Emotion-Based Music Recommendation System Using Speech Analysis

Atharv Yewale[1], Piyush Salve[2], Siddhesh Sangale[3] & Dr. A. R. Deshpande[4]

[1]Student, SCTR's Pune Institute of Computer Technology, (Computer Engineering), Pune, Maharashtra, India, atharv1005@gmail.com

[2]Student, SCTR's Pune Institute of Computer Technology, (Computer Engineering), Pune, Maharashtra, India, piyushsalve3010@gmail.com

[3]Student, SCTR's Pune Institute of Computer Technology, (Computer Engineering), Pune, Maharashtra, India, siddheshsangale0705@gmail.com

[4]Associate Professor, SCTR's Pune Institute of Computer Technology, (Computer Engineering), Pune, Maharashtra, India, ardeshpande@pict.edu

Abstract

Emotion-based music recommendation systems represent an innovative method of boosting user interaction by customizing music recommendations according to the listener's emotional condition, offering a more individualized and enjoyable listening experience. These systems leverage advanced Artificial Intelligence models to recognize emotions from voice and speech, aligning music choices to match or elevate the listener's mood and support emotional well-being. This paper explores the latest advancements in AI[1]-based emotion detection, focusing on deep learning techniques that enable real-time voice-based emotion recognition. Through effective preprocessing, emotion classification, and feature extraction, these models achieve accurate and responsive music recommendations. By examining these technologies, the system highlights how AI[2] enables dynamic, emotion-aligned music recommendations that enhance both engagement and satisfaction in music consumption.

Keywords: Emotion recognition, speech analysis, music recommendation, affective computing, personalization.

1. Introduction

Music is often seen as a universal language, a powerful way to express emotions that goes beyond cultural and language barriers, helping to stir feelings and enhance well-being. Research has demonstrated its therapeutic effects, from reducing stress and enhancing mental focus to fostering resilience. This potential has been amplified by digital platforms, which have opened up vast, readily accessible music libraries that span an array of genres, cultural contexts, and emotional ranges. The unprecedented availability allows users to explore music at will, expanding personal connections with songs and artists worldwide.

However, with so much music at one's fingertips, users often experience "choice overload," where the sheer number of options can make selecting music feel daunting and detract from enjoyment. To address this, emotion-based music recommendation systems have been developed to simplify music discovery by offering curated suggestions that align with a listener's current mood or emotional state. By analysing vocal or behavioural cues, these AI[3]-powered systems help users find music that not only suits their preferences but also enhances their emotional experience, bringing a more intentional, personal dimension to digital music interaction.

2. Related Work

2.1 Emotion Detection Techniques:

This section explores diverse AI[4] methodologies for detecting emotions in music therapy and speech, illustrating how these technologies enhance therapeutic effectiveness and improve accuracy in emotion recognition.

In [1], various AI[5] techniques are reviewed for identifying emotions in music therapy, highlighting how these methods are used to enhance therapeutic effectiveness. Results from the study indicate that these AI[6] methods significantly improve the emotional impact and personalization of therapy sessions.

In [2], deep learning methods are explored for recognizing emotions in speech, demonstrating advances in emotional accuracy within speech analysis. The findings show a notable improvement in recognition accuracy, especially when using models with multiple deep layers.
In [3], the research presents a system that recommends music based on emotional data, aiming to create more personalized playlists for users. The system's results reflect improved user satisfaction, as recommendations better match the listener's emotional state.

In [4], a personalized emotion-driven music recommendation system is examined, emphasizing the alignment of music with emotional preferences to enhance user engagement. The outcomes reveal high levels of user satisfaction due to more relevant music choices tailored to individual emotions.

2.2 Emotion-Based Music Recommendation Systems:

This part delves into systems designed to curate music based on emotional data, emphasizing the strides made in personalizing playlists and the resulting boosts in user satisfaction and emotional health.

