Emotion Based Music Recommendation System

B Shraddha Shetty[1], Dr Jithendra P R Nayak[2], Chandana B[3], Lavanya A[4], Kavya[5]
[1,3,4,5]*Member, Srinivas Institute of Technology*
[2]*Associate professor, Srinivas Institute of Technology*
doi.org/10.64643/IJIRTV12I12-206746-459

Abstract*—Music has a profound impact on human emotions, often prompting individuals to select songs that either reflect or alter their current mood. Most current music recommendation platforms depend on factors like genre, popularity, or past listening habits, but they frequently overlook the listener's emotional state. This project introduces an Intelligent Emotion-Based Music Recommendation System that identifies emotions from speech and suggests appropriate music in response. The system processes voice input or uploaded audio using techniques such as noise reduction, normalization, and feature extraction, including MFCCs and spectrograms. A deep learning model combining CNN and RNN architectures is employed to classify emotions such as happy, sad, angry, calm, and neutral. Based on the detected emotion, the system automatically generates personalized music recommendations by integrating APIs like Spotify and YouTube. This approach improves user experience by adapting to real-time emotional changes and supports mental well-being through mood-aligned suggestions. The system uses reinforcement learning to further personalize recommendations based on user feedback, and real-time playback via APIs ensures compatibility across platforms. By focusing on emotional context, the system delivers adaptive, emotion-aware music experiences tailored to individual users

Index Terms*—Speech Emotion Recognition, Music Recommendation System, Deep Learning, MFCC, CNN-RNN, Audio Processing.

I. INTRODUCTION

Music is a universal language that deeply influences human emotions, often serving as a source of comfort, motivation, or relaxation. People naturally seek songs that match their current mood or help them shift to a desired emotional state. However, most existing music recommendation systems focus on genre, popularity, or listening history, often overlooking the emotional context of the user. This limitation highlights the need for more intelligent systems that can adapt to users' real-time emotions and deliver personalized experiences Traditional recommendation platforms lack the ability to understand the nuanced emotional states of listeners. While some systems use facial expressions or text sentiment, speech provides a richer and more accessible channel for emotion detection due to its natural variability in pitch, tone, and intensity. Despite this, accurately classifying emotions from speech remains challenging due to overlapping features, background noise, and speaker differences. Addressing these challenges requires robust audio processing and advanced machine learning models. This project introduces an Emotion-Based Music Recommendation System that leverages voice analysis to detect user emotions and recommend suitable music in real time. The system employs deep learning models, specifically a combination of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), to classify emotions from user voice input or uploaded audio files. By extracting features such as Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms, the system captures both spectral and temporal patterns essential for emotion recognition. Modern music recommendation systems typically depend on genre, popularity, or a user's listening history to suggest tracks. While these methods can be useful, they often overlook the emotional context of the listener, which plays a crucial role in music selection. People frequently choose songs that match their current mood or help them transition to a desired emotional state. Accurately detecting emotions—especially from speech—remains a challenge due to overlapping acoustic features, background noise, and differences between speakers. As a result, most existing systems cannot deliver personalized, emotion-aware recommendations that adapt to real-time user needs. By understanding the listener's emotional context, the system can deliver recommendations that not only

match the current mood but also help shift it if desired. This capability supports user well-being and creates a more engaging music experience. The integration of advanced audio processing and deep learning techniques allows the system to adapt and improve recommendations over time, offering a dynamic and responsive solution. Existing platforms often lack the ability to capture and respond to users' emotional states in real time, limiting their effectiveness for mood-based music selection. This project addresses that gap by focusing on emotion detection from speech and providing tailored recommendations. By leveraging deep learning models and robust feature extraction methods, the system offers a more intelligent and adaptive approach to music recommendation.

II. LITERATURE REVIEW

Several research studies have explored the development of efficient data pipelines, automated preprocessing techniques, and scalable analytics frameworks.

[1] Kranthi Kiran et al. (2024) proposed a facial emotion-based music recommendation system using Haar Cascade for face detection and VGG16 for emotion classification. Emotions such as happy, sad, angry, and neutral were mapped to predefined Spotify playlists, achieving about 80% accuracy. While effective for visual emotion detection, the system depends entirely on facial input and performs poorly in low-light conditions or when the face is not visible. It also lacks speech-based analysis and long-term personalized recommendation features.

