JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 7, pp. 33-39, July 2025
https://doi.org/10.9708/jksci.2025.30.07.033

Design and Implementation of an AI-Based Emotion Analysis System through Music and Vocal Separation

Byong-Kwon Lee*

*Professor, School of media contents, Seowon University, Chungbuk, Korea

[Abstract]

The technology for analyzing emotions in music has recently emerged as an important research topic in the field of affective computing. This study designed and implemented an AI-based system for music emotion analysis. Using Spleeter, vocals were separated from the original music, and audio features were extracted with Librosa. The extracted features were then fed into a pre-trained emotion classification model (SVM) to predict emotions at 5-second intervals. The prediction results were visualized through graphs and pie charts, clearly illustrating the temporal flow and distribution of emotions. Experimental results confirmed that both the number of predictions and the average probability are critical variables for emotion determination. In the future, the system can be expanded by refining emotion categories, incorporating instrument-based analysis, and applying deep learning techniques. The significance of this research lies in demonstrating the practicality and scalability of music emotion analysis technology.

▸ **Key words**: Music Emotion Analysis, Music and Vocal Separation, Audio Feature Extraction, Deep Learning, Emotion Visualization

[요 약]

음악에서 감정을 분석하는 기술은 최근 감성 인식 분야에서 중요한 연구 주제로 떠오르고 있다. 본 연구는 인공지능 기반의 음악 감정 분석 시스템을 설계하고 구현하였다. Spleeter를 이용해 음악에서 보컬을 분리하고, Librosa를 통해 오디오 특징을 추출하였다. 추출된 특징은 사전 학습된 감정 분류 모델(SVM)에 입력되어 5초 단위로 감정을 예측한다. 예측 결과는 그래프와 파이차트로 시각화되어 감정의 시간적 흐름과 비율을 명확히 제시했다. 실험 결과, 감정판단에는 예측 횟수와 평균 확률 모두가 중요한 변수임이 확인되었다. 향후에는 감정 세분화, 악기 기반 분석, 딥러닝 적용 등으로 시스템을 확장할 수 있고, 음악 감정 분석 기술의 실용성과 확장 가능성을 확인한 데 의의가 있다.

▸ **주제어**: 음악 감정 분석, 음악과 보컬 분리, 오디오 특징 추출, 딥러닝, 감성 시각화

• First Author: Byong-Kwon Lee, Corresponding Author: Byong-Kwon Lee
*Byong-Kwon Lee (sonic747@daum.net), School of media contents, Seowon University
• Received: 2025. 05. 30, Revised: 2025. 06. 20, Accepted: 2025. 06. 23.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

음악은 인간의 감정을 표현하는 중요한 매개체로, 오랜 역사 동안 감정을 전달하는 도구로 사용되어 왔다. 최근에는 디지털 음원과 스트리밍 서비스의 발전으로 인해, 음악을 통한 감정 분석의 중요성이 크게 증가하였다[1][2]. 특히, 감성 분석 기술은 음악 추천 시스템, 음악 치료, 감정 기반 상호작용 시스템 등 다양한 분야에서 활용되고 있으며, 음악의 감정을 정확히 파악하는 것이 중요한 연구 주제로 떠오르고 있다. 하지만 기존의 연구는 종종 전체 오디오 신호에서 감정을 분석하는 방식으로, 음악과 보컬의 요소가 결합된 상태에서 감정을 예측하였다[3]. 이에 따라 음악과 보컬을 분리하여 각각의 감정을 독립적으로 분석하는 방법에 대한 연구는 상대적으로 부족하였다. 본 연구는 음악과 보컬을 분리한 후, 각 요소에 대한 감정을 분석하는 시스템을 설계하였다. 첫 번째 단계는 오디오 파일에서 음악과 보컬을 분리하는 것이다. 이를 위해 Spleeter라는 오픈소스 라이브러리를 활용하여, 음악과 보컬을 각각 추출한다[4]. 두 번째 단계는 각 요소에서 특징을 추출하는 것이다. Librosa를 사용하여 MFCC, chroma, mel 스펙트로그램, 스펙트럴 대비, 톤넷 등 다양한 음향적 특징을 추출하고, 이를 감정 예측 모델에 입력하여 감정을 분류한다. 본 연구에서는 감정 예측을 위해 사전 학습된 딥러닝 모델을 사용하며, 예측된 감정은 시간 간격별로 시각화된다. 이 시각화에는 감정 변화 그래프와 확률 분포 그래프가 포함되며, 감정별 빈도와 평균 확률을 각각 파이 차트로 표시하여, 사용자가 감정 분석 결과를 직관적으로 이해할 수 있도록 한다. 본 연구에서 제안한 시스템은 음악과 보컬을 분리하여 각각의 감정을 분석하는 방법론을 제시한다. 이를 통해 음악의 감정적 흐름을 더 정확하게 추적하고, 음악의 특성에 따른 감정 변화를 보다 명확히 알 수 있다. 이 시스템은 음악 치료와 같은 분야에서 사용자 맞춤형 감정 분석을 제공할 수 있으며, 음악 추천 시스템에서도 사용자의 감정 상태에 맞는 곡을 추천하는 데 활용될 수 있다. 또한, 감성 기반 대화형 인공지능 시스템이나 감정 분석을 해야 하는 다양한 분야에서도 유용한 도구가 될 것이다. 본 연구는 음악과 감정 분석 기술의 융합을 통해, 향후 더 나은 감성 인터페이스 및 서비스를 개발하는 데 기여할 것이다.

