 모델은 해당 표현 공간 내에서만 음향적 패턴을 학습하게 된다. 이로 인해 다양한 녹음 환경이나 음향 조건이 혼재된 실제 환경에서는 환경음이 지니는 감정적·맥락적 특성을 충분히 반영하는데 제약이 발생할 수 있다. 이러한 특성은 환경음 분석에서 CNN 구조의 장점을 유지하면서도, 실제 환경에서의 불확실성과 출력 해석 문제를 함께 고려할 필요성을 시사한다[11].

본 연구는 멜-스펙트로그램 기반 CNN의 구조적 안정성과 환경음 분류에서의 검증된 성능을 바탕으로, 이를 감정 분석 서비스에 적용하는 방향으로 접근하였다. 환경음의 음향적 특성을 감정 범주로 해석할 수 있는 구조를 통해 단순 분류 결과를 넘어 감정적 맥락을 반영한 출력이 가능하도록 하였다. 이러한 방식은 실제 환경에서 발생하는 다양한 음향 입력에 대해 보다 안정적인 감정 예측을 가능하게 하였다.

2.2.2 WavLM

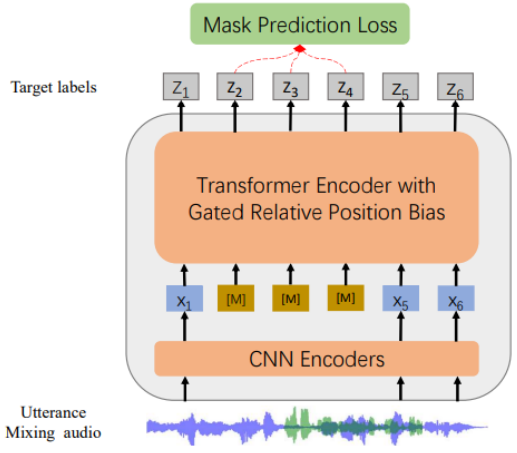
최근 음성 감정 인식(Speech Emotion Recognition, SER) 분야에서는 대규모 음성 데이터를 활용한 자기지도학습(Self-Supervised Learning, SSL) 기반 사전학습 모델이 핵심적인 연구 흐름으로 자리 잡고 있다. 감정 레이블은 주관성이 높고 대규모 수집이 어려운 특성을 지니기 때문에, 라벨이 없는 음성 데이터로부터 일반화 가능한 음성 표현을 학습하는 SSL 접근법은 SER 문제에 특히 적합한 방식으로 평가된다. 이러한 모델들은 음성의 음향적 구조와 화자 특성을 사전학습 단계에서 학습한 뒤, 감정 분류와 같은 하위 과제에 전이학습 방식으로 활용된다.

WavLM(Wav2vec-based Large Model)[12]은 이러한 SSL 기반 연구 흐름을 대표하는 모델로, 기존 wav2vec 2.0과 HuBERT 계열 모델을 확장하여 보다 현실적인 음성 환경을 고려한 학습 구조를 제안한다. WavLM은 마스킹된 음성 신호를 복원하는 masked speech prediction 과제와 더불어, 잡음이 포함된 입력으로부터 깨끗한 음성 표현을 학습하는 denoising 과제를 동시에 수행한다. 이와 같은 이중 학습 목표는 실제 환경에서 빈번하게 발생하는 배경 소음, 중첩 발화(overlapped speech), 채널 왜곡 상황에서도 견고한 음성 표현을 학습할 수 있도록 설계되었다.

이러한 학습 과정에서 WavLM은 발화의 언어적 내용뿐만 아니라 화자의 음색, 억양, 발화 리듬, 감정적 강세와 같은 비언어적(paralinguistic) 특성을 효과적으로 포착한다. 감정 인식은 음성의 의미 정보보다 이러한 비언어적 요소에 크게 의존하기 때문에, WavLM은 SER 과제에 특히 적합한 음성 표현 모델로 평가된다. 실제로 WavLM은 SUPERB 벤치마크를 포함한 다양한 음성 처리 평가의 감정 인식, 화자 분리, 음성 향상 등 다수의 하위 과제에서 기존 HuBERT 및 wav2vec 2.0 대비 일관된 성능 향상을 달성하였다[5]. 이러한 결과는 WavLM이 범용 음성 표현 학습 모델로서 높은 활용 가능성을 지님을 보여준다.

[Fig. 1]은 WavLM의 자기지도학습 기반 사전학습 구조를 나타낸다[5]. 입력 음성 신호는 CNN 인코더를 통해 저수준 음향 특징으로 변환된 후, Transformer 인코더를 통해 시계열 음성 표현으로 학습된다. 이 과정에서 WavLM은 마스킹된 음성 구간을 예측하는 masked speech prediction 과제와 잡음이 포함된 음성으로부터 견고한 표현을 학습하는 denoising 과제를 동시에 수행함으로써 실제 환경에서도 강인한 범용 음성 표현을

학습하도록 설계되었다.



