학사학위논문

실시간 표정 인식을 이용한 감정 기반 TTS 보컬 합성 시스템 연구

제 출 자 : 이 종 하

지도교수 : 구 자 환

2025 년 11 월

뉴뮤직학부

뮤직테크놀러지&컴퓨터음악작곡 전공

단국대학교 예술대학

실시간 표정 인식을 이용한

감정 기반 TTS

보컬 합성 시스템 연구

이 논문을 학사학위논문으로 제출함.

2025 년 11 월

단국대학교 예술대학

뉴뮤직학부

뮤직테크놀러지&컴퓨터음악작곡 전공

이 종 하

이종하의 학사학위 논문을

합격으로 판정함

심 사 일 : ○○○○. ○○. ○○.

심사위원장 _____ 인

심 사 위 원 _____ 인

심 사 위 원 _____ 인

심 사 위 원 _____ 인

심 사 위 원 _____ 인

단국대학교 예술대학

(국문 초록)

실시간 표정 인식을 이용한 감정 기반 TTS 보컬 합성 시스템 연구

단국대학교 뉴뮤직학부

뮤직테크놀러지전공

이종하

지도교수: 구자환

본 연구는 컴퓨터 비전 기반 표정 인식(Facial Expression Recognition, FER)을 통해 사용자의 감정 상태를 실시간으로 추정하고, 이를 Flask(Python Web Framework) 기반 감정 매핑 서버와 결합하여 OpenAI Text-to-Speech(TTS, “GPT-4o-mini-tts”)보컬 합성에 반영하는 새로운 방법을 제안한다. MediaPipe 기반 얼굴 검출과 Convolutional Neural Network(CNN) 모델을 이용한 표정 인식 모듈을 구현하고, 지수이동평균(Exponential Moving Average , EMA) 및 다수결 투표 기반 안정화 기법을 적용하여 신뢰도 높은 실시간 감정 추정을 수행하였다. 추정된 감정은 JavaScript Object Notation(JSON) 형식으로 Flask 서버에 전달되며, 서버는 이를 OpenAI TTS Application Programming Interface(API) 호출 시 프롬프트 입력으로 직접 반영하여 감정에 따른 보컬 합성을 실시간으로 수행한다. 또한, 웹 기반 User Interface(UI)를 구축하여 사용자가 텍스트를 입력할 때 실시간 표정 감정에 따라 문장 종결부마다 감정 태그가 자동 삽입되도록 설계되었다. 본 시스템은 창작자와 일반 사용자 모두에게 표정만으로 감정 특성을 제어할 수 있는 직관적인 음악 및 사운드 제작 환경을 제공하며, 음악 콘텐츠의 감정 전달력을 높이는 기반 기술로 기능할 수 있음을 시사한다.

목 차

I. 서론

1. 연구 배경

음악은 단순한 청각적 자극을 넘어 인간의 정서적 경험을 형성하고 조절하는 핵심 매개체로 기능한다(남옥선, 2007). 특히 보컬은 음악에서 감정 정보를 가장 직접적이고 풍부하게 전달하는 수단으로, 발화자의 억양, 호흡, 미묘한 뉘앙스를 통해 청자의 정서적 반응을 증폭시킨다(이은혜 & 김유정, 2020). 이러한 이유로 보컬의 감정 표현은 음악적 몰입감과 감정 전달력의 핵심 요소로 간주되어 왔다(송미연, 2019).

기존 보컬 합성 연구들은 피치(pitch)[1], 타이밍(timing)[2], 아티큘레이션(articulation)[3]과 같은 발성의 물리적 특성을 정밀하게 모델링하는 데 집중하여 음질과 자연스러움 측면에서는 괄목할 만한 성과를 거두었다(이교구 & Cremer, 2009). 그러나 인간 가창이 지닌 감정 표현의 섬세함과 즉흥성을 충분히 재현하지는 못하였다(김선영, 2025). 특히 기계적으로 합성된 보컬은 기술적으로는 정확할 수 있으나, 감정의 미묘한 차이를 구현하는 데 있어 청자에게 인위적인 인상을 남기는 한계가 존재한다(유은정, 2020).

최근 딥러닝 기반 감정 기반 TTS 기술이 등장하며 감정성을 가미한 합성 보컬 연구가 활발히 진행되고 있다(유은정, 2020.). 이러한 모델들은 특히 감정 태그(행복, 슬픔, 분노 등)를 입력으로 주어 해당 감정 특성이 반영된 합성 음성을 생성할 수 있다(Ten et al., 2023). 그러나 대부분의 연구가 정적으로 정의된 감정 입력에 의존하고 있어, 사용자의 실시간 감정 변화나 상호작용적인 음악 제작 환경을

[1] Pitch: 음의 높낮이를 나타내는 소리의 기본 특성.

[2] Timing: 음의 길이와 발성 시점을 조절하는 요소.

[3] Articulation: 음악 연주나 가창에서 음을 연결하거나 구분하는 표현 기법.

지원하는 데에는 한계가 있다(박중희 & 한광희, 2023). 따라서 실제 사용자의 감정 상태를 실시간으로 반영할 수 있는 새로운 방식의 보컬 합성 시스템에 대한 필요성이 증대되고 있다(안단태, 2024).

2. 연구 필요성

최근 TTS 분야에서는 스타일 변환(style transfer), 운율 제어(prosody control)[4]와 같은 기술을 통해 감정 표현을 조절하는 연구가 활발히 이루어지고 있다(Lee & Kim, 2019). 그러나 대부분의 감정 기반 TTS 모델은 여전히 사전에 정의된 감정 클래스나 스타일 벡터[5]를 수동으로 입력 받는 방식에 의존한다. 이러한 접근은 사용자의 감정 상태가 시간에 따라 역동적으로 변화하는 실제 상호작용 환경에서는 한계를 보인다(Lee & Kim, 2019).

