

메타버스 대화의 몰입감 증진을 위한 대화 감정 기반 실시간 배경음악 시스템 구현

김기락¹ 이상아¹ 김나현¹ 정문열^{12*}

¹서강대학교 아트&테크놀로지학과, ²메타버스전문대학원

{kirak, ivory, tinahyeon, moon}@sogang.ac.kr

Real-time Background Music System for Immersive Dialogue in Metaverse based on Dialogue Emotion

Kirak Kim¹ Sangah Lee¹ Nahyeon Kim¹ Moonryul Jung^{12*}

¹Department of Art & Technology, ²Graduate School of Metaverse, Sogang University

요약

메타버스 환경에서의 배경음악은 사용자의 몰입감을 증진시키기 위해 사용된다. 하지만 현재 대부분의 메타버스 환경에서는 사전에 매칭시킨 음원을 반복 재생하며, 이는 빠르게 변화하는 사용자의 상호작용 맥락에 어울리지 못해 사용자의 몰입감을 저해시키는 경향이 있다.

본 논문에서는 보다 몰입감 있는 메타버스 대화 경험을 구현하기 위해 1) 한국어 멀티모달 감정 데이터셋인 KEMDy20을 이용하여 발화로부터 감정을 추출하는 회귀 신경망을 구현하고 2) 음원에 arousal-valence 레벨이 태깅되어 있는 DEAM 데이터셋을 이용하여 발화 감정에 대응되는 음원을 선택하여 재생한 후 3) 아바타를 이용한 실시간 대화가 가능한 가상공간과 결합하여 몰입형 메타버스 환경에서 발화의 감정에 어울리는 배경음악을 실시간으로 재생하는 시스템을 구현하였다.

Abstract

To enhance immersive experiences for metaverse environments, background music is often used. However, the background music is mostly pre-matched and repeated which might occur a distractive experience to users as it does not align well with rapidly changing user-interactive contents.

Thus, we implemented a system to provide a more immersive metaverse conversation experience by 1) developing a regression neural network that extracts emotions from an utterance using KEMDy20, the Korean multimodal emotion dataset 2) selecting music corresponding to the extracted emotions from an utterance by the DEAM dataset where music is tagged with arousal-valence levels 3) combining it with a virtual space where users can have a real-time conversation with avatars.

키워드: 메타버스, 인공지능, 가상현실, 대화 내 감정 인식, 배경음악

Keywords: Metaverse, Artificial Intelligence, Virtual Reality, Emotion Recognition in Conversation, Background Music

*corresponding author: Moonryul Jung/ Sogang University (moon@sogang.ac.kr)

1. 서론

메타버스는 사회적인 상호작용이 일어나는 가상 공간으로 현시대 새로운 소통의 창구이다. 메타버스 내에서 사용자들은 텍스트 및 음성 채팅 뿐만 아니라 가상 캐릭터를 이용한 표정, 동작 등 비언어적 요소를 통해 다른 사용자들과 소통한다. 더욱 몰입감 있는 소통 경험을 위해 메타버스 플랫폼들은 시청각 자극을 활용하기도 한다. 예를 들어, 메타버스 플랫폼 중 하나인 계더타운은 사용자들이 가상의 맵에 배경 음악을 추가할 수 있는 기능을 넣어 공간 분위기를 조성하여 몰입도를 증가시킬 수 있도록 돕는다. 가상공간 체험에서 적절한 배경 음악은 사용자의 몰입감을 증진시킨다[1].

하지만 현재 메타버스 내에서 사용되는 배경 음악은 대부분 사전에 정해진 음악만을 반복 재생하여 변화하는 메타버스 환경의 맥락에 어울리지 않는 경우가 생긴다. 그리하여 본 시스템에서는 사용자의 발화 감정에 따라 변화하는 실시간 배경 음악 매칭 기능을 구현하여 메타버스 내에서 소통 시 보다 몰입감 있는 사용자 경험을 제공하고자 하였다.

