International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.043 ISSN: 2581-4621

Emotion-Aware AI Feedback System for Songs: A Multi-Modal Analysis of Audio and Lyrics

Aarthi Mai A S[1], Rajeswari N[2]
[1]*Assistant Professor, Department of Computer Applications, PSG College of Technology, Coimbatore*
[2]*Assistant Professor (Senior Grade), Department of Computer Applications, PSG College of Technology, Coimbatore*

***Abstract*—** Music is a powerful medium for emotional expression yet providing objective and meaningful feedback on songs remains a challenging task. This paper presents an Emotion-Aware AI Feedback System that analyzes both audio signals and song lyrics to identify emotional and mood-related characteristics. By integrating low-level audio descriptors such as MFCCs, chroma vectors, and spectral contrast with high-level semantic features extracted from lyrics using transformer-based language models, the system generates context-aware, constructive feedback for composers and musicians. A multi-modal fusion strategy is employed to align emotional cues from audio and text, resulting in improved accuracy and interpretability. Experimental analysis demonstrates that the proposed system offers richer and more reliable feedback compared to single-modal approaches.

Keywords—*Emotion recognition, music analysis, multi-modal learning, audio features, lyrics analysis, AI-based feedback system*

INTRODUCTION

Music conveys emotions through melody, harmony, rhythm, timbre, and lyrical content. Human listeners intuitively perceive these emotional cues, but translating them into structured feedback for musicians is non-trivial. Traditional music analysis tools focus on technical correctness rather than emotional impact. Recent advances in artificial intelligence, particularly in deep learning, have enabled machines to model complex emotional patterns in audio and text. This work aims to bridge the gap between emotional intent and perceived emotion by developing an AI-based feedback system that evaluates songs holistically.

The proposed system is particularly useful for independent artists, music students, and composers who seek objective feedback during the creative process. By combining audio-based emotion recognition with lyric-based sentiment understanding, the system delivers actionable insights such as emotional consistency, mood clarity, and expressive alignment.

Related Work

Earlier studies in music emotion recognition relied primarily on hand-crafted audio features and classical machine learning algorithms. While effective to some extent, these approaches often ignored lyrical semantics. Recent research has explored deep neural networks, including CNNs and RNNs, for audio-based emotion classification. Parallel work in natural language processing has leveraged transformer models like BERT for sentiment and emotion analysis of text. However, limited work has focused on integrating both modalities for feedback generation rather than mere classification. This paper extends existing research by emphasizing interpretability and feedback usefulness through multi-modal fusion.

System Architecture

The proposed system consists of five main modules: preprocessing, feature extraction, feature analysis, multi-modal fusion, and feedback generation.

1. Input Module: Accepts audio files (WAV/MP3) and corresponding song lyrics.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.043 ISSN: 2581-4621

2. Preprocessing Module: Normalizes audio, remves noise, segments signals, and cleans lyrical text.
3. Feature Extraction Module: Extracts acoustic and textual features relevant to emotion.
4. Multi-Modal Fusion Module: Combines audio and lyric features into a unified emotional representation.
5. Feedback Generation Module: Produces descriptive and prescriptive feedback for the user.

The following fig 3.1 represents the architecture diagram. The architecture ensures independent yet complementary analysis of audio and lyrics, followed by fusion to improve robustness and interpretability.

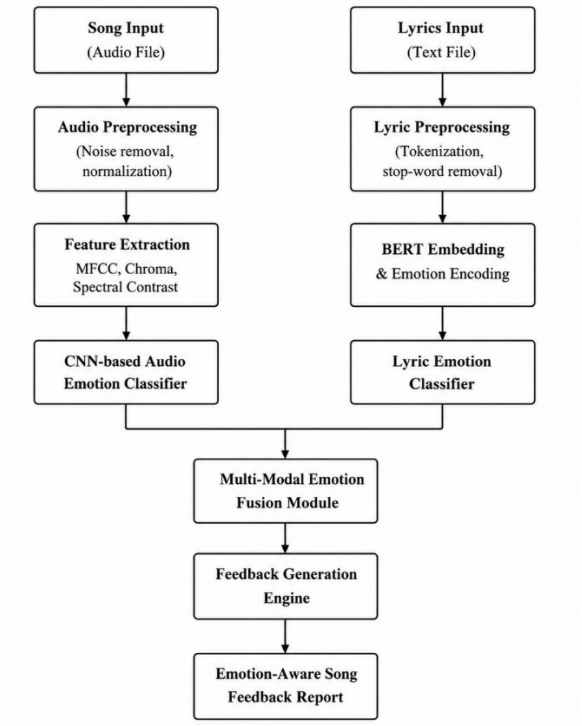

Fig 3.1 System Architecture Diagram

Input Module

The Input Module serves as the entry point of the system. It accepts audio files in common formats such as WAV or MP3 along with the corresponding song lyrics in text form. This dual-input design enables the system to capture both musical and linguistic expressions of emotion. By processing audio and lyrics together, the system ensures that emotional cues from melody, rhythm, and words are considered simultaneously, providing a holistic understanding of the song.