[Fig. 1] Self-Supervised Pre-training Architecture of WavLM

3. 소리 기반 감정 분석 AI 모델 구현

본 연구에서는 음성과 비언어적 환경음을 기반으로 사용자의 감정을 자동으로 분석하고 기록할 수 있는 소리 기반 감정 분석 모델을 제안한다. 제안하는 모델은 사용자가 일상에서 녹음한 소리를 입력으로 받아, 인공지능 기반 감정 인식 과정을 통해 해당 소리에 포함된 감정 정보를 추출하고 이를 기록 시스템에 활용하는 것을 목표로 한다. 이를 위해 본 연구에서는 대화 음성과 환경음이라는 서로 다른 특성을 지닌 소리를 각각의 분석 방식으로 처리하여 감정을 분류하는 구조를 설계하였다. 본 장에서는 제안하는 감정 분석 모델의 전체적인 처리 흐름과 각 구성 요소의 역할을 중심으로 설명한다.

3.1 음성 데이터 감정 분석 모델

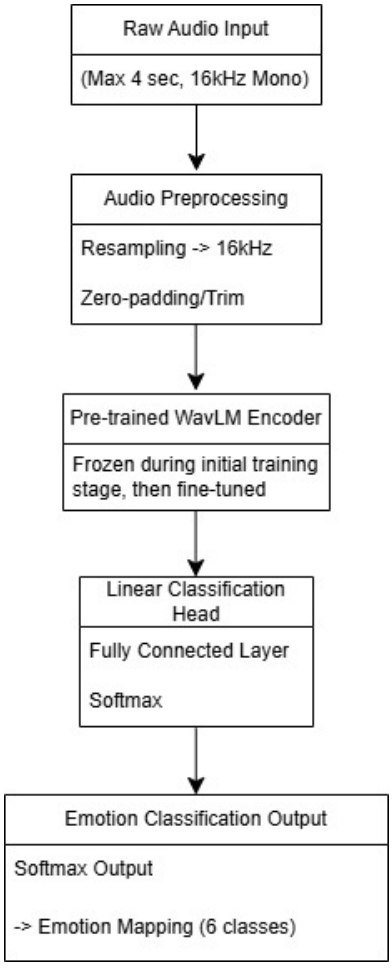
3.1.1 데이터 전처리 및 라벨 구조

본 연구에서는 대화 음성 기반 감정 인식을 위해 AI-Hub에서 제공하는 '감정 분류를 위한 대화 음성 데이터셋'을 사용하였다[13]. 해당 데이터셋은 일상 대화를 기반으로 수집된 음성 데이터로 구성되어 있으며, 원본 데이터에는 '슬픔', '행복', '중립', '분노', '놀람', '두려움', '혐오'의 감정 레이블이 포함되어 있다. 감정 회상을 목적으로 한 서비스 특성을 고려하여, '혐오' 감정을 유사한 부정적 정서를 지닌 '두려움' 범주에 통합하였으며, 최종적으로 '슬픔', '행복', '중립', '분노', '놀람', '두려움'의 총 6개 감정 클래스로 재구성하였다. 모든 음성 데이터는 16kHz 샘플링 레이트로 통일한 후, 모델 입력의 일관성을 유지하기 위해 최대 4초 길이로 자르거나 부족한 경우 패딩을 적용하였다. 전처리된 데이터는 전체 데이터셋의 80%를 학습용, 20%를 검증용으로 분리하여 사용하였다.

3.1.2 모델 선택 및 구조

초기 실험 단계에서는 멜-스펙트로그램을 입력으로 하는 CNN 기반 감정 분류 모델을 대화 음성 데이터에 적용하였다. CNN 모델은 음성 신호를 주파수 영역으로 변환하여 국소적인 시간-주파수 패턴을 학습하는 데 효과적인 구조로, 비교적 단순한 구현과 안정적인 학습이 가능하다는 장점을 지닌다. 그러나 대화 음성의 경우 억양 변화, 발화 리듬, 화자별 발화 습관과 같은 장기적인 시계열 정보와 화자 특성이 감정 인식에 중요한 역할을 한다. 이러한 특성으로 인해, 국소 패턴 중심의 CNN 기반 접근은 대화 음성에 포함된 고수준 비언어적 정보를 충분히 반영하는 데 한계를 보였다. 실제 실험 결과, CNN 기반 모델은 검증 데이터셋에서 약 47.7%의 정확도를 기록하며 대화 음성 감정 분류 성능에 제한이 있음을 확인하였다.