2. 배경 지식 및 관련 연구

2.1 배경 지식

2.1.1 전이학습 (transfer learning)

기계 학습(machine learning) 분야에서 전이학습이란 신경망이 특정 문제를 풀기 위해 학습한 정보를 다른 분야의 문제를 풀기 위해 이용하는 것을 말한다[2]. 주로 특정 문제를 해결하기 위한 데이터가 부족할 때, 다른 문제를 풀기 위해 풍부한 데이터로 학습된 신경망의 가중치를 이용하는 식으로 이뤄진다. 이때 두 문제는 서로 공유하는 부분이 있어 한 문제를 풀 때 쓰이는 정보가 다른 문제를 풀 때에도 유용해야 전이학습이 성공적으로 이용될 수 있다. 대표적인 전이학습의 예시로 사물의 종류를 분류하는 신경망의 가중치를 이용하여 적은 양의 신호등 사진 데이터셋 만으로도 신호등의 상태를 분류하는 신경망을 만든 경우가 있다.

2.1.2 Arousal-valence Model

Arousal-valence model은 arousal과 valence의 두 값으로 감정을 나타내는 모형이다[3]. Arousal 값은 감정이 일으키는 흥분의 정도를 나타내며 값이 낮을수록 차분한 감정이고 값이 높을수록 흥분되는 감정이다. Valence 값은 감정의 긍/부정도를 나타내며 값이 낮을수록 부정적인 감정이고 값이 높을수록 긍정적인 감정이다. Arousal-valence model은 감정을 수치화 할 수 있기 때문에 1980년 Russell에 의해 제안된 이후 현재까지도 감정을 분석하는데 널리 쓰이는 모형이다.

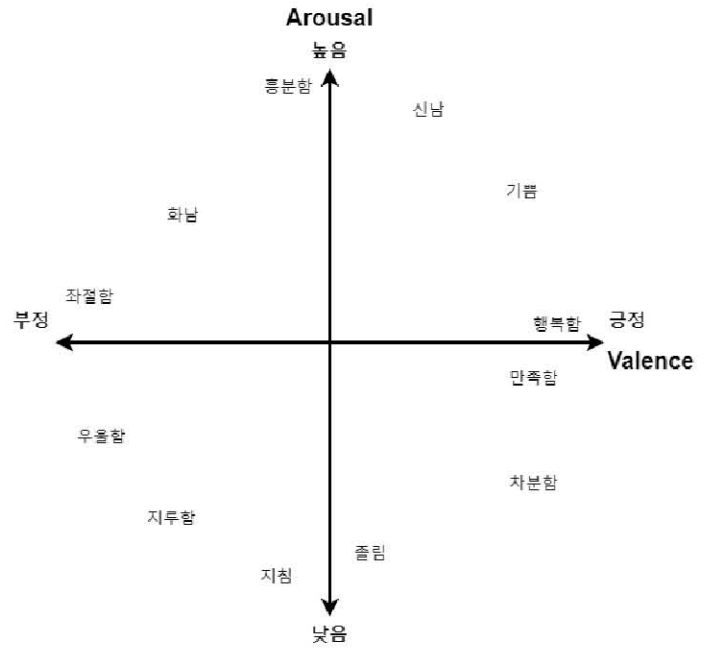


Figure 1: Emotion examples on an arousal-valence model

2.2 관련 연구

2.2.1 ETRI 한국어 기반 멀티모달 감정 데이터셋

ETRI 한국어 기반 멀티모달 감정 데이터셋은 한국어 발화에 다양한 멀티모달 데이터가 태깅된 데이터셋으로 성우의 상황극을 대상으로 수집된 KEMDy19와 일반인 자유발화를 대상으로 수집된 KEMDy20으로 구성되어 있다[4]. 수집된 데이터로는 발화 음성, 발화의 문맥적 의미, 생리반응 신호-피부전도도, 맥박관련 데이터, 손목 피부온도, 발화 별 카테고리 감정 레이블(기쁨, 놀람, 분노, 중립, 혐오, 공포, 슬픔), 발화 별 5단계(1~5)의 arousal과 valence 값이다.