II. Preliminaries

2.1 Emotion Analysis through Music and Vocal Separation

음악과 보컬은 각각 다른 감정적 정보를 담고 있으며, 이를 분리하여 분석하는 방식은 감정 인식의 정확도를 높이는 데 효과적이라는 연구가 다수 존재한다[5][6].

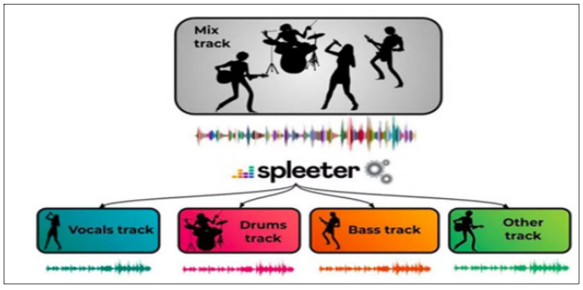


Fig. 1. Spleeter by Deezer

그림1은 오픈소스 소스 분리 라이브러리인 Spleeter를 통해 음악 신호에서 보컬과 반주를 효과적으로 분리할 수 있음을 보였으며, 이러한 분리는 감정 기반 오디오 분석의 전처리 과정에서 매우 유용하게 활용된다. 또한, 보컬을 중심으로 한 감정 분석 연구에서는 보컬의 억양, 속도, 음높이 등의 변화가 감정 인식한다.

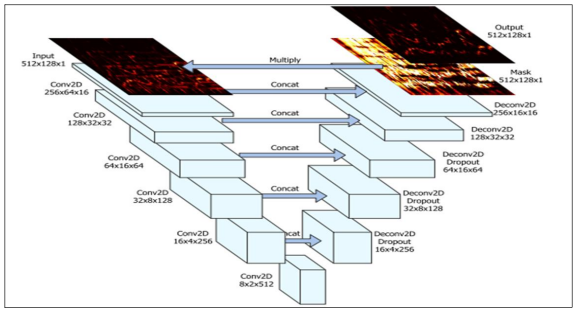


Fig. 2. Network Architecture of Spleeter

그림2는 U-Net 기반의 오디오 소스 분리 네트워크 구조이다[7][8]. 왼쪽 입력은 512x512 크기의 오디오 스펙트로그램이며, 인코더에서는 Conv2D 계층을 통해 특징을 추출한다.

중간에서는 각 단계의 특징 맵이 skip connection을 통해 디코더와 연결하고, 오른쪽 디코더에서는 Deconv2D 계층을 사용해 원래 해상도로 복원하면서 필요한 오디오 소스를 생성한다. 최종 출력은 특정 소스를 추출한 마스크(mask)를 곱해 분리된 오디오 스펙트로그램을 생성한다.