이러한 특성을 고려하여, 해당 연구에서는 대화 음성 감정 인식 모델로 대규모 음성 데이터로 사전학습된 자기지도 학습 기반 모델인 WavLM을 채택하였다. WavLM은 사전학습 과정에서 장기적인 시계열 문맥과 화자 특성을 함께 학습하므로, 대화 음성에 포함된 억양과 감정적 강세를 보다 효과적으로 표현할 수 있다. 본 연구에서는 Hugging Face Transformers 라이브러리에서 제공하는 microsoft/wavlm-base-plus[14] 사전학습 모델을 기반으로, 감정 분류를 위한 선형 분류기를 추가하여 fine-tuning을 수행하였다. 이때 WavLM의 자기지도학습 기반 사전학습 과정은 재현하거나 재수행하지 않았으며, 사전학습된 가중치를 유지한 상태에서 감정 분류 과제에 맞는 미세조정만을 적용하였다. 입력 음성은 WavLM feature encoder를 통해 잠재 표현(latent representation)으로 변환되며, 이후 분류기를 통해 각 감정 클래스에 대한 확률값을 출력한다. [Fig. 2]는 WavLM 기반 음성 감정 분석 모델의 fine-tuning 구조를 나타낸다.



[Fig. 2] WavLM-based Speech Emotion Recognition Fine-tuning Architecture

3.1.3 학습 설정

모델 학습은 Google Colab 환경에서 PyTorch 프레임워크를 사용하여 수행하였다. 손실 함수로는 클래스 간 데이터 불균형 문제를 완화하기 위해 가중치가 적용된 CrossEntropyLoss를 사용하였으며, 최적화 기법으로는 AdamW 옵티마이저를 적용하였다. 학습 초반의 안정화를 위해 WavLM의 feature extractor는 초기 2000 step 동안 고정한 뒤, 이후 단계부터 점진적으로 미세 조정을 수행하였다. 학습률은 3×10^{-5}, 배치 크기는 8, 학습 에포크 수는 5로 설정하였다. 모델 성능 평가는 Accuracy와 Weighted F1-score를 기준으로 수행하였으며, 검증은 10,000 step마다 진행하였다. 〈Table 1〉은 모델을 학습시킬 때 설정한 항목들을 요약한 표이다.

〈Table 1〉 Configuration of the WavLM-based Speech Emotion Recognition Model

Parameter	Value
Input	Raw speech waveform
Sampling rate	16kHz
Max duration	4 seconds
Base model	WavLM-base-plus
Training strategy	Fine-tuning (no pre-training)
Feature extractor	Frozen → partially unfrozen
Optimizer	AdamW
Loss Function	Weighted CrossEntropyLoss
Evaluation metrics	Accuracy, Weighted F1-score

3.1.4 학습 결과 및 비교 분석

총 5 epoch의 학습 후, WavLM 기반 음성 감정 인식 모델은 검증 데이터셋에서 71.5%의 정확도와 71.4%의 Weighted F1-score를 기록하였다. 이는 CNN 기반 모델이 약 47.7%의 정확도에 머문 것과 비교하여 유의미한 성능 향상이다. 이러한 결과는 사전학습된 음성 표현을 활용함으로써 대화 음성에 포함된 억양과 감정적 강세를 보다 효과적으로 반영할 수 있었기 때문으로 판단된다. 이를 통해 SSL 기반 음성 모델이 대화 음성 감정 인식 과제에서 기존 CNN 기반 접근법보다 우수한 성능을 제공함을 확인하였다.

3.2 환경음 데이터 감정 분석 모델

3.2.1 데이터 전처리 및 라벨 구조

환경음 감정 분석 모델의 학습에는 TAU Urban Acoustic Scenes 2022 Mobile Development 데이터셋을 사용하였다[15]. 〈Table 2〉는 장면-감정 매핑 규칙을 나타낸다. 메타데이터(meta.csv)에 포함된 장소(scene) 라벨을 기반으로 각 오디오 파일에 10개의 장소 클래스를 부여하고, 이를 CNN 모델의 분류대상으로 설정하였다. 이후 모델의 예측 결과는 사전에 정의된 규칙에 따라 6개의 감정 범주(neutral, happy, surprise, anger, fear, sadness)로 변환된다.

본 연구의 scene-emotion mapping은 환경음이 사용자의 감정을 직접적으로 표현한다기보다, 특정 장소가 제공하는 상황적·정서적 분위기를 감정 태그로 변환하기 위한 규칙 기반 매핑으로 설계하였다. 즉, 환경음 분석 모델은 사용자의 내적 감정을 직접 예측하는 것이 아니라, 입력된 환경음으로부터 장소(scene)를 먼저 추정하고, 해당 장소가 일반적으로 연상시키는 정서적 맥락을

감정 범주로 변환한다. 예를 들어 park와 shopping_mall은 여가, 휴식, 활동성과 관련된 공간으로 간주하여 happy로 매핑하였고, street와 tram은 혼잡도, 이동 스트레스, 소음 가능성을 고려하여 anger로 설정하였다. metro와 metro_station은 폐쇄적 공간, 혼잡, 지하 이동 환경이 줄 수 있는 긴장감을 고려하여 fear로 매핑하였다. 또한 bus_stop은 이동 전 대기, 정체, 일시적 고립감과 같은 상황적 특성을 고려하여 낮은 정서적 활성을 갖는 부정 감정 범주인 sadness로 설정하였다. 그리고 unclassified_sound는 CNN 모델의 scene 분류 결과 중 최대 softmax 확률값이 사전에 설정한 임계값보다 낮아 특정 장소 클래스로 확정하기 어려운 입력을 의미한다. 즉, 이는 독립적인 학습 클래스라기보다 모델의 예측 불확실성이 높은 환경음을 처리하기 위한 보조 범주이다. 이러한 입력은 특정 장소의 음향적 특성과 명확히 대응되지 않는 비정형 환경음으로 간주하였으며, 본 연구에서는 이를 낯설거나 예상하지 못한 소리에 대한 반응을 나타내는 surprise 감정으로 매핑하였다.