특히 음악 창작이나 실시간 퍼포먼스와 같은 응용 맥락에서는 창작자의 순간적인 감정 변화를 직관적으로 반영할 수 있는 합성 보컬 시스템이 필요하다(Sivaprasad et al., 2021). 기존 방식처럼 정적인 태그를 미리 지정하는 구조로는 이러한 즉흥성을 지원하기 어렵다. 따라서 사용자의 감정을 실시간으로 자동 인식하고 이를 합성 보컬에 즉각 반영할 수 있는 기술이 요구된다(AITimes, 2024). 본 연구는 이러한 필요성에 대응하기 위해 표정 기반 감정 인식과 TTS 보컬 합성을 결합한 새로운 시스템을 제안하고자 한다.

[4] Prosody control: 발화의 리듬, 강세, 억양 등 운율적 특성을 조절하는 기술.

[5] 스타일 벡터: 딥러닝 모델에서 화자 특성이나 감정 스타일을 수치화한 벡터 표현.

3. 연구 목적

본 연구의 목적은 사용자의 표정을 기반으로 실시간 감정을 추론하고, 이를 TTS 시스템에 직접 반영하여 감정 표현이 가능한 보컬 합성 구조를 제안하는 것이다. 이를 위해 표정 인식 모듈을 통해 추출된 감정 상태를 TTS 감정 태그로 자동 변환하는 감정 매핑 방식을 설계하고, 합성된 보컬 음성에 감정 특성을 반영한다. 이러한 접근은 창작자가 별도의 복잡한 제어 과정을 거치지 않고도 표정만으로 감정을 입력할 수 있는 직관적인 인터페이스를 제공하며, 결과적으로 음악 및 사운드 제작 환경에서 감정 전달력을 강화할 수 있는 기술적 기반을 마련하는 데 그 목적이 있다.

4. 연구 기여

본 연구의 주요 기여는 다음과 같다. (1) MediaPipe[6] 기반 얼굴 검출과 CNN 분류기를 결합하고 지수이동평균[7] 및 다수결 투표 기법[8]을 적용하여 안정적인 감정 추정을 실시간으로 수행하는 표정 기반 감정을 추론 모듈을 구현하였으며, (2) 추론된 감정을 JSON 형식으로 Flask 서버에 전달한 뒤, OpenAI TTS API 호출 해석 가능한 태그로 자동 변환하는 감정-TTS 매핑 구조를 설계하였고, (3) Flask 서버 및 웹 기반 UI 를 통해 사용자의 표정 감정을 문장 단위 합성 과정에 반영함으로써 음악 창작 및 사운드 제작 환경에서 직관적이고 즉흥적인 감정 제어 가능성을 제시하였다.

---

[6] MediaPipe: 구글이 개발한 오픈소스 라이브러리. 얼굴, 손 추적 등 다양한 머신러닝 기능을 제공.

[7] 지수이동평균: 최근 값에 더 큰 가중치를 주어 변화에 민감하게 반응하는 계산법.

[8] 다수결 투표 기법: 여러 예측 중 가장 많이 나온 결과를 최종 출력으로 선택하는 방법

Ⅱ. 본론

1. 음악에서 감정 표현의 역할과 뮤직테크놀러지적 배경

음악은 인간의 감정에 직접적으로 작용하는 예술 매체로서, 청자의 정서적 반응을 유도하는 기능을 갖는다(Juslin & Västfjäll, 2008). Juslin 과 Västfjäll(2008)은 감정 반응을 유발하는 음악적 메커니즘을 생리적 반응, 평가적 조건화, 정체성 연관 등 여섯 가지로 제시하며, 음악이 단순한 청각적 자극을 넘어 복합적인 감정 유도 도구로 기능함을 밝혔다. 이처럼 음악은 듣는 이의 심리적 상태에 변화를 일으킬 수 있는 정서적 자극으로 간주된다.

특히 보컬 음악은 가사, 억양, 음색 등을 통해 감정 표현의 가능성이 더욱 풍부해진다(김혜란 & 송은성, 2021.). 보컬 퍼포먼스에서 감정 표현은 주로 속도, 강세, 피치 범위, 음질 등 음성적 요소를 통해 이루어지고, 이러한 음성적 단서는 청자가 가창자의 감정 상태를 인식하는 데 필수적인 정보를 제공한다. 이처럼 감정 표현은 음악 감상의 몰입도와 만족도를 높이는 핵심 요소로 작용한다(김혜련 & 송은성, 2021.).

그러나 보컬 합성(Singing Voice Synthesis, SVS)은 기술의 발전에도 불구하고, 감정 표현의 섬세함에서는 여전히 한계가 존재한다(김선영, 2025.). Luo 등(2020)은 기존 SVS 시스템이 음정과 타이밍과 같은 정량적 요소는 효과적으로 구현하지만, 슬픔의 느린 발성이나 기쁨의 밝은 음색과 같은 정서적 감정 특성은 충분히 반영하지 못한다고 지적하였다(Luo et al, 2020). 이는 기존 기술이 발성의 물리적 측면에는 집중해 온 반면, 감정 표현이라는 정서적 차원을 자동화하는 데에는 상대적으로 미흡했음을 보여준다.

이러한 한계를 극복하기 위해 최근 뮤직테크놀러지 분야에서는 감정 표현을 위한 감정 기반 TTS 또는 감정 전이(emotion transfer) 기술의 통합이 시도되고 있다.(유은정, 2020) ICASSP 2023 에서 FastSpeech2[9] 기반의 감정 음성 합성 시스템에서 감정 레이블과 pitch/duration 인코더(encoder)를 함께 조건화 함으로써 보다 자연스러운 감정 발화를 구현하였고, 이는 보컬 합성에도 직접적으로 응용될 수 있는 가능성을 보였다(AITimes, 2023). 감정 정보의 세밀한 조작이 가능해지면서, 창작자는 음향 요소만으로 감정 표현을 보다 정확히 통제할 수 있는 기반을 갖추게 된다.