2.2 Deep learning-based voice emotion classification model

음성 신호의 감정을 분류하는 데에는 전통적인 기계학습 기법에서 딥러닝 기반 접근 방식으로 연구가 빠르게 전환되고 있다. 특히 CNN, LSTM, 그리고 최근의

Transformer 기반 모델들은 음성 특징(MFCC, Chroma, Spectral Contrast 등)을 입력으로 하여 높은 정확도의 감정 분류를 실현하고 있다. 화자의 음성만을 활용하여 감정을 인식하는 딥러닝 기반 방법으로, 음성 데이터를 멜 스펙트로그램으로 변환하여 시간-주파수 특성을 추출한다. 이후 CNN으로 특징을 벡터화하고, Bi-Directional LSTM을 통해 시간 흐름에 따른 감정 변화를 모델링한다. 완전 연결 계층(Fully Connected Layer)을 통해 최종 감정을 6가지(Anger, Excitement, Fear, Happiness, Sadness, Neutral)로 분류한다[9].

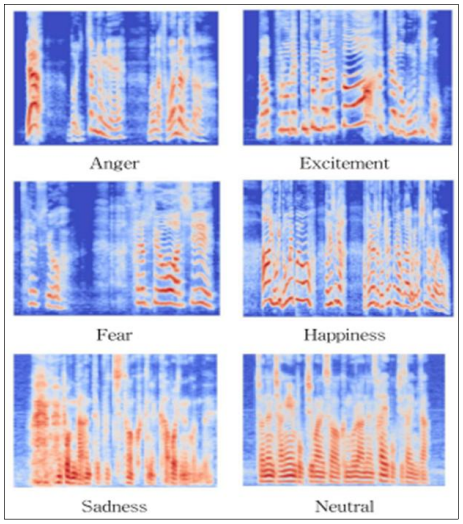


Fig. 3. Mel Spectrogram by Emotion

2.3 Emotion Visualization and User Interface Research

감정 분석 결과를 시각적으로 표현하는 연구도 활발히 이루어지고 있다. 감정 변화의 흐름을 시간 축에 따라 시각화하는 방식은 사용자로 하여금 음악의 감정적 구조를 보다 직관적으로 이해하게 해준다[10-12]. 그림 4는 감정 분포와 변화를 시각적으로 표현하는 대시보드 연구이다. 그래프는 음악 감정을 Scary, Happy, Sad, Peaceful의 네 가지 범주로 분류한 실험 결과를 보여준다. 각 감정 범주는 상단의 음악적 단서(Cue levels) 조합에 따라 구성되었다. 가로축은 감정 유형, 세로축은 참가자들의 평균 감정 평점(0~7점)을 나타낸다. 색상별 막대는 각각의 감정이 해당 음악에서 얼마나 강하게 인식되었는지를 시각화한다. 결과적으로 감정별 음악 예시는 해당 감정을 주로 유발하며, 그에 따라 뚜렷한 평점 차이를 보인다. 네 가지 감정 예시는 서로 뚜렷하게 구분되지만, 일부 감정 간에는 관련성이 나타난다. 예를 들어, 행복과 슬픔 예시는 평온함에서도 중간 정도의 점수를 받았고, 평온한 음악은 슬픔에서도 유사한 점수를 보였다. 그러나 95% 신뢰구간을 통해 볼 때, 이러한 중첩은 감정 혼동이 아닌 관련성으로 해석된다. 또한 극단적인 단서뿐 아니라 중간 수준의 단서들도 감정 표현에 활용된 것이 확인된다.

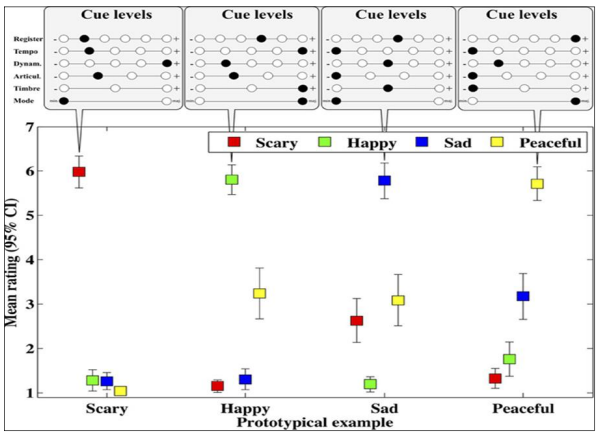


Fig. 4. Emotional expression in music

III. The Proposed Scheme

본 절에서는 절차와 실험환경은 시험에 사용한 음악은 외국 보컬 음악인 “ Becky G, KAROL G. MAMIII, Gnarls Barkley-Crazy, Rema, Calm Down” 3개의 음악에 대하여 감정 “Sad, happy, angry, neutral”에 대하여 분석했다.