〈Table 2〉 Scene-Emotion Mapping Rules

Scene	Emotion
airport	happy
bus	neutral
bus_stop	sadness
tram	anger
metro	fear
metro_station	fear
park	happy
public_square	neutral
shopping_mall	happy
street	anger
unclassified_sound	surprise

오디오 입력은 librosa를 이용하여 16 kHz로 리샘플링한 뒤, 오디오를 3초 단위로 잘라 고정된 입력 길이를 유지하였다. 각 신호는 멜-스펙트로그램으로 변환하고 로그 스케일로 정규화하여 시간-주파수 영역의 음향적 특성을 입력으로 사용하였다. 이러한 전처리 과정은 환경음 데이터의 특성을 유지하면서 CNN 기반 모델이 학습하기에 적합한 입력 형태를 구성하기 위한 것이다.

3.2.2 모델 선택 및 구조

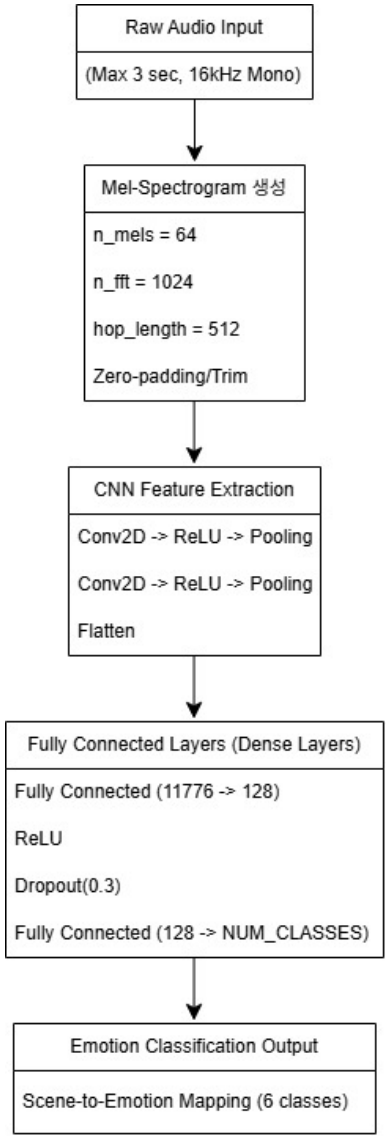
환경음 분류를 위해 멜-스펙트로그램을 입력으로 하는 CNN 기반 모델을 설계하였다. [Fig. 3]은 CNN 기반 환경음향장 분류 모델을 나타낸다. 모델은 두 개의 합성곱(Convolution) 계층과 완전연결(Fully Connected) 계층으로 구성된 2층 CNN 구조로 구현하였다. 각 합성곱 계층 뒤에는 ReLU 활성화 함수와 최대풀링(MaxPooling) 계층을 적용하여, 환경음에 포함된 국소적인 시간-주파수 패턴을 효과적으로 추출하고 특징의 안정성과 일반화 성능을 향상시켰다.

첫 번째 합성곱 계층에서는 저차원의 음향 특징을 추출하고, 두 번째 합성곱 계층에서는 보다 추상화된 시간-주파수 패턴을 학습하도록 구성하였다. 이후 특징 맵을 평탄화(flatten)한 뒤 완전연결 계층을 통해 10개 장소 클래스에 대한 분류 결과를 출력한다. 최종적으로 예측된 장소 분류 결과는 감정 분석 단계에서 감정 태그로 매핑되어 활용된다.

〈Table 3〉 Training Configuration and Experimental Settings

Item	Configuration
Development Environment	Google Colab
Framework	PyTorch
Input Data	Mel-spectrogram
Loss Function	Cross-Entropy Loss
Optimizer	AdamW (learning rate = 0.001)
Batch Size	16
Data Split	Train / Validation / Test
Training Hardware	GPU(CUDA)
Training Epochs	5

〈Table 3〉은 실험 환경을 나타낸다. 모델 학습은 Google Colab 환경에서 PyTorch 프레임워크를 기반으로 수행하였다. 손실함수로는 교차 엔트로피 손실(Cross-Entropy Loss)을 사용하였으며, 옵티마이저로는 AdamW(learning rate = 0.001)를 적용하였다. 데이터셋은 학습(train), 검증(validation), 테스트(test) 세 부분으로 분할하여 모델의 일반화 성능을 평가하였고, GPU(CUDA) 환경에서 총 5 epoch 동안 학습을 진행하였다. 학습된 모델 가중치는 체크 포인트 형태로 저장하여 추론 단계에서 재사용하였다.