ETRI 한국어 기반 멀티모달 감정 데이터셋은 2021년에 공개된 이후 여러 한국어 발화 감정 인식 연구에 사용되고 있다. 그 대표적인 예시로 MLP-Mixer 구조를 이용하여 한국어 대화에서 멀티모달 감정을 인식한 연구가 있으며[5], 사전학습된 언어모델과 음향모델을 이용하여 발화의 arousal 및 valence를 예측하는 모델을 제안한 연구가 있다 [6].

2.2.2 Wav2Vec2

Wav2Vec2는 Facebook에서 2020년 공개한 음성 인식을 위한 사전학습 모델이다[7]. 발화에서 음성적 특징 (acoustic feature)를 추출하는 모델로, 영어 발화 및 텍스트 쌍으로 구성되어 있는 Librispeech 데이터셋을 이용하여 마스킹 기법을 사용한 자기지도학습으로 훈련된 모델이다. Wav2Vec2.0은 여러 오디오 관련 태스크에서 적은 양의 라벨링 된 데이터로만 미세조정 되어도 SOTA를 달성하였다.

2.2.3 XLSR-Wav2Vec2

XLSR-Wav2Vec2(Cross-Lingual Representation Learning Wav2Vec2)는 Wav2Vec2.0 모델을 여러 언어의 음성으로 사전학습시켜 여러 언어에 공통적으로 드러나는 음성적 특징을 학습한 모델이다[8]. 2020년에 Facebook이 공개하였으며, Wav2Vec2.0 모델을 기반으로 마스킹을 이용한 자기지도학습과 대조적 학습 (contrastive learning)을 이용하여 훈련된 모델이다. 모델은 이산화된 음성적 토큰을 서로 다른 언어의 발화에 적용하도록 학습하여 언어 간 유사점을 찾는다. XLSR-Wav2Vec2 여러 오디오 관련 태스크에서 SOTA를 달성하여 단일 언어로 사전학습을 진행하는 것 보다 여러 언어로 사전학습이 진행되는 것이 여러 태스크에서 더 높은 성능을 낼 수 있다는 것을 보였다. Facebook은 논문과 함께 53개의 언어로 사전학습된 wav2vec2-large-xlsr-53 모델을 공개하였다. 학습에 사용된 데이터셋은 다양한 언어의 발화에 텍스트가 태깅되어 있는 데이터셋들로 MLS: Multilingual Librispeech, CommonVoice, Babel이다.

2.2.4 DEAM (The MediaEval Database for Emotional Analysis of Music)

DEAM은 음원에 arousal-valence 값이 태깅되어 있는 데이터셋이다[9]. DEAM은 음악이 어떤 감정을 표현하는지 분석하기 위해서 제작되었으며, 음악가들이 라벨을 태깅하였다. DEAM은 45초 길이의 1802개의 음원에 1) 음원의 구간별 arousal-valence 값, 2) 음원 전체의 arousal-valence 값, 3) 음성 특징 추출 프로그램인 openSMILE로 추출한 음원의 다양한 음성적 특징이 (Loudness, MFCC, Chroma 등) 태깅되어 있다. 음원은 락, 팝, 컨트리 등의 다양한 장르의 서양 대중 음악으로 구성 되어 있다. 음원의 출처는 freemusicarchive.org (FMA), jamendo.com, medlyDB 데이터셋이다.

3. 시스템 설명 및 구현 방법

본 시스템의 프레임워크는 1) 신경망을 이용하여 발화로 부터 감정 (arousal, valence) 레벨을 추출하고 2) 추출된 감정에 어울리는 음악을 선택하는 두 단계로 나뉜다.