Preprocessing Module

The Preprocessing Module prepares raw inputs for effective feature extraction. For audio signals, this module performs normalization to maintain consistent amplitude levels, noise reduction to eliminate background disturbances, and segmentation to divide the signal into short frames suitable for analysis. For lyrics, preprocessing includes removing punctuation, stop words, and irrelevant symbols, as well as tokenization and text normalization. These steps reduce redundancy and noise, allowing the model to focus on emotionally relevant patterns.

Feature Extraction Module

In the Feature Extraction Module, meaningful representations are derived from both audio and lyrics. Acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, and spectral contrast are extracted to capture timbre, harmony, and energy variations associated with emotion. Simultaneously, textual features are obtained from lyrics using transformer-based

DOI:10.46647/ijetms.2026.v10i03.043 ISSN: 2581-4621

language models, which encode semantic meaning and emotional context. These extracted features form the foundation for accurate emotion recognition.

Multi-Modal Fusion Module

The Multi-Modal Fusion Module integrates audio-based and lyric-based features into a unified emotional representation. Since emotions may be conveyed differently through music and words, this module aligns and combines features using a fusion strategy such as weighted averaging or late fusion. This approach enhances robustness by balancing the influence of both modalities and resolving inconsistencies between lyrical sentiment and musical expression.

Feedback Generation Module

The Feedback Generation Module translates the fused emotional representation into human-readable feedback. Instead of providing only emotion labels, this module generates descriptive insights such as dominant mood, emotional intensity, and alignment between lyrics and melody. It also offers prescriptive suggestions to improve emotional clarity and expression. This feedback-oriented design makes the system particularly valuable for musicians, composers, and students seeking constructive guidance.

Feature Extraction

Audio Feature Extraction

Audio emotion cues are captured using the following features:

- MFCC (Mel-Frequency Cepstral Coefficients): Represent timbral texture and vocal characteristics.
- Chroma Vectors: Capture harmonic and pitch class information related to mood.
- Spectral Contrast: Reflects energy distribution and dynamic variations.

MFCC (Mel-Frequency Cepstral Coefficients)

MFCCs are one of the most widely used features in audio and speech analysis for capturing timbral characteristics of sound. They model how humans perceive sound by mapping frequencies onto the Mel scale, which emphasizes lower frequencies where human hearing is more sensitive. In music emotion analysis, MFCCs effectively represent vocal tone, instrument texture, and articulation, making them highly useful for identifying emotional qualities such as warmth, harshness, softness, or brightness in a song.

Chroma Vectors

Chroma vectors, also known as pitch class profiles, represent the distribution of energy across the twelve musical pitch classes (C, C#, D, etc.) irrespective of octave. These features capture harmonic and melodic information that is closely linked to emotional perception. Major and minor tonal structures, chord progressions, and key changes reflected in chroma vectors play a significant role in conveying moods such as happiness, sadness, tension, or calmness within a musical piece.

Spectral Contrast

Spectral contrast measures the difference between peaks and valleys in the frequency spectrum, highlighting the relative distribution of energy across different frequency bands. This feature captures dynamic variations and timbral richness, which are important indicators of emotional intensity. Higher spectral contrast often corresponds to energetic or aggressive emotions, while lower contrast is associated with softer, calmer moods. Thus, spectral contrast contributes to distinguishing emotional dynamics within a song.

These features are computed over short-time frames and aggregated using statistical measures.

Lyrics Feature Extraction

Lyrics provide explicit semantic and emotional cues that play a vital role in conveying the intended mood of a song. In the proposed system, lyrical content is analyzed using a transformer-based language model such as BERT (Bidirectional Encoder Representations from Transformers), which

DOI:10.46647/ijetms.2026.v10i03.043 ISSN: 2581-4621

is capable of understanding complex language patterns, contextual meaning, and emotional nuance. Unlike traditional sentiment analysis techniques that rely on keyword frequency or predefined sentiment lexicons, BERT captures the meaning of words based on their surrounding context, making it particularly suitable for poetic, metaphorical, and expressive language commonly found in song lyrics.

The lyrics are first processed through tokenization, where the text is divided into sub-word units that enable the model to handle rare words, slang, and creative word forms often used in music. These tokens are then transformed into contextual embeddings, which are dense vector representations encoding both semantic meaning and emotional tone. Each word embedding reflects not only the individual word meaning but also its relationship with neighboring words, allowing the system to detect subtle emotional expressions such as irony, longing, emotional contrast, or intensity across different song segments.

Based on these contextual embeddings, the model infers emotional polarity and discrete emotion categories such as joy, sadness, anger, calm, and excitement. The output is represented as a probability distribution across these emotion classes, indicating both dominant and secondary emotions present in the lyrics.