[Fig. 3] CNN-based environmental sound scene classification model

3.2.3 예측 및 확신도 기반 분류 로직

모델의 출력은 장소 클래스에 대한 확률 분포로 계산되며, softmax 함수를 통해 각 클래스의 예측 확률을 산출하였다. 이 중 가장 높은 확률값을 해당 예측의 확신도(confidence)로 정의하였다. 확신도가 사전에 설정한 임계값(70%) 미만인 경우에는 특정 장소로 판단하지 않고 '미분류'로 처리하였다.

확정된 장소 예측 결과는 사전에 정의된 장소감정 매핑 규칙을 통해 최종 감정 범주로 변환된다. 이때 '미분류'로 판단된 입력은 특정 장소의 음향적 특성과 일관되게 대응되지 않는 예측 불확실성이 높은 비정형 환경음에 해당한다. 이러한 환경음은 일상적인 환경에서 반복적으로 관찰되는 소리라기보다, 평소에 자주 접하지 않는 음향적 특성을 포함하는 경우가 많다. 이러한 이유로 '미분류' 상태를 익숙하지 않은 소리에 대한 반응을 나타내는 감정 범주인 surprise로 매핑하였다. 이러한 확신도 기반 분류 로직은 실제 서비스 환경에서 발생할 수 있는 잡음이나 혼합 신호로 인한 오분류가 감정 출력에 직접적으로 반영되는 문제를 완화하기 위한 것이다.

3.2.4 다단계 라벨 매핑 설계

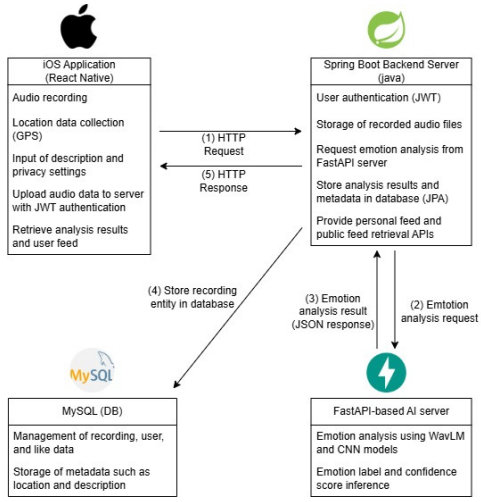
환경음은 개인의 순간적인 감정을 직접적으로 표현하기보다는, 특정 환경에서 반복적으로 형성되는 정서적 분위기나 경향을 반영하는 경우가 많다. 이러한 특성을 고려하여 본 연구에서는 환경음을 감정으로 직접 매핑하는 방식 대신, 장소 정보를 중간 단계로 활용하는 다단계 라벨 매핑 구조를 적용하였다.

환경음으로부터 먼저 장소를 추정한 뒤, 해당 장소가 일반적으로 동반하는 환경적 특성을 감정 범주로 연결함으로써 환경음에 포함된 맥락 정보를 감정 해석에 반영하고자 하였다. 이와 같은 접근은 개인의 주관적 감정을 정확히 추정하기보다는, 환경이 유발하는 정서적 경향을 안정적으로 표현하는 데 목적을 둔다.

4. 소리 기반 감정 회상 플랫폼 구현

본 연구에서는 앞서 제안한 소리 기반 감정 인식 시스템을 실제 사용자 환경에서 적용하기 위해 구현한 모바일 애플리케이션의 구조와 주요 기능을 설명한다. 인공지능 감정 인식 모델은 플랫폼 내에서 자동 감정 태깅 기능으로 연동되며, 사용자는 녹음된 소리를 기반으로 감정을 기록·공유·탐색할 수 있다. 본 장에서는 모바일 애플리케이션의 전체 구조와 주요 화면별 기능 구현을 중심으로 설명한다.

[Fig. 4]는 전체 시스템의 구조를 나타낸다. 플랫폼은 iOS 환경을 대상으로 구현되었으며, 사용자는 원하는 순간에 소리를 녹음하고, 모델을 통해 분석된 감정 결과를 일상 기록의 형태로 관리할 수 있다.



[Fig. 4] Overall System Architecture of the Audio-based Emotion Recording Platform

4.1 전체 피드 페이지 구현

전체 피드 페이지는 사용자의 감정 기록과 친구로 추가된 계정의 기록을 함께 모아볼 수 있도록 구성하였다. [Fig. 5]는 전체 피드 페이지 인터페이스를 나타낸다. 공개 범위는 상호 친구 관계로 제한되며, 불특정 다수에게 노출되거나 추천되는 구조를 배제함으로써 사용자가 신뢰하는 관계 내에서 일상의 감정과 상황을 공유할 수 있도록 설계하였다. 사용자는 전체 피드에서 특정 계정을 선택해 해당 사용자의 기록만을 모아보는 방식으로도 감정 기록을 탐색할 수 있다.