3.1 감정 추출 신경망

발화로부터 감정을 추출하는 회귀 신경망은 XLSR-Wav2Vec2 모델에 회귀 헤드를 추가하여 구현하였다. 이 모델을 특정 언어의 데이터셋을 이용하여 미세조정하면 해당 언어의 발화의 음성적 특징을 추출하는 신경망이 된다. 본 연구의 시스템에서는 XLSR-Wav2Vec2.0을 한국어 음성 데이터셋인 Zeroth-Korea를 사용하여 미세조정한 kresnik/wav2vec2-large-xlsr-korean 모델을 백본으로 사용하였다*. 한국어 ASR * <https://huggingface.co/kresnik/wav2vec2-large-xlsr-korean>

R (Automatic Speech Recognition) 태스크에서 해당 모델의 공개된 WER 기댓값은 4.6% 이다.

wav2vec2-large-xlsr-korean에서 나온 발화의 음성적 특징을 이용하여 감정 레벨 분석을 하는 헤드의 레이어 구조는 그림2와 같다.

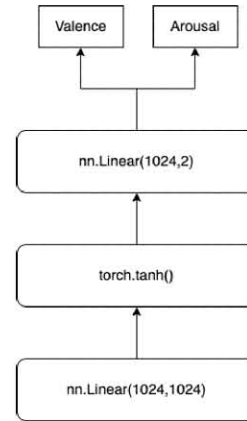


Figure 2: Regression Head Layers

학습에 사용된 데이터셋은 ETRI 한국어 기반 멀티모달 감정 데이터셋 중 일반인들의 자유발화를 대상으로 수집된 KEMDy20이다. 해당 데이터 셋에서 음성 파일과 이에 대응되는 arousal-valence 값만을 학습에 사용하였다. 해당 데이터셋의 여러 모달리티의 데이터 중 발화 음성만을 신경망의 입력으로 이용한 것은 1) 메타버스 대화 환경에서 음성만을 수집하여 실시간으로 적용될 수 있는 모델을 구현하기 위해서이고, 2) XLSR-Wav2Vec2가 한국어 음성으로부터 추출하는 음성적 특징이 감정 분석에서 얼마나 유용하게 쓰일 수 있는지를 파악하기 위해서이다. KEMDy20 데이터 셋의 13462개의 발화 샘플을 60%, 20%, 20%로 나누어 train set, evaluation set, test set으로 사용하였다. Batch size 4로 총 20000 steps 동안 학습 되었으며, loss function으로는 MSE (Mean Squared Error) loss를 사용하였다.

3.2 감정 기반 음악 재생

발화로부터 분석된 감정에 맞는 음악을 재생하는 것은 DEAM 데이터셋의 음원과 음원 전체의 arousal-valence 값을 이용하였다. 추출된 발화의 arousal-valence 레벨과 DEAM 데이터셋의 모든 음원의 arousal-valence 레벨의 L2 distance를 구한 후 그 값이 제일 작은 음원을 선택하는 Python 코드를 구현하였다.

3.3 가상 환경 구축

본 시스템의 몰입형 가상공간 환경은 Unity 3D로 구축되었다. Unity 3D의 XR Interaction Toolkit 패키지를 사용하여 가상공간을 생성하였고 사용자들이 해당 공간에서 실시간으로 만나 대화를 할 수 있도록 멀티 플레이어 네트워킹 서

비스 Photon을 사용하였다. PUN2 (Photon Unity Network 2) 패키지를 사용하여 사용자 객체들을 동기화 하였고 Photon Voice2 패키지를 사용하여 HMD (Head Mounted Display)를 착용한 두 사용자 간에 원거리 음성 통신이 가능하도록 구현하였다. 한 사용자가 다른 사용자를 인식할 수 있도록 머리와 손 모델을 이용하여 아바타를 구현하여 대화 상대의 머리와 손의 움직임을 인식 가능하게 하였다. 가상 환경에서 실시간 대화 중 사용자의 발화를 30초 마다 10초간 녹음하여 wav 파일로 저장하였다. 이 음성 파일은 발화의 감정을 분석하여 적절한 배경 음악을 선택하는 시스템의 입력으로 쓰인다.