This detailed emotional representation enables the system to analyze lyrical mood progression, identify emotional shifts between verses or choruses, and evaluate the consistency between lyrical emotion and musical expression. Consequently, lyrics feature extraction significantly enhances the overall accuracy, depth, and interpretability of the emotion-aware AI feedback system.

Multi-Modal Fusion Strategy

To effectively align emotional cues derived from both audio signals and lyrical content, the proposed system employs a late fusion strategy, which combines high-level emotion representations extracted independently from each modality. In this approach, audio-based emotion recognition and lyrics-based emotion analysis are performed separately using their respective feature extraction and classification models. Each modality produces an emotion probability vector, representing the likelihood of different emotional states such as joy, sadness, anger, calm, and excitement.

These modality-specific emotion vectors are then integrated using weighted averaging, where each modality is assigned a weight that reflects its relative contribution to accurate emotion prediction. The weights are learned and optimized based on validation performance, enabling the system to adapt dynamically to different types of songs. For instance, in instrumental-heavy compositions where lyrics are minimal or absent, greater importance is given to audio-based emotional cues. Conversely, in lyric-driven songs where textual expression dominates emotional meaning, higher weight is assigned to lyrics-based emotion predictions.

The late fusion strategy offers significant advantages in terms of flexibility, robustness, and interpretability. By combining decisions at a higher semantic level rather than raw features, the system reduces the impact of noise and modality-specific errors. This fusion mechanism also helps resolve conflicts that may arise when musical emotion and lyrical sentiment convey different emotional signals, such as upbeat melodies paired with melancholic lyrics. The resulting fused emotional representation provides a more balanced and reliable understanding of the song's overall emotional character, thereby enhancing the quality and consistency of the generated feedback.

Feedback Generation

Rather than limiting the system output to discrete emotion labels, the proposed approach focuses on generating qualitative, meaningful, and actionable feedback that can support musicians during the creative process. The goal of the feedback generation module is to translate complex model outputs into insights that are easily understandable and practically useful for composers, singers, and music students. This human-centered design ensures that the system goes beyond classification and acts as an intelligent assistant for emotional refinement in music.

DOI:10.46647/ijetms.2026.v10i03.043 ISSN: 2581-4621

The system analyzes the fused emotional representation to identify emotional consistency between lyrics and melody, highlighting whether both modalities reinforce the same mood or convey contrasting emotions. It also determines the dominant emotional state of the song along with its intensity level, allowing musicians to understand how strongly a particular emotion is expressed. In addition, the system provides prescriptive suggestions aimed at enhancing emotional clarity, such as adjusting vocal expression, tempo, harmonic structure, or lyrical phrasing to better align with the intended mood.

To generate this feedback, the system employs a combination of rule-based templates and model confidence scores. Rule-based templates structure the feedback in natural language, while confidence scores guide the selection and emphasis of feedback points based on prediction reliability. This hybrid strategy ensures that the feedback remains both consistent and context-aware. As a result, the generated feedback is human-readable, interpretable, and emotionally informative, making it accessible even to users without technical expertise and enhancing the overall usability of the emotion-aware AI feedback system.

Experimental Evaluation

The system was evaluated on a dataset of annotated songs spanning multiple genres. Performance was measured using emotion classification accuracy and qualitative user feedback. Results indicate that the multi-modal approach outperforms audio-only and lyrics-only baselines, particularly in complex emotional categories such as bittersweet and nostalgic.

Conclusion

This paper presents an Emotion-Aware AI Feedback System that integrates audio and lyrical analysis to deliver meaningful emotional insights for songs. By leveraging multi-modal learning and interpretable feedback generation, the system supports musicians in refining emotional expression. Future work includes real-time feedback, genre-specific modeling, and user-adaptive personalization.

References

[1] *Y. Kim, "Convolutional Neural Networks for Sentence Classification," EMNLP, 2014.*
[2] *A. Vaswani et al., "Attention Is All You Need," NeurIPS, 2017.*
[3] *J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, 2019.*
[4] *T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," ICLR, 2013.*
[5] *P. Tzanetakis and G. Cook, "Musical Genre Classification of Audio Signals," IEEE Transactions on Speech and Audio Processing, 2002.*
[6] *S. Hershey et al., "CNN Architectures for Large-Scale Audio Classification," ICASSP, 2017.*
[7] *D. P. W. Ellis, "Classifying Music Audio with Timbral and Chroma Features," ISMIR, 2007.*
[8] *B. McFee et al., "librosa: Audio and Music Signal Analysis in Python," Proceedings of the Python in Science Conference, 2015.*
[9] *A. Krizhevsky et al., "ImageNet Classification with Deep Convolutional Neural Networks," NeurIPS, 2012.*
[10] *Aarthi Mai A S and Rajeswari N, "An AI-Driven Approach for Comprehensive Song Feedback: From Audio Processing to Quality Enhancement",International Journal of Science and Technology, Volume 16,2025.*