Table 1. Environmental conditions

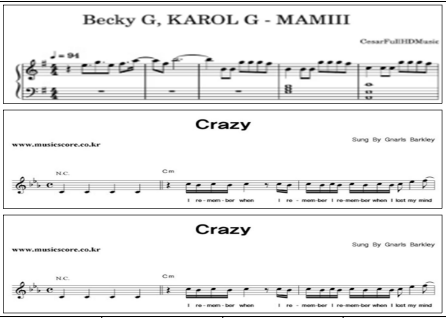
Div	Environmental conditions			
Title	1. Becky G, KAROL G. MAMIII 2. Gnarls Barkley-Crazy 3. Rema, Calm Down			
Music Sheet				
Emotion emoticon		happy		Sad
		angry		neutral
Algorithm	Spleeter, Librosa			
Tools	Visual Studio Code, Python, Pytoch			

표1는 연구 분석에 사용한 실험환경이다.

Table 2. Vocal emotion analysis procedure

step	Descripion	Function
1 Input Setup	Define the input audio file and prepare it for processing	input_audio_path = 'music.mp3' librosa.load()
2 Vocal Separation	Separate vocals from the full audio using Spleeter (2stems)	Separator.separate_to_file()
3 Feature Extraction	Extract audio features (MFCC, Chroma, etc.) to form a 180-dimension vector	extract_features()
4 Model Loading	Load the pre-trained emotion classification model (.pkl file)	joblib.load("emotion_model.pkl")
5 Emotion Prediction	Predict emotion by inputting feature vector into the model	model.predict() model.predict_proba()
6 Visualization & Output	Visualize predicted emotions over time and show prediction confidence	process_audio_in_chunks() using matplotlib

표2는 음악(보컬) 기반 감정인식 시스템은 오디오에서 감정 흐름을 분석하기 위해 총 6단계의 절차로 구성된다. 첫 번째 단계는 입력 오디오 파일을 지정하고 처리 가능한 형태로 로딩한다(MP3포맷). 두 번째 단계에서는 Spleeter 라이브러리를 활용하여 원본 음악에서 보컬 신호를 분리한다. 세 번째 단계는 분리된 보컬 신호로부터 MFCC, Chroma 등의 오디오 특징을 추출하고, 이를 고정된 길이(예: 180차원)의 벡터로 정규화하는 과정이다. 네 번째 단계에서는 다수의 레이블된 음성 데이터를 수집하여 특징을 추출하고, 이 특징 벡터와 감정 레이블(happy, sad, angry, neutral 등)을 기반으로 머신러닝 모델(SVM, RandomForest 등)을 학습한다. 이렇게 학습된 모델은 joblib을 이용해 .pkl 형식으로 저장되며, 이는 이후 감정 예측에 활용된다. 다섯 번째 단계에서는 위에서 저장한 .pkl 모델을 로드하여 예측을 위한 준비단계이다. 여섯 번째 단계에서는 오디오를 일정 시간 단위로 분할하여 각 구간마다 특징을 추출하고, 이를 기반으로 모델을 통해 감정을 예측하고, 예측 확률(confidence score)도 함께 도출한다. 이로써 시간 축에 따른 감정 변화와 예측 신뢰도를 시각화하는 과정하고, matplotlib을 활용해 결과를 출력한다. 이와 같은 전체 프로세스는 보컬 신호에 내재 감정 정보를 디지털 추출한다. 특히, .pkl 생성 과정은 감정 분류기의 핵심으로, 신뢰도 높은 예측을 위해 다양한 감정 레이블과 균형 잡힌 학습 데이터셋이 향후 연구로 정확도 향상에 필요하며, 본 연구에서는 4개의 감정과 데이터를 사용해 진행했다. 표3은 감정예측 모델학습으로 오디오 감정 인식을 위해 레이블이 포함된 음성 데이터를 수집하고, MFCC 등의 특징을 추출해 고정된 길이로 정규화한다. 이후 감정 레이블을 숫자로 변환하고, 분류 모델을 학습시켜 .pkl 형식으로 저장해, 생성된 .pkl 파일은 추후 감정 예측에 사용한다.