상호작용 기능으로는 '좋아요' 기능만을 제공하고, 댓글 및 메시지 기능은 의도적으로 포함하지 않았다. 이는 감정 기록을 공유하는 과정에서 타인의 시선이나 반응을 의식해 스스로 표현을 제한하는 상황을 줄이고, 사용자가 보다 솔직하게 자신의 감정 상태를 기록하고 정리하는 데 집중할 수 있도록 하기 위한 설계 선택이다. 대신 '좋아요'에 따른 알림 기능을 제공하여, 기록이 공유되었음을 인지할 수 있는 최소한의 반응만을 전달하도록 구성하였다.

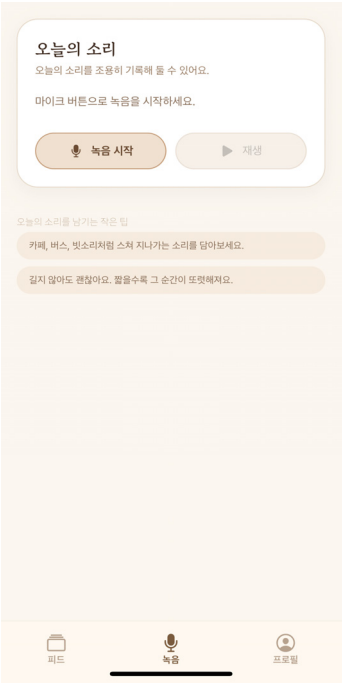
[Fig. 6]은 알림(Notification) 센터 인터페이스를 나타낸다. "오늘의 감정을 남겨보는 건 어떨까요? (How about capturing today's emotions?)", 친구 수락 등의 알림 정보를 나타낸다.

[Fig. 5] Overall feed page interface

[Fig. 6] Notification center interface

4.2 녹음 페이지 구현

녹음 페이지는 사용자가 특정 순간의 소리를 직접 기록할 수 있도록 실제 녹음 방식만을 지원한다. [Fig. 7]과 [Fig. 8]은 녹음 페이지 인터페이스, 오디오 업로드 인터페이스를 나타낸다. “오늘의 소리(Today's Sound)”를 녹음하여 저장할 수 있고 소리 기록을 남길 수 있다. 기존에 저장된 오디오 파일을 선택적으로 업로드하는 기능은 제공하지 않으며, 이는 감정이 발생한 순간의 맥락을 보다 직접적으로 반영하기 위한 설계이다. 녹음이 완료되면 사용자는 업로드 페이지를 통해 감정 분석 결과를 확인할 수 있다.



[Fig. 7] Recording page interface

녹음된 오디오는 음성 존재 여부에 따라 서로 다른 감정 분석 모델로 전달된다. 사람의 음성이 포함된 경우에는 음성 감정 분석 모델이 적용되며, 음성이 감지되지 않은 경우에는 환경음 감정 분석 모델이 사용된다. 분석 결과는 자동으로 제시되지만, 사용자가 이를 수정할 수 있도록 하여 자동 분석 결과에 대한 보완 가능성을 함께 제공한다. 또한 사용자는 감정 기록의 공개 여부와 위치 정보 저장 여부를 직접 선택할 수 있다. 이를 통해 개인 기록과 공유 기록을 구분하여 관리할 수 있도록 하였으며 감정 데이터의 활용 범위를 사용자 스스로 통제할 수 있도록 설계하였다.

[Fig. 8] Audio upload page interface

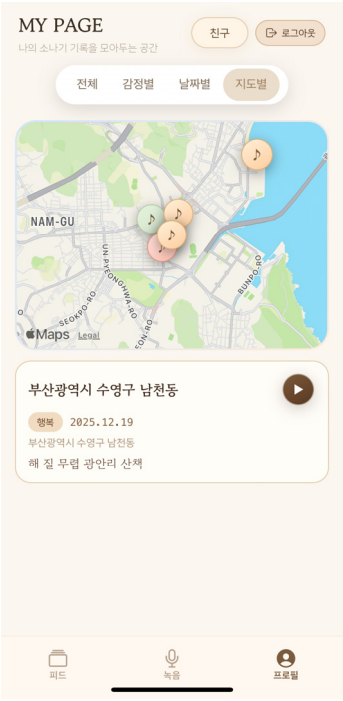
4.3 개인 피드 페이지 구현

개인 피드 페이지는 사용자의 감정 기록을 장기적으로 관리하고 회상할 수 있도록 설계되었다. [Fig. 9]는 감정 기반 피드 페이지 인터페이스를 나타낸다. 사용자는 자신의 기록을 감정, 날짜, 위치 정보를 기준으로 탐색할 수 있으며, 이를 통해 시간적·공간적 맥락 속에서 감정 변화의 흐름을 확인할 수 있다.