[2] Basharirad and Moradhaseli (2017) reviewed key Speech Emotion Recognition techniques, covering features such as MFCCs, LPCCs, and prosodic measures, along with classifiers like SVM, HMM, GMM, and KNN. They emphasized challenges such as speaker variability, noise sensitivity, and the lack of standardized evaluation benchmarks. The study highlighted the need for more robust and generalized SER models using deep learning and hybrid approaches.

[3] Loan Trinh Van et al. (2022) evaluated CNN, CRNN, and GRU architectures for emotional speech recognition using Mel-spectral features on the IEMOCAP dataset. The GRU model performed best with nearly 97.5% accuracy, supported by data augmentation techniques. However, the study was limited to a single dataset, lacked cross-corpus testing, and did not explore multimodal or real-world applications such as music recommendation.

[4] Esteves de Andrade and Buchkremer (2023) developed a lightweight SER system for edge devices like Raspberry Pi, using optimized CNN and SVM models for real-time prediction. The system achieved 70–77% accuracy while maintaining low latency and power efficiency. Despite its practicality, the accuracy remained moderate, and testing across diverse datasets and real-world conditions was limited.

[5] Rezapour Mashhadi and Osei-Bonsu (2023) compared 1D CNN and Random Forest models for SER using datasets such as RAVDESS, CREMA-D, TESS, and SAVEE. They used diverse audio features like MFCCs, Chroma, and Mel-spectrograms with data augmentation for robustness. Misclassification of similar emotions remained a challenge, and the system focused only on speech without considering multimodal approaches.

[6] Ahmad et al. (2024) introduced a CNN-based audio emotion detection system for business applications such as sentiment analysis and advertising feedback. Using spectrogram inputs and visualization tools, the system provided practical emotional insights but lacked evaluation on standardized datasets. It relied solely on CNN spectrogram analysis and did not incorporate advanced or hybrid deep learning models.

[7] Hao et al. (2025) presented a multimodal music emotion recognition system combining audio, visual, and physiological signals using a Bi-LSTM model. The multimodal approach improved accuracy and responsiveness on datasets like DEAP and AMIGOS. However, the system required high computational resources and was tested mainly in controlled environments, limiting real-world applicability.

[8] Kambham et al. (2025) developed a facial emotion-based music recommender using DeepFace, mapping webcam-captured emotions to predefined playlists.

The system was responsive and interactive but limited to facial input only. It lacked speech or multimodal emotion detection, adaptive learning, and personalized playlist generation.

[9] Liyanarachchi et al. (2023) surveyed multimodal MER approaches involving audio, visual, and physiological features, highlighting advancements in deep learning and emotion modeling (categorical and dimensional). Key challenges included the lack of large annotated multimodal datasets, limited interpretability, and difficulty deploying systems in real-time applications like personalized music recommendation.

[10] Patil et al. (2025) reviewed ML and DL techniques for music emotion recognition, emphasizing multimodal fusion of audio, lyrics, and physiological data. CNN, LSTM, and GAN-based models showed improved performance. However, the absence of standardized multimodal datasets and reliance on single-modality methods limited generalization and practical system deployment.

III. REQUIREMENTS

Requirements describe what a system needs to do and how it should work. They help developers understand the goals of the project and ensure the final system meets user needs. Requirements are usually divided into functional requirements, which explain the system's tasks, and non-functional requirements, which describe qualities like performance and security.

I. HARDWARE REQUIREMENTS

High accuracy and real-time performance require reliable computing resources, preferably a quad-core processor such as Intel i5 or AMD Ryzen 5 with at least 8GB RAM for smooth model execution. A minimum of 2GB storage is needed for datasets, model files, and logs. Audio input can be captured using a laptop microphone or a high-sensitivity external USB mic to ensure clear voice samples. While GPU support such as an Nvidia GTX 1650 can speed up training, the system can still run on CPU for inference. For IoT or edge deployment, the model can be executed on devices like Raspberry Pi 4B+ with at least 4GB RAM and an external microphone, enabling flexibility across desktop, mobile, and embedded platforms.

II. SOFTWARE REQUIREMENTS

The SER system requires a flexible software stack centered around Python, chosen for its strong machine learning support and rapid development capabilities. Deep learning models such as CNN, GRU, and hybrid architectures are built and trained using TensorFlow and Keras, while Librosa handles audio preprocessing and feature extraction, with openSMILE available for advanced acoustic features. Data manipulation is supported by Numpy and Pandas, and visualization is performed using Matplotlib and Seaborn. Development can be carried out in Jupyter Notebook for interactive testing or in IDEs like Visual Studio Code and PyCharm for production scripting. For web-based deployment, frameworks such as Flask or Django enable backend integration with browser-based clients, and Git is used for version control and collaborative development.