Figure 3: A captured image of the implemented virtual environment, users can communicate with each other using their avatars

4. 시스템 평가

4.1 감정 추출 모델 성능 평가

모델의 성능 평가를 위해 추출된 arousal-valence 레벨 각각의 예측 정확도 (CCC; Concordance Correlation Coefficient)를 계산하였다. 본 논문에서는 표 1과 같이 모델 1, 2를 구현하였다. 모델 1은 위에서 설명한 XLSR-Wav2Vec2를 한글 음성 데이터셋으로 미세조정된 모델을 백본으로 사용하였다. 모델 2는 모델 1과 회귀 헤드의 구조와 하이퍼파라미터 세팅이 동일하지만, XLSR-Wav2Vec2를 영어 발화 데이터셋인 Common Voice 6.1로 미세조정된 모델을 백본으로 사용하였다**. 그리고 모델 3은 [6]의 감정 분류 신경망으로, 본 연구와 동일하게 KEMDy20 데이터셋을 이용하여 학습되었지만 음성 및 텍스트를 모두 이용하는 멀티모달 감정인식 모델이다. 모델 3은 음성에서 특징을 추출하기 위해 Wav2V

** <https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english>

ec2를 사용하였다.

#	Model	CCC	
		Arousal	Valence
1	wav2vec2-large-xlsr-korean + regression head	0.815	0.636
2	wav2vec2-large-xlsr-53-english + regression head	0.812	0.604
3	김준우 et al. (Cat)	0.582	0.670

Table 1. CCC of models

모델 1, 2의 경우 백본 모델을 서로 다른 언어의 데이터셋으로 미세조정 하였음에도 불구하고 arousal-valence 예측 정확도가 유사했다. 이것은 XLSR-Wav2Vec2가 다양한 나라의 언어로 사전학습 되었기 때문에 미세조정에 사용되지 않은 언어의 발화에서도 감정 분석에 필요한 음성적인 특징을 잘 찾아내기 때문인 것으로 보인다.

음성과 텍스트를 모두 입력으로 이용하는 모델 3과 비교하였을 때 모델 1, 2는 발화 음성만을 입력으로 이용함에도 불구하고 arousal에 대한 예측 정확도가 0.2 이상 더 높았다. 이것은 미세조정된 XLSR-Wav2Vec2 신경망이 발화로부터 추출한 음성적 특징이 arousal을 추측하는데 효과적으로 쓰일 수 있다는 것을 나타낸다.

4.2 시스템 평가 실험 방법

VR 경험이 10회 이상 있는 10명의 참가자(여성 6명, 남성 4명, 평균 나이 29세)를 대상으로 평가를 실시하였다. 평가를 위한 실험 디자인 방식은 참가자 내 설계 (within-subject design)를 사용하였다. 실험에서 조작하는 독립 변수는 배경음악 재생 시스템이다. 첫 번째 배경음악 재생 시스템은 사전에 설정된 중립 (neutral) 감정에 해당하는 음악이 고정적으로 재생되는 기존 시스템이다. 두 번째 배경음악 재생 시스템은 본 연구에서 개발한 시스템으로 발화의 감정에 맞는 배경음악이 실시간으로 재생되는 시스템이다. 각 재생 시스템의 배경음악이 사용자가 메타버스 환경에서 대화할 때 어떤 효과를 주는지 실험하였다. 참가자 중 무작위 다섯 명에게는 기존 시스템을 경험하게 한 후, 본 시스템에서 재생한 음악을 경험하게 하는 순서로 실험을 진행하였다. 나머지 다섯 명에게는 역순으로 진행하였으며 이는 시스템 경험 순서가 실험 결과에 끼치는 편향을 막기 위함이다. 실험 참가자들은 Oculus Quest 2 HMD를 착용하여 가상 환경 안에서 실험 진행자와 일대일로 대화를 3분 동안 나누도록 지시 받았다. 미리 설정된 음악이 재생되는 기존 시스템 실험 환경을 A로 칭하고 실시간 배경음악 매칭 시스템이 적용된 본 시스템 실험 환경을 B로 칭하였다. 각 환경의 실험이 끝날 때마다 1분 동안 같은 문항의 설문지를 총 2회 작성하도록

하였다. 설문지는 총 10문항으로 리커트 척도 (Likert Scale) 로 표기되었다 (5 점 = 매우 동의함, 1점=매우 동의하지 않음). 설문지 문항은 아래와 같다.