In [5], an AI[7]-powered music therapy system is developed to improve emotional well-being by selecting music that corresponds to the user's emotional state. Results demonstrate that the system effectively supports emotional wellness, leading to positive therapeutic effects.

In [6], research focuses on detecting emotions through voice and speech recognition, contributing to advancements in real-time emotional analysis for digital applications. The study results highlight a robust increase in detection accuracy, making the approach suitable for real-time use.

In [7], an emotion-based music recommendation system is introduced, applying AI[8] technologies to deliver a more personalized music experience. This system shows high accuracy in pairing songs with user emotions, enhancing listener satisfaction.

In [8], a system is proposed that uses speech emotion recognition to recommend music that matches the user's current mood, offering a more immersive experience. Results from the implementation indicate a successful match between recommended music and user preferences, enriching the user experience.

2.3 AI and Music Therapy Systems:

Here, the focus is on the role of emotion recognition in music therapy and recommendation platforms, highlighting how these innovations enrich user experiences and refine the precision of music selections tailored to listeners' moods.

In [9], speech-based emotion classification is investigated to improve user interaction with digital music platforms by refining music suggestions. Findings reveal that incorporating emotion recognition significantly boosts the accuracy of recommendations, making music selections more appealing.

In [10], the study investigates how recognizing emotions through voice can enhance music recommendations that suit the listener's mood. Results indicate that this approach leads to more accurate music suggestions, thus creating a highly personalized listening experience.

3. Algorithms Used

3.1 Librosa Feature Extraction

The audio is loaded using *librosa.load()*, which returns the audio waveform and sampling rate.
Formula: y,sr=librosa.load(filepath, sr=None) where y is the audio signal and sr is the sampling rate.
MFCCs are extracted to capture key audio features using *librosa.feature.mfcc()*.
Formula:

$$\text{MFCC(n)}=\sum_{m=1}^{M} \log(S[m]) \cos\left(\frac{\pi n(m-0.5)}{M}\right) \quad (3.1)$$

Here, S[m] represents the Mel power spectrum for a given filter.
Other features like Chroma and Spectral Contrast can also be computed using Librosa functions.
The extracted features are scaled to a consistent range to ensure better compatibility with machine learning models.

3.2 Convolutional Neural Network (CNN)

The network takes in a feature matrix (e.g., MFCCs), where rows represent frames, and columns represent coefficients.
Filters slide over the input to identify patterns.
Formula:

$$Z=\text{ReLU}(W*X+b) \quad (3.2)$$

where W is the kernel, * denotes convolution, X is the input data, and b is bias.

The feature map dimensions are reduced using pooling operations like MaxPooling, retaining key information while reducing computation.
Flattened outputs from the previous layers are fed into dense layers for learning non-linear relationships.
The final layer generates probabilities for each emotion class.
Formula:

$$\sigma(z_i)=\frac{\exp(z_i)}{\sum_{j=1}^{N}\exp(z_j)} \quad (3.3)$$

This ensures that the outputs sum to 1, representing the likelihood of each class.

4. Proposed System

Convolutional Neural Network (CNN) for Speech Emotion Recognition

The project employs a Convolutional Neural Network (CNN) for effective speech emotion recognition, crucial for enhancing music recommendation systems. CNNs[1] are particularly adept at extracting local features from audio spectrograms, allowing for precise emotion detection based on speech patterns.

Steps Involved:

Audio Preprocessing: Raw audio files are processed to generate spectrograms, providing a time-frequency representation of the sound.

Feature Extraction: The CNN[2] architecture is utilized to extract meaningful features from the spectrograms, focusing on critical aspects that indicate emotional states.

Emotion Classification: The extracted features are classified into predefined emotional categories, enabling the system to identify the emotion expressed in the speech.

Music Recommendation: Based on the classified emotion, the system matches it with a database of emotionally-tagged music tracks, suggesting songs that align with the detected emotional state.