한편, 뮤직테크놀러지에서는 감정 표현을 위해 TTS, Musical Instrument Digital Interface(MIDI) 제어, 음색 조절, 얼굴 모션 캡처 등 타 분야 기술의 융합이 점점 일반화되고 있다(권새봄 & 김주섭, 2019). 이는 '감정을 어떻게 디지털 환경에서 정량화하고 전송할 것인가' 라는 문제로 귀결되며, 특히 실시간 인터페이스와 인공지능 기반 감정 추론 시스템이 중심 기술로 부상하고 있다. 음악적 감정 표현은 단지 '소리의 조작' 이 아니라, '청자의 감정 반응을 유도하는 목적 지향적 기술' 로 이해될 수 있으며, 이는 인간-기계 상호작용(Human-Computer Interaction, HCI)의 맥락에서도 중요하게 다루어진다.(김영희, 2015)

[9] FastSpeech2: Microsoft 가 제안한 딥러닝 기반 TTS 모델. 빠른 합성과 안정적인 운율 제어를 목표로 개선된 버전.

2. 감정 보컬 합성과 TTS 기술 개요

음성 합성은 텍스트 입력을 기반으로 자연스러운 발화를 생성하는 기술로, 최근에는 딥러닝 기반 모델이 주류를 이루고 있다.

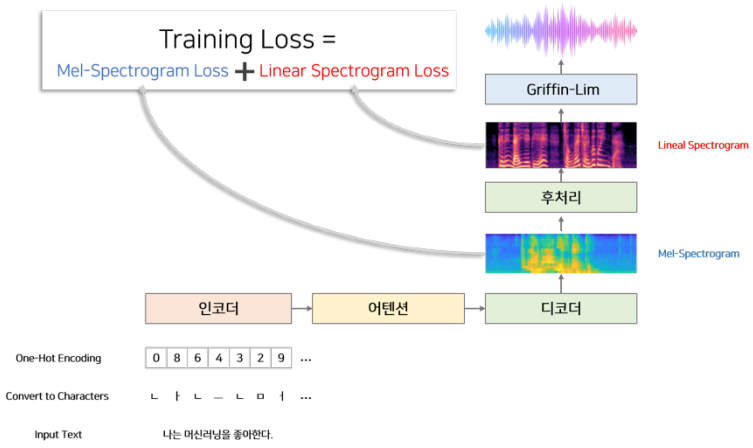

그림 1. Tractron 기반 TTS 모델의 입력과 출력 구조

Tacotron[10] 계열 모델은 텍스트를 mel spectrogram[11]으로 변환한 뒤 보코더(vocoder)를 통해 음성을 복원하는 구조를 채택하여, 자연스러운 운율과 음질을 확보하는 데 성과를 두었다.

감정 기반 TTS는 기존 TTS에 감정 표현 기능을 추가하여 억양, 속도, 강세, 음색 등 감정에 따른 음성 특성을 반영하는 확장된 형태이다. 초기 연구는 감정별 데이터 분류나 수동 레이블 입력 방식에 의존했으나, 이러한 접근은 감정의 다양성과 연속성을 충분히 반영하지 못했다는 한계가 있었다. 이를 개선하기 위해 Global Style Tokens(GST) 방식이 제안되었으며, 별도의 레이블 없이도 화자의

[10] Tacotron: 구글에서 개발한 딥러닝 기반 TTS 모델 계열.

[11] Mel Spectrogram: 음성을 시간-주파수 영역으로 변환한 시각적 표현.

말투나 감정 스타일을 embedding[12] 공간에서 학습하고 제어할 수 있는 가능성을 보여주었다.

이러한 흐름을 바탕으로, 최근 OpenAI 에서 제안한 GPT-4o-mini-tts 는 실시간 음성 합성에 최적화된 멀티모달 신경망 기반 TTS 모델로, 입력 텍스트를 자연스러운 음성으로 변환하는 과정을 엔드투엔드(end-to-end) 방식으로 수행한다. 모델의 기본 구조는 텍스트 -> 음향 특징 -> 신경망 보코더 -> 최종 음성 생성의 파이프라인을 따르며, 딥러닝 기반의 대규모 음성 데이터셋을 통해 학습되어 고품질 발화를 생성할 수 있다.

이 모델은 기존 Tacotron 계열 모델과 유사하게 mel-spectrogram 예측기를 사용하지만, 대규모 언어 모델(Large Language Model, LLM)과 결합된 통합 아키텍쳐를 채택하여 합성 속도와 표현력을 동시에 개선하였다. 특히 OpenAI TTS 는 프롬프트 기반 감정 제어 기능을 제공하여, 사전에 정의된 고정 태그에 의존하지 않고 사용자가 원하는 감정적 스타일을 자연어 지시문으로 제어할 수 있다는 점에서 차별성을 가진다. 예를 들어, 특정 감정을 “행복한 톤”, “차분한 톤” 등으로 프롬프트에 명시하면, 모델은 피치, 말하기 속도, 발화 에너지 등 음향학적 파라미터를 자동으로 조절하여 다양한 감정 발화를 합성한다.

따라서 본 연구에서 제안하는 시스템은 이러한 기술적 흐름 위에서, 표정 기반 감정 추론 결과를 TTS 감정 입력값으로 자동 변환하는 구조를 통해 보컬 합성 과정에 실시간 감정을 반영하고자 한다.

[12] Embedding: 데이터를 고차원 벡터 공간에 매핑해 의미적 유사도를 반영하는 표현 방식.

3. 표정 기반 감정 인식 모듈 구현

표정은 인간 내면의 감정을 외부로 드러내는 가장 직접적인 신호로서, 감정 인식에서 중요한 비언어적 단서로 활용된다(S. RajaA & R. Harrabi, 2021). FER 은 얼굴 영상으로부터 사용자의 감정을 자동 분류하는 기술로, 초기에는 Support Vector Machine(SVM)[13], Histogram of Oriented Gradients(HOG)[14], Local Binary Pattern(LBP)[15] 기반 전통적 방식이 주를 이루었으나 최근에는 CNN 을 중심으로한 딥러닝 접근이 주류로 자리잡고 있다.

그림 2. FER2013 데이터셋 샘플 이미지 예시

[13] SVM: 지도학습 기반의 분류 알고리즘.

[14] HOG: 이미지의 윤곽과 모양을 방향성 히스토그램으로 표현하여 객체 인식에 활용되는 기법.