Table 3. Learning emotion prediction model(pkl)

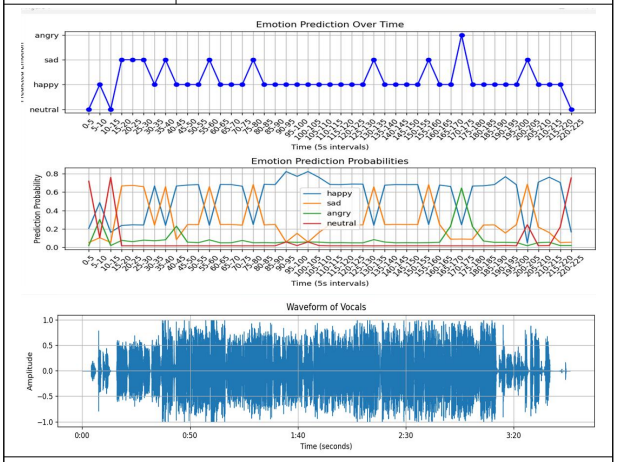
Step	Description	Function
1. Data Collection	Gather audio or vocal recordings labeled with emotion categories	Custom scripts or external datasets
2. Feature Extraction	Extract audio features such as MFCCs, Chroma, etc. from each audio sample	extract_features(audio_path)
3. Feature Normalization	Pad or truncate feature vectors to a fixed length (e.g., 180 dimensions)	pad_or_truncate(features, length=180)
4. Label Encoding	Convert string-based emotion labels (e.g., "happy", "sad") into numeric form	LabelEncoder().fit_transform()
5. Model Training	Train a classifier (e.g., SVM, Random Forest) using features and labels	model.fit(X, y)
6. Model Saving	Save the trained model to a .pklfile for future use	joblib.dump(model, 'emotion_model.pkl')

IV. Experiment and analysis

본 절에서는 대중 음악인 1. Becky G, KAROL G. MAMIII, 2. Gnarls Barkley-Crazy, 3. Rema, Calm Down에 대하여 감정요소 “Sad, happy, angry, neutral”를 실험 및 분석했다. 표4는 “Becky G, KAROL G. MAMIII” 에 대한 감정예측 결과이다. 재생시간 3분 42초를 분석한 결과 “happy”가 많이 발생했으며 평균 확률이 높은 부분이 0.75으로 “neural”이 가장 높았다. 결론적으로 전체적인 노래의 감정 형태는 “happy”하면서 “neural”하다고 판단할 수 있다.

Table 4. Emotion analysis of MMAMIII

Album	Information
	• Becky G, KAROL G. MAMIII *ref)https://www.youtube.com/watch?v=vb8wloc4Xpw* • Duration : 3minutes 42seconds • File format : MP3



- happy: 32 times, average probability: 0.69
- sad: 9 times, average probability: 0.67
- angry: 1 time, average probability: 0.65
- neutral: 3 times, average probability: 0.75