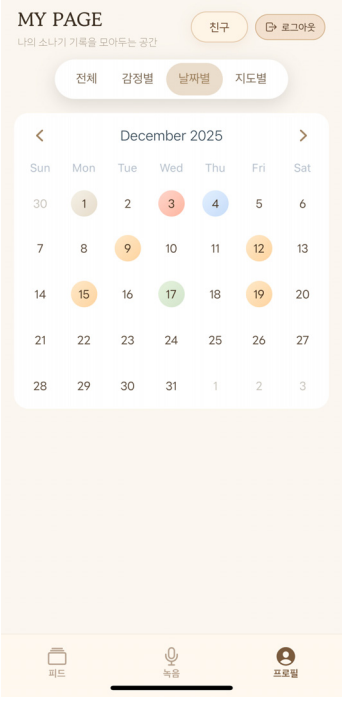
[Fig. 10]과 [Fig. 11]은 날짜 기반 피드 페이지 인터페이스와 지도 기반 피드 페이지 인터페이스를 나타낸다. 특히 지도 기반 탐색 기능은 감정 기록을 장소와 함께 시각적으로 확인할 수 있도록 하여 특정 공간에서의 감정 경험을 보다 직관적으로 회상할 수 있도록 한다. 이러한 구성은 감정 기록을 단순한 목록 형태의 데이터로 제한하지 않고, 감정 기록에 공간적 맥락을 부여하여 기록의 의미를 확장하는 데 기여한다.

[Fig. 12]는 친구 관리 페이지 인터페이스를 나타낸다. 개인 피드 페이지에는 친구 관리 기능이 함께 포함되어 있으며, 사용자는 해당 페이지에서 친구 검색, 친구 신청 및 수락·거절, 친구 목록 확인과 친구 삭제를 수행할 수 있다. 또한 친구 신청이나 수락과 같은 주요 상호작용이 발생할 경우 알림을 통해 이를 확인할 수 있도록 구성하였다.

[Fig. 9] Emotion-based feed page interface



[Fig. 11] Map-based feed page interface



[Fig. 10] Date-based feed page interface

[Fig. 12] Friend management page interface

4.4 사용자 인증 및 서비스 운용 구조

플랫폼은 사용자 계정을 기반으로 운영되며 회원가입 및 로그인 기능을 제공한다. 이를 통해 사용자별 감정 기록, 공개 범위 설정, 친구 관계가 계정 단위로 관리되도록 구성하였다.

로그인 성공 시 서버로부터 발급된 토큰을 클라이언트 로컬 저장소에 저장하고, 이후 앱 재실행 시 해당 토큰을 재사용하는 방식으로 로그인 상태를 유지하였다. 저장된 토큰은 사용자 정보 조회 요청을 통해 유효성을 검증하며, 검증 실패 시 토큰을 삭제하여 잘못된 세션 상태가 유지되는 것을 방지하였다. 이러한 자동 로그인 구조를 통해 반복적인 로그인 과정을 최소화하여 감정 기록 서비스의 지속적인 사용을 고려한 운용 구조를 마련하였다.

5. 결론

본 연구에서는 음성과 환경음을 활용하여 사용자의 감정을 기록하고 회상할 수 있는 소리 기반 감정 기록 플랫폼을 제안하고, 이를 모바일 애플리케이션 형태로 구현하였다. 제안한 플랫폼은 사용자가 특정 순간의 소리를 직접 녹음하면, 입력 오디오의 음성 포함 여부에 따라 WavLM 기반 음성 감정 인식 모델과 멜-스펙트로그램 기반 CNN 환경음 분석 모델을 분기 적용하여 감정 태그를 생성한다. 이를 통해 기존 텍스트, 사진, 영상 중심 기록 방식이 충분히 반영하기 어려운 청각적 맥락을 감정 기록에 활용할 수 있음을 보였다.

구현된 플랫폼은 자동 감정 태깅 기능과 사용자 수정 기능을 함께 제공하여 인공지능 모델의 분석 결과를 사용자가 보완할 수 있도록 설계하였다. 또한 감정 태그와 함께 녹음 시점의 시간 및 위치 정보를 저장하고, 감정별·날짜별·위치별 기록 탐색 기능을 제공함으로써 사용자가 자신의 감정 변화를 시간적·공간적 맥락 속에서 회상할 수 있도록 하였다. 아울러 상호 친구 관계를 기반으로 한 제한적 피드 구조와 최소한의 상호작용 요소를 적용하여, 감정 기록 공유 과정에서 발생할 수 있는 사회적 부담을 줄이고 개인적인 감정 기록에 집중할 수 있도록 구성하였다.