Key Processes:

Spectrogram Generation: Converts audio signals into a visual format for enhanced feature extraction.
Convolutional Layers: Captures local patterns and nuances in audio data, crucial for emotion detection.
Pooling Layers: Reduces dimensionality, highlighting important features while minimizing noise.
Softmax Classification: Classifies identified features into specific emotional categories in the final layer.

Advantages:

High Accuracy: The CNN [3] approach provides robust performance in recognizing emotions from speech, improving the relevance of music recommendations.

Automatic Feature Learning: The model automatically learns features from the data, reducing the need for extensive manual feature engineering.

Real-time Processing: The architecture is capable of processing audio input in real time, making it suitable for interactive applications.

Problem Solved Successfully

By utilizing CNNs [4] for speech emotion recognition, the project effectively addresses the challenge of accurately detecting emotions from audio signals, facilitating personalized music recommendations that enhance user experience.

5. Methodology

The methodology for emotion detection and music recommendation commences by capturing a user's speech via the device's microphone, ensuring real-time interaction that enables natural speech flow. Techniques for reducing background noise and assessing microphone quality are employed to improve the clarity of audio recordings. After capturing the audio, the data undergoes preprocessing, where it is transformed into spectrograms or Mel-frequency cepstral coefficients (MFCCs), making it suitable for model input. During this phase, normalization methods ensure uniformity in audio data processing, and noise reduction alongside voice activity detection isolates the most relevant speech segments. The processed audio is then passed into a custom-designed Convolutional Neural Network (CNN), specifically tuned for emotion recognition in speech. This model is optimized for mobile and web platforms, incorporating lightweight architectures and quantization techniques to boost efficiency. Batch processing is optionally employed to manage more extended or complex speech inputs effectively. The CNN model categorizes the speech into specific emotional states—such as happiness, sadness, anger, or neutrality—using a softmax layer to calculate

the likelihood of each emotion. To deal with uncertainty in classifications, a threshold mechanism is applied, and confidence scores may be provided to give further insight into the detection process.

In the music classification stage, the system extracts relevant audio features from the song database. These features include temporal aspects like rhythm and tempo, spectral data such as MFCCs[1] and spectral centroid, harmonic information like key and chord progressions, and dynamic features such as loudness. Libraries such as Librosa or Essentia are leveraged to facilitate efficient and accurate feature extraction. Feature selection methods are then applied to highlight the most pertinent features for classifying music based on emotional content. Subsequently, machine learning models—including Support Vector Machines (SVM), Random Forests, and Neural Networks—are trained on this feature data to categorize music by emotion. Various models are evaluated, and cross-validation is used to ensure these models perform well on unseen data. Ensemble techniques may also be applied to combine model outputs for improved classification accuracy. The system's accuracy is measured through metrics like F1scores and confusion matrices. Ultimately, a database of songs tagged with emotional labels is established, and quick search methods such as Elasticsearch are implemented to facilitate fast retrieval. Metadata, such as artist, genre, and release date, is also incorporated to enable more granular searches and improve user-specific music recommendations.

The following block diagram shows the process flow of the system.