[15] LBP: 영상의 국소 영역에서 밝기 패턴을 이진수로 변환해 질감 특징을 추출하는 방법.

감정 클래스	샘플 수	비율(%)
Angry	4,953	14.2%
Disgust	547	1.6%
Fear	5,121	14.7%
Happy	8,989	25.5%
Sad	6,077	17.4%
Surprise	4,002	11.4%
Neutral	6,198	17.2%
총합	35,887	100%

표 1. FER2013 데이터셋 구성

본 연구에서는 FER2013 데이터셋을 기반으로 CNN 분류 모델을 학습하여 face_emotion.h5 를 구축하였다. FER2013 은 48x48 픽셀 흑백 얼굴 이미지와 7 개 감정 클래스(angry, disgust, fear, happy, sad, surprise, neutral)로 구성되며, 총 35,887 장의 이미지가 포함되어 있다. 학습 과정에서는 데이터를 (N, 48, 48, 1) 형태로 정규화하고, 픽셀 값을 0-1 범위로 스케일링(scaling)하였다. 라벨은 원-핫 인코딩(one-hot encoding)[16]을 적용하였으며, 데이터 증강(회전, 이동, 좌우 반전 등)[17]을 통해 과적합을 방지하였다.

[16] One-hot encoding: 각 클래스를 한 요소로만 1 로 하고 나머지는 0 으로 하는 방식.

[17] 데이터 증강: 회전, 이동, 반전 등으로 학습 데이터의 다양성을 인위적으로 늘리는 방법.

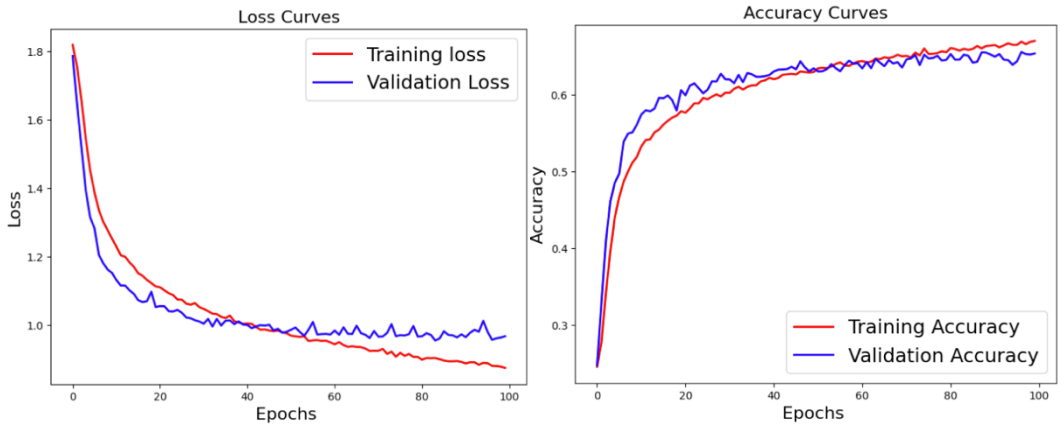

그림 3. FER2013 데이터셋 손실(loss) 곡선(왼쪽) 및 학습 정확도(accuracy) 곡선(오른쪽)

모델 구조는 CNN 기반 다층 합성곱 신경망으로 설계하였다. 입력 영상은 Conv2D-ReLU-MaxPooling-Dropout[18] 블록을 거치며 고차원 특징을 추출하고, Flatten 후 Dense(512, ReLU)를 거쳐 최종적으로 Dense(7, Softmax) 층에서 7 개 감정을 분류한다. 최적화에는 Adam 옵티마이저와 categorical crossentropy[19] 손실함수를 사용하였다. 학습은 batch size=256, epoch=100 으로 진행되었으며, 학습/검증 곡선 상에서 안정적인 수렴을 확인하였다(그림 3).

학습 완료된 모델은 face_emotion.h5 파일로 저장되었으며, 추론 단계에서는 OpenCV 및 MediaPipe 기반 얼굴 검출 모듈과 결합하였다. 웹캠 입력에서 얼굴 영역을 검출한 뒤 48x48 형태로 전처리하여 CNN 모델에 입력하면, 각 감정에 대한 확률 벡터를 산출한다. 본 연구에서는 이 확률 분포를 기반으로 지수이동평균과 다수결 투표 기법을 적용하여 단기적 노이즈를 억제하고 감정 분류의 안정성을 확보하였다. 또한 감정 변화의 순간적인 진동을 줄이기 위해 일정 시간(100ms) 이상 유지된 감정만 최종 출력으로 반영하는 Hold-Time(동일한 감정 상태가 100ms 이상 유지될 때만 최종 감정으로 반영) 전략을 도입하였다.

[18] Conv2D, ReLU, MaxPooling, Dropout: CNN 을 구성하는 주요 연산 블록.

[19] Categorical Crossentropy: 다중 클래스 분류에 쓰이는 대표적인 손실 함수.