다만 본 연구에는 몇 가지 한계가 있다. 첫째, 동일한 소리나 장소라 하더라도 개인의 경험, 상황, 시간대에 따라 서로 다른 감정으로 해석될 수 있으나, 본 연구의 환경음 분석은 사전에 정의한 scene-emotion mapping 규칙에 기반하므로 개인별 정서 차이를 충분히 반영하지 못한다. 둘째, 음성과 환경음이 혼재된 복합적인 실제 환경에서는 음성 감정과 환경음 맥락이 서로 다르게 해석될 수 있어 감정 판단의 경계가 모호해질 수 있다. 셋째, 환경음 기반 감정 태깅은 사용자의 실제 내적 감정을 직접 예측한다기보다 장소적·상황적 맥락을 감정 범주로 변환한 결과이므로, 이에 대한 사용자 평가와 정량적 검증이 추가로 요구된다.

향후 연구에서는 사용자 설문과 장기 사용 데이터를 바탕으로 scene-emotion mapping의 타당성을 검증하고, 개인별 감정 해석 차이를 반영할 수 있는 개인화 모델을 도입할 필요가 있다. 또한 음성과 환경음이 동시에 포함된 복합 오디오 상황을 보다 정교하게 처리하기 위한 멀티모달 또는 다중 음원 분석 구조를 적용하고, 감정 분류 체계의 확장과 장기적인 감정 변화 분석을 통해 플랫폼의 활용 가능성을 더욱 높이고자 한다.

REFERENCES

[1] Y.S.Kim, K.S.Kim, Y.H.Ahn, D.Y.Kim, "Analysis of the Marketability of Emotional Diary Applications: Based on Keyword Trends and Media Coverage," Advanced Industrial Science, Vol.4, No.6, pp.179-194, 2025.

[2] L.Jancke, "Music, memory and emotion," Journal of Biology, Vol.7, No.21, 2008.

[3] P.Janata, "The neural architecture of music-evoked autobiographical memories," Cerebral Cortex, Vol.19, No.11, pp.2579-2594, 2009.

[4] G.Oleksik, L.M.Brown, "Sonic Gems: Exploring the Potential of Audio Recording as a Form of Sentimental Memory Capture," People and Computers XXII Culture, Creativity, Interaction (HCI), pp.163-172, 2008.

[5] G.Oleksik, L.M.Brown, "Enhancing Speech Emotion Recognition with Hybrid Graph Neural Networks: A GCN-GAT Framework," The Journal of Korean Institute of next generation computing, Vol.20, No.4, pp.7-20, 2024.

[6] Y.J.Nam, "Speech Emotion Recognition in Noisy Environments Based on a Denoising Convolutional Neural Network," Journal of Korea Institute of Information and Communication Engineering, Vol.27, No.6, pp.772-781, 2023.

[7] J.H.Yang, H.S.Choi, N.M.Moon, J.A.Kim, "Transformer Based Korean Emotion Recognition Model through Multi-domain Fusion," The Transactions of the Korea Information Processing Society, Vol.14, No.6, pp.459-467, 2025.

[8] J.J.Seo, T.I.Kang, I.Y.Kwak, "Pre-trained models and ensemble technique for speech emotion recognition,"

Journal of the Korean Data & Information Science Society, Vol.35, No.4, pp.445-459, 2024.

[9] N.Shao, R.Zhou, P.Wang, X.Li, Y.Fang, Y.Yang, X.Li, "CleanMel: Mel-Spectrogram Enhancement for Improving Both Speech Quality and ASR," IEEE Transactions on Audio, Speech and Language Processing, Vol.33, pp.3202-3214, 2025.

[10] S.Karam, S.J.Ruan, Q.M.Haq, L.P.Li, "Episodic memory based continual learning without catastrophic forgetting for environmental sound classification," Journal of Ambient Intelligence and Humanized Computing, Vol.14, pp.4439-4449, 2023.

[11] K.J.Piczak, "Environmental Sound Classification with Convolutional Neural Networks," IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), 2015.

[12] S.Chen, C.Wang, Z.Chen, Y.Wu, S.Liu, Z.Chen, J.Li, N.Kanda, T.Yoshioka, X.Xiao, J.Wu, L.Zhou, S.Ren, Y.Qian, Y.Qian, J.Wu, M.Zeng, X.Yu, F.Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," IEEE Journal of Selected Topics in Signal Processing (JSTSP), Vol.16, No.6, pp.1502-1518, 2022.

[13] AI Hub, Conversational speech dataset for emotion classification[Internet], https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=263.

[14] Hugging Face, WavLM-Base-Plus[Internet], https://huggingface.co/microsoft/wavlm-base-plus.

[15] Zenodo, TAU Urban Acoustic Scenes 2022 Mobile, Development dataset[Internet], https://zenodo.org/records/6337421.

윤 혜 림(Hye-Rim Yoon) [준회원]

■ 2021년 3월 ~ 2026년 2월 : 국립부경대학교 컴퓨터·인공지능 공학부

〈관심분야〉
사물인터넷, 인공지능(AI), 사용자 인터페이스(UI/UX)

이 현 승(Hyun-Seung Lee) [준회원]

■ 2021년 3월 ~ 2026년 2월 : 국립부경대학교 컴퓨터·인공지능 공학부

〈관심분야〉
사물인터넷, 데이터베이스(DB)

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·인공지능 공학부 교수

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