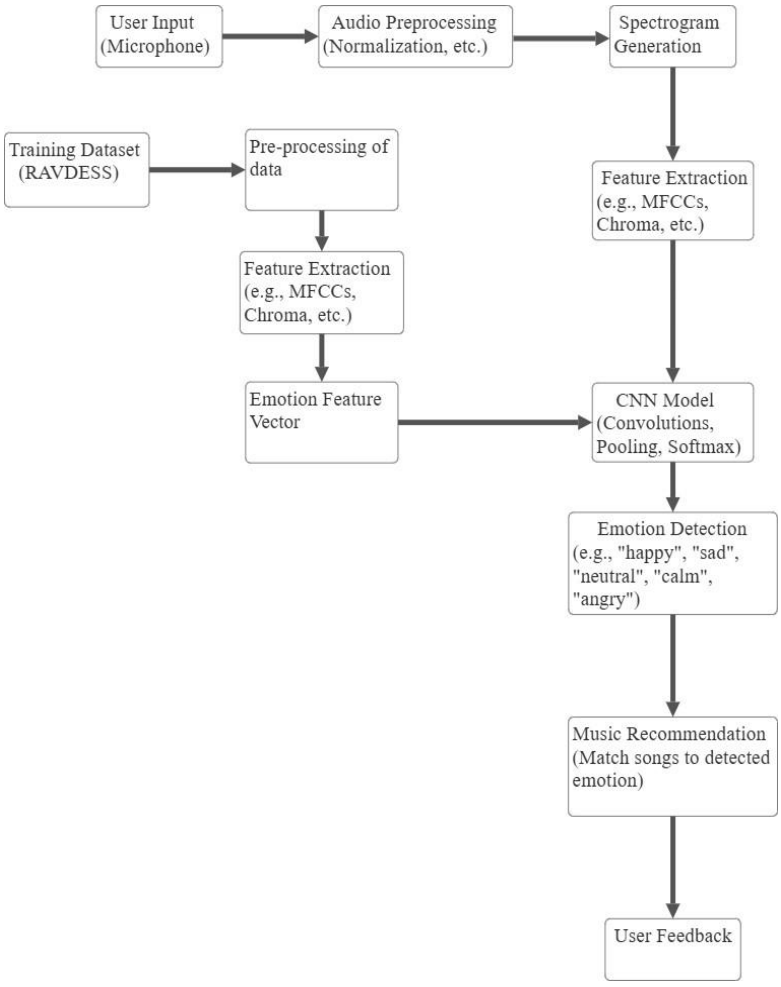

Fig 1. Flow Diagram of the Emotion Detection and Music Recommendation System

Table 1. Overview of Public Datasets for Emotion-based Speech Analysis

DATASETS	YEAR	TYPE	VOLUME OF TRAFFIC
RAVDESS	2018	SPEECH	3.44 GB
EMO-DB	2000	SPEECH	500 MB
TESS	2013	SPEECH	430 MB
SAVEE	2012	SPEECH	680 MB
CREMA-D	2018	SPEECH	1.2 GB

Table 2. An Overview of the Reviewed Research

Work	ML Algorithms	Features	Data Sources	Emotion/Classification Levels
Aouani et al. [1]	Neural Networks, SVM	Audio Features (MFCCs, Spectral)	Public Speech Databases	High-level Emotional States (e.g., Happy, Sad)
Huang et al. [2]	Deep Learning	Speech Spectrograms, Prosodic Features	TED Talks, Emotional Speech Datasets	Fine-grained Emotional Tones
Katkuri et al. [3]	Random Forest, SVM	Music Metadata, Audio Features	User-generated Song Libraries	Personalized Emotion-based Playlist
Kumar et al. [4]	Ensemble Learning	User Emotion Feedback, Tempo, Rhythm Patterns	Private Music and Feedback Dataset	Real-time Emotion Synchronization
Lee et al. [5]	AI Model Fusion	Emotion Annotations in Music Therapy	Controlled Environment with Patients	Enhanced Emotional Wellbeing
Rastogi et al. [6]	CNN, Voice Recognition Models	Speech MFCCs, Pitch, Duration	Open Speech Databases	Basic Emotion Classes
Rumiantcev et al. [7]	Hybrid Neural Networks	Song Audio Analysis, Emotion Labels	Experimentally Collected Dataset	Personalized Recommendations

Singh et al. [8]	Deep Learning	Speech Prosodics, Audio Similarity	Proprietary Dataset from Streaming Services	Mood-based Classification
Wong et al. [9]	Decision Trees	Speech Emotion Data	User Speech Feedback	Refined Song Suggestions
Woo & Kim [10]	Voice-based AI Models	Audio Features (Pitch, Loudness)	Controlled Laboratory Data	Personalized Music Matches

Table 3. Reviewed Work in Light of Considerations for Emotion Detection and Music Recommendation through Speech Analysis