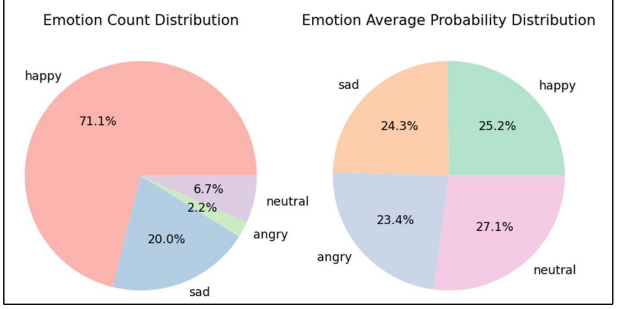
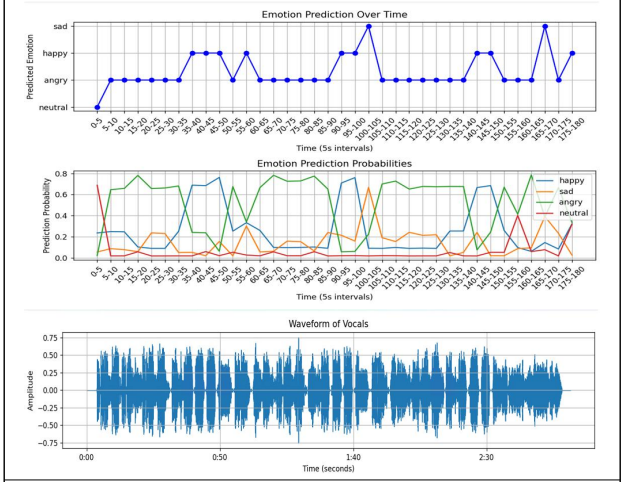


Table 5. Emotion analysis of Crazy

Album	Information
	• Gnarls Barkley-Crazy *ref)https://www.youtube.com/watch?v=oBmmJB4V5CI* • Duration : 2minutes 57seconds • File format : MP3



- happy: 9 times, average probability: 0.62
- sad: 2 times, average probability: 0.53
- angry: 24 times, average probability: 0.68
- neutral: 1 time, average probability: 0.69

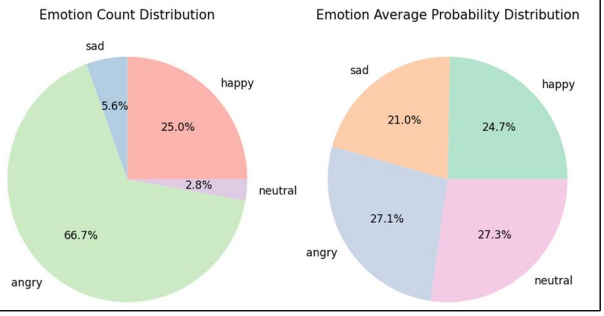


표5는 “Gnarls Barkley-Crazy” 에 대한 감정예측 결과이다. 재생시간 2분 57초를 분석한 결과 “angry”가 많이 발생했으며 평균 확률이 높은 부분이 0.75으로 “neural”이 가장 높았다. 결론적으로 전체적인 노래의 감정 형태는 “angry”하면서 “neural”하다고 판단할 수 있다.

표6는 “Rema, Calm Down” 에 대한 감정예측 결과이다. 재생시간 3분 40초를 분석한 결과 “angry”가 많이 발생했으며 평균 확률이 높은 부분이 0.74으로 “sad”이 가장 높았다. 결론적으로 전체적인 노래의 감정 형태는 “angry”하면서 “sed”하다고 판단할 수 있다.

Table 6. Emotion analysis of MMAMIII

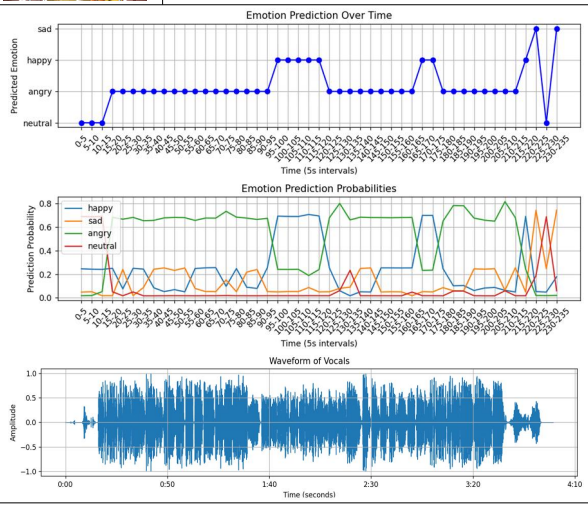
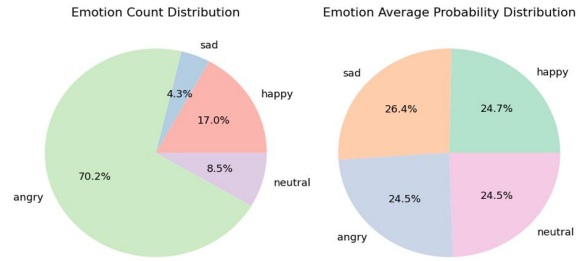
Album	Information
	• Rema, Calm Down *ref)https://www.youtube.com/watch?v=hmeOm9NmBg8* • Duration : 3minutes 40seconds • File format : MP3
	• happy: 8 times, average probability: 0.69 • sad: 2 times, average probability: 0.74 • angry: 33 times, average probability: 0.69 • neutral: 4 times, average probability: 0.69
	