기준	질문
Q1	대화를 듣는 동안 이야기에 쉽게 몰입할 수 있었습니까?
Q2	대화에 배경음악은 적절하게 어울렸습니까?
Q3	대화를 진행하는 동안 배경음악은 대화의 목적을 이해하는 데 도움이 되었습니까?
Q4	배경음악이 대화의 분위기를 긍정적으로 유지하는데 도움이 되었습니까?
Q5	대화를 하는데 적극적으로 참여했습니까?
Q6	상대방의 발언에 대해 나의 대답이나 반응이 적절하게 이루어졌습니까?
Q7	대화의 주제에 대해 흥미를 느꼈습니까?
Q8	해당 시스템을 사용하면서 편리함을 느꼈습니까?
Q9	해당 시스템을 추후에 메타버스 대화 환경에서 사용할 의향이 있습니까?
Q10	해당 시스템에 대해 전반적으로 만족했습니까?

Table 2. Question list

4.3 실험 결과

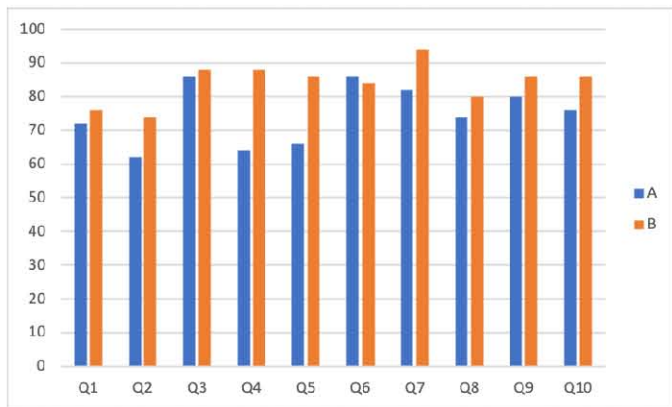


Figure 4: Experiment result, overall satisfaction of system B is higher.

본 연구에서는 기존 시스템 A와 본 시스템 B의 사용자 경험에 차이가 있는지 알아보기 위해 SPSS 프로그램을 활용하여 독립표본 t 검정을 하였다. 응답의 표준화를 위해 설문에서 사용된 5점 척도를 100도로 환산하였다. 시스템 A의 만족도 평균은 74.8이고 표준편차는 8.8이며, 시스템 B의 만족도 평균은 84.2이고 표준편차는 5.6이었다. t 검정 실시 전 정규성 검정과 등분산성 검정을 통해 데이터가 정규성과 등분산성을 만족한다는 것을 확인하였다. 독립 표본 t 검정 결과 p 값이 0.0079로 나왔고 이것이 0.05 미만이므로, 시스템 A와 시스템 B의 만족도의 차이가 통계적으로 유의미하다. 이를 통해 본 시스템 B가 시스템 A에 비해 사용자들에게 보다 전반적으로 만족스러운 사용 경험을 제공한다는 것을 알 수 있었다.

5. 결론 및 향후 과제

5.1 결론

본 연구는 기존 메타버스 환경에서 고정된 배경음악 재생

의 문제점을 보완하기 위해 사용자 발화의 감정에 어울리는 음악을 선택하여 재생하는 시스템을 제안하였다. 본 연구를 통해 해당 시스템은 기존 배경음악 시스템에서의 사용자 간 대화에 비해 몰입도, 참여도 및 만족도 측면에서 긍정적인 영향을 미친 것을 확인할 수 있었다. 이러한 결과는 감정 기반의 배경음악 매칭 시스템이 가상 환경에서 의사소통을 개선하고 사용자 경험을 향상시키는데 도움이 될 것임을 보여준다.