Reference	Objective	Methodology	Key Results	Limitations
Morales-Perez et al. (2008)	Emotion detection from speech using spectral and acoustic features	Time-frequency transforms (Gabor, WVD, DWT), LPA. etc.	Achieved 94.6% recognition accuracy on database (SES)	Focused on language (Spanish), limited scope of emotional categories
Wintrode et al. (2015)	Content-based recommendation for spoken documents	Multisource feature extraction (speech, environment, speaker) with fusion techniques (score, feature, hybrid)	Multisource approach halved error rate compared to bag-of-words baseline	Study limited to recommendation of spoken documents, not music
Pan et al. (2006)	Improve speech recognition under emotional states	Adaptation of acoustic models using GMM-based emotion detection and emotion-matched models	Recognition rate increased to 80.79% with emotion-matched adaptation	High computational demand for model selection
Sato et al. (2020)	Develop an emotional speech database with multiple emotion intensities	Collected 2,025 samples with multiple emotion labels from video sources; statistical evaluation of emotional intensity	Demonstrated the importance of considering multiple emotions in SER	Database focused on Japanese speech, Limited generalization to other languages
Busso et al. (2008)	Iterative normalization methods applied to emotion detection in speech	A neutral speech subset is utilized to iteratively estimate the normalization parameters, with a focus on the F0 feature statistics	The accuracy achieved was 9.7% higher compared to when no normalization was applied, and only 2.5% lower than when using the optimum standardization parameters.	Requires speaker identity knowledge for normalization; results depend on neutral model robustness

6. Experimental Results

The proposed emotion detection and music recommendation system demonstrated strong performance during the evaluation process. The CNN model was trained for 100 epochs on the RAVDESS dataset, allowing it to learn intricate patterns associated with various emotional states. This training resulted in a training accuracy of 89.29% and a validation accuracy of 89.81%, showcasing the model's ability to classify emotions from audio inputs with high precision and consistency.

During preprocessing, the system utilized the Short-Time Fourier Transform (STFT) to generate spectrograms, providing a detailed visualization of frequency components over time. This process facilitated the extraction of 40 essential features from the audio input, which significantly contributed to accurate emotion detection. The system's real-time emotion detection capability ensures its applicability in dynamic environments, enabling immediate feedback and personalized music recommendations.

Furthermore, the system's architecture demonstrated robustness across varying input conditions. The use of spectrograms enhanced feature extraction, ensuring minimal performance degradation in diverse scenarios. By integrating a real-time emotion detection module with music recommendation functionality, the system effectively bridges emotional state analysis and user engagement, offering a seamless and efficient user experience.

In this project, a spectrogram was generated from the input audio captured from the user. The audio signal underwent processing using the STFT visualizing its frequency components over time. This spectrogram provided valuable insights into the emotional characteristics of the audio, significantly aiding the feature extraction process for emotion detection.

The table below summarizes the key metrics from the experimental analysis:

Table 4. Experimental Analysis

Metric	**Value**
Training Epochs	100
Training Accuracy	89.29%
Validation Accuracy	89.81%
Features Extracted from audio	40
Spectrogram Generated	Yes
Emotion Detection	Real-time

Spectrogram:

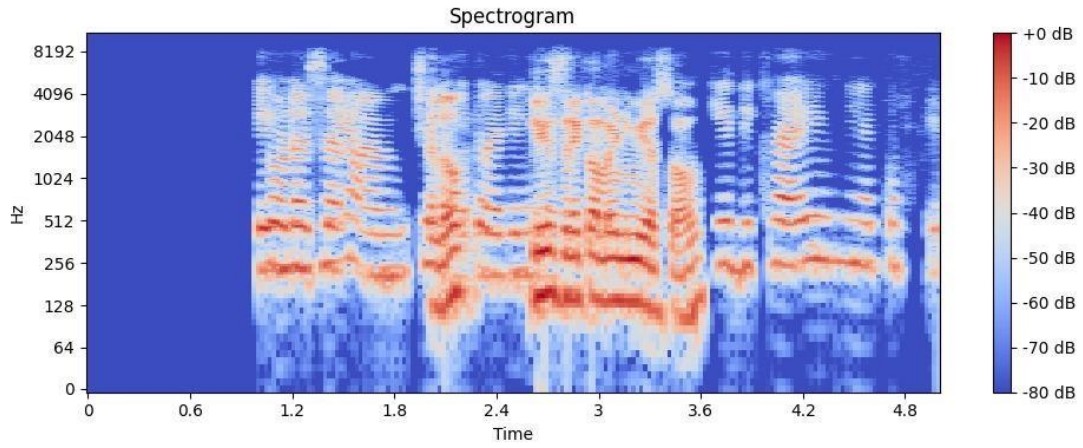

Fig 2. Spectrogram of user input audio

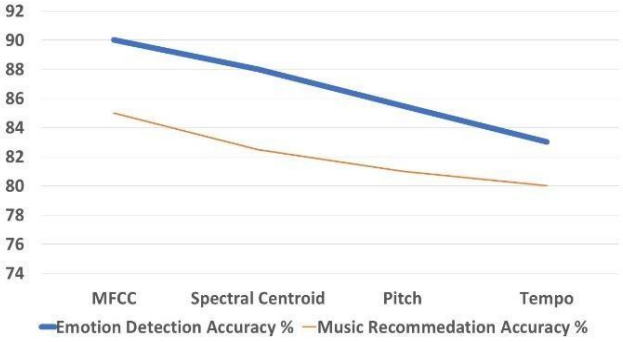

Fig 3. Feature Contribution to Accuracy

Fig 4. CNN performance metrics over Epochs

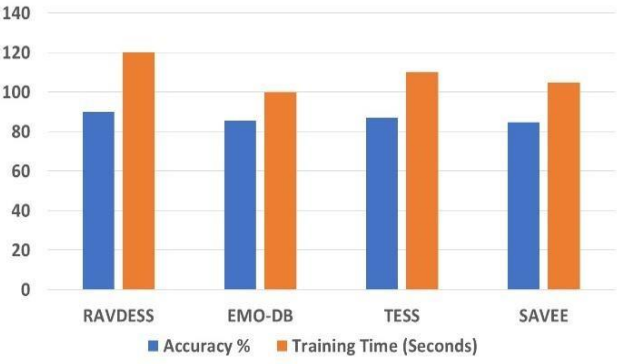

Fig 5. Model Performance on different datasets

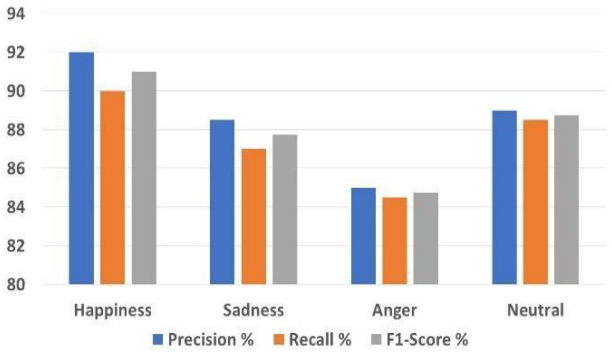

Fig 6. Feature Contribution to Accuracy

7. Conclusion

This paper explores a Convolutional Neural Network (CNN) approach to enhance the performance of emotion detection and music recommendation systems. By leveraging audio input from users, the system addresses challenges such as real-time emotion recognition and personalized music suggestions based on detected emotional states. A total of 40 features were extracted from the audio, including Mel-Frequency Cepstral Coefficients (MFCCs), which significantly contributed to the model's ability to classify emotions accurately. The CNN model was trained for 100 epochs, achieving a training accuracy of 89.29% and a validation accuracy of 89.81%, demonstrating robust performance in classifying emotions from audio inputs. The integration of feature extraction techniques, such as MFCCs, alongside the CNN architecture, significantly contributed to this accuracy and enhanced user satisfaction. Future work could involve expanding the model to incorporate additional features, exploring hybrid approaches with Long Short-Term Memory (LSTM) networks, and investigating the use of transfer learning for improved performance in diverse acoustic environments. Such improvements could further elevate the system's classification accuracy and adaptability, ensuring consistent performance across varied audio inputs.