표 4, 5, 6의 실험 결과를 통해 음악의 감정은 '예측 횟수'와 '평균 확률'이라는 두 가지 조건에 의해 결정된다는 점을 확인할 수 있었다. 예측 횟수가 많을수록 해당 감정이 주요 감정으로 판단될 가능성이 높지만, 평균 예측 확률이 낮다면 해당 감정에 대한 신뢰도는 떨어질 수 있다. 예를 들어, "angry" 감정이 가장 많이 예측되었더라도 그에 대한 확률값이 낮다면 실제로는 그 감정을 대표한다고 보기 어렵다. 따라서 신뢰도 높은 감정 분석을 위해서는 반복 횟수와 평균 확률 두 요소가 모두 충분히 충족되어야 함을 확인하였다.

V. Conclusions

최근 인공지능 기술의 발전과 함께 감성 인식 기술은 음악, 헬스케어, 콘텐츠 큐레이션 등 다양한 분야에서 주목받고 있다. 특히 음악 감정 분석은 인간의 감정 상태를 이해하고, 감성 중심의 서비스를 제공하는 핵심 기술로 대두되고 있다. 본 연구는 이러한 시대적 흐름에 부응하여, 오디오로부터 음악과 보컬을 분리하고, 이를 기반으로 감정을 분석하는 시스템을 설계하고 구현하였다. 본 시스템은 Spleeter를 활용한 보컬 분리, Librosa 기반의 음성 특징 추출, 사전 학습된 머신러닝 감정 분류기(SVM 기반)를 결합하여 구성되었다. 5초 간격으로 오디오를 분할하고 각 구간의 감정을 예측함으로써, 시간 축에 따른 감정 흐름을 정밀하게 파악할 수 있도록 설계하였다. 예측된 감정은 시각적으로 표현되었고 정보를 제공하며, 두 개의 파이차트를 통해 감정별 발생 빈도와 확률을 명확히 구분하였다. 실험은 총 세 곡의 대중 음악을 대상으로 수행되었으며, 감정 레이블은 happy, sad, angry, neutral 네 가지로 제한하였다. 실험 결과, 감정의 판별에는 반복 횟수와 예측 확률 모두가 중요한 요소임을 확인하였다. 특정 감정이 많이 예측되었더라도 그 확률이 낮다면 신뢰성이 낮을 수 있으며, 반대로 적게 나타난 감정이라도 확률이 높다면 더 신뢰할 수 있는 감정 판단이 가능하다. 또한, 모델 학습 시 사용한 .pkl 파일 생성 과정은 본 시스템의 핵심이며, 감정 예측의 정확도를 좌우한다. 다양한 감정 레이블과 감성셋에 대한 많은 학습 진행하는것이 향후 정확도 향상을 위한 필수 요소로 판단된다. 본 연구의 결과는 음악 감정 흐름을 디지털로 정량화하고 시각화하는 새로운 방향성을 제시하였으며, 감성 기반의 콘텐츠 분석 및 추천, 음악 치료 분야에 적용 가능성을 보였다. 향후 연구에서는 감정 레이블의 세분화, 보컬 외 악기 감정 분석, 실시간 스트리밍 기반 감정 분석 기능 확장 등이 필요하다. 또한, 딥러닝 기반의 모델로 확장하여 예측 성능을 강화하고, 실제 사용자 피드백을 반영한 감정 평가 시스템과의 통합도 고려할 수 있다.