IV. METHODOLOGY

The methodology describes the step-by-step process used to detect emotions from speech and recommend suitable music. It explains how the system collects audio, processes it, extracts feature, classifies emotions, and finally provides personalized music suggestions.

I. AUDIO INPUT COLLECTION

In the first step, the system collects audio input from the user through a microphone or an uploaded voice file. This voice sample serves as the main data source for detecting emotions, so it must be captured clearly and without major interruptions. The system ensures that the audio format is supported and ready for further processing.

II. PREPROCESSING

Once the audio is collected, it undergoes preprocessing to improve quality and remove unwanted noise. This includes noise reduction, normalization, silence trimming, and segmentation. These steps help eliminate background disturbances and ensure that only useful speech information is passed to the next stage for accurate analysis.

III. FEATURE EXTRACTION

After preprocessing, the system extracts important acoustic features from the audio. Techniques such as

MFCCs, Chroma features, and Mel-spectrograms are used to capture pitch, tone, frequency patterns, and other emotional cues. These features convert the raw audio into a structured form that the model can learn from effectively.

IV. EMOTION CLASSIFICATION

In this stage, the features extracted from the user's speech are input into a pre-trained deep learning model, typically combining Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU). The CNN layers are responsible for identifying spatial and spectral patterns in the audio, such as pitch, tone, and energy variations. Meanwhile, the GRU layers analyze temporal dependencies, tracking how these features evolve over time. By taking into account both spectral and temporal characteristics, the model achieves a more accurate understanding of the emotional context in the speech.

Once processed, the system categorizes the audio into distinct emotion labels—such as happy, sad, angry, calm, or neutral—based on patterns learned from a large dataset of labeled emotional speech. This reliable detection of the user's emotional state serves as the basis for generating personalized music recommendations

V. MUSIC RECOMMENDATION

Based on the detected emotion, the system selects songs that match or enhance the user’s current mood. Uses a pre-built music database with tagged emotion for each track. Recommends playlists or individual songs in real-time. Ensures smooth transitions and personalized suggestion for better user experience. Continuously updates recommendations based on user feedback or repeated interactions. All operations are executed instantly and logged within the version control system.

V. DESIGN

I. ARCHITECURE

The Emotion-Based Music Recommendation System consists of three main stages. During training, the system uses the RAVDESS dataset with 1,440 audio files and 8 emotions, performing data preprocessing, feature extraction, and training a CNN/LSTM/Hybrid emotion classification model.

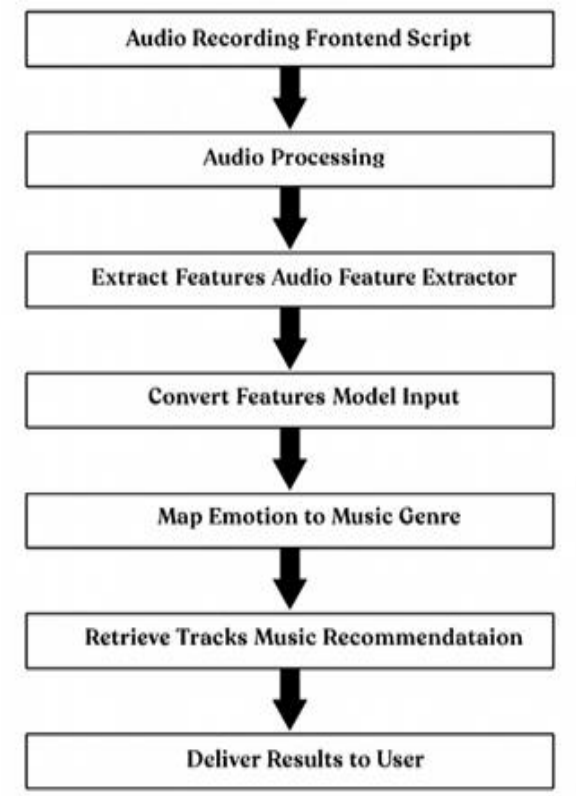

Fig1: Process Flow

In the Runtime Phase, when a user uploads or records audio, the system processes it—reducing noise, normalizing, and extracting features—and classifies the emotion using the trained model.