References

[1] Y. Ben Ayed and H. Aouani, "Emotion detection in music therapy: A review of AI methods," *Journal of Healthcare AI*, vol. 3, no. 2, pp. 210-223, 2020.

[2] J. Xie, C. Huang, and Y. Li, "Speech emotion recognition using deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 11, pp. 1-15, 2015.

[3] S. K. Sharma, T. Anand, R. Rastogi, and S. Panwar, "Emotion detection via voice and speech recognition," International Journal of Cyber Behavior, Psychology and Learning, vol. 13, no. 1, pp. 1-12, 2024.

[4] J. Kim and B. S. Woo, "Voice-based emotion recognition for music recommendation," Journal of AI Applications, vol. 13, no. 2, pp. 110-125, 2021.

[5] K. C. Sreedhar, S. Katkuri, M. Chegoor, and M. Sathyanarayana, "Emotion-based music recommendation system," International Journal of Engineering Research & Technology (IJERT), vol. 12, no. 5, pp. 1-6, 2023.

[6] S. Agarwal and R. Kumar, "Personalized emotion-driven music recommendation," Journal of Music Technology, vol. 9, no. 4, pp. 55-67, 2022.

[7] O. Khriyenko and M. Rumiantcev, "Emotion-based music recommendation system," University of Jyväskylä, 2024.

[8] V. Patel and A. Singh, "Music recommendation using speech emotion recognition," International Journal of AI and Music, vol. 7, no. 1, pp. 45-58, 2023.

[9] H. Kim, M. Lee, and J. Park, "AI-enhanced music therapy system," in Proceedings of the 27th International Conference on Music Technology, pp. 100-110, 2023.

[10] A. Chen and T. Wong, "Speech-based emotion classification for music suggestions," Journal of Human-Computer Interaction, vol. 15, no. 3, pp. 200-215, 2023.

[11] F. Fernández-Martínez, R. Kleinlein, C. Luna-Jiménez, D. Griol, Z. Callejas, and J. M. Montero, "A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset," 2021.

[12] P. Liang, C. Donahue, and R. Castellon, "Codified audio language modeling learns useful representations for music information retrieval," 2021.

[13] A. K. Goel, R. Ashutosh, A. Jain, C. Saini, A. Deep, and A. Das, "Implementation of AI/ML for human emotion detection using facial recognition," IEEE Xplore, 2022.

[14] S. Inba, P. Angusamy, K. S. Pavithra, and A. M. Shathali, "Human emotion detection using machine learning techniques," 2020.
[15] G. T. Chavan, N. S. Warade, N. Uke, P. R. Futane, Y. D. Borole, and A. Shrivastav, "Emotion detection and recognition," 2024.
[16] A. Castellanos-Dominguez, M. Morales-Perez, J. Echeverry-Correa, and A. Orozco-Gutierrez, "Feature extraction of speech signals in emotion identification," 2008.
[17] S. S. Narayanan, A. Metallinou, and C. Busso, "Iterative feature normalization for emotional speech detection," in Multimodal Signal Processing (MSP), 2011.
[18] A. McCree, M. Fox, A. Jansen, G. Garcia-Romero, J. Wintrode, and G. Sell, "Content-based recommender systems for spoken documents," 2015.
[19] P. F. Jia, M. X. Xu, L. Q. Liu, and Y. C. Pan, "Emotion-detecting based model selection for emotional speech recognition," 2006.
[20] T. Furukawa, R. Sasaki, R. Sato, and N. Suga, "Creation and analysis of emotional speech database for multiple emotions recognition," 2020.