REFERENCES

[1] C. -F. Huang and C. -Y. Huang, "Emotion-based AI Music Generation System with CVAE-GAN," 2020 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE), Yunlin, Taiwan, 2020, pp. 220-222, doi: 10.1109/ECICE50847.2020.9301934.

[2] D. Zhang, X. Li, D. Lu, Y. Tie, Y. Gao and L. Qi, "Multitrack Emotion-Based Music Generation Network Using Continuous Symbolic Features," 2024 IEEE International Conference on Multimedia and Expo (ICME), Niagara Falls, ON, Canada, 2024, pp. 1-6, doi: 10.1109/ICME57554.2024.10688343.

[3] V. S. Narayanan and A. Tarafdar, "Music Therapy-Driven Mood-Based Music Recommendation System Integrating User Emotion, Song Lyrics, and Health Reflections," 2025 AI-Driven Smart Healthcare for Society 5.0, Kolkata, India, 2025, pp. 19-24, doi: 10.1109/IEEECONF64992.2025.10963042.

[4] P. Kalansooriya, G. A. D. Ganepola and T. S. Thalagala, "Affective gaming in real-time emotion detection and Smart Computing music emotion recognition: Implementation approach with electroencephalogram," 2020 International Research Conference on Smart Computing and Systems Engineering (SCSE), Colombo, Sri Lanka, 2020, pp. 111-116, doi: 10.1109/SCSE49731.2020.9313028.

[5] R. Orjesek, R. Jarina, M. Chmulik and M. Kuba, "DNN Based Music Emotion Recognition from Raw Audio Signal," 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), Pardubice, Czech Republic, 2019, pp. 1-4, doi: 10.1109/RADIOELEK.2019.8733572.

[6] Z. Zhang, L. Xie and Z. Zhao, "Temporal Integration of Emotion Perception for Cross-Cultural and Multi-Emotion Music," 2021 International Conference on Culture-oriented Science & Technology (ICCST), Beijing, China, 2021, pp. 198-203, doi: 10.1109/ICCST53801.2021.00050.

[7] Jansson, A., Humphrey, E., Montecchio, N. view all authors (2017). Singing voice separation with deep U-Net convolutional networks. Paper presented at the 18th International Society for Music Information Retrieval Conference, 23-27 Oct 2017, Suzhou, China.

[8] SangHyeuk Yoon, Dayun Jeon, Neungsoo Park,“Speech emotion recognition based on CNN - LSTM Model”, ACK 2021, p939~p941

[9] Eerola Tuomas , Friberg Anders , Bresin Roberto, “Emotional expression in music: contribution, linearity, and additivity of primary musical cues”, Frontiers in Psychology, Volume 4 - 2013, doi:10.3389/fpsyg.2013.00487

[10] J. Ohene-Djan, A. Sammon and R. Shipsey, "Colour Spectrum's of Opinion: An Information Visualisation Interface for Representing Degrees of Emotion in Real Time," Tenth International Conference on Information Visualisation (IV'06), London, UK, 2006, pp. 80-88, doi: 10.1109/IV.2006.34.

[11] N. Wagener, J. Schoning, Y. Rogers and J. Niess, "Letting It Go: Four Design Concepts to Support Emotion Regulation in Virtual Reality," 2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Shanghai, China, 2023, pp. 763-764, doi: 10.1109/VRW58643.2023.00224.

[12] Y. -S. Chen, L. -H. Chen, T. Yamaguchi and Y. Takama, "Visualization system for analyzing user opinion," 2015 IEEE/SICE International Symposium on System Integration (SII), Nagoya, Japan, 2015, pp. 646-649, doi: 10.1109/SII.2015.7405055.

Authors

Byong-Kwon Lee received the B.S., M.S. and Ph.D. degrees in Computer Science and Engineering from Hanbat, Hannam and Chungbuk University Korea, in 2000, 2003 and 2007, respectively.

My main areas of interest are embedded systems, virtual and augmented reality(VR.AR), and artificial intelligence(AI). The field currently being studied is the construction of an exhibition hall using virtual reality. It is a technology that combines AI with cultural uniform restoration technology as a future research field.