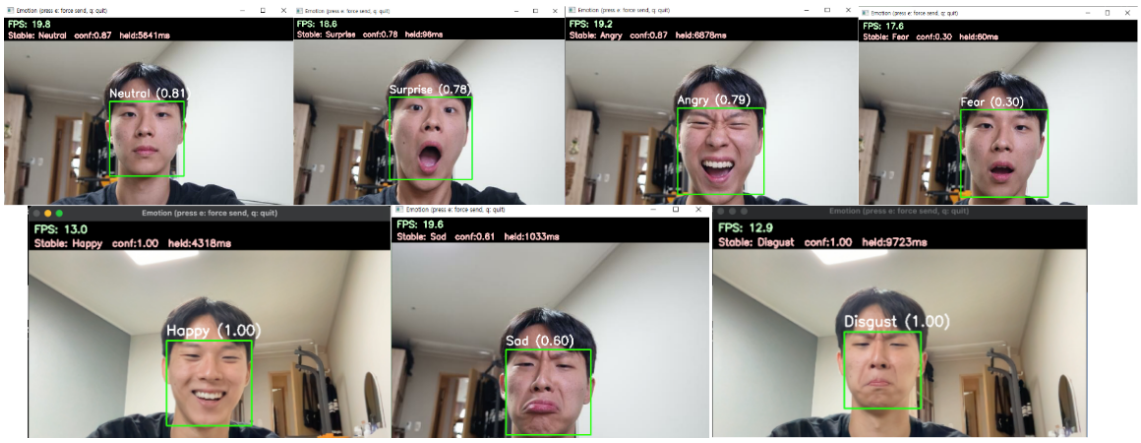

그림 4. FER 모듈을 활용한 실시간 표정 기반 감정 분류 결과 예시

그림 4 는 학습된 모델을 사용하여 실시간 웹캠 환경에서 추론한 감정 분류 결과를 보여준다. 7 가지 감정 클래스(Neutral, Surprise, Angry, Fear, Happy, Sad, Disgust)에 대해 모델이 산출한 확률과 함께 예측 결과를 시각화 하였다.

최종적으로 FER 모듈은 실시간으로 안정적인 감정 분류 결과를 산출하며, 이는 후속 단계에서 TTS 감정 조건부 입력 값으로 자동 변환된다. 즉, 사용자의 표정은 CNN 기반 감정 인식 과정을 거쳐 감정 태그로 추출되고, 해당 태그는 곧바로 보컬 합성 과정에 반영된다. 이러한 모듈은 본 연구 시스템에서 실시간 감정-TTS 연동을 가능하게 하는 핵심 구성 요소이다.

4. 실시간 감정 기반 TTS 통합 시스템 구현

본 연구에서는 표정 기반 감정 인식 모듈과 감정 매핑 구조를 OpenAI TTS 에 연동하여, 사용자의 얼굴 표정을 실시간으로 반영하는 감정 보컬 시스템을 구현하였다. 전체 파이프라인은 (1) 얼굴 표정 기반 감정 추출, (2) 감정 벡터-TTS

매핑 구조, (3) 실시간 처리 및 상호작용, (4) 뮤직테크놀러지 적용 가능성의 네 단계로 구성된다.

4.1 얼굴 표정 기반 감정 추출

얼굴 표정 기반 감정 추출은 CNN 기반 FER 모델(face_emotion.h5)을 활용하여 실시간으로 사용자의 표정을 분석하고 감정 라벨 및 신뢰도(confidence)를 산출하였다. 모델 학습 과정은 제 3 장에서 기술한 바와 같이 FER2013 데이터셋을 이용하여 수행되었으며, 최종적으로 7 개의 감정 클래스(angry, disgust, fear, happy, sad, surprise, neutral)를 분류한다.

실시간 감정 추출 파이프라인은 OpenCV 와 MediaPipe 를 기반으로 구현되었다. 웹캠으로부터 입력되는 영상 프레임에서 얼굴 영역을 탐지한 뒤, 이를 48x48 픽셀의 흑백 이미지로 전처리하여 학습된 CNN 모델의 입력으로 제공한다. 모델의 출력은 softmax 확률 벡터 형태로 계산되며, 각 감정 클래스별 신뢰도를 포함한다.

예를 들어, 사용자가 중립적인 표정에서 웃는 표정으로 변화할 때 모델이 산출하는 실시간 로그는 다음과 같다.

```
[send] {'t': 1755939388738, 'label': 'Disgust', 'confidence': 0.9999}
[send] {'t': 1755939415117, 'label': 'Happy',  'confidence': 0.9920}
```

위 로그에서 알 수 있듯이, 시스템은 약 2.6 초 이내에 표정 변화를 인식하고 “Disgust” 에서 “Happy” 로 감정 상태를 갱신한다. 이와 같이 모델은 매 프레임 단위로 감정 확률을 산출하며, 본 연구에서는 감정 추출의 안정성을 확보하기 위해 세 가지 안정화 기법을 적용하였다. 지수이동평균, 다수결 투표 기법, Hold-Time 전략을 적용하였다. 최종적으로 FER 모듈은 실시간으로 안정적인 감정 라벨과 신뢰도 값을 산출하며, 이 결과는 Flask 서버를 통해 JSON 포맷(“emotion” :” happy” ,” confidence” :0.9920)으로 전달된다.

4.2 감정 벡터-TTS 매핑 구조

본 연구에서는 얼굴 표정 기반 감정 인식 모듈에서 추출한 감정 라벨을 실시간으로 음성 합성에 반영하기 위해 OpenAI GPT-4o-mini-tts 모델을 활용하였다. GPT-4o-mini-tts 는 텍스트 입력을 자연스러운 음성으로 변환하는 멀티모달 TTS 모델로, 기본적으로 감정 태그를 직접 지원하지 않기 때문에 감정별 발화 스타일, 속도, 음색을 제어할 수 있는 감정 프리셋을 설계하였다. 감정 프리셋은 각 감정에 따라 발화 스타일을 정의한 instruction, 음성 톤을 설정하는 voice, 발화 속도를 조절하는 speed 로 구성되며, FER 모듈의 실시간 감정 라벨을 기반으로 동적으로 선택된다.

4.2.1 감정 프리셋 매핑 설계

각 감정별 프리셋은 OpenAI TTS 합성 시 전달되는 핵심 파라미터로 표 2 와 같이 설정하였다. 예를 들어, “행복” 감정이 선택되면, 경쾌한 리듬과 높은 에너지를 지시하는 instruction 과 speed= 1.22 가 적용되어 활기찬 톤이 합성된다.