The identified emotion is then used in the Recommendation Phase to query the music database.

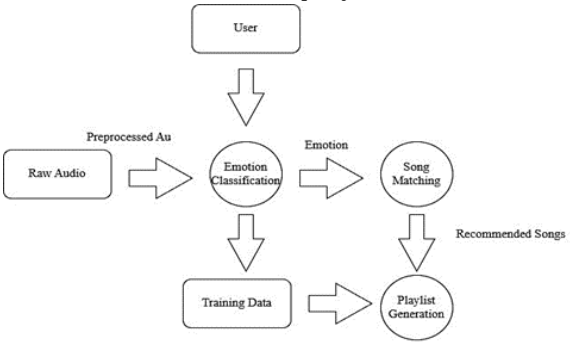

Fig2: Data Flow Diagram

The flow of data through the system's processing pipeline is depicted in the data flow diagram. The preprocessing part receives the user's audio input and does noise reduction and normalization. After the audio has been cleaned, it is sent to the feature extraction module, which produces MFCC, spectral,

pitch, and energy characteristics. The trained machine learning model uses these features as input to produce an emotion label. The recommendation engine then uses the sensed emotion to find appropriate songs by consulting the music database. The flow is completed by passing the final playlist back to the frontend. This DFD makes it easier to comprehend dependencies, data transfers, and system operations.

II. SYSTEM ARCHITECTURE

Fig3: System Architecture

The structural elements that underpin the Emotion-Based Music Recommendation System are described in the system architecture. Users can examine the final playlist and capture live inputs or upload audio using the Frontend Interface. Communication with the recommendation engine, feature extraction, emotion classification, and audio processing are all handled by the Backend API. Song metadata, genre information, and emotional tags are all stored in the Music Database and are necessary for precise playlist construction. Python, TensorFlow, Flask/FastAPI, and React/HTML5 are among the contemporary technologies used in the system's backend operations. This architecture guarantees effective performance, scalability, and modularity.

III. MODULE DESIGN

- Audio Input Module:

Handles audio data from users, either via uploaded files or real-time recording. Ensures the audio is captured in a compatible format for processing.

- Preprocessing Module:

Cleans the audio by removing noise and normalizing volume levels. Extracts relevant features such as MFCCs, chroma, spectral contrast, and tonality.

- Emotion Classification Module:

Uses a trained deep learning model (CNN, LSTM, GRU, or Hybrid) to analyse extracted features. Classifies the audio into emotions like happy, sad, angry, calm, or neutral.

- Music Database Module:

Stores songs with metadata including emotion tags. Provides a structured collection of tracks for recommendation.

- Recommendation Engine Module:

Maps detected emotions to suitable songs using an emotion-music mapping algorithm. Generates personalized playlists based on the current emotional state.

- User Interface Module:

Displays recommended songs and playlists to the user.

- Integration and Backend Module:

Connects all modules to work seamlessly. Handles data flow, model inference, and communication between front-end and back-end.

VI. RESULTS

The Emotion Based Music Recommendation System successfully identifies the user's emotional state through voice analysis and provides suitable musical suggestions. The system supports both audio upload and live voice recording, ensuring flexibility for users. The interface was tested for usability, responsiveness, and accuracy of emotion detection, and all core features performed as expected.

I. HOME PAGE

Displays the main interface of the Emotion Detection System. The page allows users to either upload an audio file or record their voice directly. The interface is simple, user-friendly, and visually appealing with gradient styling. The Choose File option lets users select a pre-recorded audio file, while the Start

Recording button activates the microphone for live audio capture. Once the audio is ready, the Detect Emotion button processes it and identifies the user's emotional state.

Fig4: Home Page

II. RECORDING IN PROGRESS

Shows the interface when live voice recording is ongoing. Once recording starts, the Start Recording button changes to Stop Recording, indicating that audio capture is in progress. A microphone notification appears in the browser, confirming system access to the user's microphone.

This visual change informs the user clearly about the recording status, ensuring that they maintain control over the process. The interface remains consistent with gradient styling, helping maintain visual continuity across the application.