5.2 향후 과제

본 시스템의 배경음악으로 쓰인 DEAM 데이터셋은 배경음악에 특화된 데이터셋이 아니기 때문에 노랫말이 들리는 등의 문제로 대화의 방해 요소가 될 수 있는 한계가 있다. 이는 차후 연구에서 배경음악으로 더 적절한 데이터셋을 이용하는 것으로 해결할 수 있을 것이다. 또한 본 연구는 KE MDy20의 다양한 모달리티의 데이터 중 음성과 arousal-valence 값만 사용하였다. 음성으로부터 추출한 음성적 특징은 arousal을 추측하는데에는 효과적으로 쓰일 수 있었지만 valence 추측에서는 상대적으로 덜 효과적인 모습을 보였다. 차후 연구에서 음성 뿐만 아니라 체온 등의 다른 모달리티의 데이터를 함께 이용하여 감정을 분석한다면, 더 높은 정확도의 시스템을 개발할 수 있을 것이다.

감사의 글

이 연구는 2023년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구이자 (P001274 6, 2023년 산업혁신인재성장지원사업), 과학기술정보통신부 및 정보통신기획평가원의 메타버스 융합대학원의 연구 결과로 수행되었음 (IITP-2023-RS-2022-00156318)

References

- [1] Sanders, Timothy, and Paul Cairns. "Time perception, immersion and music in videogames," *Proceedings of HCI 2010* 24, 160-167, 2010.
- [2] Weiss, Karl, Taghi M. Khoshgoftaar, and DingDing Wang. "A survey of transfer learning," *Journal of Big data* 3.1, 1-40, 2016.
- [3] Russell, James A. "A circumplex model of affect," *Journal of personality and social psychology* 39.6, 1161, 1980.
- [4] K. J. Noh and H. Jeong, "KEMDy20," https://nanum.etri.re.kr/share/kjnoh/KEMDy20?lang=ko_KR
- [5] Na-Mo Bang, Heui-Yeen Yeen, Jee-Hyun Lee, Myoung-Wan Koo. "MMM: Multi-modal Emotion Recognition in conversation with MLP-Mixer," *한국정보과학회 학술발표논문집*, 2288-2290, 2022.
- [6] June-Woo Kim, Dong-Hyun Kim, Ju-Seong Do, Ho-Young

Jung. "Strategies of utilizing pre-trained text and speech model-based feature representation for multi-modal emotion recognition," *한국정보과학회 학술발표논문집*, 2282-2284, 2022.

[7] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems* 33, 12449-12460, 2020.

[8] Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[9] M. Soleymani, A. Aljanaki, Y. Yang, "DEAM: Mediaeval database for emotional analysis in music," <http://cvml.unige.ch/databases/DEAM/>, 2016.

< 저자 소개 >

김기락

- 2021년 연세대학교 학사
- 2022년~현재 서강대학교 석사과정
- 관심분야 : Cross-modal generation, Metaverse
- <https://orcid.org/0000-0002-2960-4583>



이상아

- 2017년 University of Minnesota-Twin Cities 심리학 학사
- 2022년 Universitat Osnabruck 인지과학 석사
- 2022년~현재 서강대학교 아트앤테크놀로지 박사과정
- 관심분야 : HCI, XR, Multimodal interaction
- <https://orcid.org/0009-0006-3058-9567>



김나현

- 2022년~현재 서강대학교 석사과정
- 관심분야 : VR/AR, HCI, AI
- <https://orcid.org/0000-0003-4566-1657>



정문열

- 1980년 서울대학교 계산통계학과 학사
- 1982년 KAIST 전산학과 석사
- 1992년 펜실베이니아 대학 전산학과 박사
- 2018년~현재 서강대학교 아트 & 테크놀로지학과/메타버스전문대학원 교수
- 관심분야 : Machine learning, VR/AR, Interactive media arts
- <https://orcid.org/0000-0003-2462-7820>