감정	음성(Voice)	발화 속도(Speed)	발화 스타일
행복	shimmer	1.22	매우 행복한 감정이 또렷하게 들리도록, 밝은 피치와 높은 에너지로, 리듬을 가볍고 경쾌하게, 호흡은 짧고 가볍게, 억양을 위쪽으로
슬픔	shimmer	0.85	깊은 슬픔이 분명하게 느껴지도록, 낮은 피치와 낮은 에너지로, 속도는 느리게, 호흡은 길고 무겁되 소리는 추가하지는 않고, 억양 폭을 줄여 평탄하게
분노	shimmer	1.15	분노가 명확하게 들리도록, 낮은 피치와 강한 어택으로 선명하게, 호흡은 짧고 단단하게, 문장 내 쉼을 줄이고 밀도 있게

놀람	shimmer	1.25	놀람이 즉시 느껴지도록, 시작을 높은 피치와 강한 어택으로, 이후 살짝 안정시키며, 호흡은 짧고 반응적이되 소리 추가는 안되게
두려움	shimmer	0.92	두려움이 분명히 느껴지도록, 낮은 피치에 약간 위축된 톤으로 작게, 속도는 조금 느리게, 단어 사이 간격을 아주 약간 늘려 긴장을 만듦
중립	shimmer	1.0	감정 개입 없이 중립적으로, 중간 피치와 표준 속도로 담담하게, 억양 과장 없이

표 2. OpenAI TTS 감정 프리셋 설정

4.2.2 OpenAI TTS API 기반 감정 매핑

FER 모듈이 결정한 감정에 대해 프리셋을 불러와 텍스트를 감정에 맞게 리라이팅한 뒤(단어는 유지, 부호/쉼/리듬만 조정), 해당 파라미터로 TTS 합성을 수행한다. 아래는 실제 호출 구조의 요약이다.

```
def tts_one(text: str, emotion: str):
    instruction, voice, speed = preset_for_emotion(emotion)  # 프리셋 선택
    perf_text = rewrite_with_instruction(text, instruction)  # 단어 불변, 리듬/부호만 조정
    resp = client.audio.speech.create(
        model="gpt-4o-mini-tts",
        voice=voice,
        speed=speed,
        input=perf_text,
        response_format=OUTPUT_FORMAT  # mp3(기본) | wav
    )
    # 스트리밍 응답을 버퍼로 수집해 바이트로 반환
    buf = BytesIO()
    try:
        for chunk in resp.iter_bytes(chunk_size=1024*1024):
            buf.write(chunk)
    except AttributeError:
        tmp = f"tts_tmp.{OUT_EXT}"
        resp.stream_to_file(tmp)
        with open(tmp, "rb") as f:
            buf.write(f.read())
        try: os.remove(tmp)
        except OSError: pass
    return buf.getvalue()
```

문장에 감정 태그가 다수 포함된 경우 문장별로 서로 다른 프리셋을 적용해 개별 합성 후, 0.25 초 무음을 삽입하여 순차 병합한다(pydub[20]/ffmpeg[21]). 태그가 없을 때는 전역 감정(기본: 중립)으로 단일 합성을 수행한다. 결과 오디오는 WAV[22]로 스트리밍 전송된다.

4.3 실시간 처리 및 상호작용

본 연구에서는 구현한 실시간 감정-TTS 연동 시스템은 Flask 기반 서버와 웹 기반 UI 를 중심으로 동작하며, 사용자의 표정 변화와 입력 텍스트를 연계하여 동적으로 음성을 합성한다. 웹캠을 통해 입력되는 영상 스트림은 FER 모듈을 거쳐 실시간으로 라벨을 산출하고, 산출된 결과는 OpenAI GPT-4o-mini-tts 모델의 입력 파라미터에 즉시 반영된다. 이를 통해 동일한 스크립트라도 사용자 표정의 변화에 따라 보컬의 억양, 발화 속도, 음색이 실시간으로 조절된다.

시스템은 문장 단위 감정 태깅 전략을 적용하였다. 입력된 텍스트는 마침표(.), 느낌표(!), 물음표(?)의 문장부호를 기준으로 분리되며, 각 문장 작성 시간 동안 사용자가 지은 표정을 추적한다. 이후 문장별로 가장 오랫동안 유지된 감정 상태를 해당 문장의 대표 감정으로 결정하여 태그를 삽입한다. 예를 들어, 사용자가 특정 문장을 말하는 동안 대부분의 시간 동안 “행복” 표정을 유지했다면 해당 문장 끝에 <행복> 태그가 자동 부착된다. 이러한 감정 태그는 TTS 요청 시 내부적으로 감정 프리셋으로 변환되어 OpenAI TTS 합성 모듈에 전달되며, 감정에 적합한 발화 스타일과 속도가 적용된다.

[20] Pydub: 파이썬 기반 오디오 처리 라이브러리.

[21] Ffmpeg: 오디오, 비디오 포맷을 처리할 수 있는 오픈소스 멀티미디어 프레임워크.

[22] WAV: 무손실 오디오 파일 포맷

또한 시스템은 실시간성을 고려하여 설계되어, 영상 입력부터 감정 분석, 음성 합성 및 재생까지의 전체 처리 과정에서 표정 변화가 빠르게 반영되도록 구현하였다. 이를 통해 표정 변화 시 합성 보컬의 감정적 특성이 즉시 갱신되어, 사용자와 시스템 간의 자연스러운 상호작용을 지원한다.

그림 5. 웹 기반 실시간 감정 태그 생성 및 음성 합성 인터페이스

그림 5는 본 연구에서 구현한 웹 기반 인터페이스 화면을 나타낸다. 사용자가 입력한 문장과 실시간 표정 변화를 반영하여 자동으로 감정 태그가 생성되며, 각 문장별로 상이한 보컬 스타일과 발화 속도가 적용되는 과정을 직관적으로 확인할 수 있다.

4.4 뮤직테크놀러지 적용 가능성

기존 음악 창작 환경에서 감정 제어는 주로 수치화된 파라미터(MIDI velocity, filter cutoff frequency)나 사전 정의된 태그 입력 방식으로 이루어져 왔다. 그러나 이러한 접근법은 창작자에게 직관성이 부족하며, 연주자 또는 보컬리스트의 순간적 감정 변화를 실시간으로 반영하기 어렵다는 한계가 있다(박중희 & 한광희, 2023).

본 연구에서 제안한 시스템은 이러한 한계를 극복하기 위해 표정이라는 비언어적 신호를 감정 입력으로 활용한다. 사용자의 표정을 실시간으로 분석하여 추출한 감정 정보를 OpenAI TTS 합성 과정에 직접 반영함으로써, 창작자는 별도의 파라미터 조작 없이도 표정을 통해 감정을 직관적으로 전달할 수 있다.