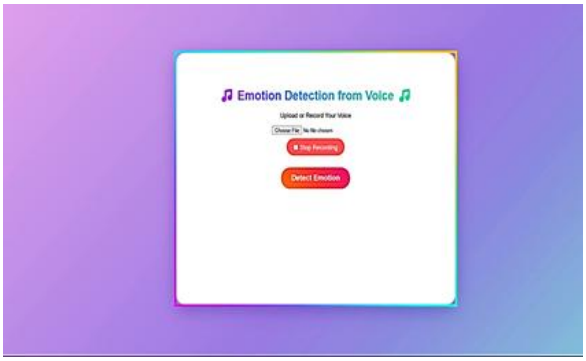

Fig5: Recording Active Interface

III. AUDIO FILE READY FOR DETECTION

Presents the interface after the system has successfully captured audio. Once the recording is completed, a confirmation message such as Audio Ready: recorded_audio.wav is displayed to notify the user that the audio file has been saved and is ready for analysis. The user can now proceed by clicking the Detect Emotion button.

Fig6: Audio Ready Status

IV. UPLOADING VOICE FILE FROM SYSTEM

The file upload feature of the application. Clicking the Choose File button opens the system's file explorer, allowing users to select an audio file manually. In the example shown, the user accesses a dataset folder to choose a .wav file from the RAVDESS dataset

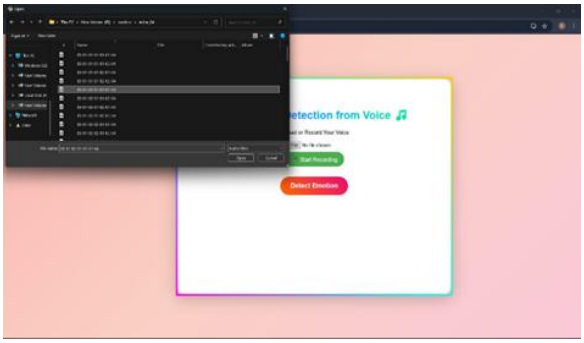

Fig7: File upload view

V. AUDIO SUCCESSFULLY UPLOADED

The screen after the user uploads an audio file. Once the file is selected, the system displays a message like Audio Ready, confirming that the audio is successfully loaded. The user can now click on Detect Emotion to proceed with emotion analysis

Fig8: Audio Upload Confirmation

VI. AUDIO ANALYSIS IN PROGRESS

The interface while the uploaded audio is being processed by the system. After the user selects a file,

the screen displays an Analyzing message, indicating that the system is actively extracting voice features from the audio. During this stage, the machine learning model calculates various acoustic parameters such as pitch, intensity, and spectral features to determine the most likely emotional category.

Fig9: Audio Analysis in Progress

IV. EMOTION DETECTION RESULT

The output screen after the system analyses the uploaded or recorded audio. The detected emotion is displayed clearly on the screen. This result is generated using the trained emotion recognition model, which processes voice features and predicts the most likely emotional state. The user can now proceed to view recommended music based on this emotion.

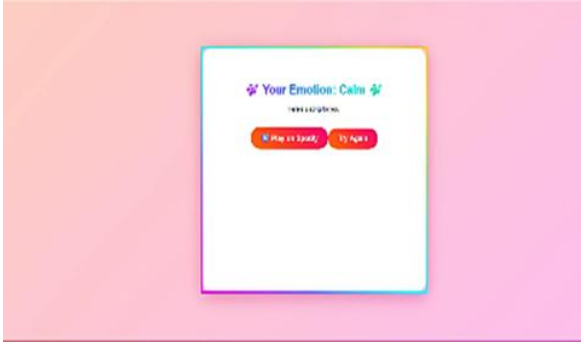

Figure 10: Detected Emotion Display

VII. CONCLUSION AND FUTURE SCOPE

The Emotion-Based Music Recommendation System developed in this project successfully demonstrates how speech-based emotion recognition can be integrated with intelligent music recommendation to deliver a highly adaptive and user-centric experience. By employing robust audio processing techniques such as Mel-Frequency Cepstral Coefficients (MFCCs), spectrogram analysis, and deep learning classification models, the system can accurately identify emotional states like happiness, sadness, neutrality, and anger solely from voice input. This approach overcomes limitations associated with visual-based recognition methods, such as lighting dependency, user privacy concerns, and non-availability of facial data.

Future enhancements of this system can focus on expanding multimodal emotion detection by incorporating facial expressions, physiological signals, and text sentiment to improve accuracy in real-world scenarios. The music recommendation engine can be further strengthened using hybrid models that combine content-based filtering, collaborative filtering, and deep learning-based music embedding techniques. Additionally, scalability improvements—such as integrating cloud services, optimizing real-time processing, and supporting larger music libraries—can make the system more robust for widespread deployment. Incorporating user profiling, contextual awareness (location, time, activity), and cultural personalization can significantly enhance recommendation relevance.

REFERENCES

[1] K. R. Scherer and H. Ellgring, “Multimodal expression of emotion: Affect programs or componential appraisal patterns?” *Emotion*, vol. 7, 2007.

[2] P. Delattre, “Les dix intonations de base du français,” French Review, vol. 40, pp. 1–14, 1966.

[3] D. K. Mac, E. Castelli, V. Aubergé, and A. Rilliard, “How Vietnamese attitudes can be recognized and confused: Cross-cultural perception and speech prosody analysis,” in Proc. Int. Conf. Asian Language Processing (IALP), Penang, Malaysia, 2011, pp. 220–223.

[4] K. R. Scherer, R. Banse, and H. G. Wallbott, “Emotion inferences from vocal expression correlate across languages and cultures,” Journal of Cross-Cultural Psychology, vol. 32, pp. 76–92, 2001.

[5] F. Daneš, “Involvement with language and in language,” Journal of Pragmatics, vol. 22, pp. 251–264, 1994.

[6] S. Shigeno, “Cultural similarities and differences in the recognition of audio-visual speech stimuli,” in Proc. 5th Int. Conf. Spoken Language Processing (ICSLP), Sydney, Australia, 1998, p. 1057.

[7] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," Electronics, vol. 10, Art. no. 1163, pp. 1–20, 2021.

[8] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, pp. 335–359, 2008.

[9] S. Chen, Q. Jin, X. Li, G. Yang, and J. Xu, "Speech emotion classification using acoustic features," in Proc. 9th Int. Symp. Chinese Spoken Language Processing (ISCSLP), Singapore, 2014, pp. 579–583.

[10] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition," in Proc. INTERSPEECH, Shanghai, China, 2020, pp. 2327–2331.

[11] R. da Silva, M. Valter Filho, and M. Souza, "Interaffection of multiple datasets with neural networks in speech emotion recognition," in Proc. 17th National Meeting on Artificial and Computational Intelligence (ENIAC), Porto Alegre, Brazil, 2020, pp. 342–353.

[12] Y. Yu and Y. J. Kim, "Attention-LSTM-attention model for speech emotion recognition and analysis of IEMOCAP database," Electronics, vol. 9, Art. no. 713, pp. 1–15, 2020.

[13] D. N. Krishna and A. Patil, "Multimodal emotion recognition using cross-modal attention and 1D convolutional neural networks," in Proc. INTERSPEECH, Shanghai, China, 2020, pp. 4243–4247.

[14] Z. Lu, L. Cao, Y. Zhang, C.-C. Chiu, and J. Fan, "Speech sentiment analysis via pre-trained features from end-to-end ASR models," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7149–7153.

[15] F. Chen, Z. Luo, Y. Xu, and D. Ke, "Complementary fusion of multi-features and multi-modalities in sentiment analysis," in Proc. 3rd Workshop on Affective Content Analysis, New York, NY, USA, 2020, pp. 82–99.

[16] R. Li, Z. Wu, J. Jia, Y. Bu, S. Zhao, and H. Meng, "Towards discriminative representation learning for speech emotion recognition," in Proc. 28th Int. Joint Conf. Artificial Intelligence (IJCAI), Macao, China, 2019, pp. 5060–5066.

[17] B. Basharirad and M. Moradhaseli, "Speech emotion recognition methods: A literature review," AIP Conference Proceedings, vol. 1891, no. 1.

[18] B. K. Kiran, P. Shanthan, K. S. Ram, and K. Usha, "Emotion-based music recommendation system using VGG16-CNN architecture," International Journal for Research in Applied Science and Engineering Technology (IJRASET), vol. 12, no. 6, 2024.

[19] L. T. Van, T. D. T. Le, T. L. Xuan, and E. Castelli, "Emotional speech recognition using deep neural networks," Sensors, vol. 22, no. 4, Art. no. 1414, Feb. 2022.

[20] D. E. de Andrade and R. Buchkremer, "Enhancing human–machine interaction: Real-time emotion recognition through speech analysis," in *Proc. Int. Conf. Human–Machine Interaction*, 2023.