이러한 방식은 단순한 음성 합성을 넘어, 보컬의 억양, 톤, 리듬을 실시간 감정 상태에 따라 동적으로 변화시킬 수 있다는 점에서 의미가 있다. 이를 통해 정서적 음악성을 강화할 수 있으며, 음악 및 사운드 제작 과정에서 창작자와 시스템 간의 자연스러운 상호작용을 가능하게 한다.

결과적으로, 본 시스템은 음악, 보컬 합성 및 인터랙티브 사운드 디자인 분야에서 실시간 감정 기반 생성 기술의 새로운 가능성을 제시하며, 감정 중심의 창작 환경을 위한 뮤직테크놀러지 발전에 기여할 수 있다.

5. 실험 설계 및 결과 분석

본 장에서는 제안한 표정 기반 감정-TTS 통합 시스템의 성능과 활용성을 검증하기 위해 수행한 실험 설계 및 결과를 기술한다. 실험은 크게 (1) 시스템 성능 평가, (2) 사용자 청취 평가 두 가지 관점에서 진행되었다.

5.1 실험 환경

실험은 Windows10 환경에서 수행되었으며, 하드웨어 사양은 AMD Ryzen 5 5600X 6-Core Processor, 32GB RAM, NVIDIA GeForce GTX 1060 6GB GPU 이다. 표정 인식 모듈은 Python 가상환경(venv)에서 구동되었으며, 주요 라이브러리는 OpenCV(4.10.0), MediaPipe(0.10.14), TensorFlow-CPU(2.13.1) 등을 포함한다. OpenAI TTS 는 openai(1.35.3), pydub(0.25.1), ffmpeg(6.1.1), Flask(3.0.3), requests(2.31.0)를 활용하였다. 또한, WebSocket[23]을 통한 실시간 인터페이스를 제공하였다.

5.2 시스템 성능 평가

제안된 시스템의 실시간성 확보 여부를 검증하기 위해, 동일한 문장을 대상으로 수동 감정 태깅 방식과 표정 기반 자동 감정 입력 방식을 비교하였다. 실험 결과, 수동 태깅의 경우 평균 약 6 초의 시간이 소요된 반면, 표정 기반 자동 입력 방식에는 약 4 초가 소요되어 전체 합성 과정이 약 2 초 단축되었다. 이는 감정 입력 절차를 자동화함으로써 전체 지연을 약 33% 감소시킨 것으로, 실시간 상호작용 환경에서 사용자 경험을 향상시키는 중요한 요소로 평가된다.

5.3 사용자 청취 평가

시스템의 감정 전달 효과를 검증하기 위하여 청취 평가를 실시하였다. 본 실험은 소규모 예비 청취 평가를 통해 시스템의 감정 전달 효과를 검증하였다. 참가자는 동일한 텍스트 문장을 기쁨, 슬픔, 분노, 중립 등 네 가지 감정 조건에서 합성하여 제시하였다. 그리고 각 샘플을 청취한 뒤 감정 전달의 명확성과 음성의

[23] WebSocket: 클라이언트와 서버 간에 양방향 실시간 통신을 가능하게 하는 프로토콜.

자연스러움을 평가하였다. 그 결과, 전반적으로 감정 구분이 비교적 명확하게 인지되었으며, 특히 기쁨과 슬픔 조건에서는 비교적 높은 감정 전달 효과가 관찰되었으나, 분노 조건에서는 감정 구분에 다소 혼동이 존재하였다. 이러한 결과는 제안된 시스템이 표정 기반 입력만으로도 감정적 변화를 유의미하게 전달할 수 있음을 시사하며, 동시에 감정 매핑 규칙과 학습 데이터의 정교화가 향후 성능 개선을 위한 과제로 남아있음을 시사한다.

5.4 논의

본 연구의 실험 결과는 제안된 표정 기반 감정-TTS 통합 시스템이 실시간성 및 감정 전달 효과 측면에서 의미 있는 진전을 보였음을 보여준다.

첫째, 감정 태깅 과정을 자동화함으로써 전체 처리 시간이 평균 6 초에서 4 초로 단축되었으며, 이는 약 33%의 효율성 개선으로 이어졌다. 이러한 성능 향상은 특히 사용자 상호작용 환경에서 중요한 의미를 가지며, 감정 기반 음성 합성의 실시간 응용 가능성을 뒷받침한다.

둘째, 청취 평가 결과에서 피험자들은 전반적으로 제안 시스템이 감정 표현의 명확성과 자연스러움에서 긍정적인 효과를 나타낸다고 응답하였다. 이는 표정 기반 입력이라는 직관적 인터페이스가 감정 표현의 몰입도를 높이는 데 기여할 수 있음을 시사한다. 다만 감정 범주 간 구분이 모호하거나, 특정 감정(놀람, 슬픔 등)의 전달력이 상대적으로 약하게 평가된 경우도 관찰되었다. 이는 감정-TTS 매핑 규칙의 정밀도 부족에서 기인할 수 있다.

셋째, 본 연구에서 구축한 시스템은 제한된 데이터셋과 환경에서 검증되었으므로, 실제 대규모 음악 퍼포먼스나 다양한 언어적 맥락에서의 활용에는 추가적인 확장이 요구된다. 특히 감정 인식 정확도를 높이기 위한 학습 데이터의 다양화, 감정

벡터와 음향 파라미터 간의 매핑 최적화, 그리고 더 많은 사용자 평가를 통한 객관적 검증이 필요하다.

종합하면, 제안된 시스템은 뮤직테크놀러지 맥락에서 감정 표현 보컬 합성의 새로운 가능성을 제시하였으며, 특히 실시간성 확보와 사용자 직관적 감정 제어라는 측면에서 실질적인 기여를 한다. 향후 연구에서는 감정 범주의 세분화, 다중 모달 입력(제스처, 생체 신호)과의 결합, 그리고 음악 창작 도구로서의 사용자 경험(UX) 최적화를 통해 본 시스템의 활용성을 더욱 확장할 수 있을 것이다.

Ⅲ. 결론

본 연구는 FER 과 감정 기반 TTS 를 통합하여, 사용자의 표정에 따라 실시간으로 감정이 반영된 보컬 합성을 구현하였다. 기존 시스템이 요구하던 수동적 감정 태깅 과정을 자동화함으로써 평균 처리 시간이 약 33% 단축되었으며, 청취 평가를 통해 감정 전달의 명확성과 자연스러움에서도 긍정적인 피드백을 확인하였다.

이러한 성과는 뮤직테크놀러지 분야에서 직관적 감정 제어 방식을 제안했다는 점에서 의의가 있다. 본 시스템은 감정을 수치화된 파라미터로 제어하는 기존 방식의 한계를 넘어, 사용자가 본능적으로 표현하는 표정을 음악 합성의 입력으로 활용할 수 있음을 보여주었다. 이는 향후 보컬 합성 기술이 단순한 음향 모사에서 벗어나, 정서적 음악성을 구현하는 방향으로 발전하는 데 기여할 수 있다.

다만 본 연구는 제한된 데이터셋과 소규모 청취 평가에 기반하여 진행되었으므로, 감정 범주의 세분화, FER 모델의 정밀도 향상, 다양한 언어와 음악적 맥락에서의 검증이 향후 과제로 남아 있다. 또한 실시간성 확보는 확인되었으나, 퍼포먼스 환경이나 대규모 사용자 인터랙션에 적용하기 위해서는 추가적인 최적화가 필요하다.

결론적으로, 본 연구는 감정 인식과 음성 합성을 통합한 실시간 보컬 합성 시스템의 가능성을 제시하였으며, 이는 음악 창작 및 뮤직테크놀러지 응용에서 감정 표현의 새로운 장을 열 수 있는 기초적 발판으로 작용할 것이다.

Ⅳ. 참고문헌

1. 남옥선. (2007). 음악활동이 영아의 사회・정서적 행동에 미치는 영향. 인간행동과 음악연구, 4(2), 18–40. https://scholar.kyobobook.co.kr/article/detail/4010028787890

2. 이은혜, & 김유정. (2020). 뮤지컬 노래의 극과 음악 분석을 통한 조 에스틸 보컬 기법(EVT)의 적용과 실제. 한국엔터테인먼트산업학회논문지, 14(6), 313–324. https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE10520484

3. 송미연. (2019). 사라 맥라클란(Sarah Mclachlan)의 보컬 플립 기법 활용 양상 연구 [석사학위논문, 상명대학교]. DBpia https://www.dbpia.co.kr/journal/detail?nodeId=T15395329

4. 이교구, & Cremer, M. (2009). 음악에서의 보컬 신호 인식을 위한 트레이닝 데이터 자동주석기법 연구. 한국음향학회지, 28(3), 201–211. https://www.semanticscholar.org/paper/9f23213ada6782aa583e01babd596986efa1128b

5. 김선영. (2025). 혁신 확산 이론에 기반한 K-POP 팬의 AI 커버 곡 선택 양상에 관한 연구. Journal of Future Society, 16(1), 133–153. https://doi.org/10.22987/jifso.2025.16.1.133

6. 유은정. (2020). TTS 시스템을 이용한 감정 합성 모델에 관한 연구. 차세대융합기술학회논문지, 4(4), 374–380. https://m.earticle.net/Article/A380080

7. Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. arXiv. https://arxiv.org/abs/2106.15561

8. 박중희, & 한광희. (2023). 감정별 음성분석을 통한 인공지능 TTS 도구들의 감정발화를 구분 짓는 SSML 문법 요인 설계. 한국콘텐츠학회논문지, 23(2), 148–157. https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE11656871

9. 안단태. (2024). 음악 분석을 통한 실시간 감성 인지 음악 생성 시스템 설계 [공학석사학위논문, 서울대학교]. 서울대학교 오픈 리포지터리. https://s-space.snu.ac.kr/handle/10371/216085

10. Lee, Y., & Kim, T. (2019). Robust and fine-grained prosody control of end-to-end speech synthesis. In ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 5911–5915). IEEE. https://ieeexplore.ieee.org/document/8683501

11. Sivaprasad, S., Kosgi, S., & Gandhi, V. (2021). Emotional prosody control for speech generation. In Interspeech 2021: 22nd Annual Conference of the International Speech Communication Association (pp. 4643–4647). ISCA. https://www.semanticscholar.org/paper/794ade10864b198f5b007f05eeb9220b219b72f8

12. 최광민. (2024, April 22). 몇 번의 학습만으로 “AI 실시간 인간 감정인식” 구현…UNIST 김지윤. AI타임즈. Retrieved September 10, 2025, from https://www.aitimes.kr/news/articleView.html?idxno=30167

13. Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. Behavioral and Brain Sciences, 31(5), 559–621. https://doi.org/10.1017/S0140525X08005293

14. 김혜란, & 송은성. (2021). 음악 구성요소의 감정 구조 분석에 기반한 시각화 연구. 한국디자인지식학회지, 53, 95–106. https://koreascience.kr/article/JAKO202119759310793.page

15. Luo, Y.-J., Hsu, C.-C., Agres, K. R., & Herremans, D. (2020). Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders. Proceedings of ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/ICASSP40776.2020.9054582

16. 권새봄, & 김주섭. (2019). 감정 전이와 실시간 얼굴 표정 리타게팅 기술 기반 음악 감상 경험 증진에 관한 연구. 디지털콘텐츠학회논문지, 20(6), 1097–1106. http://journal.dcs.or.kr/_PR/view/?aidx=20233&bidx=2220

17. 김영희. (2015). 표현적 웨어러블 테크놀로지의 확장: 웨어러블아트. 정보과학회지, 33(3), 58–66. https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE06555967

18. S. RajaA. & R. Harrabi. (2021). Facial Expression Recognition System Based on SVM and HOG Techniques.
https://www.semanticscholar.org/paper/cbfda4acb1c93f0d010f98219df27ea373310593